The University Of Sheffield.

Deep learning applications to automate phenotypic measurements on biodiversity datasets

Yichen He

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Sheffield

Faculty of Science

Department of Animal and Plant Sciences

September 2020

Acknowledgements

First, I would like to thank my first supervisor Dr Gavin Thomas. Gavin has been a super helpful and supportive supervisor to me. With Gavin's supports, I had a great PhD experience. From bottom of my heart, thank you, Gavin, you are the best supervisor! I would like to thank my second supervisor Dr Steve Maddock. Steve has given a lot of ideas and comments on my PhD projects. His expertise in computer science helps me a lot, especially my PhD is interdisciplinary. I want to shout out to Dr Christopher Cooney. Not only Chris and I are really good mates, but Chris has also given a lot of comments and ideas to my PhD. I really enjoyed working next to Chris, he is a career role model for me. I would also like to thank other people such as my lab mates, Prof. Roger Butlin, Jenny Larsson and Zuzanna Zagrodzka who have helped and supported my PhD.

Thanks to the Leverhulme Trust, which funds my PhD. I am really happy that I spent four years in the Department of Animal and Plant Sciences and the University of Sheffield. I have met so many nice people here and made a lot of friends. My lab mates are all wonderful people, here I would like to thank everyone, Alex, Angela, Chris, Emma, Frane, Joe, Jon, Joseph, Louie and Thomas, in the lab! I had a really good time in the office. There are also many amazing people and friends I met in Sheffield. I will remember all the memories we had and all the supports from them. Special thanks to Angela and Joe, we had a great time for four years, supporting each others, sharing things from life and work and doing fun things. These four years will be an unforgettable and shining memory in my life.

Finally, I want to thank my family. 谢谢爸爸妈妈在这将近三十年对我的养育之恩，在这四年没有你们的支持我是不可能完成博士的。接下来要进入新的篇章，革命尚未成功，同志仍需努力。我爱你们，我爱我们这个家。最后祝身体健康，万事如意，望早日能相见！谢谢陈艳给我的爱，今生今世我不忘怀。能一起相互支持过完这四年很不容易，祝我们携手同行 共创辉煌。

最后这篇论文我想献给我的奶奶，姥姥，姥爷，我会继续努力，然后成为一个能让你们骄傲的人！

Abstract

The growing number of digitised biological specimens brings new possibilities for the study of a wide range of evolutionary questions at broad scales. Taking measurements on digital photos often requires annotations (e.g. placing points on focal locations), and many projects label their digitised specimen datasets (commonly more than thousands of images) manually, which could take a huge amount of time. Deep learning is the state-of-the-art for many computer vision tasks. Deep learning models can be trained on a set of manually annotated images, and can make accurate predictions based on what they learned. To what extent deep learning can help to improve the measurement process on digitised collections has yet to be thoroughly explored. Here, I have applied deep learning models to three tasks on two datasets (bird specimens and *Littorina* shell images), and show that predicted labels are remarkably accurate and that downstream biological analyses using these labels generated biologically meaningful results. First, I used pose estimation models (algorithms originally designed to identify human body parts) to locate keypoints on bird specimens. The results showed high accuracy with 95% of the validation images (N=5,094) correctly predicted, and rapid generation of data (less than three days to predict keypoints on the whole dataset of >120,000 images). Colours measured by points showed that male birds tend to be more colour-diverse than females. Second, I applied deep learning models to segment the overall plumage areas on the bird dataset. More than 95% of the plumage areas were correctly segmented, and it also took less than three days to segment the whole dataset. I found that colour diversities (calculated from segmentations) among closely related birds tend to be similar across more than 7,500 bird species. Finally, I built PhenoLearn, a user friendly tool that provides functions such as manual annotation (to create training images), predicting using deep learning, and reviewing predictions. I illustrate the broader applicability of PhenoLearn in an example of morphological landmarking on *Littorina* shell images. More than 98% of the predicted landmarks were placed within the acceptable range. The methods and tools introduced here both illustrate the value of deep learning and significantly increases accessibility to deep learning approaches for non-expert biologists allowing the rapid accumulation of phenotypic datasets at large scales. Taken together, these results show that deep learning

methods have great potential for speeding up the measuring process on digitised specimens while producing accurate annotations.

Declaration

The following people were involved in this research project:

Yichen He

Gavin H. Thomas

Steve Maddock

Christopher R. Cooney

Roger Butlin

Zuzanna B. Zagrodzka

Chapter 2: A Deep Learning application on colour measurements of plumage regions on standardised avian specimen images

YH conceived the idea for the study (with data from GHT and CRC);

YH performed the analyses and wrote the manuscript;

GHT, SM and CRC commented on drafts of the manuscript.

Chapter 3: Segmenting biological specimens from photos: a comparison of classic computer vision and segmentation methods

YH conceived the idea for the study (with data from GHT and CRC);

YH performed the analyses and wrote the manuscript;

GHT, SM and CRC commented on drafts of the manuscript.

Chapter 4: PhenoLearn: A software package and workflow for annotating digital images of biological data

YH conceived the idea for the study (with data from RB and ZBZ);

YH performed the analyses and wrote the manuscript;

GHT, SM, CRC and RB commented on drafts of the manuscript.

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously been presented for an award at this, or any other, university.

CONTENTS

LIST OF MAIN FIGURES

LIST OF MAIN TABLES

Chapter 1 General Introduction

1.1 What is phenomics

The field of phenomics was originally defined as a discipline for understanding biochemical and physiological pathways by measuring phenotypes (Bilder et al. 2009). The field focuses in particular on the genetics of complex diseases (Schork 1997), including disease prediction (Manolio et al. 2009), and estimating the heritability of complex diseases. More recently, the scope of phenomics has expanded and is now more broadly defined as the acquisition of high-dimensional and large-scale phenotypic data (Houle et al. 2010) such as identifying breeding values of crop plants (Monteiro et al. 2002; Araus and Cairns 2014). A growing subfield of phenomics encompasses studying patterns and processes from biodiversity data (Klingenberg and Gidaszewski 2010; Benton 2015). A major challenge for all applications of phenomics is to develop high-throughput pipelines for large-scale data.

While generating genomic data has become progressively more cost-effective allowing compilation of complete genome databases for numerous organisms (Roach et al. 2010; Jarvis et al. 2014; Zhang et al. 2014), phenotypic databases are less well developed in part due to low efficiency in collecting phenotypic information (Lussier and Liu 2007). Measurement of large samples of organisms is particularly important in studies of biodiversity that address evolutionary and ecological questions. Natural history museums house extensive, but largely underexploited, collections of biological specimens. Measurement is often limited by access to, and quality of, specimens yet museum collections house vast numbers of specimens, estimated to be around 1.2 × 10^9 to 1.9 × 10^9 units in museum collections globally in 2010 (Ariño 2010). Collection digitisation is an important aim for many natural history museums, which can provide a rich source of phenotypic and biodiversity data (Blagoderov et al. 2012). Digitisation (including 2D photos, videos and 3D scans) and associated metadata is a common way to store permanent records of specimens as digital data while keeping their colour, shape and posture information (Rohlf 2006; Stevens et al. 2007). Some studies have built automatic phenotyping pipelines based on computational methods such as machine learning (Pearson et al. 2020; Porto and Voje 2020; Soltis et al. 2020). However, most the digitised data cannot be used in analyses directly, so

powerful, high-throughput, pipelines for getting useful information and measurements are necessary.

1.2 Mobilising data from biodiversity images

Digital photos of specimens can be labelled in numerous ways to extract or measure phenotypic information. Point placement is probably the most common annotation to measure the location on an image (Figure 1.1A). Point annotations can be used as landmarks to identify morphological features (e.g. homologous points) on specimen images (Adams, Rohlf, and Slice 2013; Bookstein 1991). Polygons are an annotation that can measure focal areas on photos, for example to identify body parts of birds to measure their colour information (Dale et al. 2015; Cooney et al. 2019). Segmentation is a commonly used method to label regions of interest in medical images like cells (Meijering 2012; Xing and Yang 2016) and organs (Balafar et al. 2010; Mharib et al. 2012). Segmentation has also been used to measure phenotypes of live plant photos (Minervini et al. 2014; Scharr et al. 2016). An example of segmenting a specimen image is shown in Figure 1.1B. Detecting (e.g. placing bounding boxes around focal specimens as shown in Figure 1.1C), identifying (e.g. classifying whether focal specimens appear as shown in Figure 1.1D) or even counting the number of specimens have been widely used on photos collected in the wild (e.g. images from camera traps; Karanth 1995; Wegge et al. 2004; Khorrami et al. 2012). Results can be used to study questions such as the density of one or multiple types of organisms. Besides digitised museum collections, here, photos collected in the wild or from live specimens should also be considered as biodiversity images or digitised specimen images. In this thesis, I aim to assess and apply deep learning models on biodiversity to improve measuring speed, focusing on digitised museum collections, but where the results are equally applicable to wildlife and live specimen photos.

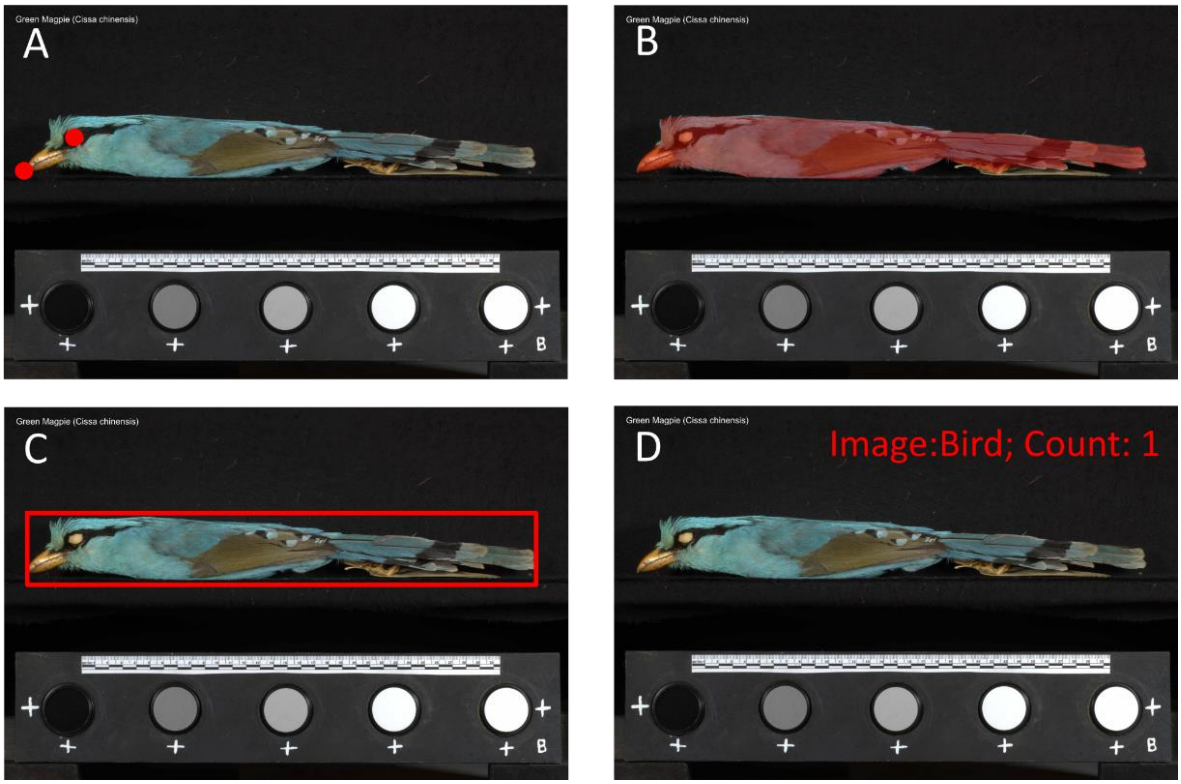


Figure 1.1. Examples of annotations that can be used on digitised images (one image of a bird specimen from Project Plumage was used here). (A) Points are used to identify the beak tip and eye of the bird; (B) The whole bird is segmented, and the remaining parts of the image were not segmented; (C) The bird is detected and located using a bounding box. (D) The image is classified as 'bird' and one bird was presented in the image.

Placement of annotations and measurements by people with expert knowledge is widely used on digitised datasets, and can produce accurate annotations. However, it can be slow when datasets are large, and the number of experts is insufficient. Crowdsourcing provides an alternative way of annotating images. Crowdsourcing can speed up the annotation process by opening up digitised datasets to a large, relatively open and often rapidly-evolving group of internet users. Crowdsourcing has been used in landmarking of thousands of 3D scans of bird beaks (www.markmybird.org) to study patterns of diversification through time across birds (Cooney et al. 2017). Other crowdsourcing applications include paying internet users to landmark fish images using an internet application based on Amazon Mechanical Turk (Chang and Alfaro 2016). There are also many citizen science projects of annotating large-scale specimen datasets

on Zooniverse (www.zooniverse.org), which is a web-based platform that allows users to create and customise their own citizen science projects. Zooniverse includes Project Plumage (placing annotations on digitised bird specimen images to measure plumage colours; see data section in this chapter, chapter 2 and 3 for detail), and many other projects related to biology that are mainly used in identifying and counting specimens in photos or on digitising specimen metadata. However, the faster accumulation of data is traded off against potentially lower accuracy of crowdsourcing compared to experts' labelling and the fluctuant engagement of citizens. Manually annotating images (both expert-only and crowdsourcing) can not meet the goal of creating a high-throughput phenotyping pipeline, therefore semi- or fully automatic and accurate measurement placement methods are crucial. Deep learning is a possible solution for producing large-scale sets of annotations on images accurately, rapidly, and automatically requiring only a small training set (comparing to the whole dataset) of manual annotations. Notably, some of the projects on Zooniverse (e.g. Zen of Dragons, www.zooniverse.org/projects/willkuhn/zen-of-dragons) aimed to build training sets with crowdsourcing for machine learning or deep learning models.

1.3 Deep Learning

Neural networks are one of the main methods in machine learning. Regular neural networks have three types of layers (input, hidden and output) as shown in Figure 1.2A. Nodes (or neurons) can be seen as values, and edges can be seen as transforming from input values (start of the edge) to output values (end of the edge). After processing inputs in the hidden layer, outputs are generated in the output layer. Initial parameters (e.g. weights and biases) of transformations would normally generate outputs that are far from the ground truth (i.e. humans' outputs). A loss function is used to show the difference between predicted outputs and the ground truth. The gradient of the loss is then used to update parameters using the gradient descent algorithm by backpropagating through each layer (Bishop 2007), aiming to make the updated network produce to generate outputs that are similar to the ground truth. This step can be seen as the network learning (or training the network).

More recently, deep learning, a newly emerged field using neural networks with multiple hidden layers, has led to great improvements in many computer vision problems such as image classification and object detection. Deep learning has been the key technique for real-world applications like facial recognition (Bulat and Tzimiropoulos 2017; Ranjan et al. 2017) and autonomous driving (Treml et al. 2016; Al-Qizwini et al. 2017). The main architecture of the deep neural networks used to solve computer vision and image problems is the convolutional neural network (CNN), which was originally proposed for handwritten digit recognition (LeCun et al. 1990).

CNNs take images as the input, treating each pixel as an input node. Because regular neural networks have fully-connected layers (nodes from adjacent layers are fully pairwise connected), there will be a large number of inputs (e.g. a 256 x 256 RGB image has 196,608 pixels), which produces too many transformation parameters and makes the learning inefficient and prone to overfitting. Convolutional layers (Conv layers) are used to reduce node parameters by sharing parameters using the convolution operation. Conv layers transform inputs (e.g. pixel values of the input images) into a three-dimensional output (imagine an image with multiple channels; dimension: width x height x depth) and normally an output has more channels (depths) than the input (see Conv layer in Figure 1.2B). During the convolution, multiple nodes (i.e. pixels) are convoluted into one output node. Pooling layers are normally applied following Conv layers and served as down-sampling results (reducing width and height) by taking the average or maximum value of selected regions (see Pooling layer in Figure 1.2B), which create outputs with suitable dimensions. After multiple layers of Conv and pooling layers, layers for generating the final outputs are added depending on the task. Fully-connected layers are common for classification problems (Krizhevsky et al. 2012; Simonyan and Zisserman 2014), generating nodes that represent scores of the classes (Figure 1.2C). Using convolution layers is another choice that produces image-like outputs that can be used in segmentation tasks (Figure 1.2D; Long et al. 2015). Similar to the neural network, backpropagation and gradient descent are used to optimise CNNs. After deep learning networks are trained, image features can then be automatically learned and extracted avoiding the need to use pre-defined features such as scale-invariant feature transform (SIFT, Lowe 1999).

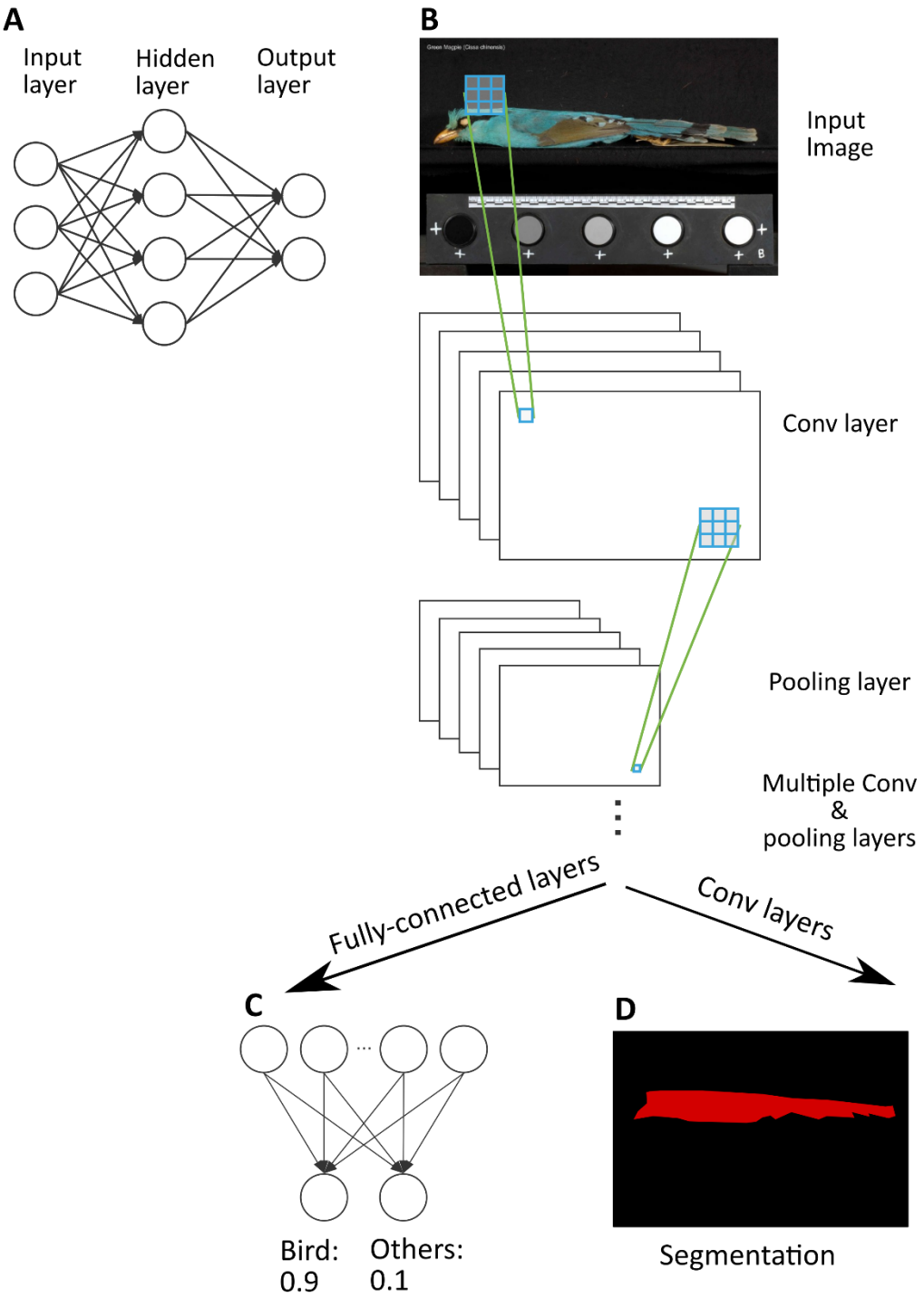


Figure 1.2. (A) A regular neural network with one hidden layer. (B) A convolutional neural network (CNN) with convolutional layers and pooling layers. Blue grids and squares show that multiple nodes (grids) from inputs are calculated into one output node in Conv and Pooling layers. Outputs of multiple Conv and pooling layers can then be used in tasks like (C) classification (the image scores 0.9 of being a bird and 0.1

of being others) and (D) segmentation (red areas are segmented as bird and black areas are segmented as the background).

The development of graphic processing units (GPU) and programming libraries in recent years has enabled parallel large-scale calculations over hundreds and thousands of threads (NVIDIA 2017) making the approach computationally efficient. In 2012, a deep learning method won the best model for the ImageNet challenge (an image classification challenge; www.image-net.org; Deng et al. 2009) for the first time (Krizhevsky et al. 2012). Since then, different architectures have been designed to improve feature extraction with deeper and more complex combinations of layers (Simonyan and Zisserman 2014; Szegedy et al. 2014; He et al. 2016).

1.3.1 Deep learning challenges

The top-performing models of many computer vision challenges that cover classification (Deng et al. 2009), object detection (Deng et al. 2009; Lin et al. 2014), pose estimation (Johnson and Everingham 2011; Andriluka et al. 2014), and semantic segmentation (Everingham et al. 2015) are all deep learning-based. These challenges aim to solve similar problems as annotating digitised specimen images (Figure 1.1 and Figure 1.3), suggesting the possible applications of using deep learning on specimen datasets.

Pose estimation uses deep neural networks to identify human body parts and joints as points on images (Figure 1.3A). Methods such as Stacked Hourglass (Newell et al. 2016) and Convolutional pose machine (CPM, Wei et al. 2016) have achieved high accuracy in identifying human posture images. Pose estimation models have been applied in other fields, such as tracking mice (Mathis et al. 2018b) and identifying fruit flies postures (Pereira et al. 2019), which have shown the potential for placing points on specimen photos.

Semantic segmentation using deep learning is the state-of-the-art for segmenting images into different classes and categories automatically (Long et al. 2015; Chen et al. 2017b), and has outperformed other classic segmentation methods (e.g. thresholding) on many tasks. An example result of semantic segmentation is shown in Figure 1.3B. Semantic segmentation

methods have achieved high accuracies on many datasets and segmentation challenges, for example, PASCAL VOC 2012 (Everingham et al. 2015) is a segmentation challenge that aims to segment images with complex background and objects into 20 classes (e.g. humans and cars). Semantic segmentation networks have also been designed for segmenting biomedical images (Ronneberger et al. 2015; Li et al. 2018). It is possible to use semantic segmentation methods to automatically segment focal areas on specimens.

Classification challenges aim to use computer algorithms to classify images into different classes (Figure 1.3C) while object detection is targeted at detecting focal objects and locating them (e.g. placing bounding boxes around detected objects) on images (Figure 1.3D). Image classification (Simonyan and Zisserman 2014; He et al. 2016) and object detection (Ren et al. 2015) algorithms have been widely used in detecting and classifying images of wildlife animals (Kellenberger et al. 2017, 2018; Norouzzadeh et al. 2018; Schneider et al. 2018). The numbers of images taken from (UAV) and camera traps are normally very large, so using deep learning can speed up the measuring step.

Digitisation normally produces images with high imaging standards (e.g. a consistent background, a fixed number of specimens per one image). Previous studies have shown that deep learning can predict accurate annotations on images that are more complex, variant and less consistent than specimen images from digitisation. This suggests that deep learning is potentially a powerful tool that could be applied to many large-scale image databases.

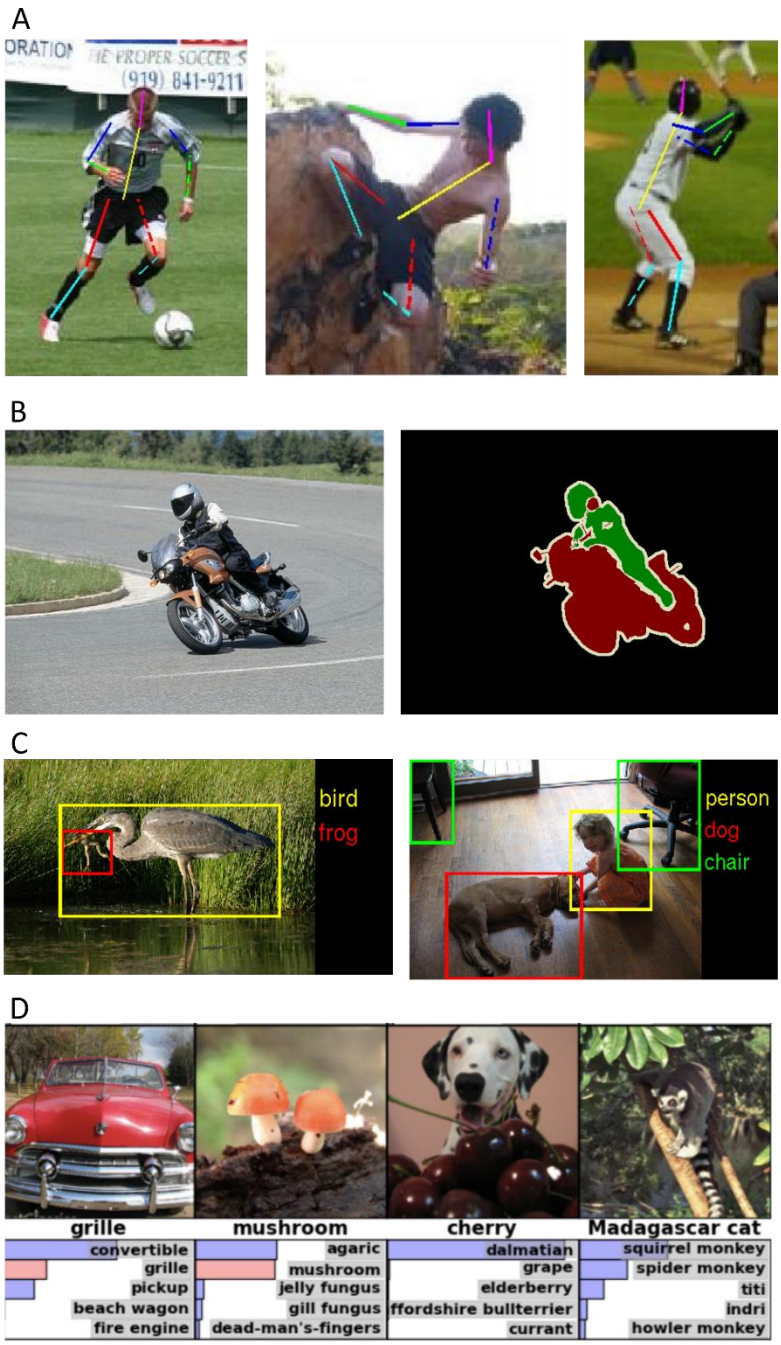


Figure 1.3. Examples of different deep learning challenges. (A) Identifying human body parts and joints to measure humans' postures from the Leeds Sports Pose Dataset (Johnson and Everingham 2011). (B) Segmenting the human and the motorcycle on an image from PASCAL VOC 2012 challenge (Everingham et al. 2015). (C) Detecting different objects on images from ImageNet Large Scale Visual Recognition Challenge (Deng et al. 2009). (D) Classification examples from the work of Krizhevsky et al. (2012; a part

of Figure 4 in Krizhevsky et al.'s paper is used here). Examples A-D are similar to annotations in Figure 1.1 A-D respectively.

1.4 Thesis objectives & data

The overall aim of this thesis is to implement a high-throughput annotation pipeline on digitised specimen images using deep learning models. To achieve this aim, I tried multiple deep learning architectures including Stacked Hourglass (Newell et al. 2016), convolutional pose machine (Wei et al. 2016) and DeepLabV3+ (Chen et al. 2018), which can predict points and segmentations. I applied these methods to three problems in two different image datasets.

1.4.1 Data

1.4.1.1 Avian plumage colour

Birds have evolved into a wide diversity of colour space (Stoddard and Prum 2011). Birds can perceive light across a wider spectrum, including parts of the ultraviolet spectrum, than humans (Goldsmith 1990; Cuthill et al. 2000). Studies have mapped birds colours into an avian-visual based tetrahedral colour space based on the four cones receptor cone types (Stoddard and Prum 2008, 2011; Cooney et al. 2019). Bird's plumage has functions from crypsis or camouflage to mate choice/attraction and social signalling (Hill et al. 2006a). Understanding how birds see their plumage colours helps to answer questions such as whether plumage colour evolution is driven by sexual selection (Dale et al. 2015; Dunn et al. 2015; Cooney et al. 2019) or natural selection (Slagsvold et al. 1995; Willink et al. 2014; Dunn et al. 2015); whether colour producing mechanisms limit the evolution of plumage colours (Stoddard and Prum 2011); how ecological factors impact on colours (Dalrymple et al. 2015).

Spectrometers, the traditional way of measuring colours on specimens (Ewen et al. 2006; Stoddard and Prum 2008, 2011), can only measure the colour information of a small point for every measurement, and it is cumbersome and time-consuming to measure an area of colour information which including many points. Measuring the actual specimens can be invasive and it will be hard to replicate the measuring procedures. Digital photos provide high-quality colour information of different wavelengths (e.g. visible or ultraviolet light) with specialised cameras

and calibrations (Stevens et al. 2007; Troscianko and Stevens 2015). These techniques enable biologists to quantify signalling traits such as colouration and pattern by placing annotations on digital images. More than tens of thousands of annotations are placed manually in studies (Dale et al. 2015; Cooney et al. 2019). Raw colour pixel values are normally transformed to fit the research goals. For example, raw pixels values (i.e. values from RGB and UV channels) from photos are transformed into a tetrahedral colour space to simulate how birds' receptor see colour (Stoddard and Prum 2011). Building a dataset of bird plumage colours that covers a wide range of avian species can be useful for analysing questions related to bird plumage colour at a large scale. Measuring colours on digital photos instead of the actual specimens has advantages of non-invasive manipulations, easy-to-reuse. Another important aspect of using digital photos is that people can apply computational methods such as computer vision and deep learning algorithms to increase the efficiency of colour measuring on a huge amount of photos.

To study avian plumage colours of global bird species, an online citizen science project Project Plumage (www.projectplumage.org) was created to measure plumage colours on digitised bird specimens. The images were taken in the bird collections at the Natural History Museum, Tring. All images follow a standardised design (see chapter 2 and section 2.2 for detail). Each image includes one specimen and a set of five Labsphere Spectralon Diffuse reflectance standards that allow images to be standardised. Specimens were imaged in both the human-visible and ultraviolet (UV) light spectra. There are a total of 122,610 images, and 121,547 images (1,063 images were excluded due to problems associated with extracting correctly calibrated colour values) were used in colour analysis which covers 8,509 species, 178 families, and 34 avian orders (more than 85% of bird orders). Citizens have to place 7 to 10 points per images (the number of points is dependent on the view of the specimen) and segment the whole bird's plumage region using polygons. One image needs to be labelled by three persons in order to reduce variance and error from annotators. Based on manual processing (time for users that are familiar with the labelling, placing points takes 1 to 3 minutes and segmenting takes 3 to 5 minutes per image) and the participation of citizens (number of images labelled per day), I estimate that it would take more than one year to label all images.

1.4.1.2 *Littorina* shell shape

Shell morphological traits have been used to study parallel evolution (Hollander and Butlin 2010; Butlin et al. 2014; Ravinet et al. 2016). *Littorina* shell images were provided courtesy of Prof. Roger Butlin (Department of Animal and Plant Sciences, University of Sheffield). The data includes images of *Littorina* specimens collected from around Europe with landmarks placed on shell images to measure shell shape. 15 landmarks are placed on the shell based on the study of Ravinet et al. (2016). The aim of placing landmark points on shells is different from placing points on Project Plumage images. Placing morphological landmarks often requires higher accuracy than using points to identify regions, as a landmark normally captures the accurate homologous location (e.g. the apex of the *Littorina* shell).

1.5 Thesis outline

In this thesis, I tested and applied deep learning networks on images from Project Plumage in the first two data chapters. Chapter 2 focuses on predicting points (pose estimation) and chapter 3 focuses on predicting regions (semantic segmentation). Deep learning networks need training sets (i.e. manually labelled images). Here, I used annotations labelled by people with expert knowledge rather than citizen-labelled annotations to ensure high-quality annotations. In chapter 4, I introduced a new software tool, PhenoLearn, that provides biologists with a pipeline from labelling training images to predicting the whole dataset of digitised specimen images.

1.5.1 Chapter 2. Point measurements on bird plumage colours

Bird plumage colour has functions from crypsis or camouflage to mate choice/attraction and social signalling (Hill et al. 2006a). Many studies have measured bird plumage on specific body regions to quantify their colours. Body region measurements (e.g. point or patch) are commonly used in extracting colours of specific body regions. Plumage colours of body regions are commonly measured using spectrometers (Stoddard and Prum 2008, 2011; Dunn et al. 2015) or (more recently) by placing points or polygons (Cooney et al. 2019; Miller et al. 2019) on digital photos with special reflectance calibrations (Troscianko and Stevens 2015). For images in Project Plumage, keypoints need to be placed for (i) measuring plumage colours in specific locations on the body and (ii) calibrating lighting variations among images.

Here, I trained two pose estimation networks, Stacked Hourglass (Newell et al. 2016) and CPM (Wei et al. 2016), on a subset of 5,094 expert-labelled images. Additionally, I trained models with different input configurations (e.g. input image resolution, training length). For the best model, predictions across 95% of the validation images were reviewed as correct by humans. I used the best model from evaluations (both geometric and colour accuracy) to predict colour for the whole Project Plumage dataset. Finally, I used point predictions to create bird plumage colour space across more than 7,000 bird species and measure colour volumes using convex hull and alpha shape (Edelsbrunner and Mücke 1994; Gruson 2020) for all and individual body regions. I also compared colour volumes between males and females to test whether males are more colour diverse than females as shown in previous studies (Cooney et al. 2019).

1.5.2 Chapter 3. Segmenting bird plumage areas

In Chapter 2 I used point measurements to measure colour on body regions. However, the information in the whole plumage area which point measurements fail to capture might be important. Segmentation is commonly used in biomedical images for segmenting focal regions such as cells, organs, bones and fossils (Aljabar et al. 2009; Baiker et al. 2010; Meijering 2012; Davies et al. 2017). It has also been used in segmenting digitised natural history datasets (Kumar et al. 2015; Unger et al. 2016). Here, I used DeepLabv3+ (Chen et al. 2018), one of the most accurate semantic segmentation networks, to learn expert-labelled images and predict the whole Project Plumage Dataset. I then compared segmentation results of DeepLabv3+ to results from four classic segmentation methods (thresholding, region growing, Chan-Vese and graph cut). I also tested the model performance on low-quality datasets and small size training sets. Thus, I can evaluate how resilient deep learning is to these two factors and provide guidance for biologists before starting projects using semantic segmentation. Over 95% of the plumage areas were correctly predicted in the best model. I then created the overall bird plumage colour space using predicted segmentations of the whole project plumaged dataset. Convex hull volume is a common metric used to measure colour diversity (as well as the size of other trait spaces) (Stoddard and Prum 2011; Tuset et al. 2014; Renoult et al. 2017). However, it may not be suitable for quantifying colour diversities measured using segmentation, due to the effect of outliers (caused by an increasing number of colour measures from segmentation) and because it does

not account for variation in the proportions of different colour hues. I designed a metric, named proportional colour diversity, which measures colour diversity while accounting for colour proportions. Finally, I visualised phylogenies of plumage colour diversities across more than 7,500 species and calculated phylogenetic signals.

1.5.3 Chapter 4. Building a tool for the deep learning pipeline

Results from the first two data chapters were very promising, as points and segmentations were predicted reliably accurately. Therefore I further explored a pipeline for using deep learning in phenotyping digital images. Image annotation (for measuring phenotypic traits) can be done on many tools such as ImageJ (Schindelin et al. 2012). Similarly, there are many software packages and tools for biologists to run phenotypic analysis (Klingenberg 2011; Adams and Otárola-Castillo 2013; Maia et al. 2013). However, deep learning normally requires coding using deep learning libraries (Abadi et al. 2016) and there are few tools that are designed for non-experts to utilise the power of deep learning on digitised images in an easy-to-use interface.

To fill this gap, I developed PhenoLearn, an open-source image analysis tool that generates annotations for digitised collections using deep learning. PhenoLearn has functions including (i) labelling images, (ii) training deep learning networks, (iii) evaluating networks, (iv) predicting annotations for the rest of the images with trained networks and (v) reviewing predictions. These functions can produce accurate measurements on digitised datasets especially large-scale datasets, including visualising all the manipulations with user interfaces and minimum requirements for deep learning knowledge on the part of the users. I then showed an example application on a digitised *Littorina* shell dataset, predicting landmark points on these images. Morphospaces from both predicted and manually labelled landmarks were built, and I evaluated morphospaces to see if deep learning landmarks can detect shell morphological differences between two ecotypes (Butlin et al. 2014; Ravinet et al. 2016).

Chapter 2 A Deep Learning application on colour measurements of plumage regions on standardised avian specimen images

Abstract

A growing number of biological specimens are being digitised, particularly from museum and herbarium collections. A key challenge in mobilising these data for scientific research is measuring the digitised data automatically where manual measurements become hugely time-consuming for large data sets. Deep learning is becoming the state-of-the-art for many computer vision tasks along with improvements in computational power. It is therefore useful to explore how deep learning can perform on measuring digitised specimen data. I used more than 120,000 digital photographs of bird specimens that cover most of the world's extant species to identify plumage colour. Bird plumage colours can be quantified by using pixel values around particular body region keypoints. I used pose estimation methods (Stacked Hourglass and Convolutional Pose Machine) from deep learning to predict these keypoints automatically. My results show that the deep learning model can produce accurate results, with 95% of predicted keypoints in the correct areas. It took approximately three days to predict the whole plumage dataset, compared to hundreds of days required for expert labelling. I applied the method to measure avian colour volume using two methods (convex hull and alpha shape). The overall avian colour volume calculated from the whole dataset shows that bird plumage colours have only evolved to occupy about a quarter to one-third of the total possible colour space. The results show that it is possible to use deep learning to implement accurate, automatic and high-throughput keypoint localisation on plumage photos. I also provide guidelines for building a workflow of digitising and measuring large-scale biological data.

2.1 Introduction

Measurement of phenotypic traits from large collections of digitised specimens is an increasingly important step in studies of biodiversity that address evolutionary and ecological questions. While generating genomic data has become progressively more cost-effective allowing the compilation of complete genome databases for numerous organisms (Roach et al. 2010; Jarvis et al. 2014; Zhang et al. 2014), phenotypic databases are less well developed in part due to low efficiency in collecting phenotypic information (Lussier and Liu 2007). A significant challenge is to develop high-throughput phenotyping pipelines for such large-scale data. Globally, natural history museums house extensive, but often underexploited, collections of biological specimens. Measurement is often limited by access to, and quality of, specimens, yet estimated 1.2 to 1.9 billion specimens in museum collections globally (Ariño 2010). Collection digitisation is a major goal for many natural history museums and can provide a rich source of phenotypic and biodiversity data (Blagoderov et al. 2012; van den Oever and Gofferjé 2012).

Digitisation (including 2D photos, videos and 3D scans) and collecting associated metadata is a straightforward way to store permanent records of specimens as digital data (Stevens et al. 2007; Kuzminsky and Gardiner 2012; Mantle et al. 2012; Goswami 2015; Hudson et al. 2015). However, most of the raw digitised data (e.g. images, scans) cannot be used in ecological and evolutionary analyses without extensive processing. To mobilise natural history data, robust, high-throughput data extraction (e.g. phenotyping or trait measurements) pipelines are necessary. Images may contain diverse information including colour, shape, and posture. Imaged specimens can be labelled in numerous ways to extract information of interest. For example, specific labelled points of the organism can be landmarked and then used in geometric morphometrics (Bookstein 1991; Conde-Padín et al. 2009; Zelditch et al. 2015; Chang and Alfaro 2016; Ravinet et al. 2016) or position tracking (Mathis et al. 2018a), segmentations can be used to extract morphological traits (Meyer and Beucher 1990; Gehan et al. 2017), polygons can be used to measure colours on specimens (Cooney et al. 2019). A problem common to any form of labelling is that it is often slow and labour intensive. Manual labelling and measuring by experts is a common but time-

consuming way of phenotyping from digitised specimens especially datasets that are becoming increasingly large.

Seeking ways that can increase the labelling speed while maintaining accuracy is, therefore, a growing challenge for biodiversity science. Crowdsourcing provides an alternative to expert-only labelling. Crowdsourcing can improve the labelling rate by opening up digitised data to a large, relatively open and often rapidly-evolving group of internet users. For example, morphometric landmarks can be collected on 2D and 3D data through crowdsourcing websites allowing downstream analyses of large-scale evolutionary trends (Chang and Alfaro 2016; Cooney et al. 2017). However, the advantages of faster accumulation of data are traded off against fluctuanting engagement of citizens and, for some tasks, potentially lower accuracy of crowdsourcing compared to experts' labelling (Kamar et al. 2012). A potentially more tractable or complementary solution to large scale labelling is the use of advanced computational techniques.

Computer vision algorithms provide powerful tools in which the main goal is to enable computers to understand images and videos as close to the way that humans do as possible. Deep learning, a subfield of machine learning, uses neural networks with tens or hundreds of layers and has led to vast improvements in many computer vision problems such as image classification, object detection, and pose estimation (Krizhevsky et al. 2012; He et al. 2016; Newell et al. 2016; Redmon et al. 2016; Wei et al. 2016; Chen et al. 2017b). Convolutional neural networks (CNN) are an important component in deep learning and were proposed for handwritten digit recognition (LeCun et al. 1998) and are now widely used in image and video problems. CNNs overcome input complexity from images or videos and the use of convolutional layers provides learning parameters that can be shared across inputs, increasing the network performance by reducing overfitting and computational costs (reducing the total number of parameters). After many convolutional layers, different levels of image features are automatically extracted from the training set. The extracted features are used in machine learning models for different vision tasks.

The use of deep learning in biological data processing and analysis has been growing rapidly. For example, leaves of 44 plant species were identified with over 90% accuracy using CNN (Lee et al.

2015). This work also has shown that features extracted by deep learning could achieve more accurate prediction than hand-engineered image features. Image classification algorithms have been used to identify species from digital data including identifying animal species and extracting numbers of animals from camera-trap images (one species per image) and from underwater photos (Kellenberger et al. 2017; Rathi et al. 2017; Norouzzadeh et al. 2018; Schneider et al. 2018). Classification accuracies from these studies were higher than 90%. These results have shown that deep learning can provide fast, automated and accurate solutions in a range of biological applications. However, extraction of phenotypic data from digitised collections often requires more complex tasks than recognition. For example, identifying and placing landmarks on biological digital data can be used to capture morphometric data, behaviour and other biological information. Although few studies have been conducted, deep learning approaches, and landmark detection in particular, are particularly promising for keypoint localisation from natural history collections. For example, recent studies used deep learning to classify and localise different features (leaf tips, bases, ear tips and ear bases) of wheat from photos (Pound et al. 2017). Similarly, pose estimation methods have been used to detect landmarks and track behaviours of both *Drosophila* and mice from videos (Mathis et al. 2018a; Pereira et al. 2019).

Here, I focused on applying deep learning to extract colour information from images of birds. Birds have evolved into a wide diversity of colour space (Stoddard and Prum 2011). Birds also see colour differently from humans and can perceive light across a wider spectrum. This is because birds have four receptor cones, compared to three in humans, and are sensitive to ultraviolet (UV) light (Goldsmith 1990; Cuthill et al. 2000). The total extent of perceivable colour expressed in bird plumage has been referred to as the avian plumage colour gamut (Stoddard and Prum, 2011). The breadth and diversity of the colour gamut are expected to reflect the many functions that plumage colour serves from crypsis or camouflage to mate choice/attraction and social signalling (Hill et al. 2006a). Measuring and understanding the avian colour gamut is therefore important in the study of a wide range of questions linking plumage evolution to natural and sexual selection (e.g. Dale et al. 2015; Dunn et al. 2015; Gomes et al. 2016; Cooney et al. 2019).

Measuring colour from all of the birds of the world (10,000 bird species) is only possible by using natural history collections. I utilise part of an extensive set of photos of bird specimens

representing >85% of all of the world's bird species taken under controlled lighting conditions at the Natural History Museum, Tring. Body region measurements (e.g. point or patch) are commonly used in extracting colours of specific body regions. Colours can be measured directly from specimens using spectrometers (Stoddard and Prum 2008, 2011; Dunn et al. 2015) or placing points or polygons (Cooney et al. 2019; Miller et al. 2019) on digital photos with special reflectance calibrations (Troscianko and Stevens 2015). The primary goal is therefore to find an automated landmark detection algorithm that places point labels with sufficient accuracy to measure colours that are comparable with expert identification of body regions.

Pose estimation uses deep neural networks to identify human body parts and joints as points on images. Methods such as Stacked Hourglass (Newell et al. 2016), DeepCut (Pishchulin et al. 2016) and Convolutional Pose Machine (Wei et al. 2016) have predicted accurately on pose estimation datasets such as MPII Human pose dataset (seven points for seven body parts per human. Andriluka et al. 2014) and Leeds Sports Pose (LSP, 14 points for 14 joint locations per one human. Johnson and Everingham 2011). For the MPII dataset, 90.9% of the predicted points from Stacked Hourglass were located within half of the head lengths (individual heads are used rather than the average head) from their ground truth. The result is considered very accurate. Here, I explore the use of pose estimation to measure colour from specific body regions. Previous studies have used deep learning algorithms to detect landmarks to track animals' positions or postures in videos (Mathis et al. 2018a; Pereira et al. 2019). These studies were mainly applied to a small number of animals on many frames of videos. Here, I used fixed images where the subjects have variant looks but similar postures, and tested whether pose estimation can place multiple accurate points on digitised bird specimen photos.

If colours measured by point predictions are accurate then the colour information dataset can be used to study evolutionary and ecological questions about bird plumage, especially plumages among body regions. I calculated the size of the avian colour gamut as a use case for the dataset. The size of the avian colour gamut can be measured in numerous ways but typically requires consideration of the perception of the receiver. Mapping colour into a tetrahedral colour space using avian visual models can be used to describe how birds see plumage colour (e.g. Stoddard and Prum 2011). The axes of the tetrahedral colour space are defined in relation to the four

receptor cone types (ultraviolet (u), shortwave (s), mediumwave (m) and longwave (l)) in the avian visual system. Despite the theoretical extent of colour that can be expressed within the avian tetrahedral colour space, the evidence suggests that only a limited range of colours are expressed in bird plumage. For example, Stoddard and Prum (2011) estimated a colour space of bird plumage colour containing 111 species from 55 families in 18 avian orders (Stoddard and Prum 2011). This avian colour space filled only 26-30% of the total available avian colour space, suggesting that there are limitations on birds ability (i.e. colour producing mechanisms) to produce certain colours or colours that birds avoid to evolve, despite retaining the ability to perceive them.

Stoddard and Prum (2008, 2011) used the convex hull volume to measure colour space volume. Convex hull volume is widely used as a colour volume metric (Stoddard and Prum 2008, 2011; Cooney et al. 2019). However, it may overestimate the volume due to its convex property. For example, a set of points that has a shape similar to a star and its best fit shape should be a star-like shape. But its convex hull is a pentagon-like shape, which has larger areas than the star-like shape. Despite the fact that birds are the most colour-diverse terrestrial vertebrate, this implies that the avian colour gamut may be smaller than previous estimates suggest. Alpha shape is an alternative method of estimating the volume of a colour space, and can produce non-convex shapes (Edelsbrunner et al. 1983; Edelsbrunner and Mücke 1994). A brief definition of the alpha shape is written in the appendix along with examples that are shown in Supplementary Figure 6.1.1. It, therefore, is a potentially more precise way to measure volume and that reduces the extent of overestimation associated with the convex hull volume (Cholewo and Love 1999; Gruson 2020).

For this paper, I aimed to find deep learning networks from the pose estimation that can predict points that are accurate enough to extract colour information on plumage images at a level comparable to expert labelling. The neural network is a black-box model, so it is necessary to test different network configurations and hyperparameters (i.e. configurations for the training process, such as training steps and learning-rate) to find the best configuration. I trained and validated a range of different pose estimation networks and training configurations. I then evaluated the results with respect to (i) accuracy of point placements and (ii) accuracy of colour

measurement. Additional experiments and evaluations were used to further assess the robustness and the network performance (e.g. test how colour diverse images affect the model performance). I used the best-performed model to extract colour measurements from more than 7000 species. I then constructed a tetrahedral colour space based on avian visual models. The colour space was used to generate a new estimation of the avian colour gamut and assess how the gamut varies among species and body patches. Finally, I discussed the potential for deep learning approaches to replace manual labelling in similar phenotypic datasets.

2.2 Data and Methods

2.2.1 Imaging and raw data

The images and labels used in this study were collected as part of a broader study of bird diversity and form part of the online citizen science project Project Plumage (www.projectplumage.org). The images were taken in the bird collections at the Natural History Museum, Tring. All images follow a standardised design as described by Cooney et al. (2019). I repeat the main protocols here for convenience. Each image includes one specimen and a set of five Labsphere Spectralon Diffuse reflectance standards (2%, 40%, 60%, 80% and 99% reflectance arranged left to right in each image (referred as Standard 1-5) photographed against a black background under controlled lighting conditions (two Bronocolor Pulso G 1600 J lamps with UV filters removed and powered by a Broncolor Scoro 1600S Power Pack). Specimens are placed with heads on the left and tails on the right in images where possible. Due to variation in size and shape of different species (e.g. exceptionally long neck or legs) some museum specimens are arranged in non-standard ways (e.g. fold necks to fit specimens in the camera). Photos were taken from three views (back, belly and side) for each specimen and each view is imaged twice, once in the human-visible and once in the ultraviolet (UV) light spectra. All photos were taken using a Nikon D7000 DSLR camera and a Nikon 105mm f/4.5 UV Nikkor lens. The camera was modified (by Advanced Camera Service, Norfolk; http://advancedcameraservices.co.uk/) to allow both human visible and ultraviolet (UV) wavelengths of light to be recorded. Pairs of images were taken in the human-visible or UV spectrum by using either a Baader UV/IR Cut filter / L filter (transmits light in the human visible range 400-680nm) or a Baader U-Venus-Filter (transmits light in the UV range 320–380 nm). The

same camera settings were used for all photographs (1/250 sec, f/16.0, 'Daylight' white balance, RAW photo format), with the exception that ISO was for human visible (ISO 100) UV images (ISO 1000) differed to achieve correct exposure. The human-visible image and UV image of the same specimen with the same view are identical (e.g. bird settings, backgrounds) except the pixel values. So when referring to image numbers, human-visible and UV images of the same specimen with the same view were counted as one image (i.e. a specimen was taken from three views for both spectra, it is counted as three images). Images were saved in RAW format at a resolution of 4,948 x 3,280 pixels. The full Project Plumage data set consists of 122,610 images but here I used a subset of 5,094 images to test the performance of pose estimation methods and demonstrate the utility of machine learning in data extraction. Then I applied the best-performed pose estimation method to the whole Project Plumage dataset and built a bird plumage colour space using keypoint predictions.

2.2.2 Image labelling (keypoints for pose estimation)

Pose estimation requires the definition of keypoints. In human pose estimation applications, these are typically joints (e.g. shoulder, elbow). 15 labels (keypoints) were used to capture colour information from different bird body regions. Each of the three views outlined above includes labels placed on specific body regions (Figure 2.1). Photos of all three views have five labels at the centre of each of the reflectance standards. Body region labels were placed in the centre of the corresponding body region. Bird body regions may be occluded in some images, and these regions were not labelled by experts. The image labels are used to localise each body region and, in downstream analyses, can be used to measure colours of this region (e.g. using points to create patch measurements like Cooney et al's 2019 work). A total of 5,094 photos representing three views of 1698 bird species were labelled manually by one expert (YH). After accounting specific body regions for views and occluded body regions, the sample sizes across keypoints were: N(Standard 1-5)=5,094, N(Throat, Breast, Belly, Flight feathers)=1,698, N(Mantle)=1,697, N(Coverts)=1,696, N(Crown, Nape)=1,695, N(Tail)=1,678 and N(Rump)=1,422. The sample of 1,698 bird species encompass representatives of 81% bird genera and 27 bird orders, so the labelled images capture much of the extent of variation in plumage colour, patterns, and body

shape among birds. These expert-labelled images were used to train and validate the deep learning models.

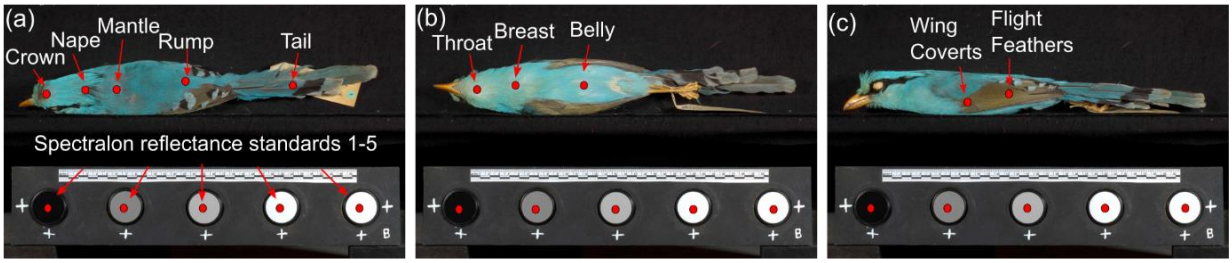


Figure 2.1. Examples of keypoints on the Project Plumage images, which capture (a) five reflectance standards and five body regions (crown, nape, mantle, rump and tail) of the back view; (b) Three body regions (throat, breast and belly) of the belly view; (c) Two body regions (wing coverts and flight feathers) of the side view.

2.2.3 Deep Learning workflow overview

After expert keypoint labelling, the core workflow involves four steps: a) image preparation, b) model training, c) model predictions on the trained network, and d) evaluating model performance. This workflow is summarised in Figure 2.2. Image preparation (pre-processing) includes resizing images or labels so that they can be fed into the network (Figure 2.2a). I here used 5 fold cross-validation and split data into training and validation sets with an 80:20 ratio. Cross-validation can provide an accurate estimate of model performance by averaging performance for different partitions (5 partitions for 5 fold cross-validation) of training and mutually exclusive validation sets. A common approach in a deep learning pipeline is to split data into the training set, validation set, and test set where the test set is used to provide the final benchmark (e.g. ImageNet; Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016). I used only the training and validation sets so that every image from the labelled dataset (covering a wide range of extant bird species) can have a predicted keypoint from the same data partition routine. This allows the relationship between bird taxonomy and network performance to be evaluated (i.e. to assess whether performance varies among groups of bird species due to broad differences in size, shape and colouration of specimens).

The model is trained with a training set under pre-defined network hyperparameters (i.e. configurations for the training process, such as training steps and learning-rate). For each training step, the network generates predictions with input images, from which a loss function that represents differences between ground truth (expert labels) and predictions can be calculated. Gradient descent is then applied to optimise the network parameters and decrease the value of the loss function using its gradient (Ruder 2016). The goal is for the network to generate predictions that are iteratively closer to the ground truth in the next step (Figure 2.2b). When the network converges or training finishes, the ground truth data from the validation set is used to evaluate the precision and accuracy of the validation set result from the trained network (Figure 2.2c).

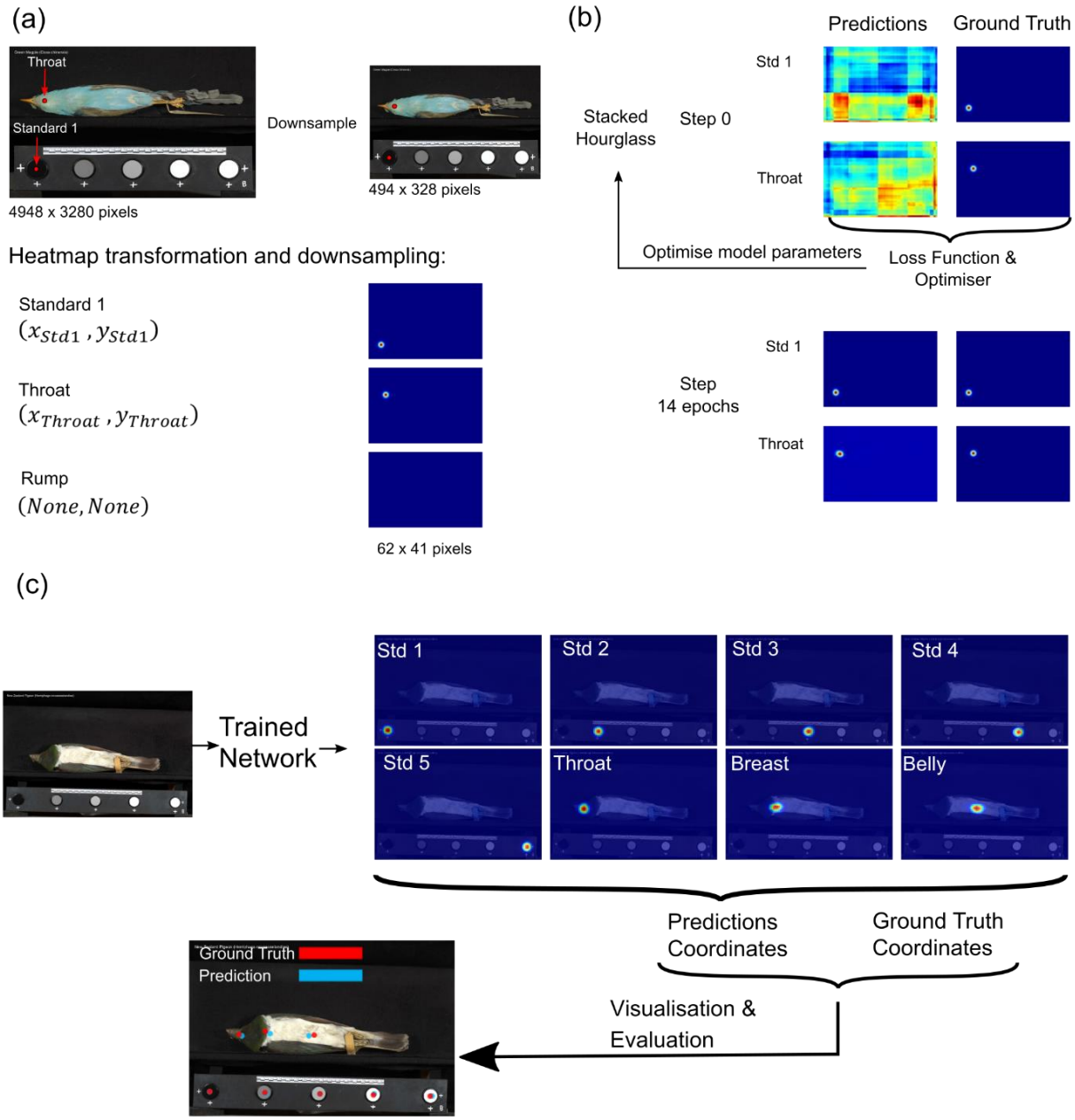


Figure 2.2. Workflows for applying the Stacked Hourglass (Newell et al. 2016) to predict keypoints on an example image from Project Plumage. (a) The preparation (pre-processing) step resizes images and transfers coordinates into heatmaps. (b) The pre-processed training data is used to train the network. (c) The trained network is used to generate predictions of validation images. Then post-processing (e.g. transfer heatmaps back to coordinates) and evaluations are applied on predictions.

In this section, I describe the workflow using the Stacked Hourglass neural network architecture (one of the most accurate models for identifying single human posture; Newell, Yang, and Deng 2016) with a fixed set of hyperparameters and input images. In section 2.2.4 I describe experimental manipulations to compare with an alternative neural network architecture (the convolutional pose machine, CPM - an earlier deep learning model which influenced the Stacked Hourglass model; Wei, 2016). I further assess the effects of manipulating the model hyperparameters, and the input image (e.g. image resolution and different ways of pre-processing images).

2.2.3.1 Image preparation

Here, I used 5,094 expert-labelled images (1698 images per view) with 15 labels. Images were first converted to JPG from RAW format, which can be easily read and saved by numerous software tools (e.g. Python, R and MATLAB). Only the human-visible visible spectrum versions were used in training, therefore a pixel has values from R, G and B channels, ranging from 0 to 255. I split the images into training (80% of images) and validation sets (20% of images). Due to the memory of the graphics processing unit (here an NVIDIA GTX 1080Ti with 12GB GPU memory) and the model complexity, it was necessary to reduce the input resolution. While the original Stacked Hourglass paper uses 256 x 256 pixels as the input resolution (Newell et al. 2016) the resolution of the raw project plumage images is 4,948 x 3,280 pixels. I down-sampled images 10-fold to 494 x 328 pixels using bilinear interpolation from OpenCV (a computer vision library; Bradski 2000). I have tested that this resolution can be trained on the GPU without any difficulties while keeping as much information from the original images as possible.

2.2.3.2 Model training and application

Stacked hourglass and CPM outputs heatmaps instead of coordinates. Outputting heatmaps has been shown to outperform other output formats in pose estimation in many studies (Newell et al. 2016; Pishchulin et al. 2016; Wei et al. 2016). The CNN reduces the input resolution to extract image features due to convolution and pooling layers (LeCun et al. 1990; Krizhevsky et al. 2012), heatmaps generated from Stacked Hourglass (using 494 x 328 pixels as the input resolution) have a resolution of 62 x 41 pixels which is 8 times smaller than the input resolution. Ground truth

heatmaps are generated as 2-dimensional Gaussian distributions around each keypoint. I applied a Gaussian peak with a standard deviation of 2 to the location of scaled coordinates to create each of the 15 ground truth heatmaps (Figure 2.2a). Where a body region did not appear in an image (i.e. it was completely occluded or did not belong in the view), an empty heatmap was used (Figure 2.2a). After pre-processing, the original data and labels were turned into input images with a resolution of 494 x 328 pixels and ground truth heatmaps with a resolution of 61 x 41 pixels, which were ready for training.

The data set was divided into batches of four images, and one batch per training step was fed into the model to generate prediction heatmaps. Using batches balances the memory usage of the GPU and the optimisation of each step (Hinton et al. 2012). In this method, a loss function for the network, which shows the difference between predictions and ground truth, was calculated as the mean squared error between prediction heatmaps and ground truth heatmaps (the heatmap dimension: 62 x 41 x 15). To minimise the loss function, model parameters were updated using the ADAM optimiser (Kingma and Ba 2014) along with the gradient of the loss function, as shown in Figure 2.2b. 0.01 was used as the initial learning rate. Through the training process, the learning rate was cosine decayed and restarted at the initial value after reaching zero, which increases the possibility to reach a better local optimum (Loshchilov and Hutter 2016). The length of the first period of decay-restart was set to one epoch (an epoch is defined as one pass of the full training set for the network). After each period, the new period was two times longer than the previous one (i.e. the second period takes two epochs to decay to zero, the third period takes four epochs and so on). 15 epochs were chosen to have the model trained for four complete decay-restart periods, which the model converged (i.e. the loss has stopped decreasing).

After the training finished, validation images were fed into the trained network to generate prediction heatmaps. Heatmaps were resized back to the same resolution (4948 x 3280 pixels) as the original images. The coordinate of one prediction was generated using the location of the maximum value (or the average of locations if there are multiple maximum values) of the corresponding heatmap (Figure 2.2c). The network generated 15 keypoints for each image without checking and excluding possible occluded points. Keypoints that were not in the view

were discarded (e.g. seven keypoints were retained for a side view image, see Figure 2.1c). To make predictions and ground truth comparable, occluded points in the expert labels were not evaluated. The model training and predicting were implemented using Python 3 and Tensorflow 1.12 (Abadi et al. 2016), a deep learning library, on one NVIDIA GTX 1080Ti GPU (12GB GPU memory).

2.2.3.3 Evaluation

I evaluated the model performance at the original image resolution (4,948 x 3,280 pixels) and used metrics that describe the accuracy of predictions of the validation set from both geometric and colour perspectives. The pixel distance is the most straightforward metric to use for comparing geometric accuracy, and is simply the Euclidean distance (measured in pixels) from the input ground truth keypoint coordinates to the predicted coordinates (the pixel location with the maximum heatmap value). Pixel distance can be used to assess each keypoint individually and as the average for all keypoints within an image.

An alternative geometric measure is the Percentage of Correct Keypoints (PCK) which is a commonly used metric in pose estimation (Newell et al. 2016; Wei et al. 2016). PCK is the percentage of predictions that have pixel distances below a given threshold (Andriluka et al. 2014). I used PCK with a threshold of 100 pixels (PCK-100). PCK-100 is suitable for evaluating the accuracy of the five reflectance standard predictions because the minimum radius of reflectance standard circles is slightly greater than 100 pixels and ground truth points of standards were always placed in the centre of standards. The areas of body regions were variable (e.g. the crown is usually smaller than the breast or belly) and in some cases, the body region could be smaller than 100 pixels across at its minimum, however, PCK-100 provides a starting point for standardising the evaluation across all 15 keypoints.

For some use cases point accuracy may be critical (e.g. if the goal is to place landmark points for geometric morphometric analyses of shape). However, the ultimate goal of labelling points on the Project Plumage dataset is to extract colour information for regions and to do so automatically and accurately enough to replace human labels. It is therefore necessary to evaluate similarity in the colour information between predictions and ground truth. I extracted

pixels for colour comparison by placing an area around the body region keypoint using a fixed-size bounding box measuring 20 x 20 pixels (I refer this extraction method as Bbox-20). This is small enough to account for the smallest body regions within the dataset, such as the flight feathers of hummingbirds. I used the averaged RGB values of extracted pixels (i.e. one mean RGB value for one body region) as simple metrics of colour. I further calculated the average normalised RGB and lightness, both derived from raw RGB. The normalised RGB is obtained by dividing the original RGB by the sum of RGB ($R' = \frac{R}{R+B+G}$, $G' = \frac{G}{R+B+G}$, $B' = \frac{B}{R+B+G}$). Normalised RGB provides an estimate on chromatic information (e.g. hue). Lightness is the average of the sum of maximum RGB and minimum RGB. To evaluate colour similarity between ground-truth and predicted values, Pearson's correlation coefficients (R) were calculated for all and individual extracted region across images.

Taken together, these evaluation metrics capture the precision, accuracy, and biological relevance (i.e. plumage colours) of the predicted points.

2.2.4 Experimental manipulations to increase the model performance

The workflow described above was based on a Stacked Hourglass neural network with 15 training epochs and input resolution of 494 x 328 pixels. Neural network architectures, input image resolutions, length of the training process and other factors may be critical to the model accuracy. I applied a series of manipulations to assess how they affect the accuracy of model predictions relative to the expert ground truth. All configurations were trained and cross-validated 5-fold for a robust validation result (all 5,094 images have predictions).

2.2.4.1 How do architecture & resolution affect the performance

I compared two network architectures (Stacked Hourglass and CPM) with different image resolutions. Specifically, I used resolutions that were 10, 15, and 20 times lower than the input images (i.e. 494 x 328, 329 x 218 and 247 x 164 pixels). This image resolution manipulation gives five comparisons against the benchmark configuration (Stacked Hourglass, 494 x 328 pixels) described in section 2.2.3.2.

I used this experiment to identify the best combination of the architecture and image resolution as the basis for subsequent manipulation. I found that the Stacked Hourglass method outperformed CPM in all cases and that the highest resolution input images (494 x 328) consistently gave the best results (see 2.3.1 for full details). Therefore later manipulations all used Stacked Hourglass as the network and 494 x 328 pixels as the resolution.

2.2.4.2 How does training duration affect the performance

To ensure the model converged within 15 epochs, I assessed the effects of training durations on network performance by manipulating the number of epochs for the Stacked Hourglass architecture with 494 x 328 pixels images. I compared the 15 epochs network with a longer network consisting of 31 epochs which will allow the model to train 5 periods of decay.

2.2.4.3 How does image pre-processing affect the performance

Pre-processing images can improve network performance (Krizhevsky et al. 2012; Chen et al. 2016). I manipulated different aspects of the input images to assess their effects on network performance. First, background and reflectance standards take up large proportions of each image (e.g. Figure 2.1). Increasing the area filled by the specimen may provide more specimen information for the model to learn. To assess this, I cropped specimens with bounding boxes to create specimen-only images. Cropped images were padded and scaled into a uniform resolution (1024 x 256 pixels). Second, the RGB histograms of Project Plumage images are mostly left-skewed because the image background is black with images typically having more dark pixels than bright pixels. I applied the histogram equalisation (effectively stretching the histogram) using OpenCV (Bradski 2000) to increase the contrast within each image.

Three kinds of pre-processed datasets were generated: (i) specimen-only images, (ii) applying histogram equalisation on the original images, and (iii) applying histogram equalisation on specimen-only images. Examples of pre-processed images are shown in Supplementary Figure 6.1.2. These datasets were then used in training and evaluation.

2.2.4.4 How do image augmentation and subsetting the model affect the performance

For the final manipulation, I assessed the effects of image augmentation (equivalent to the term data augmentation in some papers) and splitting the model per view. Image augmentation is a technique that increases the size of the training set by altering the existing images (Perez and Wang 2017) and it was used in the original work of Stacked Hourglass (Newell et al. 2016). In this project, the augmented training set consisted of the images from the original training set and images which were randomly rotated (-8° to -1°, 1° to 8°), translated in both x and y axes (100 to 500 pixels) and scaled (0.8 to 1.25). Image variations among views may limit the model learning, I trained one model per view separately with the augmented data rather than one all-view model.

2.2.4.5 Methods of selecting pixels for heatmaps

Extracting pixels for each body region is a key step for measuring different aspects of the birds plumage colour. Using bounding boxes can generate fast (i.e. measuring pixels in a rectangle or square is fast on computers) and uniform colour measures (i.e. colours were extracted under a fixed-size bounding box) across results from tested models. It is good for comparing across results due to its fast and uniform features. However, using the fixed-size bounding box may not be the optimum approach to measure the colour information for the final result (i.e. predictions from the best model). Body region sizes vary and the fixed-size bounding box may not adequately capture colour variation within regions. Although ground truth heatmaps have uniformed sizes (i.e. Gaussian peaks with a standard deviation of 2 on the heatmap scale), prediction heatmaps can have different sizes. Supplementary Figure 6.1.3 shows prediction heatmap examples of the crown and rump on an image. The rump (Supplementary Figure 6.1.3b) has a larger heatmap than the crown (Supplementary Figure 6.1.3a). Prediction heatmaps are distributions of probabilities for the location of the focal keypoint and provide an alternative to using fixed bounding boxes for the pixel extraction. Heatmaps were interpolated to the range of 0 to 100, where a pixel value then correspond to how likely the keypoint is located in the pixel. I used 90 as the threshold to determine which pixels should be included within each body region (I refer this extraction method as Heatmap-90), which created regions that were non-uniform in size and

shape to measure colours (number of pixels inside predicted heatmaps (based on the resolution of 4,948 x 3,280 pixels), mean: 6,484; minimum: 1,600; maximum: 30,560, which are approximately equal areas of squares with lengths of 80, 40 and 174). Colour information correlations were evaluated on the best result from the experimental models described above and were compared to the same metrics extracted by Bbox-20.

2.2.5 Post-hoc tests on the model performance

2.2.5.1 Performance based on the expert evaluation

Predictions from the best configuration selected from the metrics introduced above were manually checked by two experts (YH and CRC) and classified as correct or incorrect. For one photo, if all keypoints are placed somewhere inside and not around the border of the correct corresponding body regions, the labels of the image are correctly predicted. If at least one point is placed outside its body region, the image is considered incorrectly predicted. The boundary of a body region can be subjective, and I want to minimise the human variance and error, so predictions were cross-checked by the two people (YH and CRC). In addition, the manual checking error rate of every order was calculated to evaluate whether error rates are similar across taxonomical groups. The accuracy of the manual check result provides further verification on how well the deep learning model can predict labels on the Project Plumage data.

Training sets for many deep learning applications are generated by humans who may place keypoints differently on the same image. If differences among human labellers are similar to the error of the deep learning predictions then it is possible to say that the predicted result of deep learning is good enough. I took 300 images (100 images per view) sampled from the expert-labelled dataset to be labelled by another two people with expert knowledge (CRC and GHT). The pixel distances between the original expert points (YH) and the points placed by CRC and GHT were used to quantify human variability in image labelling. Then, pixel distances were calculated pair-wise from four results (YH, CRC, GHT and predictions). Pixel distances were categorised into three groups by datasets used for comparing (i) predictions vs trainer (i.e. between predictions and YH), (ii) predictions vs non-trainer (i.e. between predictions and CRC or GHT), (iii) between

experts. Statistical tests (ANOVA and Tukey test) were applied to quantify how close the predicted coordinates were to those from different experts using R.

2.2.5.2 Performance with low-quality data

Images from Project Plumage were taken in a highly consistent manner by controlling the placement of the specimen, light environment and background (Blagoderov et al. 2012; Hudson et al. 2015). Not all datasets are likely to be so consistent. I therefore tested whether greater variability in data quality could limit performance by generating lower quality datasets. To do this I applied a series of affine transformations to the images and their labels as well. Four datasets were created with different transformations applied: (i) rotation (angles between -45° to 45°), (ii) translation on both x and y axes (-500 to 500 pixels), (iii) horizontal flip 50% images randomly, (iv) the combination of all three transformations. 45 degrees rotation and 500 pixels translation give images large transformations while keeping all keypoints inside the image. In contrast to not manipulating the validation set in image augmentation, here transformations were applied to both training images and validation images. The transformed datasets are trained and evaluated with the model of Stacked Hourglass, the input resolution of 494 x 328 pixels and the training duration of 15 epochs.

2.2.5.3 Effects of within specimen colour variability

Specimens vary greatly in the diversity of colour. Some birds have body regions that are polychromatic whereas others are entirely monochromatic. In the context of colour data, it is important to assess the performance of the deep learning methods on specimens with different degrees of colourfulness. I quantified the colour variability of each specimen by calculating the average pairwise colour distance between body regions. Three types of colour variability were used based on colour distances of RGB (both chromatic and achromatic information), normalised RGB (hue information) and lightness (achromatic information). Colours were extracted using the ground truth label with the Bbox-20 extraction method. I then evaluated the correlations between pixel distances and three colour variability measures. The colour information metrics were split by quartiles of each of the three colour variabilities to see if colour variabilities affect colour measuring accuracies.

2.2.6 Measuring bird plumage colour volume

2.2.6.1 Building bird plumage colour space

I applied the best configuration (the model of Stacked Hourglass, the input resolution of 494 x 328 pixels, the training duration of 15 epochs and using images without any manipulations. See Result section for detail) to generate predictions for the rest of the images and combining 5,094 predictions in the validation step. Together predictions of 122,610 images were generated. Occluded point checking was not applied to predictions, as no accurate and automatic method is able to do the checking. The proportion of occluded body regions was low in the expert-labelled dataset (Rump has the highest occluded region rate which is about 16%, the rest of the body regions have rates less than 1.1%). The colour space should be similar to the one excluding occluded regions based on the occluded region rate.

The average pixel values from the deep-learning predicted heatmap regions (Heatmap-90) were used as the raw colour information for the corresponding body regions. In total 405,155 plumage colour points from 8,509 species were generated using 121,547 photos (a small number of photos were excluded from the initial dataset, contains 122,610 images, due to problems associated with extracting correctly calibrated colour values) and their keypoints. The average RGB colour value of each predicted heatmap was converted into u, s, m, l receptor cone stimulation values using the avian UVS visual model in pavo (Endler and Mielke 2005; Stoddard and Prum 2008; Maia et al. 2013). These values were then mapped into a tetrahedral colour space. Each species has up to six specimens (three males, three females), and each specimen has three views that contain 10 body region patches. I took the average colour across specimens for each species, patch, and sex. This reduced the raw data to 157,349 colour points. To ensure that male and female colour spaces are comparable, I only kept species which have both male and female data. The final colour data has 143,932 colour points (N(Male), N(Female)=71,966), which covers 7200 species, 174 families, and 34 avian orders.

2.2.6.2 Quantifying colour volumes

I used the convex hull volume and alpha shape volume to estimate colour volumes of the overall data and across individual patches for both sexes, male and female. Both volumes were calculated using pavo 2.5.0 (Maia et al. 2019; Gruson 2020).

Convex hull volume has been used in many studies for measuring the colour diversity of colour spaces (Renoult et al. 2017). In contrast to the convex hull, where the same set of data points always have the same convex hull volume, the alpha shape volume is positively correlated with its shape parameter: the α value. α* is defined as the smallest α (resulting the smallest volume) while all data points are included in the volume calculation (Cholewo and Love 1999; Gruson 2020), and can be seen as the optimal α for measuring the volume of a set of data points. Section 6.1.1 and Supplementary Figure 6.1.1 in the appendix give detail and an example of how α values affect the alpha shape and how to calculate α*.

α* varies among different sets of data points so that, for example, α* for colour measures from the crown may not be the same as α* measured from the belly. Using respective α* for every alpha shape may make volume comparisons inaccurate and inconsistent. Section 6.1.1 explains the impact of using respective α* on the volume estimation and Supplementary Figure 6.1.4 shows an example.

To solve this, I estimated a global constant α. To ensure that every alpha shape includes all data points, the smallest constant α is defined as the largest α* among all calculated patches and sexes. This α value is 0.2262 for this dataset, which is the α* for male breast (Supplementary Figure 6.1.5b). A qualified α should meet the condition that the female and male volumes should be smaller than the combined-sexes volume across patches. If there are multiple values that meet this condition, then I defined the smallest α as the optimal one. I tested three α values (0.2262, 0.5 and 1.0) along with patches respective α* values. Only the α=0.2262 satisfied the condition (see Supplementary Table 6.1.1). Alpha shape volumes have similar trends to convex hull volumes across patches (see Supplementary Figure 6.1.5a), the bigger α is, the higher volume similarity is. I therefore used the minimum possible consistent α (0.2262) to estimate alpha shape volumes.

I aimed to test whether the difference between male and female volume (convex hull and alpha shape) can explain bird colour diversity between sexes. For overall and each patch data points, they were randomly split into a pair of equal-sized groups and 1000 pairs of groups were sampled. I then compared volume differences of the null 1000 paired groups to the observed male-female differences. If the male-female difference is significantly different from random sample differences, it shows that the male-female difference is associated with sex differences in colour volume rather than arising by chance alone.

2.3 Results

2.3.1 Effects of architecture & resolution on the performance

I compared input image sizes and network architecture and I found that there were significant effects on pixel distances from architectures and resolutions (Supplementary Table 6.1.2). The Stacked Hourglass method with the highest pixel resolution had the best performance (Figure 2.3a, Supplementary Figure 6.1.6). Pixel distances of the Stacked Hourglass model were significantly better (i.e. smaller) than CPM under the same input resolution (except for rump, see Supplementary Figure 6.1.6a). The mean differences of pixel distances (of all and individual keypoints) between two networks have no values larger than 55 pixels under the resolution of 4,948 x 3280 pixels (about 1.6% of the image height, 3.280 pixels).

The input resolution was positively related to performance. An image resolution of 494 x 328 pixels has the lowest pixel distance and 247 x 164 pixels has the highest (Supplementary Figure 6.1.6b). Overall, among the six trained configurations, the CPM network with 247 x 164 pixels input images had the largest pixel distance between the ground truth and model predictions and the Stacked Hourglass with 494 x 328 has the lowest pixel difference. The best configuration inferred by pixel distance(the Stacked Hourglass with 494 x 328) can predict all standards inside reflectance standards as PCK-100 of the reflectance standards 1-5 are 100% (the blue line in Figure 2.3b) and ground truth points were always placed in standard centres and the minimum radius was larger than 100 pixels. Using colour metrics (colour extraction method: Bbox-20), the Stacked Hourglass with 494 x 328 pixels images was again the best performing method, having

the highest overall and per-region RGB correlation coefficient (Figure 2.3c). Correlations are consistently higher for hue (normalised RGB) than for lightness (Table 2.1).

Accuracies varied for different keypoints (Figure 2.3). As noted above, the reflectance standards were consistently identified accurately. In contrast, even the best model configuration was less reliable for certain body regions, most notably the tail, rump, flight feathers, and coverts. These regions were sometimes not present in the training images (rumps were frequently obscured or partially occluded, Figure 2.6c shows an example) or were small and indistinct (coverts) to the human eye (Figure 2.6h shows an example). However, correlation coefficients for colour remained above 0.92 for normalised RGB and above 0.75 for lightness when using the Stacked Hourglass with 494 x 328 even for these body regions (Table 2.1). This suggests that the deep learning algorithm provides accurate estimates of chromatic colour.

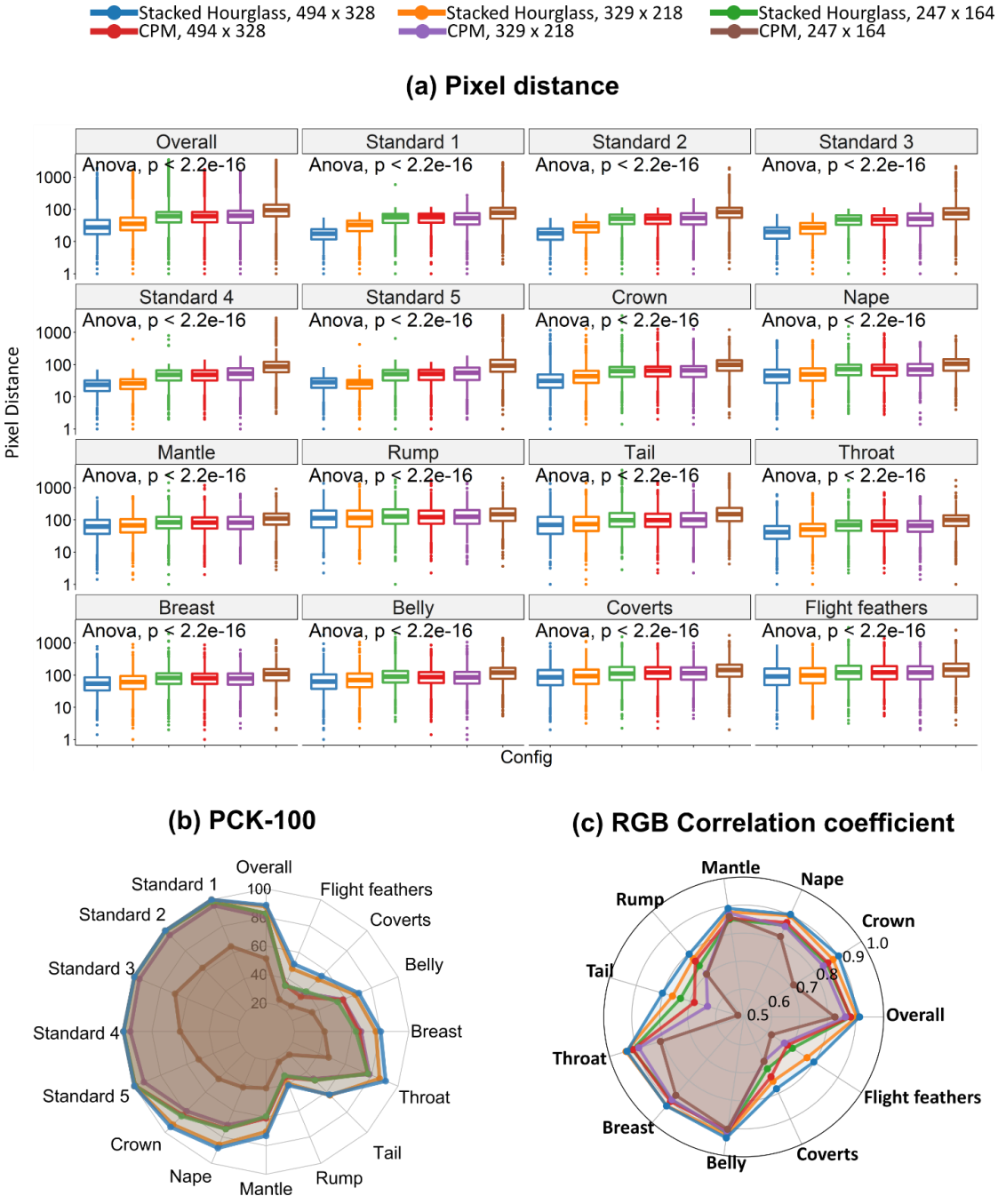


Figure 2.3.Evaluation results of models with different network architecture and input resolution. The plots show comparisons of model performance by comparing metrics from the ground truth data with the model prediction: (a) Pixel distances of all and individual keypoints. The p values of ANOVA are displayed in plots. Details of ANOVA tests can be found in Supplementary Table 6.1.2. Tukey test results can be found in Supplementary Figure 6.1.6. (b), PCK-100 of all and individual keypoints; (c) RGB Colour correlations of all and individual keypoint-defined body regions (colour extraction method: Bbox-20).

Table 2.1. Tables of evaluation results of the best configuration (Stacked hourglass with the input resolution of 494 x 328 pixels, the training duration of 15 epochs and using the unmanipulated images and labels).

	Pixel distance	**Pck-100**	**R (RGB; Bbox-20)**	**R (Norm RGB; Bbox-20)**	**R (Lightness; Bbox-20)**	**R (RGB; Heatmap-90)**	**R (Norm RGB; Heatmap-90)**	**R (Lightness; Heatmap-90)**
Overall (N=42,145)	47.3	89.3	0.914	0.949	0.910	0.941	0.965	0.938
Standard 1 (N=5,094)	18.2	100						
Standard 2 (N=5,094)	18.7	100						
Standard 3 (N=5,094)	20.4	100						
Standard 4 (N=5,094)	23.9	100						
Standard 5 (N=5,094)	28.4	100						
Crown (N=1,695)	42.7	95.0	0.903	0.957	0.899	0.944	0.976	0.94
Nape (N=1,695)	56.4	89.1	0.903	0.955	0.901	0.93	0.967	0.929
Mantle (N=1,697)	77.5	74.8	0.892	0.951	0.885	0.93	0.969	0.925
Rump (N=1,422)	147.4	44.6	0.797	0.921	0.776	0.834	0.938	0.815
Tail (N=1,678)	106.7	65.0	0.802	0.941	0.785	0.835	0.962	0.819
Throat (N=1,698)	52.1	91.2	0.935	0.956	0.932	0.957	0.974	0.955
Breast (N=1,698)	67.8	81.5	0.921	0.965	0.913	0.948	0.975	0.945
Belly (N=1,698)	85.6	72.2	0.937	0.963	0.93	0.955	0.973	0.95
Coverts (N=1,696)	111.5	57.8	0.782	0.922	0.766	0.853	0.949	0.839
Flight feathers (N=1,698)	123.0	54.2	0.798	0.945	0.773	0.857	0.961	0.839

2.3.2 Effects of training duration, image pre-processing, data augmentation and workflow subsetting on the performance

I applied image pre-processing methods, longer training epochs, and image augmentation with model subsetting based on the best network and resolution configuration tested above. Only body regions were evaluated because the reflectance standards can be predicted perfectly and because they are absent from the specimen-only images (See Supplementary Figure 6.1.2). So the overall pixel distance and PCK-100 only included body regions. There were significant effects on pixel distances of overall and nine individual regions (Except for rump) between these experimental runs (Figure 2.4a, Supplementary Table 6.1.3). No improvements were found in pixel distance, whether averaged across all keypoints, or for each keypoint individually for any of the manipulations to training, images, data, and workflow (Supplementary Figure 6.1.7a). Manipulations except for the histogram equalised specimen-only images have significantly larger pixel distances than the original result. Similarly, there was no clear improvement for using either the PCK-100 metric or the colour correlation (colour extraction method: Bbox-20) as shown in Figure 2.4b and Figure 2.4c.

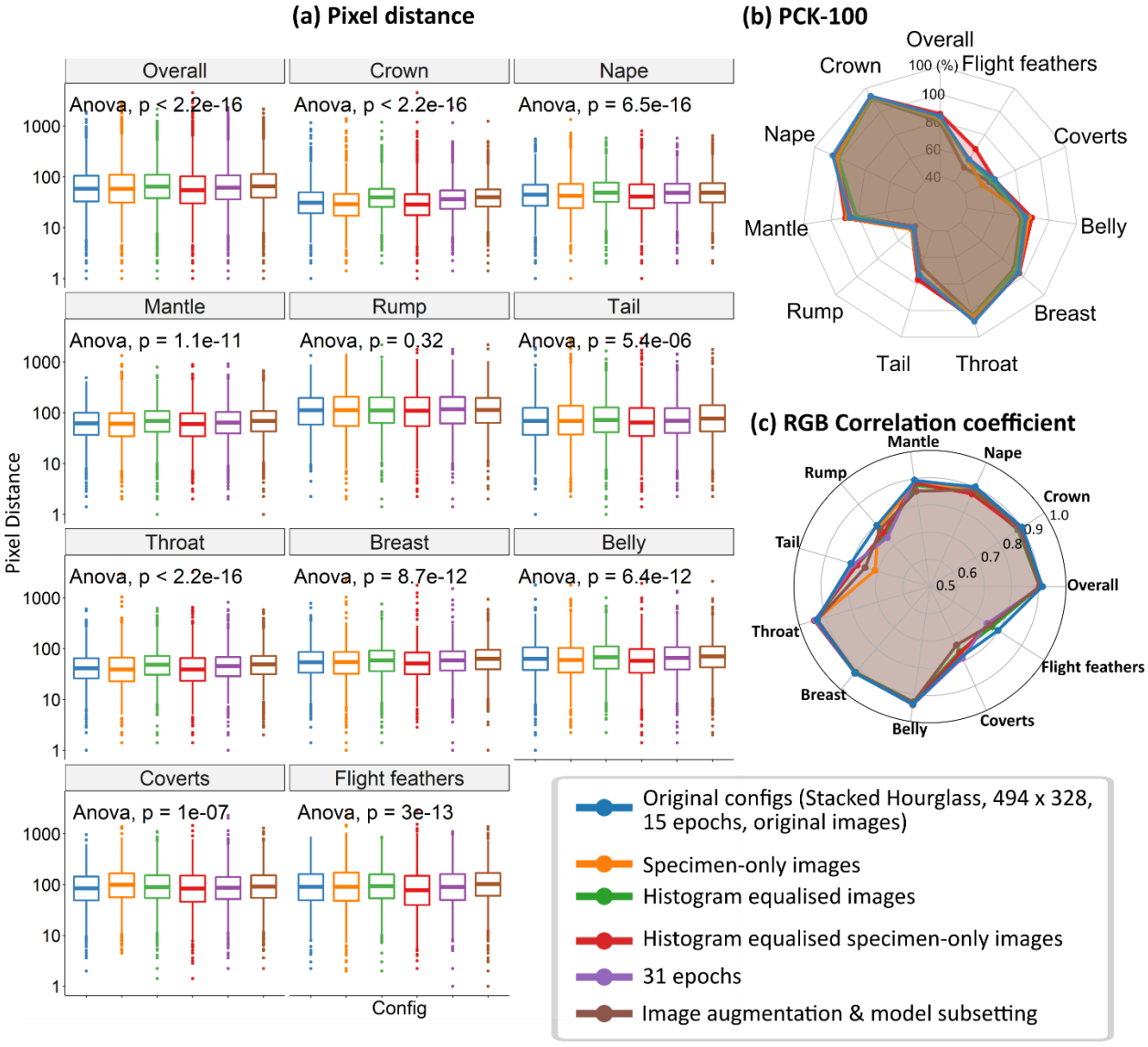


Figure 2.4. Evaluations of the best model from section 2.3.1 and applying it with image pre-processing, longer training duration, and data augmentation and workflow subsetting. The plots show comparisons of model performance by comparing metrics from the ground truth data with the model prediction: (a) Pixel distances of all and individual body region keypoints; Y-axis is the pixel distance, X-axis is configurations. Details of ANOVA tests can be found in Supplementary Table 6.1.3. Tukey test results can be found in Supplementary Figure 6.1.7. (b) PCK-100 of all and individual body region keypoints; (c) Correlations of all and individual keypoint-defined body regions (colour extraction method: Bbox-20).

2.3.3 Methods of selecting pixels for heatmaps

After testing different configurations that cover different networks, input resolution, training steps, pre-processing, image augmentation and model subsetting, Stacked Hourglass with 494 x 328 pixels resolution without any extra manipulations provided the best configuration. I therefore used predictions of all 5,094 expert-labelled images with this model in analyses and evaluations hereafter. In section 2.2.4.5, I tested using prediction heatmaps with a threshold of 90 to extract colour information. This approach had the potential to better capture the region around the keypoint based on the data and model, rather than relying on an arbitrary, fixed (in size and shape) box. Lower threshold values could be used to extract larger areas at the cost of a lower probability of being in the correct body region. The threshold results are similar to those with the Bbox-20: colours were correlated better for hue (normalised RGB) than lightness (Table 2.1). Combining all body regions, the correlation coefficients using Heatmap-90 were 0.941 for raw RGB, 0.965 for hue, and 0.938 for lightness (Figure 2.5). Considering each body region individually, RGB correlations ranged from 0.834 (rump) to 0.957 (throat), from 0.938 (rump) to 0.976 (crown) for hue, and from 0.815 (rump) to 0.955 (throat) for lightness. As with the Bbox-20 results, rump, tail, coverts and flight feathers were the four regions with the lowest correlation coefficient of all three metrics of colour information. However, the correlation coefficients are generally higher using Heatmap-90 than using Bbox-20 (Table 2.1).

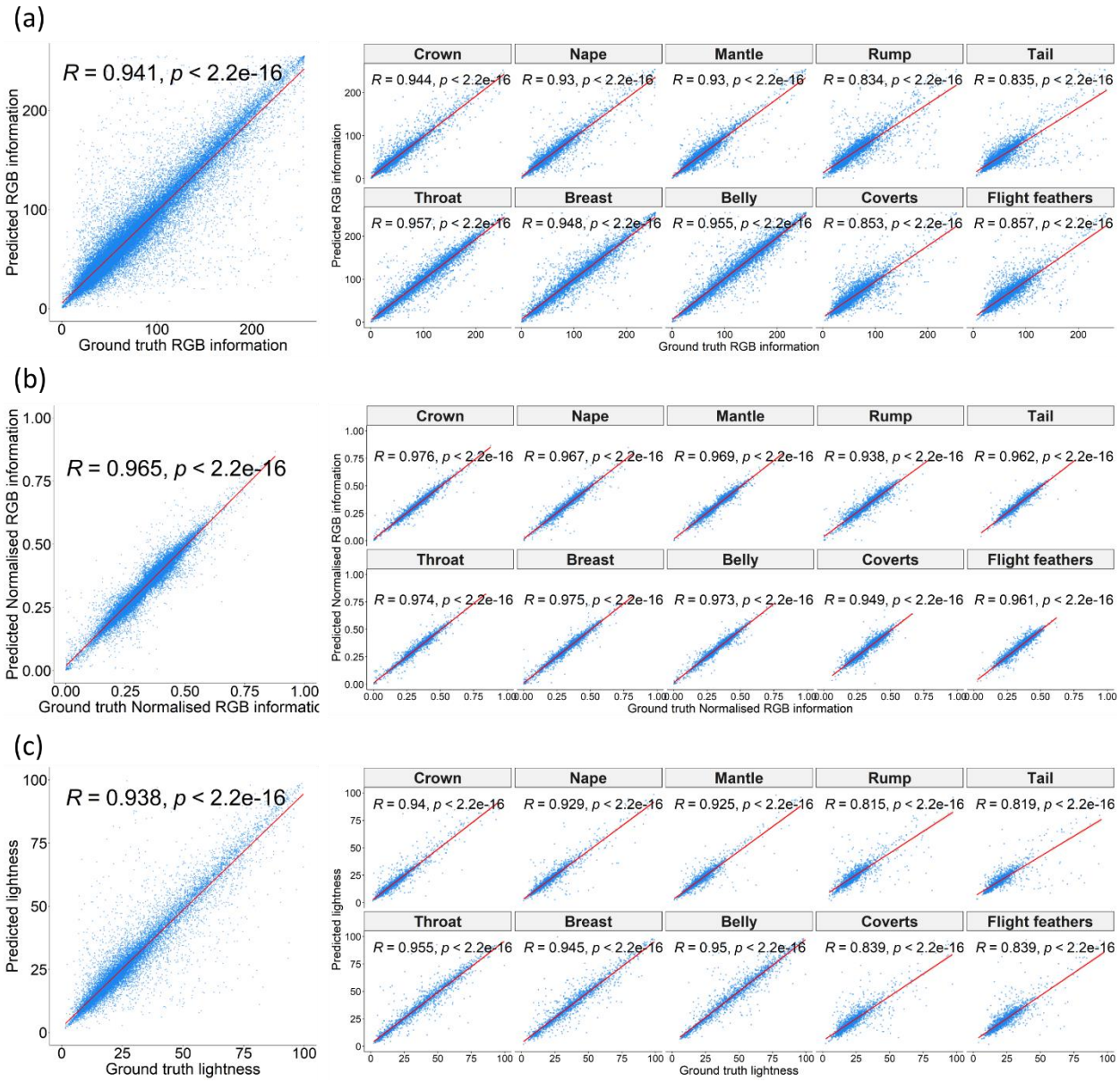


Figure 2.5. Correlations between predicted colour information (Y-axis) and ground truth colour information (X-axis) of all and per body region keypoints (Colour extraction method: Heatmap-90). (a) RGB (value ranges from 0 to 255). (b) Normalised RGB (value ranges from 0 to 1). (c) Lightness (value ranges from 0 to 100).

2.3.1 Post-hoc tests on the model performance

2.3.1.1 Performance based on the expert evaluation

In total, predictions on 234 images (back view: 97, belly view: 15 and side view: 122) out of 5,094 were manually marked as incorrect. So more than 95% of the predictions were qualified as correct predictions and can be used in future analysis. 308 regions were predicted incorrectly among the 234 images (Occurrences: 165 images had one error regions; 64 images had two error regions; 5 images had three error regions). Supplementary Figure 6.1.8a shows counts of different regions causing incorrect predictions. Flight feathers, coverts, tail, rump and crown were the top five most problematic regions, which is consistent with the higher pixel distances and lower PCK-100 and colour metrics for these body regions.

Causes of error could be broadly classified by four non-mutually exclusive features that increase difficulties for a human to place labels: (i) small area, (ii) similar colour to adjacent parts, (iii) rare posture of the specimen, (iv) partially occluded body regions. Occurrences of these features per error regions are shown in Supplementary Figure 6.1.8b. Figure 2.6 shows some correct examples and incorrect examples with their explanations in the figure legend. The dataset includes 27 orders and 1,698 genera of specimens, and the image numbers and incorrect predictions rates by bird orders are listed in Supplementary Table 6.1.4. Galliformes (11 error images out of total 15 images), Sphenisciformes (2 out of 3), Pelecaniformes (2 out of 6), Procellariiformes (1 out 3) were the top four orders in error ratio. The order Passeriformes had the most images (3405 images and about 67% of the dataset, an order of magnitude more than the second most abundant order, the Apodiformes with 363 images), the error rate of Passeriformes was only 3% and ranked 19th in 27 orders.

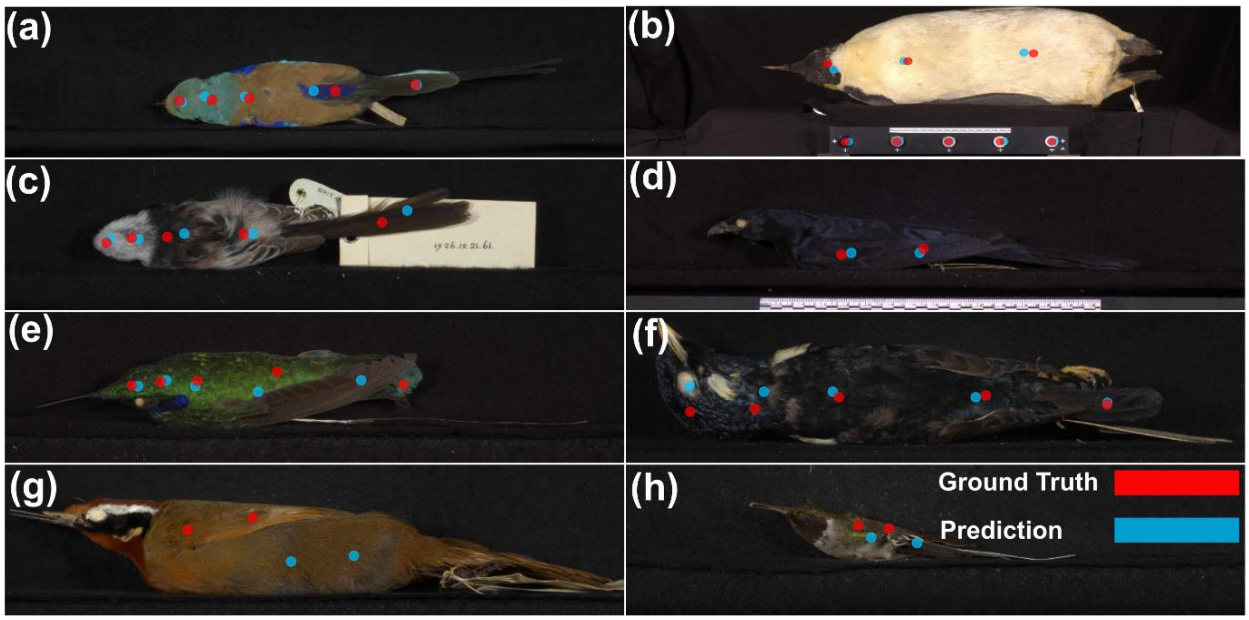


Figure 2.6. Examples (a)-(d) are correct predictions, which show that the model can place accurate labels on images with (a) a specimen with high colour diversity, (b) a big specimen and very small reflectance standards, (c) a specimen with interferential objects such as a specimen tag, (d) a specimen which its regions are similar to other parts of the image. The examples (e)-(h) are incorrect with the four characteristics of error regions: (e) The tail was incorrectly labelled to the wing, and the tail is partially occluded by a wing. (f) The eye of the bird was miss-identified as the crown, while the specimen was placed in a rare posture (twisted-head) in the image. (g) The predictions of the wing were placed on the body, and the wing has a similar colour to the rest body regions. (h) The predictions were not placed on the wing, and the wing is small. (Note: I cropped only focal parts of images to achieve better visualisation).

I further assessed variations in pixel distances across three groups (predictions vs trainer, predictions vs non-trainer and between experts). There were no average pixel distances greater than 200 pixels (Figure 2.7). There is a significant effect of the overall pixel distance among the three groups (ANOVA: $F=26.2$; $df=2.0, 14910$; $p<0.01$) and ANOVA per individual keypoints are shown in Supplementary Table 6.1.5.

The overall distance of predictions vs trainer was not significantly different from the distance between experts, suggesting that pose estimation can perform as well as expert labelling (Supplementary Figure 6.1.9a). The overall pixel distance from predictions to non-trainer,

however, was larger than the other two groups (predictions vs trainer and between experts), which shows that the model does not predict labels from non-training experts as reliably. In addition, all mean differences of pixel distances (of all and individual keypoints) are smaller than 100 (Supplementary Figure 6.1.9a).

Average pixel distances of reflectance standards between experts (range: 12.2 to 18.9 pixels) were smaller than average pixel distances for predictions vs trainer and for predictions vs non-trainer (range: 17.0 to 34.8 pixels). There were significant effects in four regions (nape, mantle, belly and coverts) when comparing the between experts group and the predictions vs trainer group, with the predictions vs trainer distance being smaller in all four regions. Covert was the only body region where the distances between experts are significantly different from the distances of predictions vs non-trainers (distances between experts are smaller).

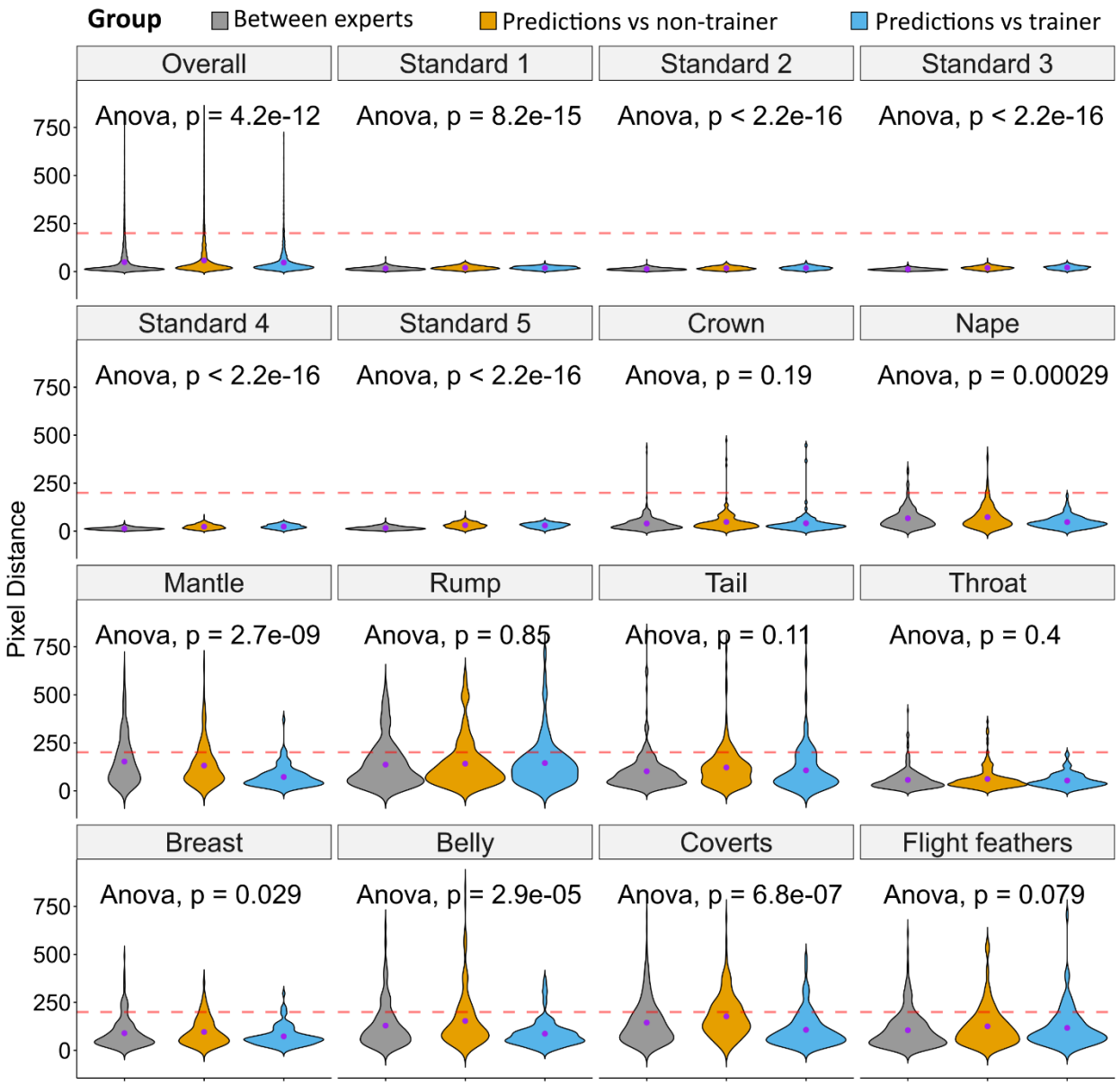


Figure 2.7. Pixel distances of overall and individual keypoints across three groups (predictions vs trainer, predictions vs non-trainer and between experts); Purple points are the mean values, red dotted lines are the pixel distance of 200. ANOVA results are listed in Supplementary Table 6.1.5. Tukey test results can be found in Supplementary Figure 6.1.9a

2.3.1.2 Performance with low-quality data

I assessed model performance using four low-quality datasets (see section 2.2.5.2). Low-quality datasets had a significant negative effect on the accuracy (ANOVA of pixel distance for all keypoints: $F=87.2$; $df=4.0, 210{,}700$; $p<0.01$. Supplementary Table 6.1.6 shows ANOVA results for individual keypoints. Supplementary Figure 6.1.9b shows results of Tukey-tests). Performances of different transformed datasets were consistently worse than the original dataset as shown in Supplementary Figure 6.1.9b and Supplementary Figure 6.1.10. Mirror and combined datasets had worse performance than translated and rotated datasets. However, all differences of mean overall pixel distance were less than 10 pixels (translation: 3.6, rotation: 4.8, mirror: 9.1, combined: 8.1), which were more accurate than using CPM (See section 2.3.1) suggesting that the effect of poor quality data on performance is minor.

2.3.1.3 Effects of within specimen colour variability

Correlations of pixel distances and colour variability are shown in Supplementary Figure 6.1.11. Pixel distance values are either positively but weakly correlated with colour variability measure by RGB (R values ranged from 0.032 to 0.21. See Supplementary Figure 6.1.11a) and lightness (R values ranged from 0.039 to 0.22. See Supplementary Figure 6.1.11c) of all and induvial body regions except for flight feather measured by RGB ($p=0.087$). No correlations of individual body regions were found between pixel distances and normalised RGB, the correlation of all body regions are negative but even weaker ($R=-0.018$, $p=0.023$. See Supplementary Figure 6.1.11b). The result shows that RGB and lightness variability of images had negative effects on the result accuracy while hue had a positive effect, though these effects were very weak.

Most of the colour information correlations decreased as the quartile increases as shown in Supplementary Figure 6.1.12. 30% of the images were divided into the same quartiles between using RGB and normalised RGB as the splitting method. 80% of the images were in the same quartile whether using RGB or lightness to define groups. Specimens with both black and white plumage were likely to be categorised in the top quartile using RGB and lightness as the measurement while specimens with colourful plumages tended to be in the top quartile using the normalised RGB measurement (Supplementary Figure 6.1.13).

2.3.2 Colour volumes of world birds

Pixels were extracted on heatmaps generated from the Stacked Hourglass with input resolution of 494 x 328 pixels, the training duration of 15 epochs and using the original images and labels. The total possible volume of avian tetrahedral colour space is 0.2165 – this represents the upper limit on the volume of colours that could theoretically be perceived by birds. The convex hull volume of all colour points in the dataset is 0.0755 (~34.9% of the total colour space), and the alpha shape (α =0.2262) volume is 0.0487 (~22.5% of the total colour space). Male volume is larger than female volume for the overall data and across individual patches as shown in Figure 2.8 and Figure 2.9 and Supplementary Table 6.1.1. Alpha shape volumes are always smaller than convex hull volumes as shown in Figure 2.8, Figure 2.9 and Supplementary Figure 6.1.5a. The mantle has the largest volume and the tail has the smallest volume (using both convex hull and alpha shape) in males. In contrast, the nape has the largest convex hull volume, whereas the breast has the largest alpha shape volume, and the tail has the smallest volume (using both convex hull and alpha shape) in females.

Detailed statistics such as volumes, hue disparity and colour span for every patch are listed in Supplementary Table 6.1.7. For all and individual patches, the convex hull and alpha shape volume differences between male and female are all significantly larger than differences between two randomly sampled groups as shown in Supplementary Figure 6.1.14. This suggests that male birds have larger volumes than female birds for every body region.

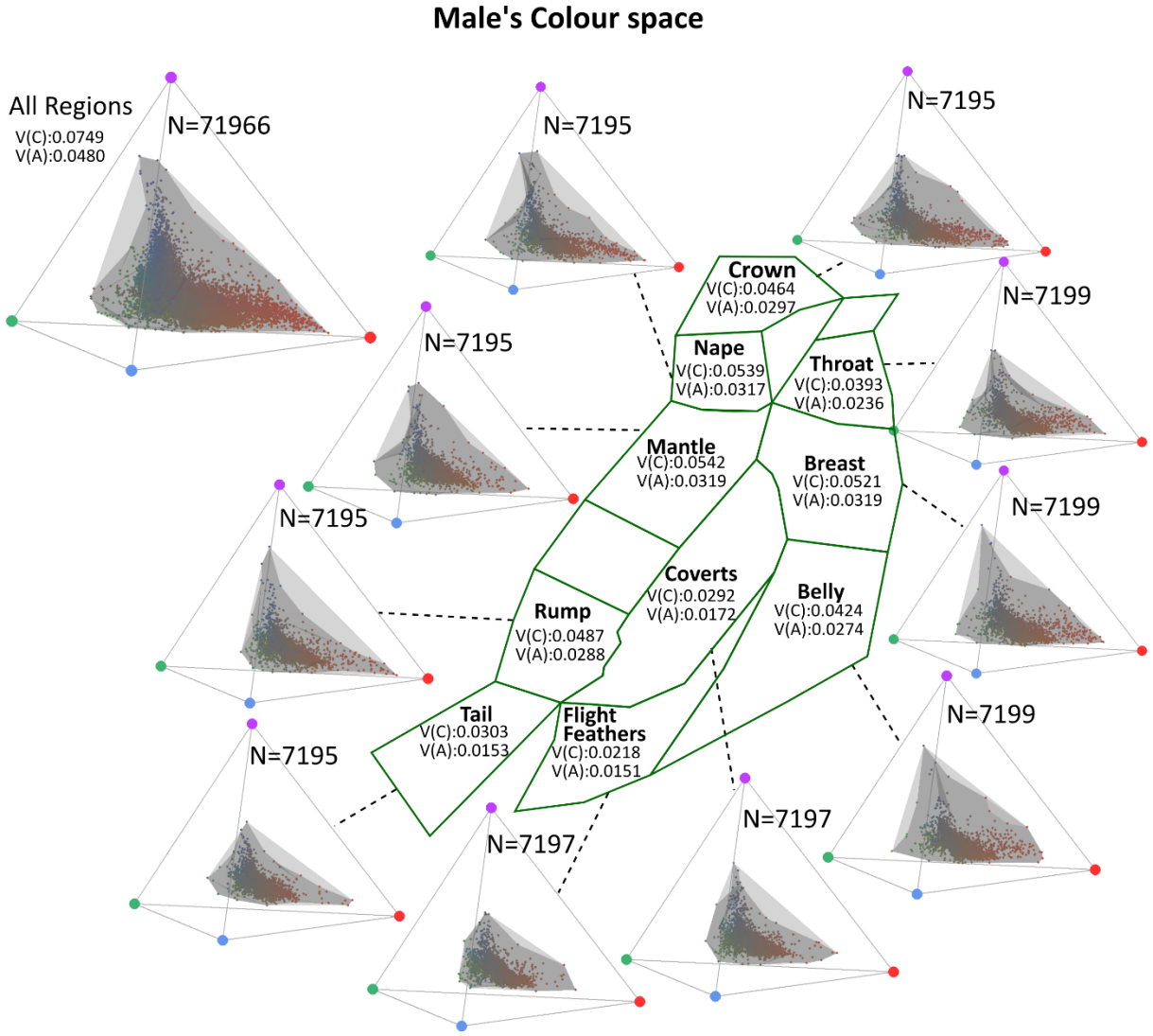


Figure 2.8. Tetrahedral colour spaces for all and individual patches for male. Convex hull (light grey) and alpha shape (dark grey) are used to calculate volumes. V(C) is the convex hull volume, V(A) is the alpha shape volume and N represents the number of colour points. Vertices of the tetrahedron are coloured as violet (u), blue (s), green (m), and red (l).

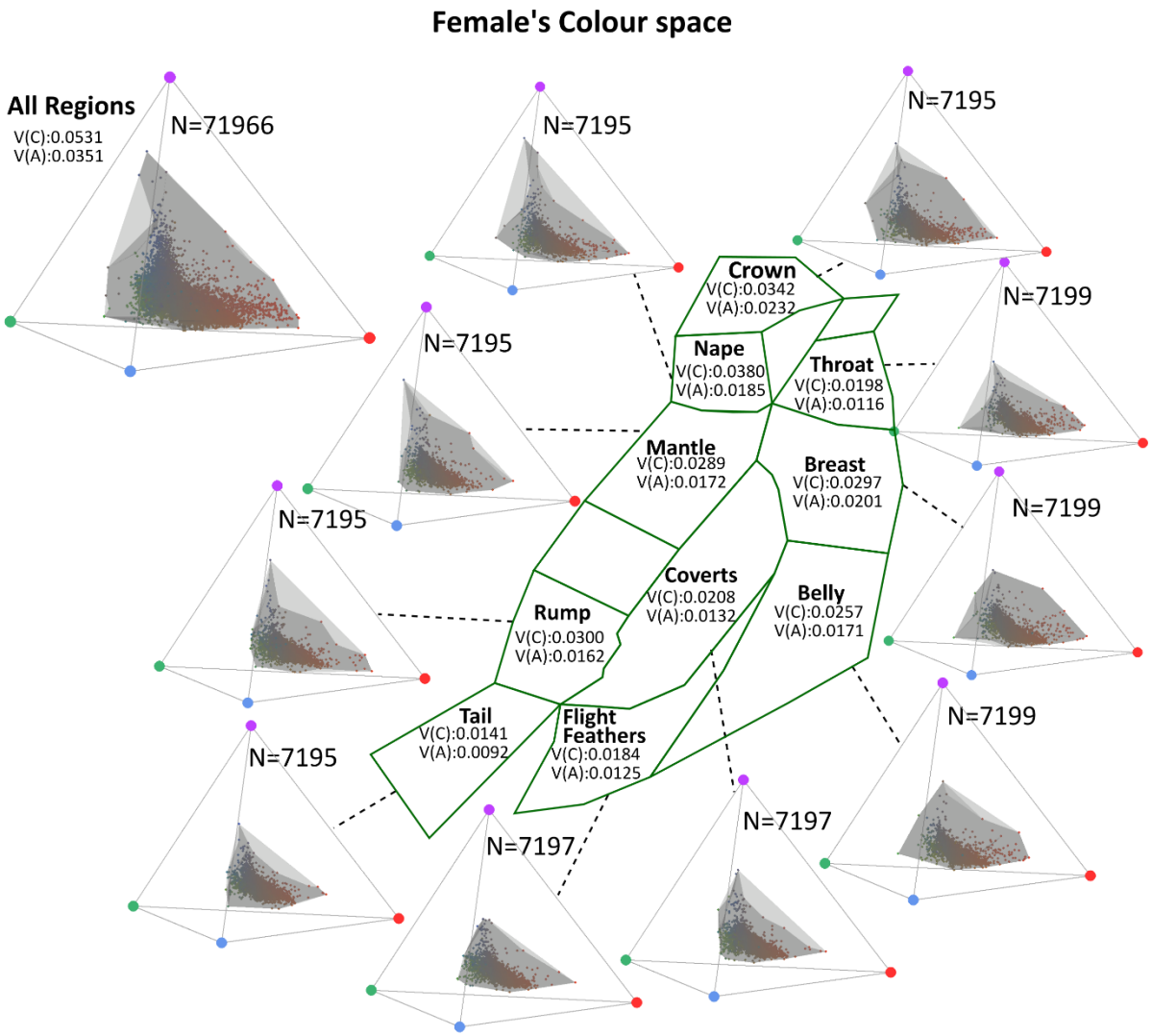


Figure 2.9. Tetrahedral colour spaces for all and individual patches for female. Convex hull (light grey) and alpha shape (dark grey) are used to calculate volumes. V(C) is the convex hull volume, V(A) is the alpha shape volume and N represents the number of colour points. Vertices of the tetrahedron are coloured as violet (u), blue (s), green (m), and red (l).

2.4 Discussion

2.4.1 Pose estimation on Project Plumage

I found that pose estimation methods can automatically locate reliably accurate keypoints on Project Plumage images. I then use colour information extracted by predicted keypoints from

more than 120,000 images to estimate the bird plumage colour space. I tested whether pose estimation networks could be used to identify keypoints on images to extract colour information from broader regions around each keypoint. While my application of pose estimation differs from its original objective (i.e. identifying human parts and joints; Andriluka et al. 2014), I found that the model performs well with output often comparable to the ground truth expert labelling. Stacked Hourglass clearly outperformed CPM and input resolutions of 494 x 328 pixels generated the best result of those tested. CPM was one of the first deep learning algorithms applied to pose estimation using heatmap output and intermediate supervision (Wei et al. 2016) to increase the model accuracy. In contrast, Stacked Hourglass (Newell et al. 2016) uses residual modules from the deep residual network (He et al. 2016), which is the state-of-the-art architecture for feature extracting in deep learning and is one of the best models for identifying single human posture on the MPII Dataset (Andriluka et al. 2014).

Pose estimation networks have been typically applied to human posture images while achieving accurate results. Human posture photos often contain varieties of human postures, clothes, and backgrounds, and, from the human perspective, seem more visually complex than the plumage images that I focus on. However, while posture can create clearly defined keypoints, the focal regions that I aim to identify are more ambiguous in the definition. This creates a different set of challenges for expert labelling that could potentially limit the performance of deep learning algorithms. In the Project Plumage data, different experts placed the same label on the same specimen in different locations but are still considered correct. Because body regions were not homologous parts (e.g. landmark points in morphology studies, Bookstein 1991), and people have different opinions on areas defining each body region, it can be difficult to place a body region point in the same location. The mantle and belly were the most variable among experts. These two regions often had larger areas than the rest of the regions, so it was easy for experts to pick different, but correct locations. On the other hand, from experts' experiences, keypoints were more challenging to place confidently for some regions than others. For example, identifying (i) crowns on folded or twisted heads (e.g. Figure 2.6f), (ii) rumps or tails that were partially occluded by wings or museum labels (e.g. Figure 2.6e), and (iii) coverts and flight feathers that had small areas (e.g. Figure 2.6h) or similar colours (e.g. Figure 2.6g) are particularly

difficult. Other regions are more straightforward, e.g. the neck, breast and belly from the belly view all have relatively large areas and are rarely occluded. Deep learning models may share similar difficulties in labelling plumage images to humans. Experts' experiences were similar to errors made by the deep learning predictions (Supplementary Figure 6.1.8a). Differences among experts can be seen as the ceiling of the model performance. Differences between experts did not surpass the difference between predicted labels and labels from the trainer expert (predictions vs trainer) across all body regions (Supplementary Figure 6.1.9a). Predicted labels were closer to the YH's labels than CRC and GHT. This is not surprising because the model was trained on the original expert labels (YH), while labels of other experts (CRC and GHT) were completely independent. The errors of predictions are generally similar to the differences between experts. Although pixel distances of five standard points from predictions vs trainer were larger than those distances from between experts (Supplementary Figure 6.1.9a), PCK-100 for five standards from predictions vs trainer were all 100, suggesting that all standard predictions were placed inside reflectance standard circles (ground truth points were placed in circle centres and the minimum radius was larger than 100 pixels). The colour information correlations show that colours extracted by deep learning predictions were highly correlated with colours extracted by ground truth labels. This is especially true for the chromatic information (normalised RGB), which is the main colour information used in creating the tetrahedral colour space, and had coefficients higher than 0.9 for all and individual body regions.

2.4.1.1 Colour volumes measured by deep learning

After using the colour information measured by deep learning to build the colour space, I found that bird plumage colours only evolved to occupy a quarter to one-third of the total possible colour space. Based on the results from expert checking (95% of images were predicted correctly) and evaluation metrics, I suggest that the predictions on the best configuration of the deep learning network are sufficiently accurate to generate usable biologically meaningful data (i.e. the bird plumage colour space). The deep learning model to the project plumage data set required less than three days to predict all photos. I used 7200 species with data for both males and females compared to 111 species for male-only (92 shared species) in Stoddard and Prum (2011). It is therefore predictable that the total colour volume is larger than found by Stoddard

and Prum (colour space proportion: 34.9% vs 26-30%) when based on the same volume metric, since adding species should increase volume by chance alone. However, the alpha shape (α =0.2262) volumes cover on average 61.3% of the convex hull volumes. The alpha shape measures the actual colour volume more precisely with the possibility of being a concave shape. This is very likely for data points that are mostly clustered together and where a small number of points are far from the cluster centre, the convex hull tends to create large areas of empty colour space to include all points, as shown in Figure 2.8 and Figure 2.9. In contrast, the edges of the alpha shape fit points closer than the convex hull edges. Colour gamuts measured by the alpha shape in this paper all have smaller volumes than when measured by the convex hull. Gruson (2020)'s result also shows that the alpha shape estimates smaller volumes than the convex hull for the tetrahedral colour space uses birds from the Nouragues rainforest, in French Guiana. Together, I suggest that previous calculations of the size of the avian colour gamut have been overestimated. The alpha shape is more accurate in measuring a single colour gamut, and the actual volume value is important (e.g. use the volume to calculate the colour gamut proportion in the overall possible volume in Stoddard and Prum (2011)). However, because the alpha shape is sensitive to the α value. When comparing alpha shape volumes across sets of colour points, the choice of the α value should be examined carefully (e.g. checking if the male or female alpha volume is larger than both sexes alpha volume) to avoid comparison inaccuracy like Supplementary Figure 6.1.4.

As with Stoddard and Prum's colour space, I found that bird plumage colours are less diverse at short wavelengths (the s cone which corresponds to blue colours) than along wavelengths associated with the other three receptor cones. Male volumes are larger than female volumes across all patches, confirming the expectation that male birds have more diverse plumage colour than female birds. Previous studies have suggested that male plumage colour is driven by sexual selection often favouring more diverse and extreme colours (Gomes et al. 2016; Cooney et al. 2019). The crown, nape and breast have relatively large volumes (measured with both convex hull and alpha shape) whereas the coverts and flight feathers have small volumes. This is consistent with patch-specific rates of evolution in Cooney et al (2019) implying that different selection pressures act on different body regions.

Overall, these results illustrate that deep learning can be successfully used to create a comprehensive and reliable bird plumage colour dataset using vastly less time and human resources than manual labelling. This dataset has great potential in understanding questions related to the bird plumage colour.

2.4.2 Experience and guide for future projects

My analyses highlight the potential of automated labelling of biological specimens with these methods. The designs and layouts of the Project Plumage dataset were highly consistent: each image contains one specimen, and the view information is always known, and the orientation of the specimen is always the same. However, breaking down the consistency by using low-quality datasets did not cause major decreases in model performance. This result is encouraging because it implies that the system developed here could be applied to other data sets. I am somewhat cautious here because some of the worst predictions were found in the more unusual specimens, for example, very large specimens relative to the size of the standards (Figure 2.6b, although it was correctly predicted), specimens with uncommon placements (Figure 2.6f), and very small specimens (Figure 2.6h). This indicates that specimen variability is a challenge and highlights the need for large and representative training sets. For example, in expert-labelled data, the order Passeriformes represents about 70% of the training set whereas other orders such as Galliformes, Sphenisciformes, Pelecaniformes and Procellariiformes are poorly represented, have more unusual shapes and are associated with the highest error rates. Expanding the training set, or training for specific groups as subsets may therefore be beneficial.

Specimen variability also poses a challenge for assessing the accuracy of predicted keypoints. In human pose estimation a variant of PCK, called PCKh (h stands for the head) is often used. PCKh uses a certain proportion of the human head length as the threshold, providing an intuitive metric to measure the accuracy of predictions of pose estimation. As the head-body ratios of humans are very similar, this approach normalises the uncertainty introduced by the depth of humans in photos. This metric has been widely used in evaluating pose estimation methods (Insafutdinov et al. 2016; Newell et al. 2016; Wei et al. 2016). However, bird specimens were variable in size and shape, it is difficult to find one meaningful PCK threshold. PCK-100 is suitable for measuring

the accuracy of predictions for the five reflectance standards. The variability of the PCK performance against different thresholds shows the problem of using fixed thresholding (Supplementary Figure 6.1.15). The five reflectance standards have steep growths, while the rump, tail, covert and flight feather have flat growths, and they are the most challenging regions as shown in the result section. However, PCK that uses a proportion of a certain body part may be an ideal metric for geometric morphometrical datasets. Width, length or dimensional values (e.g. diagonal length) of focal areas can be measured invariantly by landmarks like the height of *Littorina* shells (Ravinet et al. 2016), and the length of squirrel's mandible (Zelditch et al. 2015).

It is important to choose the metrics according to the datasets and labels. Pixel distance is a straightforward metric to compare models trained by different configurations, but it is not intuitive in the context of the data that I wish to capture (i.e. colour information). By using colour metrics that directly relate to the variables that I might wish to measure, I have provided an alternative to distance-based indices of model performance. The correlation coefficient gives an intuitive metric to show how accurately predictions capture colours. RGB, normalised RGB and lightness measure different types of colour information. The metrics used in the evaluation have a good match with the actual correctness of predictions. In the expert checked result, the belly is the view that has the least incorrect predictions. The flight feathers, coverts, tail, and rump are the top regions of erroneous predictions, which is similar to the colour correlation result of the best model (Supplementary Figure 6.1.8, Table 2.1). Comparing to use the end result (i.e. 2D coordinates) from the pose estimation to create fixed-size and -shape measurements (e.g. bounding boxes), I adapted the output heatmaps from Stacked Hourglass to create a more intuitive and size- and shape-varied colour extraction method. The visualisation (Supplementary Figure 6.1.3) and colour information correlations between two extraction methods based on the heatmap and the bounding box (Table 2.1) may suggest that using heatmaps rather than points on measuring colours on specimen body regions could be a better option. More tests examining a wider range of extraction methods than those considered here (Heatmap-90 vs Bbox-20) would be beneficial and the specific choice of extraction may be dataset dependent.

The major advantage of deep learning is that it can significantly increase the speed of data measurements. For Project Plumage, using pose estimation, hundreds of days of human work

(estimated by the progress on www.projectplumage.org) were reduced to less than three days (predicting points on 122,610 images) with the computing power of GPU. Because of the limitation of GPU memory, it is computational-expensive or even impossible to train networks on the original image resolution. Images were down-sampled into a uniform and reasonable resolution and used for training. Although lower input resolution had a negative effect on prediction accuracy (Figure 2.3), the maximum possible resolution (a 10-fold reduction in resolution) can generate reliable predictions. Therefore, I recommend that analyses are conducted at the highest manageable resolution for the available computing hardware. More powerful GPUs could increase speed further, or potentially allow the training network to handle higher resolution images.

Taken together, I propose a general workflow to label points on other large-scale biodiversity datasets, with the following steps:

i. Imaging specimens. Specimens should be digitised in a consistent setup (same orientation, background, lighting) where possible. This will reduce the complexity of the training/learning step. However, if it takes too much time to place objects centre and upright precisely, or if imaging live specimens that cannot be so easily manipulated, it is possible to have less precise placement steps.

ii. Creating the training set. The training set should be labelled or checked by experts to keep its quality. Here, I did not explore the effects of varying the size of the training set but usually the more, the better. Images used in the training set should be representative of the whole dataset. Predictions for unusual images that are poorly represented in the training set are less likely to be reliable.

iii. Training the model. When training the neural network, different configurations should be tested, including different architectures, input resolutions and other hyperparameters. Although experimental manipulations did not make the model perform better for this dataset, they are useful techniques to try on other projects. However, if it takes a rather long time to train, start with the highest input resolution on the original network without any improvement techniques.

It will provide a baseline, and it is useful to guide different combinations of hyperparameters for the subsequent test runs.

vi. Evaluating the predictions. Appropriate metrics should be used to fit the goal of a project. In contrast to requiring high colour accuracy, studies that aim to generate keypoints (or landmarks) for geometric morphometrics (Chang and Alfaro 2016), or for tracking animal movements or behaviours (Mathis et al. 2018a), require high spatial accuracy. Exploring examples of poor predictions and running post-hoc tests can help to understand if there are common causes of low performance and can be useful for tuning the network performance.

v. Post-processing. A good practice is to check and correct predictions manually. If the dataset size is huge, only checking predictions with high error risk from post-hoc tests is a pragmatic compromise between efficiency and accuracy.

2.5 Conclusions

With the help of deep learning neural networks and the growth of large-scale biological data, it is possible to develop a high-throughput data collection method on digitised data. Although the result of deep learning can hardly surpass the human result, this paper shows that a real-world example can provide a result that is remarkably close to expert labelling. The colour information extracted by deep learning from the Project Plumage dataset is accurate enough to use in analyses and could be further optimised with time-saving checking and correction steps. Colour volumes from more than 80% of bird species have solidified statements that bird plumage colours only occupy a proportion of the total available avian colour space and male birds tend to have more diverse plumage colour than female.

Chapter 3 Segmenting biological specimens from photos: a comparison of classic computer vision and segmentation methods

Abstract

The growing number of digitised biological specimens brings new possibilities for the study of a wide range of evolutionary questions at broad scales. For example, digital images of specimens can be used to measure biological traits such as size, area, shape and colour. Image segmentation is a common method for separating focal areas from images, such as segmenting cells from microscope images, which can then be used for further analysis. However, the performance of alternative segmentation methods on large biodiversity datasets has not been adequately tested. Here, I used digital photographs (more than 120,000 images) of birds from museum specimens that cover most of the world's extant species and a manually segmented dataset (5,094 images) to compare the performance of deep learning models with classic computer vision as tools to measure colour. I applied semantic segmentation using the DeepLab deep neural network, which is considered state-of-the-art for many segmentation tasks, on the labelled dataset. DeepLab achieved scores of over 90% in measures of accuracy and precision. Classic segmentation methods (e.g. thresholding and graph cut) were clearly outperformed by DeepLab. The results show that deep learning can segment plumage photos accurately and efficiently. I then applied DeepLab to the whole dataset (>7,500 bird species) to demonstrate the utility of the method. Specifically, I (i) generated measures of colour diversity among bird species, and (ii) assessed the phylogenetic distribution of colour diversity. I found that colour diversity is normally distributed and shows a moderately strong phylogenetic signal (closely related species share similar colour diversity). Finally, I provide guidelines for building a workflow of digitising and segmenting large-scale biological datasets.

3.1 Introduction

Globally, natural history museums house extensive, but often underexploited, collections of biological specimens. There are an estimated 1.2 to 1.9 billion specimens in museum collections globally (Ariño 2010). Collection digitisation is a major goal for many natural history museums and can provide a rich source of phenotypic and biodiversity data (Blagoderov et al. 2012, https://www.dissco.eu/). Digitised objects and specimens are often stored as images that provide a permanent record (Flemons and Berents 2012; Nelson et al. 2012; Holovachov et al. 2014; Hudson et al. 2015) but subsequent analysis typically requires additional image processing to extract usable data. A particularly important and widely used step in data extraction from specimen images is segmentation. Image segmentation is a pixel-level label that can define regions or contours of interest and has been used in many fields, including facial recognition (Saito et al. 2016), robot vision (Milioto et al. 2018), and automated driving (Treml et al. 2016; Siam et al. 2018). It is a particularly important method for processing medical images. For example, segmentation has been used to measure and visualise the morphology of cells (Meijering 2012; Xing and Yang 2016), brains (Whitwell 2009), whole-body skeletons (Baiker et al. 2010), and phenotypes of embryos (Johnson et al. 2006). Segmentation has also been used in phenotyping live plant photos (Minervini et al. 2014; Scharr et al. 2016). Scans (e.g. CT, micro-CT or MRI) of fossils require segmentation to generate the accurate scans (Ebey Honeycutt et al. 2014; Davies et al. 2017; Park et al. 2017). Segmented fossil scans are often used as the input data of morphometric and functional analyses. However, segmentation has not yet been widely used in phenotyping natural history photos. While segmentation results of natural history photos are often used as inputs for classification problems as segmentation can extract precise specimen information from images (Kumar et al. 2015; Unger et al. 2016).

One key challenge is that manually segmenting images (especially when applied to large datasets, such as those that can be taken from museum collections) is time-consuming and usually takes more time and effort than placing straightforward labels (e.g. points and bounding boxes). Automated methods often rely on forms of thresholding. Broadly, a wide variety of segmentation methods, including both classic computer vision methods and the emerging approach of deep

learning semantic segmentation, have the potential to provide accurate, high-throughput pipelines to phenotype natural history datasets. Popular methods and some applications in natural history datasets are introduced below.

Table 3.1 summarises commonly used segmentation methods. Methods such as thresholding and watershed are fast and do not need much prior information. Other methods including region growing, active contour, and graph cut, need seeds (i.e. starting locations) as input. With the help of seeds, these methods can make predictions based on spatial information. Many interactive segmentation methods, such as lazy-snapping (Li et al. 2004) and GrabCut (Rother et al. 2004) were developed based on graph cut. Atlas-based segmentation and statistical shape models (SSM) use prior shape knowledge learnt from a training set to segment images. And they have been widely used in facial recognition and medical image segmentation (Cootes et al. 2000; Pohl et al. 2006; Aljabar et al. 2009; Heimann and Meinzer 2009).

Table 3.1. Overview of some commonly used and classic segmentation methods

METHOD	DESCRIPTION	PRIOR SHAPE KNOWLEDGE?	SEEDS FOR INITIALISATION?
THRESHOLDING	Segments an image by allocating each pixel to either the foreground or the background based on a pre-defined value. Can be set either manually or automatically calculated based on image features such as the image histogram or entropy (Sezgin and Sankur 2004).	No	No
WATERSHED	Uses the concept of flooding waters from low-intensity regions to high-intensity regions, and boundaries are edges between fields of waters (Vincent and Soille 1991).	No	No
REGION GROWING	Segments neighbour pixels of initial seeds which have intensity difference less than a specific range. Finishes when no new pixel is segmented (Adams and Bischof 1994).	No	Yes
ACTIVE CONTOUR	Developed by Kass, Witkin and Terzopoulos in 1988. A closed contour is placed on the image, it then transforms to minimise the sum of the internal and external energy.	No	Yes

	The converged contour is used as the segmentation.		
CHAN-VESE	An improved method of active contour that can capture separated regions with one initial contour (Chan and Vese 2001), because it uses level sets to represent contours rather than the parametric format.	No	Yes
GRAPH CUT	Foreground and background seeds are placed. The algorithm sees an image as a graph and pixels as nodes. Each pixel has edges to its neighbour pixels and edges to a source node (foreground) and a sink node (background). Weights of edges are based on pixel intensities and identities (i.e. foreground, background or to be segmented). The minimum cut cuts the graph into two subgraphs that have the largest weighted sum (Boykov and Jolly 2001). The result is the foreground subgraph.	No	Yes
ATLAS-BASED SEGMENTATION	An atlas is first created using one scan, the average of scans or a statistical representation of a population of scans. Segments images by aligning themselves with the atlas using global (e.g. affine and rigid) and local (e.g. elastic, B-splines) transformations (Cabezas et al. 2011).	Yes	No
STATISTICAL SHAPE MODEL	Uses a point distribution model (PDM) to make landmark predictions on images. PDM consists of the mean shape and modes of shape variation, it therefore can represent a range of shapes by tuning the shape variations. Algorithms with different criterion are used to optimise the shape on given images (Cootes et al. 2000, 2001).	Yes	No

Many of the methods introduced above have been used in segmenting specimen photos. Lazy snapping (initiated with manual seeds placement) was used to segment leaves and then to extract shapes and outlines to use as features in a leaf classifier (Unger et al. 2016). Similarly, the interactive graph cut was used in segmenting animals from ~4000 photos of internet images (Kumar et al. 2015) to generate colour features and train a classifier. Henries and Tashakkori (2012) used thresholding to create an initial segmentation of plants (leaves, stems and branch), and used watershed to optimise the results. The thresholding-watershed pipeline was also used

in the pre-processing step of a digitisation software tool to segment insects from photos of museum trays (Hudson et al. 2015). Examples such as these show how computer vision-based thresholding has dominated approaches to segmentation. However, newly emerging deep learning semantic segmentation techniques have the potential to improve accuracy for high-throughput image analysis pipelines. Indeed, advances in deep learning have facilitated the extraction of phenotypes from labels automatically placed on digitised collections. My previous work, for example, automatically placed keypoints of body regions on 122,610 bird photos using deep learning in just three days (Chapter 2).

Semantic segmentation using deep neural networks has developed rapidly, as growing computational power and applications using the graphics processing unit (GPU) improve optimisation speed greatly. The convolutional neural network (CNN) is the core deep neural network architecture for feature extraction from images (Krizhevsky et al. 2012; He et al. 2016). CNNs transforming input images into predictions using convolutional and pooling layers. Networks extract images features during the training (i.e. trying to generate closer predictions to the ground truth labels). Features are then used in computer vision tasks such as image classification, pose estimation and semantic segmentation. There are many variations of CNN architectures, with the common goal of building networks with deeper and deeper layers to extract features more accurately and precisely (Szegedy et al. 2014; Ioffe and Szegedy 2015; He et al. 2016).

Early semantic segmentation studies classified superpixels, patches of images, or detected bounding boxes of objects (Farabet et al. 2012; Girshick et al. 2014; Pinheiro and Collobert 2014). The outputs of these deep neural networks were typically refined by post-processing methods (Farabet et al. 2012; Hariharan et al. 2014). However, precise predictions were challenging because these networks do not make pixel-wise predictions. The fully convolutional network (FCN) was the first network architecture to make pixel-wise predictions (Long et al. 2015). Because convolutional and pooling layers often downsample the original input image, the output of the convolutional part has a lower resolution than the original image. Backwards convolutional layers can be used to upsample the result as a backward convolutional layer simply reverses a convolution layer. The output of the FCN is a heatmap (with the resolution same as the input

image) which can represent the class of each pixel from the input image. Segmentation accuracy of the FCN, measured as the mean intersection over union (mIOU), has been shown to perform well and, for example, reached 62.2% on the PASCAL Visual Object Classes 2012 (PASCAL VOC 2012) dataset (thousands in segmented photos of 21 classes, Everingham et al. 2015), an improvement of 20% over previous methods.

More recently, a group of networks named DeepLab has been developed that adapted the pixel-wise prediction idea from the FCN while improving its network architecture. There are four versions of the DeepLab architecture. Version 1 uses a fully connected conditional random field (CRF) to post-process the result from the fully connected network and has improved the mIOU of PASCAL VOC 2012 to 71.6% (Chen et al. 2014). While keeping the CRF as the post-processing step, version 2 uses atrous convolutions with bilinear interpolation, instead of the backward convolutional layer, to save memory and time (Chen et al. 2017b). The atrous convolutions can adjust the stride used for sampling the input signal while extracting enough features to control the output resolution. Atrous convolutional layers with different rates are used to learn multi-scale features and the module is called the atrous spatial pyramid pooling (ASPP). Intermediate results from the ASPP are upsampled using bilinear interpolation. This improvement increased accuracy (mIOU of PASCAL VOC 2012: 79.7%) and was less computationally expensive. Version 3 (DeepLabv3) uses the cascaded atrous convolution module and improved ASPP (batch normalisation added) module (Chen et al. 2017a). This version can capture more features from different scales and achieved a mIOU of 85.7% on PASCAL VOC 2012 without any post-processing such as CRF. DeepLab version 3 plus (DeepLabv3+) extended DeepLabv3 by adding a decoder module which helps to refine the segmentation and has achieved 89.0% of mIOU on PASCAL VOC 2012 (Chen et al. 2018). The DeepLab family provides accurate pixel-wise semantic segmentation networks and is currently one of the most accurate semantic segmentation methods available. Despite the improvement that deep learning offers for image segmentation, it has not yet been widely applied in segmenting specimen photos. One recent exception is the application of DeepLabv3 to segment specimens from noisy backgrounds in herbarium photos (Hussein et al. 2020). In this application, the non-specimen areas were replaced by white colour to provide a

noise-free background. Processed photos were then used in an identification task. With a dataset of around 400 photos, accuracy as high as 98% mIOU was achieved.

Here, I assessed the performance of deep learning segmentation in comparison to classic computer vision methods using photos of bird specimens taken at the Natural History Museum, Tring, UK. Segmentation results of specimen photos are often used as the input for specimen classification and identification, but applications of segmentation on phenotyping medical images show the broader potential for use as a phenotyping pipeline for different types of natural history data. For example, it is possible to use photographs to objectively measure colour from digital images from calibrated cameras (Troscianko and Stevens 2015). I aim to test different methods in order to build a pipeline that can segment bird photos automatically and accurately. I used, evaluated, and compared classic and deep learning segmentation methods to segment the specimen from the background and to remove obstructions (labels, string etc.) that obscure the specimen. I then selected the most accurate segmentation method to apply on a bird specimen photo dataset that covers more than 7,500 bird species. Colour information measured from the output segmentation result was used to study the plumage colour diversity across birds.

3.1.1 Application: global variation in intraspecific colour diversity

Birds have evolved into a wide diversity of colours (Stoddard and Prum 2011). Birds can perceive light across a wider spectrum, including parts of the ultraviolet spectrum, than humans (Goldsmith 1990; Cuthill et al. 2000). Numerous studies have mapped bird colours into an avian-visual based tetrahedral colour space based on the four cones receptor cone types (ultraviolet (u), shortwave (s), mediumwave (m) and longwave (l)) (Stoddard and Prum 2008, 2011; Cooney et al. 2019). The true (i.e. from an avian perspective) colourfulness or colour diversity of bird plumage can then be measured in the avian colour space. The colour diversity of an individual bird is thought to be related to and constrained by the underlying colour producing mechanisms (Stoddard and Prum 2011). For example, melanin-based plumage colours are mainly black, grey and brown, while carotenoid-based plumage colours are mainly red, orange and yellow (Hill et al. 2006b). Stoddard and Prum (2011) suggested that colour diversity within species is partly determined by the range of colour producing mechanisms of each species. This link between

individual colour diversity and colour producing mechanism implies that colourfulness may be evolutionarily constrained (Stoddard and Prum 2011) due to conserved developmental genetic pathways. However, many studies have also shown that plumage colour diversity can be driven by both natural (Slagsvold et al. 1995; Willink et al. 2014; Dunn et al. 2015) and sexual selection (Price and Eaton 2014; Dale et al. 2015; Dunn et al. 2015) which may imply greater evolutionary lability. The extent to which colourfulness (i.e. within species diversity of colour) is phylogenetically conserved among bird species is unknown. Understanding the diversity and distribution of plumage colourfulness may also have implications for bird conservation. For example, Garnett et al. (2018) showed that humans are intrinsically interested in colourful birds and suggested that promoting birds that are colourful but not well-known to the public may be a viable approach to raise conservation awareness and attention.

I used the best-trained model to predict plumage area segmentation and extract colour measurements from more than 7,500 bird species. I then used the segmentations to build a tetrahedral colour space based on avian visual models for each species. Segmenting the whole specimen allows measurements of the entire range of colours and patterns on the bird. This contrasts with the more common approach of measuring the colour of specific body regions directly from specimens using spectrometers (Stoddard and Prum 2008, 2011) or placing points or polygons (Cooney et al. 2019, Chapter 2) on digital photos with special reflectance calibrations (Troscianko and Stevens 2015). I compared tetrahedral colour spaces between segmentation (this chapter) and patch measurements (Chapter 2). To unleash the advantage of colour information completeness provide by segmentation, I designed a metric that can estimate colour diversity and the proportions of different colours. I used this proportional colour diversity metric, and the convex hull volume, a common metric for colour diversity (Stoddard and Prum 2008; Renoult et al. 2017), to visualise the phylogenetic distribution and measure the phylogenetic signal of colourfulness for birds. Finally, I discuss the potential of using segmentation to measure phenotypes on natural history datasets.

3.2 Data and methods

3.2.1 Data

The images and labels used in this study were collected as part of a broader study of bird diversity and form part of the online citizen science project Project Plumage (www.projectplumage.org). The images were taken in the bird collections at the Natural History Museum, Tring. All images followed a standardised design (see Cooney et al. 2019 and section 2.2 of this thesis for detail). Each image had only one specimen, and one specimen was imaged from three views (back, belly and side). Each specimen and each view was imaged twice (human-visible and ultraviolet (UV) light spectra). Here, I counted human-visible and UV images as one image that has different channels (RGB and UV channels). The full Project Plumage dataset consists of 122,610 images, here I used a manually labelled subset (N=5,094) to test the performance of segmentation methods and demonstrate the utility of automated segmentation. I then applied the best segmentation method on the full dataset to generate segmentation for every image. And using colour measurements from segmentations to explore questions about the bird plumage colour diversity.

3.2.1.1 Image labelling

Polygons are placed on photos to segment plumage areas as shown in Figure 3.1. Multiple polygons can be used to capture unconnected areas (Figure 3.1b). Figure 3.1c shows an example of using nested polygons to label non-plumage areas inside plumage areas (e.g. eyes and feet). A segmentation then contains two classes, plumage and non-plumage areas. One key rule is that segmentation should not include any regions outside the plumage area, and it is preferable to segment within the focal area (i.e. to be conservative in the estimation of the plumage area). This is because the colour space should only contain plumage colour information. A total of 5,094 photos representing three views of 1,698 bird species were labelled manually by two experts. The sample of 1,698 bird species encompass representatives of more than 81% of bird genera and 27 bird orders, so the labelled images should capture a large extent of the total variance in plumage colour, patterns, and bird body shape in the whole Project Plumage dataset.

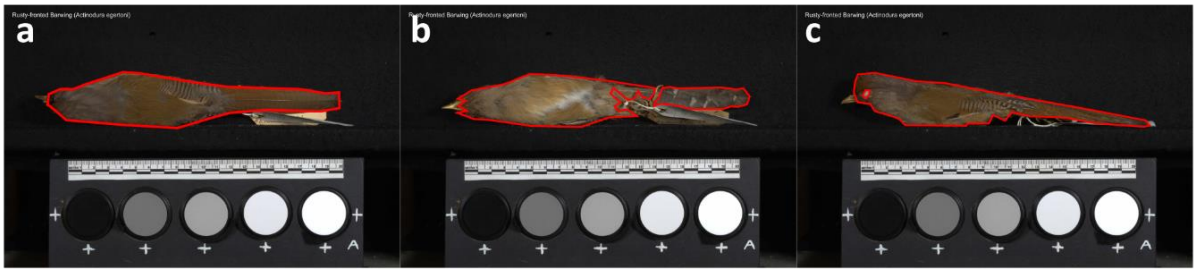


Figure 3.1. Examples of using polygons to segment plumage areas of specimens. (a) A specimen is segmented using a single polygon. (b) A specimen is segmented using multiple polygons. (c) A specimen is segmented using nested polygons as the eye is not plumage area and is excluded using a nested polygon.

3.2.2 Deep learning in segmentation.

I applied DeepLabv3+ to a segmentation workflow with three steps: i) data preparation, ii) model training, and iii) evaluation. In the data preparation step, I converted images and labels into formats which can be fed into the network. I then used 5 fold cross-validation and split data into training and validation sets with an 80:20 split. Cross-validation can provide an accurate estimate of model performance by averaging performance for different partitions (5 partitions for 5 fold cross-validation) of training and mutually exclusive validation sets. A common approach used in many studies or projects (e.g. methods in solving the ImageNet challenge; Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016) is to split data into a training set, a validation set, and a test set where the test set is used to provide the final benchmark. I used only the training and validation sets so that every image from the labelled dataset (covering a wide range of extant bird species) can have a prediction from the same data partition routine. This allows the relationship between bird taxonomy and network performance to be evaluated (i.e. to assess whether performance varies among groups of bird species due to broad differences in size, shape and colouration of specimens). After data splitting, I trained the model with the training set under the pre-defined network hyperparameters. For each training step, the network generates predictions from input images. The model optimises the loss between output heatmaps and ground truth heatmaps (converted from segmentations, see next section) by updating its parameters with the gradient of the loss function (Ruder 2016). After training, I used the validation set to evaluate model performance.

3.2.2.1 Data preparation

DeepLab outputs heatmaps at the same resolution as the input image, and with channel numbers equal to class numbers (e.g. if the resolution of an input image is 4,948 x 3,280 pixels, and there are two classes to be predicted, then the output heatmap has a resolution of 4,948 x 3,280 pixels and two channels or a dimension of 4,948 x 3,280 x 2). The output heatmap pixel value (0 to 1) of each channel represents the probability that the pixel belongs to the corresponding class. An example of the segmentation-heatmap relation is shown in Supplementary Figure 6.2.1a. The ground truth heatmaps have the same dimension (resolution and number of channels) as the output heatmaps. I converted coordinates of polygons to heatmaps, with the first channel as the non-plumage area and the second channel as the plumage area. Pixels of the non-plumage area were set to 1 for the first channel and 0 for the second channel, and vice versa for pixels of the plumage area.

DeepLabv3+ may not work using input images with excessively large resolution due to the memory limitations of the graphics processing unit (GPU) and the model complexity. I therefore downsampled images to 618 x 410 pixels (from 4,948 x 3,280 pixels) using bilinear interpolation from OpenCV (a computer vision library. Bradski 2000). This resolution is eight times smaller than the original resolution and is the largest possible resolution that could be trained on the NVIDIA GTX 1080Ti GPU (12GB GPU memory) which I used to train models without any difficulties. I also downsampled ground truth segmentations to match this resolution. Finally, I split the images into the training set and validation set with an 80:20 ratio.

3.2.2.2 Training

I divided training images into batches of four images. The model takes one batch per training step to balance the memory usage of the GPU and the optimisation at each step (Hinton et al. 2012). I used the sum of cross-entropy between pixel values of output heatmaps and ground truth heatmaps as the loss function. To minimise the loss function, I used the ADAM optimiser (Kingma and Ba 2014) and the gradient of the loss function to update model parameters. I set the initial learning rate to 0.01. Through the training process, the learning rate was cosine decayed and restarted at the initial value after reaching zero, which increases the possibility to reach a better

local optimum (Loshchilov and Hutter 2016). The length of the first period of decay-restart was set to one epoch (an epoch is defined as one pass of the full training set for the network). After each period, the new period is two times longer than the previous one (i.e. the second period takes two epochs to decay to zero, the third period takes four epochs and so on). I trained the model over 31 epochs (i.e. five complete decay-restart periods), after which the optimisation had converged (i.e. the loss has stopped decreasing). I implemented and trained the network using Python 3 and Tensorflow 1.12 (Abadi et al. 2016), a deep learning library, on one NVIDIA GTX 1080Ti GPU (12GB GPU memory). After the training process, I passed the validation images into the trained network to generate validation predictions. I resized the predicted segmentations to the same resolution (4,948 x 3,280 pixels) as the original images and used these resized predictions for the evaluation.

3.2.2.3 Evaluation

I used three metrics for the performance evaluation: the mean intersection over union (mIOU), precision, and recall. The mIOU is the average IOU of all classes (e.g. plumage area and non-plumage area for the Project Plumage dataset), and it is widely used as the evaluation metric for semantic segmentation (see section 3.1). The IOU of class i is $IOU_i = \frac{p_{ii}}{p_{ii} + p_{ij} + p_{ji}}$, where: p_{ii} are pixels of class i and classified as class i (true positive); p_{ij} are pixels of class i but classified as other classes (false negative); and p_{ji} are pixels of others classes classified as class i (false positive). IOU is a straightforward metric to measure the segmentation performance by combining aspects of both precision and recall but it can be useful to consider precision and recall separately. Precision shows the proportion of correct predictions whereas recall measures the segmentation area that the model does not predict. I used the following formulas for precision and recall of class i: $Precision_i = \frac{p_{ii}}{p_{ii} + p_{ji}}$, and $Recall_i = \frac{p_{ii}}{p_{ii} + p_{ij}}$. I used IOU and precision to measure the network performance as they both reflect well on the plumage area segmentation accuracy based on the project-specific goal of minimising the inclusion of non-plumage regions of the image. Achieving high recall is less critical but nonetheless important because I do not want results with excessively low recall (i.e. extremely conservative). Segmentations have only two classes (plumage area and non-plumage area) that are mutually exclusive, so mean metrics and

plumage area metrics are highly correlated and have similar values. I therefore report metrics based on the evaluation of the plumage area only rather than the mean (e.g. mIOU), which fits the target of this paper better (i.e. focused only on the plumage area).

I tested the effect of excluding the non-plumage area on the accuracy of colour measurements. To do this, I created segmentations using two kinds of morphological transformation: erosion and dilation (Haralick et al. 1987). Eroded segmentations have 100% precision (compared to the ground truth) and less than 100% recall. Dilated segmentations have 100% recall and less than 100% precision. Examples of these transformations are shown in Supplementary Figure 6.2.1b (erosion) and c (dilation). I created the erosion and dilation segmentations using a kernel (a structuring element) size of five applied for 1-9 iterations. The average IOU of each iteration ranged from 84.8% to 98.3%, and the average IOU difference of file-wise eroded minus dilated segmentations is -0.8%. The eroded segmentations and dilated segmentations have similar IOU, allowing the extracted colour information accuracies to be compared with respect to precision and recall.

I used the Earth mover's distance (EMD), a common metric to measure the similarity between two distributions, to calculate image similarity (Rubner et al. 2000). The EMD between the normalised histogram of a ground truth segmentation and its corresponding dilated or eroded segmentation was used to measure the colour information accuracy of each transformed segmentation. I then calculated the difference between EMD and IOU scores for eroded segmentations versus their corresponding dilated version (i.e. $EMD_{eroded} - EMD_{dilated}$ and $IOU_{eroded} - IOU_{dilated}$). Finally, I can quantify the effect of excluding the non-plumage area on the colour information accuracy by looking at the erode-dilated difference distribution of all segmentations.

3.2.3 Experimental manipulations to increase the model performance

The workflow described in section 3.2.2 uses unmodified RGB images with the largest possible resolution (618 x 410 pixels). I used workflow manipulations to test whether the following factors affect model performance: (i) image properties (the input resolution and input channels), (ii)

image augmentations, and (iii) subsetting neural networks. I trained and cross-validated all configurations to ensure robust validation results.

3.2.3.1 How does input resolution affect the performance?

DeepLab has been shown to perform better when using the original input resolution compared to resized resolutions (Chen et al. 2017b), in particular downscaled images result in lower accuracy for pose estimation and classification (Kim, Kwon Lee, and Mu Lee 2016, Chapter 2). I used resolutions that were 10 and 16 times lower than the input images (i.e. 494 x 328 pixels and 309 x 205 pixels) to test whether performance degrades at lower resolutions.

3.2.3.2 How does input channels affect the performance?

Previous studies have included non-visible light (e.g. UV and IR) information as the input in deep learning tasks, sometimes leading to better performance when compared to using only RGB channels (Basu et al. 2015; Potena et al. 2016; Milioto et al. 2018). The Project Plumage image data set includes two sets of images, one filtered to include only human visible (RGB) wavelengths and one to include only UV wavelengths, because bird plumage frequently includes UV reflecting regions. All images were taken against a black background made of theatre blackout curtains with very low reflectance of the UV light. The specimens should therefore reflect more UV light than the background. To test whether the inclusion of UV improved network performance, models were trained with i) images using UV channels only and ii) images using RGB plus UV channels.

3.2.3.3 How does image augmentation affect the performance?

Image augmentation is a common technique that increases the size of the training set by creating new labelled training images from manipulating the existing images and their labels, and has been shown to improve the model performance of DeepLab (Chen et al. 2017b). I created an augmented training set from the original training set in which images and their segmentations were randomly rotated (-15° to -1°, 1° to 15°), translated in both x and y axes (100 - 500 pixels), scaled (0.1 – 1.1). I used the augmented dataset to train the model with evaluation performed on the original validation set.

3.2.3.4 How does subsetting models affect the performance?

I used one model on images from all views (all-views model) for the model in section 3.2.2 and previous experiments, but image variations of different views may introduce difficulties for the network to learn. I therefore additionally trained and validated separate models for each of the three image views (back, belly and side). This reduces the input data for each model run to 1,698 images (compared to 5,094 images).

3.2.4 Applying classic methods to segment the Project Plumage dataset

In addition to examining different DeepLabv3+ configurations, I also assessed and compared the performance of classic computer vision techniques. Specifically, I used thresholding, region growing, Chan-Vese, and Graph cut from the OpenCV library on the expert-labelled dataset. I used image smoothing as the global pre-processing step, which is a commonly used pre-processing step for many classic segmentation methods. Images were converted to greyscale for thresholding. Along with segmenting the plumage area, thresholding will inevitably segment parts of the reflectance standards (standards cover the whole range of the greyscale values). I therefore selected the most upper connected component (the specimen is always placed above the reflectance standards and I assumed the segmented plumage area is not connected with other segmented parts) as the plumage area segmentation. I tested whether using the modal pixel value of the image with a positive offset of 15 performs better than Otsu's method (Otsu 1979) and adaptive thresholding methods, and I used this to threshold images. Many classic computer methods require spatial information as starting values (see Table 3.1). These are usually points within the focal region. I used body region predictions (2D points that are placed on specific bird body regions) as initial spatial information using data from my previous chapter (see Chapter 2). Region growing segments the initial pixels' neighbour pixels if the neighbour pixel values are within a certain range of the initial pixels' values, and it iterates the same procedure by examining the neighbour pixels of newly segmented pixels until no more pixels can be segmented (Adams and Bischof 1994). I tried 150 ranges from different upper (even numbers from 2 to 30) and lower (even numbers from 2 to 20) boundaries for region growing. I found that

the best combination for the highest IOU is a lower boundary of 6 and an upper boundary of 30 (see section 3.3.3 and Supplementary Figure 6.2.2). I made the initiated area for the Chan-Vese algorithm from squares with 20 pixels length around body region predictions and applied the algorithm for 100 iterations (Chan and Vese 2001). For the Graph Cut method (Boykov and Jolly 2001), I set the body region predictions as the foreground. The consistent setup for imaging specimens means that specimens would not be placed near the top, bottom, left and right boundaries, and would always be placed above the reflectance standards. I therefore set pixels within 20 pixels of the top, left and right edges and below the standard predictions as background. I applied the morphological close (close segmentation holes) and open (remove segmentation noises) to all results as a global post-processing step (Haralick et al. 1987). I then evaluated and compared the segmentation results from each computer vision model with the best result from DeepLabv3+.

3.2.5 Post-hoc tests on the model performance

3.2.5.1 Quality of the training data

Images from Project Plumage were taken in a highly consistent manner by controlling the placement of the specimen, light environment and background (Blagoderov et al. 2012; Hudson et al. 2015). However, not all datasets are likely to be so consistent due to practical limits (e.g. inadequate lightings). I tested whether greater variability in data quality could limit performance by generating lower quality datasets. To do this, I applied a series of image manipulations in which (i) images were rotated (angles between -45 to 45), translated (-500 to 500 pixels on x and y axes) and scaled (scale ratio from 0.8 to 1.2), (ii) 50% of images were randomly horizontally flipped, (iii) images were given new contrast and brightness (α from 0.5 to 2 and β from -50 to 50) using brightness and contrast adjustment functions in OpenCV (Bradski 2000; Bradski and Kaehler 2008), and (iv) a combination of manipulations from (i), (ii) and (iii). I applied these operations to both training images and validation images (in contrast to image augmentation outlined above where I did not manipulate the validation set). I trained and evaluated the transformed datasets under the same protocol as described in section 3.2.2.

3.2.5.2 The size of the training data

Another of my goals was to quantify the impact on performance when using smaller training sets. Previous studies suggest that larger training set sizes may improve the performances of deep learning models (Joulin et al. 2016; Hestness et al. 2017; Sun et al. 2017). I used a subset of 1,018 images (20% of the dataset) as the only validation set for every result in this section. The training set (4,076 images) was randomly sampled five times for one proportion selected from 15 proportions (1%, every 5% from 5% to 50% and every 10% from 50% to 90%). These training subsets and the original training set were used to train models under the same training protocol as described in section 3.2.2. IOU, precision and recall were then used to evaluate the results.

3.2.5.3 Effects of orders and plumage area properties

The expert-labelled dataset includes images representing the majority of bird genera and so includes diverse sizes, shapes and colours. I studied the relationship between the taxonomic information and model performance by testing for differences in the accuracy of predictions between orders (N=27). I also explored how other plumage area properties, including contrast and colour variability, affected performance. A plumage area with high contrast to its surrounding non-plumage area may be easier to segment in DeepLabv3+ than low contrast plumage areas (and many classic methods normally perform well on high contrast images). To test this I first calculated the mean of absolute Laplacian derivatives of images. A large Laplace derivative of a pixel indicates it is likely to be the part of an edge (high contrast around this pixel), which has been widely used for image edge detection (Berzins 1984; van Vliet et al. 1989). Then I used the pixels around the plumage area borders by using pixels of the difference between a dilated and eroded (both are applied with a kernel size of five for one iteration) segmentation. The mean derivatives of these pixels were used to represent the contrast between the foreground and the background. A large value means that the plumage area is very different from its surrounding non-plumage area and vice versa.

Plumage can be mono- or multi-coloured. I hypothesised that colour variability introduces difficulties for the model to segment the complete plumage area. Classic methods, like thresholding and region growing, can perform poorly on high colour variability images (see

section 3.3.3). To test this I used metrics from Chapter 2 to measure the colour variability: average pair-wise colour distances between body regions using RGB (both chromatic and achromatic information), normalised RGB (hue information), or lightness (achromatic information). I then evaluated correlations between metrics (IOU, precision and recall) of segmentations and plumage area properties.

3.2.6 Plumage colour diversity of world birds

3.2.6.1 Building bird plumage colour space

Using the segmentation pipeline evaluated earlier in this chapter, I applied the best-trained DeepLabv3+ to the whole Project Plumage dataset (122,610 images) and generated plumage colour data. The way of measuring colour depends on how pixels are sampled from the segmentation. Using all segmented pixels for every segmentation can create extremely large data sets (the average segmented pixel number of segmentations is 1,079,619). To make the data computationally tractable, I used data subsetted to 200 grid cells (provided by CRC), evenly placed on segmented pixels for each photo. The average pixel value within a grid cell was used as the colour information for one grid cell. Each photo has 200 colour points, and one specimen has 600 points (three photos per specimen). Since, for each species, there are up to six specimens (three per each sex), the number of colour points per species therefore ranged from 600 to 3,600.

In total, the segmentation data generated 24,309,400 plumage colour points (i.e. grid cells) from 121,547 photos (a small number of photos were excluded from the initial dataset due to problems associated with extracting correctly calibrated colour values), which covers 8,509 species, 178 families, and 34 avian orders. Among this data, there are 62,474 male specimens (7,873 species) and 57,013 female specimens (7,513 species). Colour data points (RGB value) were converted into u, s, m, l receptor cone stimulation values using the avian UVS visual (Endler and Mielke 2005; Stoddard and Prum 2008) model in pavo (Maia et al. 2013). These values were then mapped into a tetrahedral colour space.

3.2.6.2 Calculating and visualising colour diversity of bird plumage

I used the convex hull volume to measure the colour diversity of avian visual model colour spaces (using the segmentation data) across species for both sexes, male and female (Stoddard and Prum 2008; Renoult et al. 2017). Other measurements of the volume, like alpha shape (Cholewo and Love 1999; Gruson 2020), can reduce the possible overestimation associated with using convex hulls. However, alpha shapes of different sets of data points require data-specific parameter-tuning to make their volumes comparable (see Chapter 2). Here I focused on comparing the volumes across thousands of species and therefore only used the convex hull to measure the volume.

Using the segmentation allows the measurement of colours across the whole specimen. Therefore, I developed a flexible new metric called the proportional colour diversity which quantifies the colour diversity while accounting for the proportions of every colour. Specifically, the proportional colour diversity is the mean of average Euclidean distances to the centroid across all octants (using the centroid as the origin of octants). Octants are eight divisions of a 3D coordinate system (The equivalent term in 2D is quadrant). The colour points in tetrahedral colour space from pavo have 3D cartesian coordinates (x, y, z). I defined the equation of the proportional colour diversity as:

$$\frac{\sum_{i=1}^{8} \frac{\sum_{j=1}^{N_{Octant\_i}} Distance_{pt_j\ to\ centroid}}{N_{Octant\_i}}}{8}$$

The 2D version is defined as:

$$\frac{\sum_{i=1}^{4} \frac{\sum_{j=1}^{N_{quadrant\_i}} Distance_{pt_j\ to\ centroid}}{N_{quadrant\_i}}}{4}$$

Where N_{Octant_i} and $N_{quadrant\_i}$ are defined as the number of points in octant or quadrant i. Figure 3.2 shows an example of applying the proportional colour diversity in 2D.

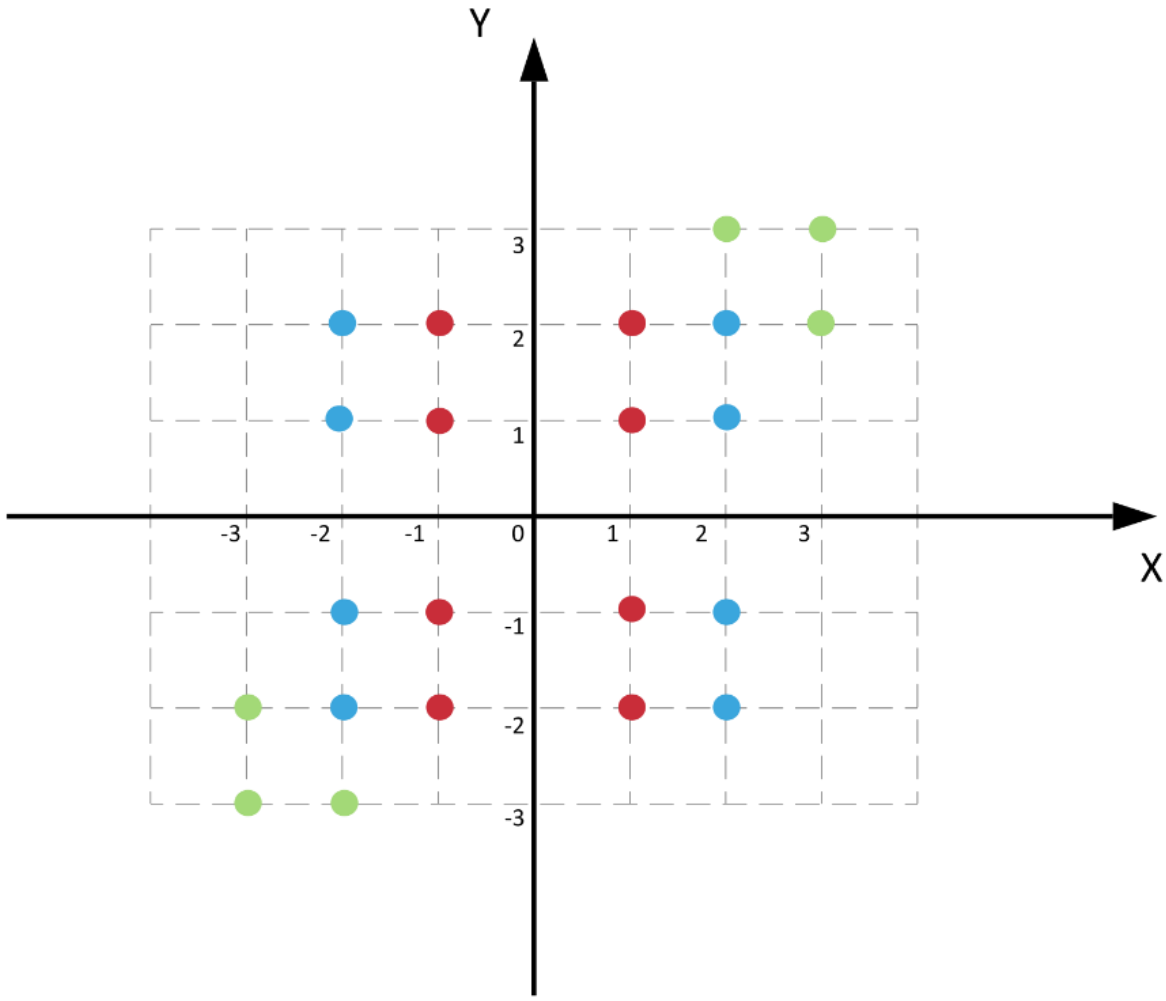


Figure 3.2. An example of calculating the proportional colour diversity for three groups of points (red, blue and green) in 2D. These three groups of points share the same centroid (0, 0). Based on the 2D proportional colour diversity equation, the result of the red group is: $(\frac{\sqrt{2}+\sqrt{5}}{2}+\frac{\sqrt{2}+\sqrt{5}}{2}+\frac{\sqrt{2}+\sqrt{5}}{2}+\frac{\sqrt{2}+\sqrt{5}}{2})/4 \approx 1.83$;The result of the blue group is: $(\frac{\sqrt{8}+\sqrt{5}}{2}+\frac{\sqrt{8}+\sqrt{5}}{2}+\frac{\sqrt{8}+\sqrt{5}}{2}+\frac{\sqrt{8}+\sqrt{5}}{2})/4 \approx 2.53$; The result of the green group is: $(\frac{\sqrt{18}+\sqrt{13}+\sqrt{13}}{3}+0+\frac{\sqrt{18}+\sqrt{13}+\sqrt{13}}{3}+0)/4 \approx 1.91.$

The proportional colour diversity takes account of octants estimates colour diversity rather than the colour span. This property is shown in Figure 3.2, of which the green group has smaller proportional colour diversity than the blue group. It has the advantage that it uses the average distance to centroid estimates diversity based on the distribution of the colour points rather than extreme values which can drastically affect the convex hull volume.

To explore the colour diversity in bird plumage colouration, I visualised variation in the convex hull volume and the proportional colour diversity across species for both sexes, male and female on a global phylogeny of birds. The tree (containing 9,993 species) used in this paper represents a composite maximum clade credibility tree produced by combining the backbone tree from Prum et al. (2015) with species-level trees from Jetz et al. (2012) (downloaded from www.birdtree.org). For full details see Cooney et al. (2017). The phylogeny was then pruned into sub-trees that match species in this data (both sexes: 8,509 species, male: 7,873, female: 7,513). Ancestral states were estimated for visualisation purposes only and are based on a Brownian motion model. Phylogenetic signals of the convex hull volume (colour diversity) and proportional colour diversity for both sexes, males-only, and females-only were calculated using Pagel's lambda (Pagel 1999). The R package phytools was used to estimate ancestral states and phylogenetic signal (Revell 2012). The phylogeny plots were made using the ggtree (Yu et al. 2017) and ggplot2 (Wickham 2011).

3.2.6.3 Comparing the patch and segmentation colour data

To show the differences between the segmentation colour data and the patch colour data (see Chapter 2), I compared widely used colour measures (the convex hull volume, colour span and hue disparity) in previous studies (Stoddard and Prum 2008; Maia et al. 2013) between the two datasets across species for both sexes, male and female. Because the proportional colour diversity metric was designed for the segmentation colour data (i.e. for data sets with many colour data points, rather than the 10 colour points from the patch data), it was not used in the comparison here. I calculated ratios of the convex hull volume, colour span and hue disparity from the patch data to those from the segmentation data ($\frac{Colour\ measure_{patch}}{Colour\ measure_{segmentation}}$).

3.3 Results

3.3.1 Accuracies of the DeepLab model

Across all three views, the original model achieved 93.1% IOU (per view, back: 94.6%; belly: 91.9%; side: 92.9%), 96.3% precision (per view, back: 96.8%; belly: 95.7%; side: 96.4%) and 96.6% recall (per view, back: 97.6%; belly: 95.8%; side 96.2%) in the evaluation (Table 3.2).

Table 3.2. IOU, precision and recall of predictions from the DeepLabV3+ model with the original configuration.

	MEAN	SD	MIN	MAX
IOU OVERALL (N=5094)	93.1	3.2	53.6	98.5
BACK (N=1698)	94.6	2.8	67.6	98.5
BELLY (N=1698)	91.9	3.4	53.6	97.2
SIDE (N=1698)	92.9	2.9	58.3	97.2
PRECISION OVERALL (N=5094)	96.3	2.4	70.1	99.9
BACK (N=1698)	96.8	2.5	70.2	99.9
BELLY (N=1698)	95.7	2.5	70.1	99.8
SIDE (N=1698)	96.4	2.0	70.9	99.9
RECALL OVERALL (N=5094)	96.6	2.5	53.6	99.9
BACK (N=1698)	97.6	2.0	72.0	99.9
BELLY (N=1698)	95.8	2.7	53.6	99.6
SIDE (N=1698)	96.2	2.4	58.3	99.4

88.8% of the segmentations (4,525 out of 5,094) had IOU higher than 90%. The lowest IOU is 53.6%. Four out of the worst five segmentations were caused by low recalls and all have precision higher than 85%. 97.9% of the segmentations (4,985 out of 5,094) had precision higher than 90%. The lowest precision was 70.0%. No segmentation had recall lower than 50%. Less than 0.2% of the results (7 out of 5,094, per view, back:1; belly:4; side:2) had recall lower than 75%, and less than 1.8% of the results (89 out of 5,094, per view, back:13; belly:43; side:33) had recall lower than 90%.

3.3.2 Experimental manipulations to increase the model performance

3.3.2.1 Effects of input resolution on the performance

I compared input image resolutions of 618 x 410, 494 x 328 and 309 x 205 pixels (8, 10 and 16 times lower than the original resolution). There was a significant effect of input image resolution on IOU (ANOVA: $F=1361.0$; d.f=2, 15279; $p<0.01$), precision (ANOVA: $F=1069.2$; d.f=2, 15279; $p<0.01$) and recall (ANOVA: $F=456.3$; d.f=2, 15279; $p<0.01$). The IOU and recall of 618 x 410 pixels and 494 x 328 pixels were not significantly different from each other, while the rest of the accuracies (IOU, precision and recall) were positively related to the input resolution (Supplementary Figure 6.2.3a, Supplementary Figure 6.2.4a). The image resolution of 618 x 410 pixels had the best overall performance, while the 309 x 205 pixels had the worst performance.

3.3.2.2 Effects of input channels on the performance

Images with UV and RGB + UV channels were also used to train models to compare the effect of different input channels (Supplementary Figure 6.2.3b). There was a significant effect of input channels on IOU (ANOVA: $F=395.6$; d.f=2, 15279; $p<0.01$), precision (ANOVA: $F=184.9$; d.f=2, 15279; $p<0.01$), and recall (ANOVA: $F=236.6$; d.f=2,15279; $p<0.01$). RGB was consistently better than UV and RGB+UV although the effects tended to be small (evaluation results of UV and UV+RGB were <2% worse than RGB as shown in Supplementary Figure 6.2.4b).

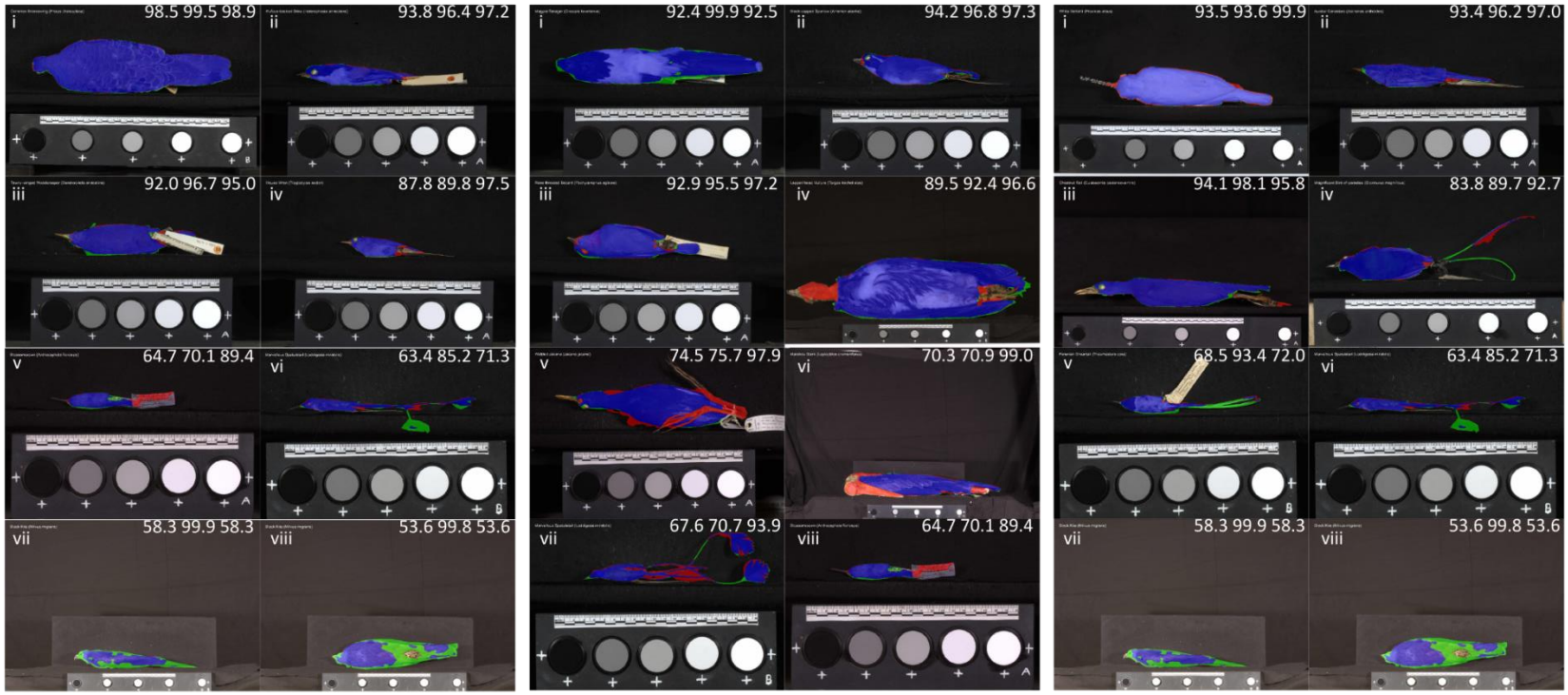
3.3.2.3 Effects of image augmentation on the performance

I then tested the effect of image augmentation (Supplementary Figure 6.2.3c). IOU was significantly higher ($t(10186)=5.90$, $p<0.05$) for the original dataset (Mean=93.1, Standard deviation (SD)=3.24) than the augmented dataset (Mean=92.7, SD=3.35). I also found a significant difference in the precision ($t(10186)=6.63$, $p<0.05$) and again the original dataset (Mean=96.3, SD=2.38) outperformed the augmented dataset (Mean=96.0, SD=2.56). However, there was no significant difference in the recall ($t(10186)=1.81$, $p=0.07$).

3.3.2.4 Effects of subsetting models on the performance

Supplementary Table 6.2.1 and Supplementary Figure 6.2.3 (d) shows that subsetting models per image view (i.e. back, belly, side) was significantly worse than using the all-views model, except for recall on the side view ($t(10186)=1.43$, $p=0.15$). The back view had the largest IOU difference (the all-views model has 0.7 higher IOU than divided models) and recall difference (recall of the all-views model is 0.5 higher), while the side view had the largest precision difference (precision of the all-views model is 0.4 higher).

Overall, none of the experimental manipulations improved model performance compared to the original model. I therefore used predictions from the original model as the benchmark result for comparison with classic models (see next section). Many examples correctly classified eyes and labels as non-plumage area (e.g. Figure 3.3a.ii, a.iii and a.iv). 3 out of 4 (Figure 3.3a.vi, a.vii and a.viii) of the worst IOU segmentations were caused by the low recall problem (large green areas). The two worst recall examples (Figure 3.3c.vii and c.viii) had many plumage areas un-detected, and these images have light black backgrounds and long camera distances (due to the large size of the specimens). Other low recall examples failed to detect complete tails as these tails are extremely thin or irregular (Figure 3.3c.v and c.vi). Thin tails can also cause low precision: the model misclassified background around thin tails as plumage area (Figure 3.3b.vii). Legs that are placed on the top of the plumage area can be hard for the model to exclude (Figure 3.3b.v). Figure 3.3b.vi shows an example of misclassifying an irregular beak as the plumage area.


i 98.5 99.5 98.9
ii 93.8 96.4 97.2
iii 92.0 96.7 95.0
iv 87.8 89.8 97.5
v 64.7 70.1 89.4
vi 63.4 85.2 71.3
vii 58.3 99.9 58.3
viii 53.6 99.8 53.6
i 92.4 99.9 92.5
ii 94.2 96.8 97.3
iii 92.9 95.5 97.2
iv 89.5 92.4 96.6
v 74.5 75.7 97.9
vi 70.3 70.9 99.0
vii 67.6 70.7 93.9
viii 64.7 70.1 89.4
i 93.5 93.6 99.9
ii 93.4 96.2 97.0
iii 94.1 98.1 95.8
iv 83.8 89.7 92.7
v 68.5 93.4 72.0
vi 63.4 85.2 71.3
vii 58.3 99.9 58.3
viii 53.6 99.8 53.6

a b c

Figure 3.3. Images of the best, 50th, 75th and 95th percentile (ranked by metrics from high to low; from i to iv) and 4 worst predictions (from v to viii) based on a) IOU, b) precision and c) recall. The IOU, precision and recall (from left to right) are displayed on the top right corner of each image. Blue is correctly predicted by the model (True positive); Red is the non-plumage area that has been classified as plumage area by the model (False positive); Green is the plumage area that has been classified as non-plumage area (False Negative).

3.3.3 Accuracies of classic methods on segmenting the Project Plumage dataset

I compared the results of the best DeepLab configuration with four classic computer vision segmentation methods (thresholding, region growing, Chan-Vese and GraphCut). I used the region growing result with the lower boundary of 6 and the upper boundary of 30, which achieved the best IOU (Supplementary Figure 6.2.2). IOU varied significantly among segmentation methods (ANOVA: $F=3141.3$; $d.f=4, 25465$; $p<0.01$), as did precision (ANOVA: $F=1678.6$; $d.f=4, 25465$; $p<0.01$) and recall (ANOVA: $F=1989.6$; $d.f=4, 25465$; $p<0.01$). DeepLab had a superior performance for IOU, precision and recall, as well as the lowest variance in performance metrics, compared to classic methods (Figure 3.4). Specifically, DeepLab outperformed classic results by at least 23.4% on IOU, 6.4% on precision and 9.5% on recall (Supplementary Figure 6.2.4c). Graph cut had the best IOU among tested classic methods, while Chan-Vese had the best precision and Thresholding had the best recall, suggesting Graph cut was the overall best classic method in plumage images, Chan-Vese segmented area conservatively, and thresholding tend to segment lots of non-plumage regions.

The worst examples from classic methods were clearly far worse than those from DeepLabv3+. Examples in Figure 3.5 show that dark plumage, high plumage colour variability and museum labels can be obstacles for classic methods whereas DeepLabv3+ predicted accurately on the same images.

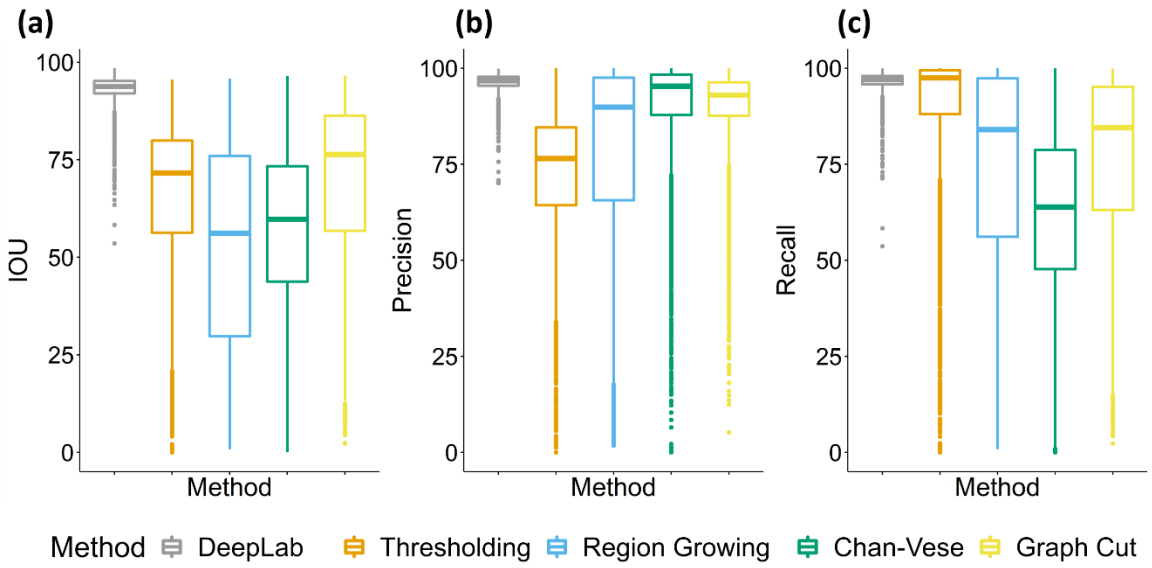


Figure 3.4. The performance of predictions (N=5,094) from DeepLabv3+ and tested classic methods (thresholding, region growing, Chan-Vese and Graph cut) for (a) IOU, (b) Precision, and (c) Recall.

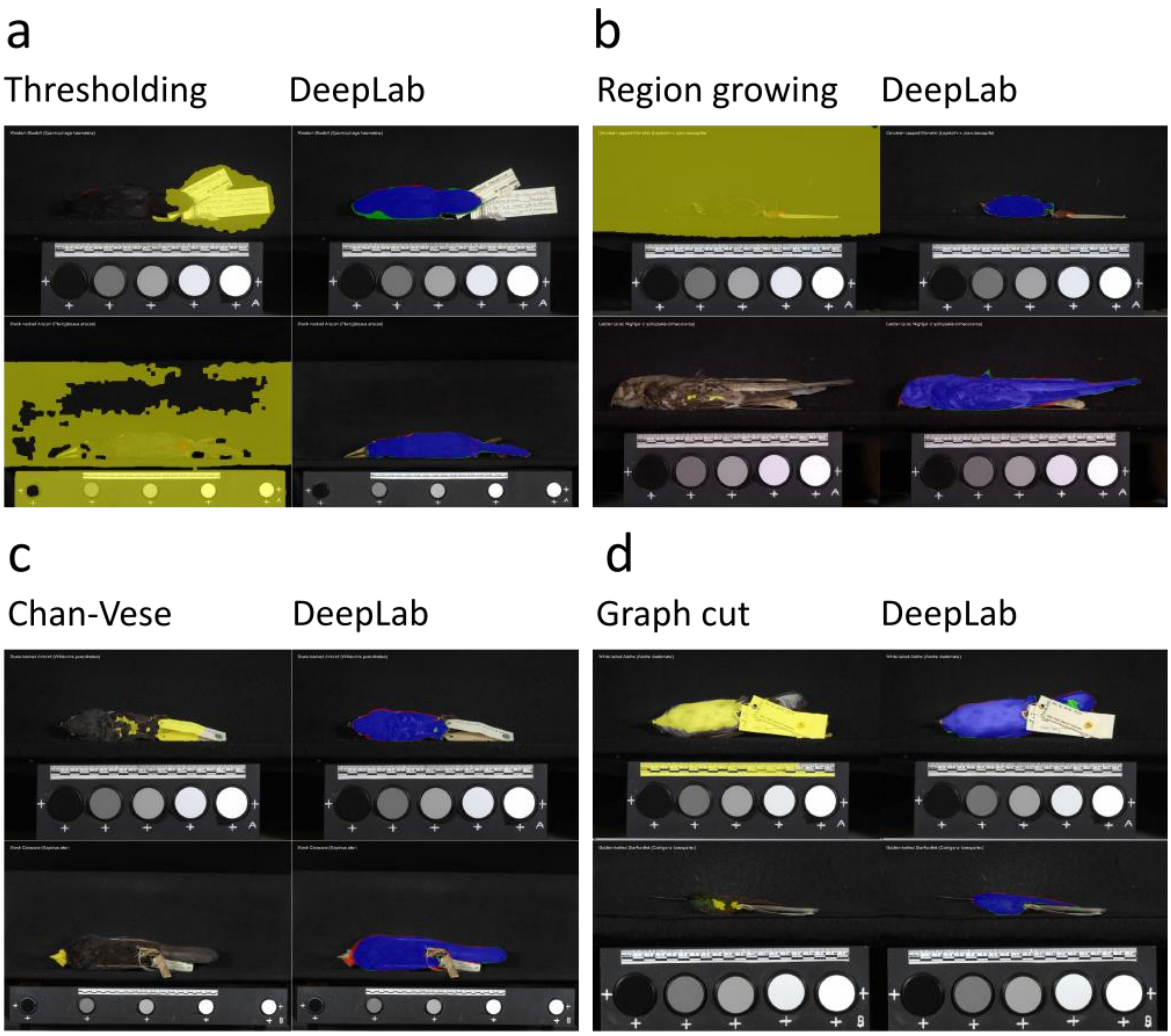


Figure 3.5. Examples of poorly segmented results from classic methods as well as deep learning predictions of the corresponding images. (a) Thresholding; (b) Region growing; (c) Chan-Vese; (d) Graph Cut. Yellow is the segmentation from classic methods. Deep learning results are represented in blue, red and green as defined in Figure 3.3.

3.3.4 Post-hoc tests on the model performance

3.3.4.1 Results of eroded and dilated segmentations

The distribution between IOU difference and EMD difference is shown in Supplementary Figure 6.2.5. IOU difference (Mean=-0.8, SD=0.99, $t(50939)=-186.7$, $p<0.05$) and EMD difference (Mean=9.2E-5, SD=1.2E-4, $t(50939)=-178.8$, $p<0.05$) were both significantly smaller than 0. The result shows that dilated segmentations had slightly higher IOU than the corresponding eroded

ones. The mean and median of EMD difference ($EMD_{erode} - EMD_{dilated}$) were smaller than zero which show that eroded segmentations have more similar frequency distributions to the ground truth than the corresponding dilated ones do.

3.3.4.2 Quality of the training data

I assessed how DeepLabv3+ performed on four low-quality datasets (see section 3.2.5.1). Low-quality datasets had a significant negative effect on the IOU (ANOVA: F=205.3; df=4.0, 25465; p<0.01), precision (ANOVA: F=132.8; df=4.0, 25465; p<0.01) and recall (ANOVA: F=88.0; df=4.0, 25465; p<0.01). The original dataset produced more accurate results than low-quality datasets (Supplementary Figure 6.2.6). The 4th dataset (the combination of translation, rotation, scale, horizontal flip and manipulations of brightness and contrast) had the worst performance and examples of its predictions are shown in Supplementary Figure 6.2.7. The 4th dataset was 1.9%, 1.2% and 0.9% worse than the original dataset on IOU, precision and recall (Supplementary Figure 6.2.4d).

3.3.4.3 Size of the training data

Model performance was positively related to the training set size following an approximately logarithmic pattern (Supplementary Figure 6.2.8). At least 10% of the dataset was required to attain IOU higher than 90%, at least 5% of the dataset to get precision and recall higher than 90%, and 15% of the dataset for precision and recall higher than 95%. With 100% of the dataset used for training, the model achieved 93.3% for IOU, 96.3% for precision and 96.8% for recall.

3.3.4.4 Effects of orders and plumage area properties

Supplementary Table 6.2.2 shows the average IOU, precision and recall per order. The Galliformes had the lowest IOU (84.6%) and recall (88.8%), whereas Ciconiiformes had the lowest precision (88.2%). However, the expert-labelled dataset had only 15 (~0.3% of the dataset) Galliformes and 9 (~0.2%) Ciconiiformes specimens. Columbiformes had the highest IOU (95.2%) and precision (97.8%). The most abundant order was the Passeriformes with 3408 images (~66.9% of the total, compared to 7.1% for the second most abundant order), having IOU of 93.3% (15th of 27 orders), precision of 96.4% (14th of 27 orders), and recall of 96.6% (15th of 27 orders).

Examples with the highest and lowest plumage area contrast and colour variability scores are shown in Supplementary Figure 6.2.9. IOU, precision and recall generally declined with increasing plumage area contrast (Supplementary Figure 6.2.10a) and increasing colour variability (Supplementary Figure 6.2.10b,c,d), but the correlations were weak and appear to be driven in part by lower sample sizes at higher contrasts and greater colour variability values.

3.3.5 Plumage colour diversity of world birds

3.3.5.1 Calculating and visualising colour diversity of bird plumage

Supplementary Table 6.2.3-Supplementary Table 6.2.5 show the top 200 bird species ranked by convex hull volume based on the segmented data. The top 200 includes 34 families (top family: Psittacidae; 47 species), 14 orders (top order: Passeriformes; 97 species) for both sexes; 35 families (top family: Psittacidae; 50 species), 12 orders (top order: Passeriformes; 108 species) for male; 43 families (top family: Psittacidae; 59 species), 17 orders (top order: Psittaciformes; 59 species) for female. The top three species ranked by male species colour volume using the segmentation data are the Paradise Tanager (*Tangara chilensis*), Red-legged honeycreeper (*Cyanerpes cyaneus*) and Mrs. Gould's sunbird (*Aethopyga gouldiae*) (Figure 3.6a). The top ranked species by female colour volume is the Paradise Tanager (*Tangara chilensis*), but the second and third ranked species - black-crowned pitta (*Pitta ussheri*) and Garnet pitta (*Pitta granatina*) - differ from male ranking (Figure 3.6b).

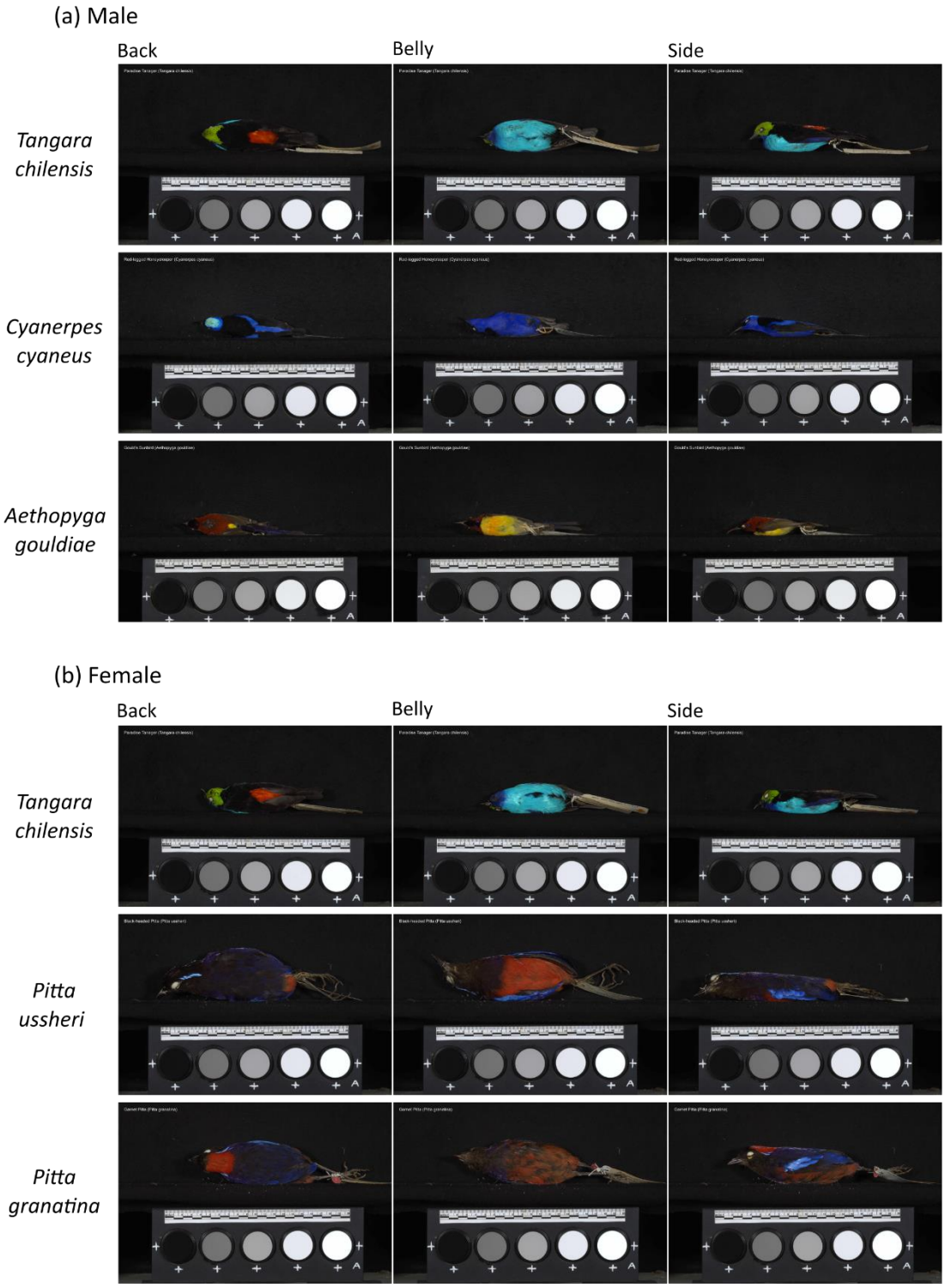


Figure 3.6. (a) Back (first column), belly (second column) and side (third column) images of the top 3 convex hull volume male species (row 1-3: *Tangara chilensis, Cyanerpes cyaneus and Aethopyga gouldiae*)

and (b) the top 3 convex hull volume female species (row 1-3: *Tangara chilensis, Pitta ussheri and Pitta granatina*).

Supplementary Table 6.2.6-Supplementary Table 6.2.8 show the top 200 species ranked by proportional colour diversity. The top 200 includes 30 families (top family: Psittacidae; 59 species), 5 orders (top order: Passeriformes; 106 species) for both sexes; 34 families (top family: Psittacidae; 53 species), 6 orders (top order: Passeriformes; 118 species) for male; 32 families (top family: Psittacidae; 70 species), 9 orders (top order: Psittaciformes; 70 species) for female. The top three species ranked by proportional colour diversity in males are the Red-legged honeycreeper (*Cyanerpes cyaneus*), Purple honeycreeper (*Cyanerpes caeruleus*) and Collared lory (*Phigys solitarius*) (Figure 3.7a). In contrast, the top 3 female species are Papuan Lorikeet (*Charmosyna papou*), Collared lory (*Phigys solitarius*) and Scarlet macaw (*Ara macao*) are shown in Figure 3.7b. There are 62 species (both sexes list), 71 species (male list) and 82 species (female list) shared between the top 200 convex hull volume and proportional colour diversity lists.

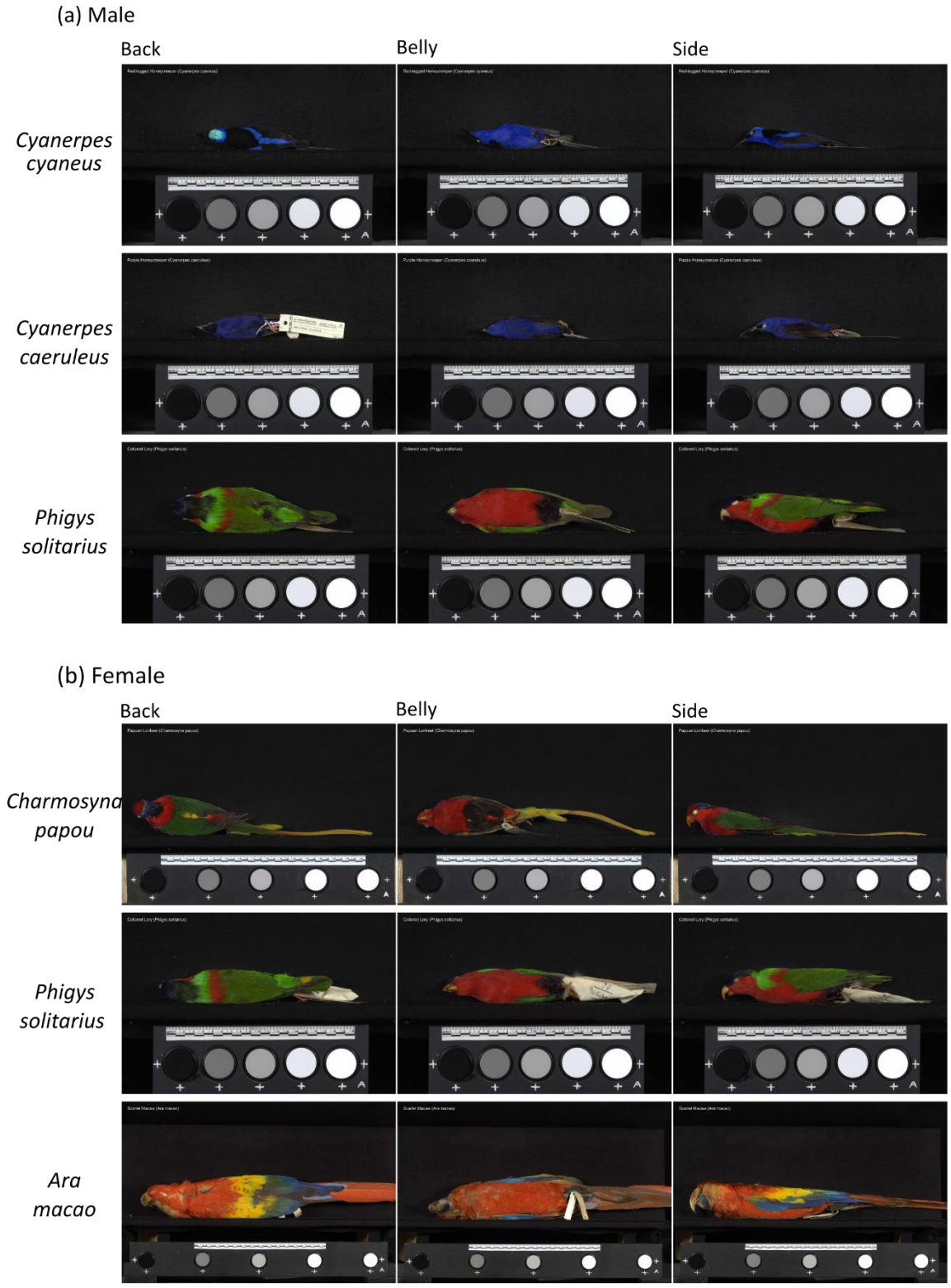
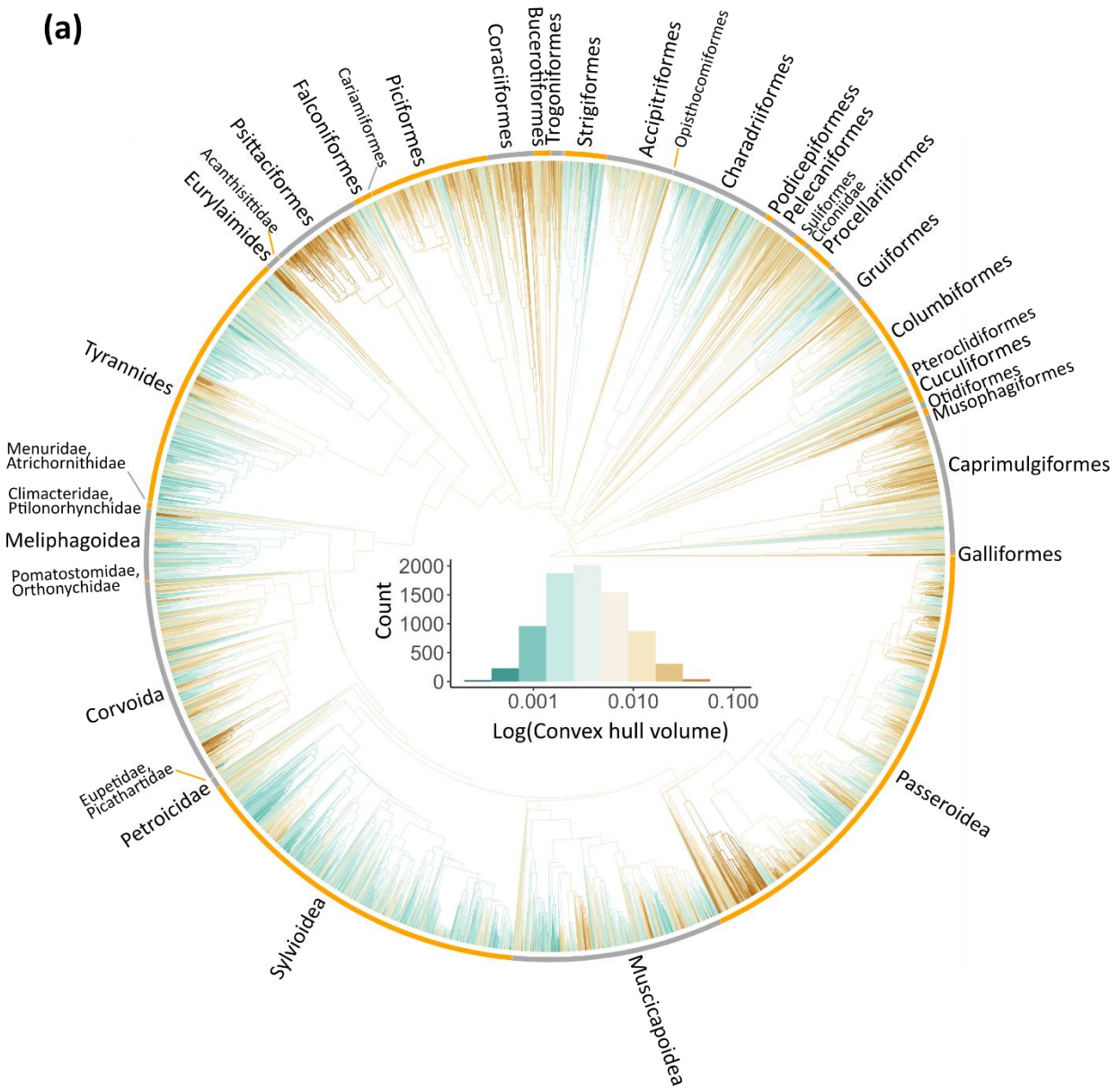


Figure 3.7. (a) Back (first column), belly (second column) and side (third column) images of the top 3 proportional colour diversity male species (row 1-3: *Cyanerpes cyaneus*, *Cyanerpes caeruleus and Phigys*

solitarius) and (b) the top 3 proportional colour diversity female species (row 1-3: *Charmosyna papou, Phigys solitarius and Ara macao*).

Phylogenetic visualisations of the convex hull volume in plumage colour evolution across male and female species are shown in Figure 3.8. This highlights that high convex hull volume is conserved within certain orders (Psittaciformes and Coraciiformes). Pagel's Lambda for the convex hull volume across the whole phylogeny are 0.743 (both sexes; lambda=0: likelihood ratio (LR)=4555.89; p<0.001.), 0.789 (males; lambda=0: LR=4958.65; p<0.001.) and 0.719 (females; lambda=0: LR=3659.7; p<0.001.). Phylogenetic visualisations of proportional colour diversity in plumage colour evolution across male and female species are shown in Figure 3.9. Species in orders of Psittaciformes and Coraciiformes also generally have high proportional colour diversity. Pagel's Lambda for the proportional colour diversity across the whole phylogeny are 0.876 (both sexes; lambda=0: LR=6499.31; p<0.001.), 0.883 (males; lambda=0: LR=5892.02; p<0.001.) and 0.838 (females; lambda=0: LR=4807.65; p<0.001.). The result indicates that both colour volume and proportional colour diversity have strong correlations to phylogenetic signals but with significant departures from a Brownian motion null model. Phylogenetic signals in proportional colour diversity are higher than colour volume signals, and male colour has a higher signal than female colour within each colour diversity measure.


(a)
Coraciiformes
Bucerotiformes
Trogoniformes
Strigiformes
Accipitriformes
Opisthocomiformes
Charadriiformes
Podicepiformes
Pelecaniformes
Suliformes
Ciconiidae
Procellariiformes
Gruiformes
Columbiformes
Pteroclidiformes
Cuculiformes
Otidiformes
Musophagiformes
Caprimulgiformes
Galliformes
Passeroidea
Muscicapoidea
Sylvioidea
Petroicidae
Eupetidae,
Picathartidae
Corvoida
Pomatostomidae,
Orthonychidae
Meliphagoidea
Climacteridae,
Ptilonorhynchidae
Menuridae,
Atrichornithidae
Tyrannides
Eurylaimides
Acanthisittidae
Psittaciformes
Falconiformes
Cariamiformes
Piciformes
Count
2000
1500
1000
500
0
0.001
0.010
0.100
Log(Convex hull volume)

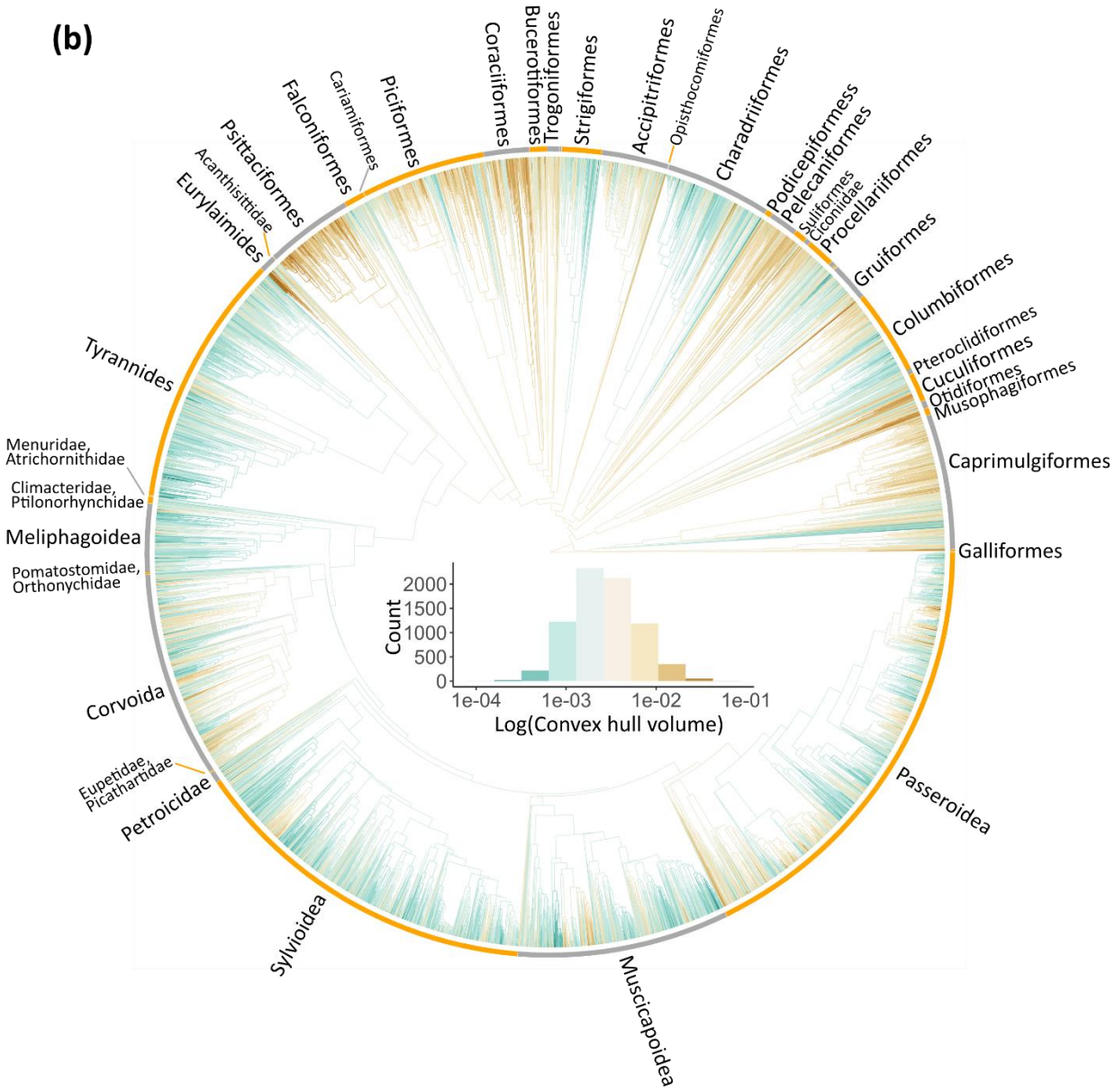
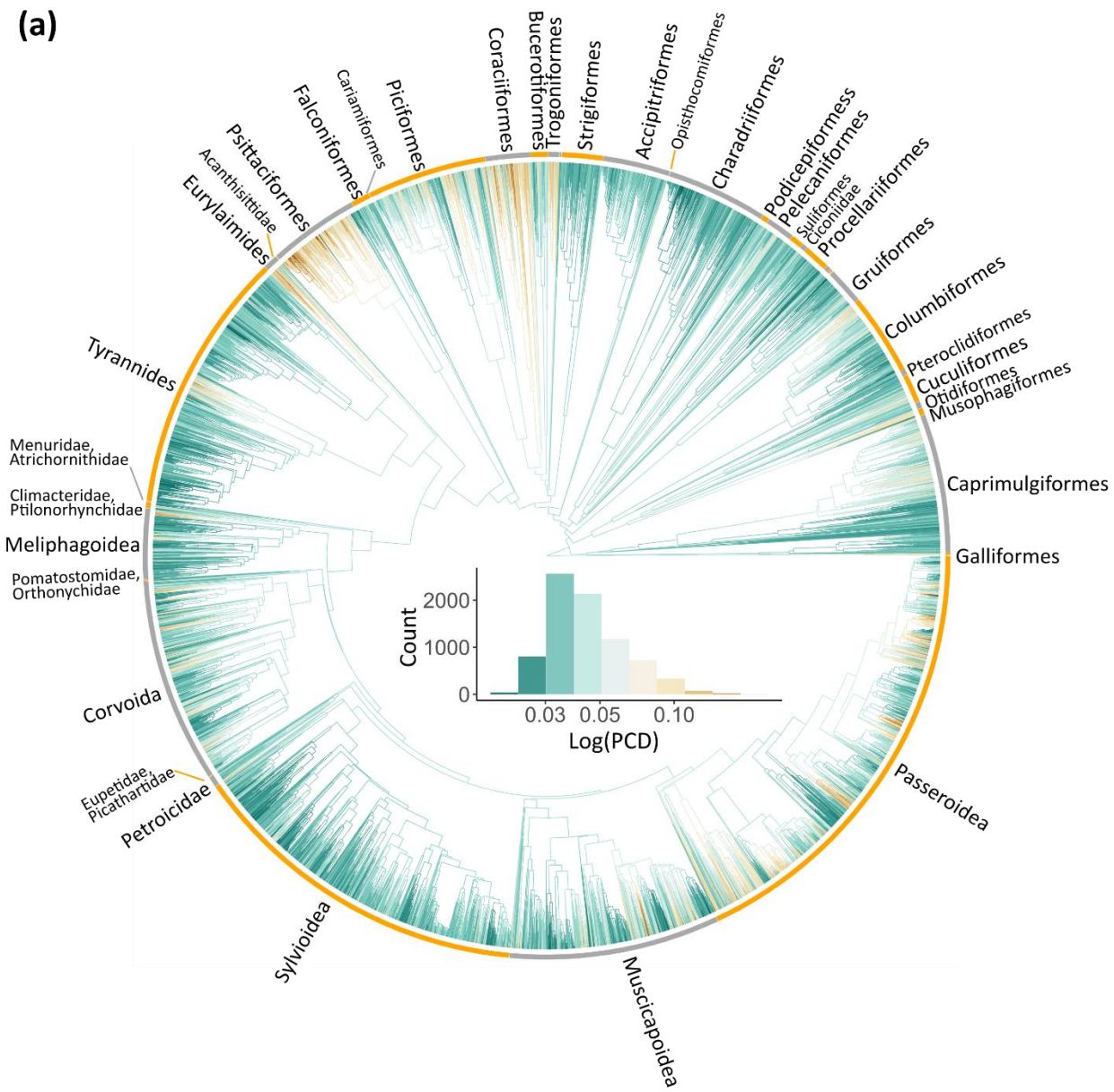


Figure 3.8. Phylogenetic visualisation of convex hull volume (log transformed) in plumage colour of (a) males (across 7,873 species) and (b) females (across 7,513 species) from the segmentation data. Histograms (inset) show the convex hull volume distribution (log transformed)


(a)
Coraciiformes
Bucerotiformes
Trogoniformes
Strigiformes
Accipitriformes
Opisthocomiformes
Charadriiformes
Podicipediformes
Pelecaniformes
Suliformes
Ciconiidae
Procellariiformes
Gruiformes
Columbiformes
Pteroclidiformes
Cuculiformes
Otidiformes
Musophagiformes
Caprimulgiformes
Galliformes
Passeroidea
Muscicapoidea
Sylvioidea
Petroicidae
Eupetidae,
Picathartidae
Corvoida
Pomatostomidae,
Orthonychidae
Meliphagoidea
Climacteridae,
Ptilonorhynchidae
Menuridae,
Atrichornithidae
Tyrannides
Eurylaimides
Acanthisittidae
Psittaciformes
Falconiformes
Cariamiformes
Piciformes
Count
2000
1000
0
0.03
0.05
0.10
Log(PCD)

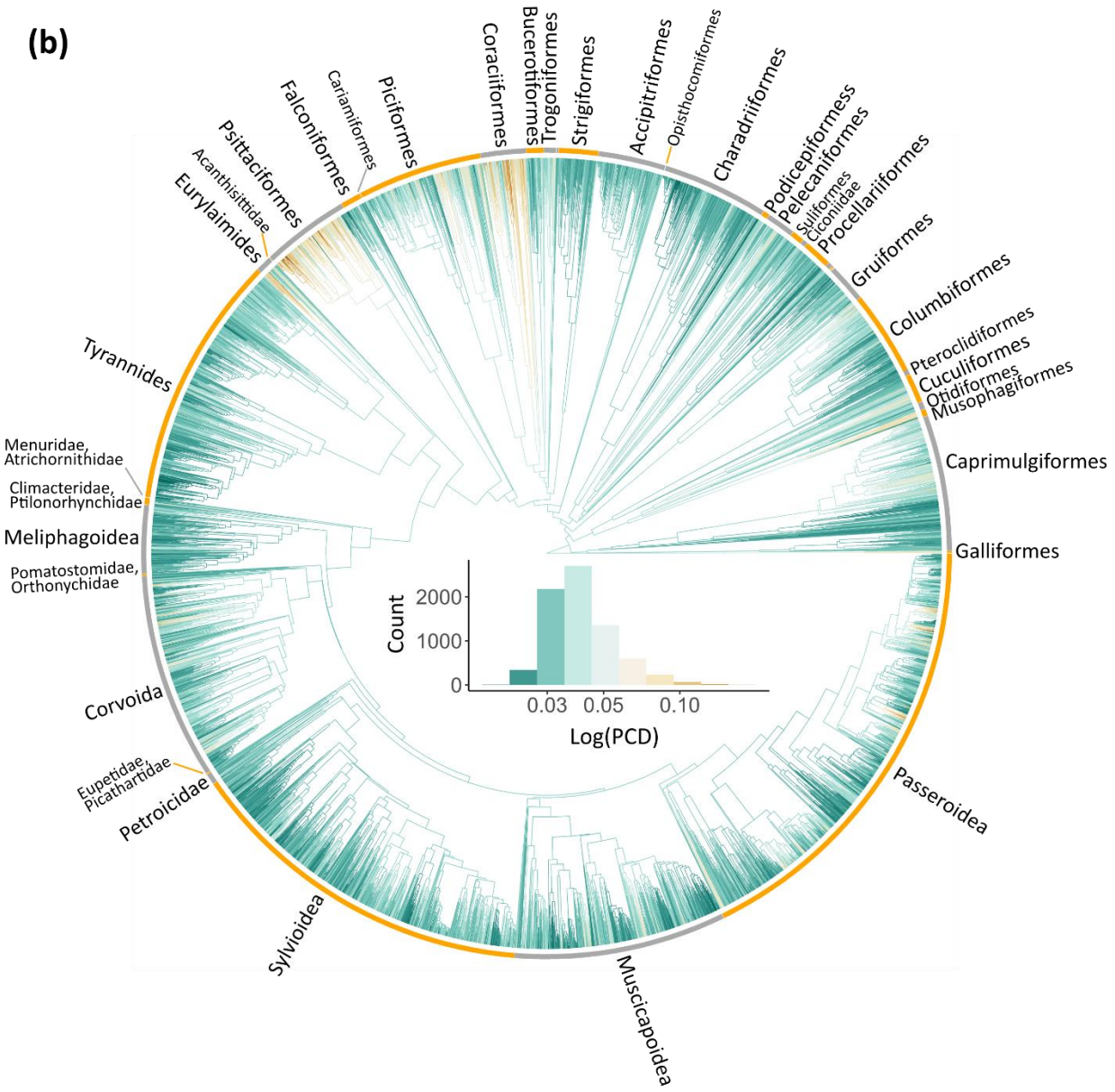


Figure 3.9. Phylogenetic visualisation of the proportional colour diversity (log transformed) in plumage colour of (a) males (across 7,873 species) and (b) females (across 7,513 species) from the segmentation data. Histograms (inset) show the proportional colour diversity distribution (log transformed)

3.3.5.2 Comparing the patch and segmentation colour data

For both sexes, male and female, all ratios (convex hull volume, colour span and hue disparity) were significantly smaller than 1 as shown in Supplementary Figure 6.2.11. The convex hull volume ratio was the smallest among the three tested measures. 8,499 species (99.9% out of

8,509) for both sexes, 7,866 (99.9% out of 7,873) species for male, and 7,510 species (99.9% out of 7,513) for female had larger convex hull volumes based on the segmentation data than the patch data. 6,275 species (73.7% out of 8,509) for both sexes, 5,564 (70.7% out of 7,873) species for male, and 5,694 species (75.8% out of 7,513) for female had larger colour spans based on the segmentation data than the patch data. 6,474 species (76.1% out of 8,509) for both sexes, 5,780 (73.4% out of 7,873) species for male, and 5,901 species (78.5% out of 7,513) for female had great hue disparity based on the segmentation data than the patch data.

3.4 Discussion

3.4.1 Segmentation methods

DeepLab, a semantic segmentation method using deep learning, can automatically segment generally accurate bird plumage areas from other parts across more than 120,000 Project Plumage image. The results show that segmentation using DeepLab strongly outperformed all classic computer vision methods. Indeed segmentations from classic methods are frequently so poor that they would often be unusable for downstream analyses of colour. Of the classic methods, Graph cut had the best average plumage area IOU but was 23.4% worse than the average IOU from DeepLabv3+. In contrast to the DeepLab predictions, images with dark birds and prominent label tags could not be reliably segmented using classic methods. Dark birds were normally under or over segmented, and label tags were included as plumage area (e.g. Figure 3.5a). Besides deficiencies shown in these examples, setting starting parameters for classic methods, for example choosing threshold values for thresholding and region growing by hand-crafted image features, is a troublesome task (Chang and Li 1994; Fan et al. 2001). They are suitable for segmenting objects that are clearly different from the backgrounds. Deep learning models learn to detect image features automatically from training and so require expert labelling only for a subset of images. Classic methods such as Chan-Vese and Graph cut require some starting spatial information for all images limiting their use for large data sets of thousands of images such as Project Plumage. Here, I used the spatial information derived from deep learning point predictions (Chapter 2) to seed classic methods because there is not an automatic and simple way to generate accurate spatial information for this data, other than from manual or

deep learning. However, other projects may not have sufficient digitisation information. Based on accuracy and image complexity, I conclude that deep learning is far more suitable than classic methods in segmenting the plumage area.

I tested how reliably DeepLabv3+ segments data from Project Plumage. Experimental configurations were evaluated to identify configurations that maximised model performance. I found that input image resolution had positive effects on performance, as expected and previously reported for DeepLabv2 (Chen et al. 2017b). Image augmentation, additional channels and subsetting models did not improve the performance. In my experimental runs, the best performance overall was achieved with DeepLabv3+ and an input resolution of 618 x 410 pixels. The IOU was about 93% which is higher than DeepLabv3+'s performance (mIOU: 89.0%) on the standard PASCAL VOC 2012 data set (Chen et al. 2018). This higher level performance may be due to the lower complexity of the Project Plumage dataset compared to PASCAL because (i) the Project Plumage dataset has only two classes (plumage and non-plumage) while PASCAL dataset has 21 classes (Everingham et al. 2015) and (ii) images consist of few and fixed focal objects (one) under a consistent, high resolution imaging setup. In contrast, the PASCAL images are more varied (e.g. different objects, backgrounds). These two factors may explain why no improvements were observed with image augmentation, additional channels and subsetting models as the model had already been trained well with the original dataset and configurations.

DeepLabV3+ can identify the plumage area (precision: 96.3%) and plumage area completeness (recall: 96.6%) reliably. Figure 3.3 and Figure 3.4 illustrate most of the predictions are reliable and therefore usable. Unwanted parts of the images include eyes and specimen labels. Eyes are small and labels often resemble tails in terms of position and shape - they are considered hard to be excluded by the model. However, DeepLab did surprisingly well in excluding them. It took less than 3 days to train a DeepLabV3+ model with training set size of 4,074 (80% of 5,094 images from the dataset splitting) images and trained for 31 epochs and less than five days to generate predictions for the 122,610 images that constitute the full Project Plumage dataset on one NVIDIA GTX 1080 (12GB GPU memory). Although the predictions are sometimes less accurate than expert manual segmentations, DeepLabv3+ can generate usable segmentations quickly and save a huge amount of time.

Ideas about how deep learning can perform on less perfect museum digitised datasets (small size or low-quality datasets) were tested, and the results on tested datasets were promising. Large training sets are commonly used in deep learning. DeepLab (Chen et al. 2017b) used a training set size of 1400 images in PASCAL VOC 2012 (Everingham et al. 2015) and 2975 images in Cityscapes (Cordts et al. 2016). Tasks like classification and pose estimation have used even larger datasets, such as 1.2 million training images in ImageNet classification (Deng et al. 2009) and more than 28,000 images in MPII pose estimation challenge (Andriluka et al. 2014). My results are consistent with previous studies that the training set size was positively related to the model performance (Joulin et al. 2016; Hestness et al. 2017). However, small training set sizes did not decrease the performance drastically. It is possible to use just 15% of the original dataset (~600 images) to generate segmentations with 90% IOU on 1,018 validation images. This is still much more accurate than results using classic methods' results. The highly consistent imaging layout in Project Plumage may reduce the size of training data needed to get an acceptable result from deep learning. Modern pipelines for museum collection digitization typically follow similarly consistent standards such as uniform specimen placements, background and light environment (Hudson et al. 2015; Unger et al. 2016; Hussein et al. 2020) suggesting that such data can be analysed with deep learning. However, high standard digitisation is time-consuming. I simulated varied camera distances, angles, specimen placements, orientation and light environments by manipulating original images to create less consistent images. Using low-quality datasets did not provide excessively inaccurate predictions, and the worst performance (low-quality dataset 4) was much better than classic methods' results. This result along with promising results on low consistent datasets such as PASCAL VOC 2012 shows that the DeepLab is likely to be robust on less consistent datasets.

3.4.2 Bird plumage colour space

I used predicted segmentations to assess and visualise the phylogenetic distribution of bird plumage colour diversity and found that the colour diversity is normally distributed and shows a moderately strong phylogenetic signal. I used the segmentation data to estimate and visualise plumage colour diversity that covers more than 85 % of the total 9,993 bird species (79.7% of species for males only and 75.1% of species for females only). Lambda values on both the convex

hull value and the proportional colour diversity show that there are strong phylogenetic signals in colour diversity for all three configurations (both sexes, male-only and female-only). The signals in the proportional colour diversity are stronger than signals in the convex hull volume. The result of phylogenetic signals indicates that the proportional colour diversity is consistent with the idea that colour producing mechanisms constrain the evolution of plumage colour (Stoddard and Prum 2011). Phylogenetic signals in colouration can also occur if closely related species share similar selection pressures (e.g. most birds of prey are camouflaged brown/white). Birds with limited mechanisms can evolve large overall colour volume (i.e. convex hull) by taking extreme values along a limited number of axes whereas large proportional colour diversity is achieved by evolving along multiple divergent colour axes. Under the same possible colour volume, measuring the average distance can therefore be more reliable than measuring the convex hull volume (affected a lot by extreme values or the shape of the data points) while is sensitive to whether colour points are diverse. The proportional colour diversity metric is therefore likely to be a more reliable proxy for the range of colour producing mechanisms which themselves may be phylogenetically conserved.

The proportional colour diversity can estimate proportions of colours and reveals different patterns compared to the colour volume. Supplementary Figure 6.2.12a and b show the colour spaces of *Tangara chilensis* (2nd for the convex hull volume; 140th for the proportional colour diversity) and *Buthraupis montana* (480th for the convex hull volume; 9th for the proportional colour diversity). The majority of colour points are located densely near the centroid for the *Tangara chilensis* colour space, with less dense clusters stretching to more extreme colours. In contrast, *Buthraupis montana's* colour points are distributed in two similarly dense clusters. *Buthraupis montana* therefore has higher proportional diversity than *Tangara chilensis*.

More striking differences in the inference of the volume versus proportional colour diversity metric emerge when the colour space includes outliers. The proportional colour diversity metric is robust to outlier colour points which have extreme stimulation values in the tetrahedral space. Supplementary Figure 6.2.12c shows the colour space of *Rostrhamus sociabilis* (10th for the convex hull volume; 4,979th for the proportional colour diversity). There is one large outlier in the shortwave receptor cone, which affects the convex hull volume drastically. This outlier may

be due to prediction or colour processing error. The robustness to outliers of this measure may explain why there are fewer orders and families represented in the top 200 species for the proportional diversity metric compared to the convex hull volume (see section 3.3.5.1 and Supplementary Table 6.2.3-Supplementary Table 6.2.8).

Body region measurement (e.g. point or patch) and segmentation are two types of measurements on colour information and they can be used in different situations. Body region measurement is used to measure colour information on specimens for many projects (Stoddard and Prum 2008, 2011; Dale et al. 2015; Cooney et al. 2019, see also Chapter 2). The top 5 species for males (*Charmosyna papou*, *Tangara chilensis*, *Amazona albifrons*, *Trichoglossus rubritorquis* and *Erythrura gouldiae*) in Stoddard and Prum (2011) are ranked 23rd, 1st, 224th, N/A (Project Plumage dataset does not have images for this species) and 137th in my male colour spaces using the segmentation data. While they are ranked 11th, 1st, 515th, N/A (missing species) and 33th in the male colour spaces using the patch data. It is intuitive to use body region measurement if the project focuses on the colour of body regions. Comparing to the region measurement, segmentation focuses on the whole plumage area of specimens, which captures complete colour information on specimens. The segmentation data has higher colour span, hue disparity and convex hull volume than the patch data (Supplementary Figure 6.2.11) in the project plumage photos. The result shows that the colour points of the segmentation are more diverse than colour points from patches of 10 body regions. When measuring the overall colour of specimens, using segmentation may be a better method than region measurements. Extracting colours evenly on three views of a specimen can introduce duplicate colour measurements (e.g. apart from the side view, wings may appear in the back and belly view). Better measuring methods on segmentation should be designed to reduce the duplication problem. The segmentation of the Project Plumage dataset retains spatial information of plumage colours, which could be used to further quantify plumage patterns (Van Belleghem et al. 2018) across broad avian species. This would allow, for example, studies of the evolution of camouflage patterns (Troscianko et al. 2016; Stevens et al. 2017).

An example where segmentation captures more colour information than patch-based measure is the convex hull volume rank for male *Amazona albifrons* (224th for the segmentation data,

515th for the patch data among 7,873 male species). Here, the head has four distinct colours in the human visible spectrum (blue, green, red and white). However, the patch method only measures the average value in the crown patch, which fail to capture colour diversity in the head, while the segmentation method captures the extent of the whole head.

3.4.3 Possible improvements and application suggestions

Biologists should choose methods within their knowledge to improve deep learning results. Normally, better deep learning model architectures generate more accurate results. However, designing networks requires a lot of mathematical and statistical expertise. Finding an optimal hyperparameter configuration on a wider hyperparameter space can potentially increase model accuracy but tuning hyperparameters is a tedious task. The trade-off between segmentation accuracy and expertise or time may not fit the goal of phenotyping museum collections. Ways of improving segmentation accuracy that are feasible and easy to apply for biologists include manual correction and post-processing on predictions, which are intuitive and easy to apply. A manual checking and editing workflow is a feasible step for the optimisation (see Chapter 2). Images can be prioritized for checking, for example by checking those from high error rate orders first. The difference between eroded and dilated segmentations (Supplementary Figure 6.2.5) have shown that segmenting inside the plumage tended to be more similar to the ground truth. Image erosion can be applied as a post-processing method to make more conservative segmentations. Therefore, increasing the possibility of placing segmentations inside the plumage area. However, the use of automated processes such as this should be used with caution and may sacrifice some IOU and recall to improve precision.

Globally, biological and museum collections represent a huge and relatively underused resource. My study demonstrates that segmenting images using deep learning is a powerful and promising approach for high-throughput data extraction on collections. Projects, like Project Plumage (122,610 images) require a large amount of time and resource to digitise collections often using highly standardised setups and manually labelling (5,094 images in the Project Plumage analysis presented here). The results in section 3.3.4.2 and 3.3.4.3 (results about the size and quality of the training data) implied that it is possible to lower criteria without sacrificing a great deal in the

segmentation accuracy. For setting up new projects, I therefore recommend pilot studies digitising a small number of images with flexible imaging setups followed by manually segmenting images to create the labelled dataset, then training and evaluating the model using this dataset. If deep learning results are reliable and usable, continue with relaxed pipelines, which can save time and resources. Otherwise, iterate with a higher quality standardized digitising setup and more labelled training images until predictions are accurate enough or the model performance stops increasing.

3.5 Conclusion

With the help of deep neural networks and the growth of large-scale biological data, it is possible to develop a high-throughput segmentation pipeline on digitised data. Deep learning can segment bird specimen photos accurately, while other classic segmentation methods can not, at least on the data assessed here. The colour information extracted from the Project Plumage data is accurate enough to use in analysis and could be further optimised with time-saving checking and correction steps. Here I demonstrate the utility of the segmented colour data from Project Plumage by visualising plumage colour diversity across species on a global bird phylogeny and showing that plumage colour diversity has strong phylogenetic signals.

Chapter 4 PhenoLearn: A software package and workflow for annotating digital images of biological data

Abstract

Measuring phenotypic traits on digitised data is becoming more and more common in ecological and evolutionary studies. Due to the number of museum collections and requirements from biological questions (e.g. macroevolution and comparative studies), collection digitisation can generate large scale datasets. Measuring traits on digital images normally requires annotations (e.g. landmark points). Placing annotations accurately with high-throughput pipelines is important to shorten the time span of a project. Deep learning is state-of-the-art for predicting points and segmentations on images automatically and has great potential to be applied to many digitised natural history datasets which would otherwise take extensive human effort to measure. I introduce PhenoLearn, an open-source image analysis tool to apply deep learning on digitised biological specimens. PhenoLearn currently supports the placement of points and the segmentation of objects within images. Users can use the software to annotate images and create training sets, which are used to train deep learning models. Models can be evaluated and then selected to predict annotations for whole datasets. PhenoLearn also has features to review images and predictions with high efficiency. PhenoLearn has a user interface to visualised through these functions, which reduce the expertise of deep learning to use it. Along with its deep learning based pipeline, PhenoLearn may be used to place annotations on any image dataset according to the needs. Here I describe the main features of PhenoLearn and use it to place landmarks on photos of shells of the marine gastropod genus *Littorina*. I use the landmarks to compare deep learning with manual landmarking in detecting morphological differences between two shell ecotypes (‘crab’ vs ‘wave’).

4.1 Introduction

Measuring traits on digitised specimens is increasingly used to phenotype specimens for a range of tasks. Similar to analysing medical imaging data using annotations like points and segmentations (Balafar et al. 2010; Mharib et al. 2012), studies have used annotations to measure phenotypic traits on digitised specimen data such as images (Zelditch et al. 2015; Chang and Alfaro 2016; Ravinet et al. 2016; Cooney et al. 2019) and 3D scans (Cooney et al. 2017; Felice and Goswami 2017; Giacomini et al. 2019). Digital imaging allows rapid and non-invasive measurements on natural history collections and can be used to make detailed image annotations, including points, polygons (e.g. bounding boxes), and segmentations, which can mobilise collections for biological analyses. Digitising specimens and mobilising datasets are two major goals to utilise the full potential of natural history collections (Blagoderov et al. 2012). Although there are many specimens yet to be digitised (an estimated 1.2-1.9 billion specimens in natural history collections globally from Ariño 2010), new workflows and techniques, like the tray scanning system (Blagoderov et al. 2010), whole drawer imaging (Mantle, la Salle, and Fisher 2012; Holovachov, Zatushevsky, and Shydlovsky 2014) and automatic metadata recognition (Heidorn and Wei 2008; Drinkwater, Cubey, and Haston 2014) have increased the speed of digitisation. While placing annotations on digitised collections manually can be time-consuming, especially for datasets at a macroevolutionary-scale. Robust, high-throughput data extraction (e.g. phenotyping or trait measurements) pipelines are necessary.

The most common way to place annotations is by placing them visually with the help of interactive software tools. For example, Fiji, a distribution of the opensource software ImageJ, is a powerful toolbox for biological image analysis (Rasband and others 1997; Schindelin et al. 2012). Users can annotate images along with image processing, analysis and visualisation using a straightforward user-interface. Users can also write and contribute plugins for custom methods. Other popular annotation tools such as LabelIMG (Tzutalin 2015) are easy to install, user-friendly and open-source. Web-based annotation tools, such as VGG image annotator (Dutta and Zisserman 2019), are also increasingly popular. Web-based applications make crowdsourcing annotation possible. For example, Amazon Mechanical Turk has been used to landmark images

of fish (Chang and Alfaro 2016), while citizen scientists have helped to place landmarks on thousands of 3D bird beak scans (Cooney et al. 2017) and to identify phenological states of more than 7000 thousand herbarium specimen images (Park et al. 2019).

Annotation opens up a wide range of scientific enquiry. For examples, point placement is perhaps the most commonly used annotation and can measure the exact pixel location on images. Point annotations can be used as landmarks to identify morphological features on specimens such as homologous point (Adams, Rohlf, and Slice 2013; Bookstein 1991). In Chapter 2 I used the point to locate body regions to measure colour. Examples of large scale studies (i.e. more than 10 landmarks on hundreds to thousands of images) that used points to measure phenotypic traits are listed in Table 4.1. Other studies also used polygons to measure focal regions, for example, studies have placed more than ten thousand polygons on bird images to measure plumage colours of body regions (Dale et al. 2015; Cooney et al. 2019). Placing one polygon normally takes a longer time than placing a point, as points are basic components for polygons (i.e. points need to be placed as vertices). Segmentation is a commonly used method to annotate regions of interest in medical images like cells (Meijering 2012; Xing and Yang 2016), brains (Balafar et al. 2010) and livers (Mharib et al. 2012). Segmentation has also been used in phenotyping live plant photos (Minervini et al. 2014; Scharr et al. 2016). My previous chapter (Chapter 3) has segmented the whole bird plumage area on photos to measure the plumage colour diversity of birds. Segmentation is powerful to capture region-based features (e.g. plumage areas). However, the applications and many aspects of segmenting digitised specimen images for phenotypic measurements are yet to be explored.

Table 4.1. Studies and datasets that have used point annotations to measure traits on digitised specimens.

Dataset	Annotations	Reference
741 images of lanternfish specimens taken in lateral view.	Measure the overall body shape: • 23 Landmarks	(Denton and Adams 2015)
1677 images of squirrel mandibles taken in lateral view.	Measure the mandible shape: • 14 Landmarks	(Zelditch et al. 2015)
7044 images of sigmodontine rodents (2402 of skulls taken in ventral view; 2401 of skulls taken in lateral view; 2241 of mandibles taken in lateral view).	Measure the skull shape: • Ventral view of skulls: 56 landmarks • Lateral view of skulls: 19 landmarks Measure the mandible shape: • lateral view of mandibles: 13 landmarks	(Maestri et al. 2017)
359 images of lizard heads taken in dorsal view.	Measure the head shape: • 28 landmarks	(Lazić et al. 2015)
5,631 images of bird specimens	Locate the calibration standards: • All views: 5 points	(Cooney et al. 2019)
5,094 images of bird specimens (1,698 images taken for dorsal, lateral and ventral view)	Locate the calibration standards: • All views: 5 points Locate body regions: • Dorsal view: 5 points • Lateral view: 2 points • Ventral view: 3 points	Chapter 2

The types of datasets highlighted above often need thousands or tens of thousands of annotations. The factors of time and human work may even limit the ideal dataset sizes for some studies. If it is possible to find ways to move away from completely manual annotation, the data contained within digital images can potentially be unlocked much more efficiently. The recent success of deep learning networks for many imaging processing has made the possibility of high-throughput placement of accurate annotations on digital specimen images a reality. Deep learning networks (Insafutdinov et al. 2016; Newell et al. 2016) have placed point annotations accurately to identify human joints and body parts on human pose estimation datasets (Andriluka et al. 2014). Semantic segmentation methods (Chen et al. 2017b, 2018) have generated reliable results on segmenting images into 21 categories (Everingham et al. 2015). Network architectures

have been designed to segment biomedical images (Ronneberger et al. 2015; Li et al. 2018). My previous works have applied deep learning networks to place points (Chapter 2) and segmentations (Chapter 3) on digitised bird images. Deep learning predictions were accurate for downstream analyses, and I have used predictions to calculate colour diversities of birds across more than 7,000 species for the overall bird and individual regions.

Taken together, these examples show the potential for deep learning in high-throughput phenotypic measurements on large specimen digitisation datasets. Tools such as MATLAB (MATLAB 2018) and Supervisely (https://supervise.ly/) provide users with a workflow for annotating a part of the dataset as the training set, selecting a deep learning model from a model pool for training, and predict the whole dataset using the trained model. No coding is involved and the tools are accessible but limited to the set of provided deep learning tools.

There are many computational tools and packages for phenotypic analysis. Tools like geomorph (Adams and Otárola-Castillo 2013) and MorphoJ (Klingenberg 2011) quantify, analyse and visualise morphological and shape information using landmarks as the input (Adams and Otárola-Castillo 2013). R packages like pavo (Maia et al. 2013) and patternize (Van Belleghem et al. 2018) visualise and analyse colours and patterns using pixel values from images as the input. With these computational packages, annotations can be converted to meaningful traits quickly. Therefore the step of measuring digitised datasets is crucial between digitisation and phenotypic analysis. A robust, high-throughput data extraction (e.g. phenotyping or trait measurements) pipeline can bring the whole process from the digitisation to datasets of specimen phenotypes faster.

To enable rapid annotations from large digitisation datasets (i.e. more than thousands of images), I aimed to build a standalone, open-source and user-friendly software tool for biologists, which implements an innovative phenotypic annotation workflow using deep learning. Here I introduce PhenoLearn, an image analysis tool for deep learning on digitised biological (and other) specimens. The PhenoLearn software is designed for use after the initial data digitisation (imaging) step to annotate data using either the point placement or segmentation prior to user-specific data analyses as shown in Figure 4.1. PhenoLearn's functions enable users to (i) build a training set by manual labelling (for the current version, annotations include the point and segmentation);

(ii) train deep learning models, and use the best-performed model to predict the whole dataset; (iii) review and correct the predictions to increase accuracy. I first described the core functionality of PhenoLearn. I then used PhenoLearn to locate landmarks on photos of the shells of the marine gastropod genus *Littorina*. Finally, I used the output from PhenoLearn to build a shell-shape morphospace and tested for environmental variation in shell shape relating to the dominance of selective processes of ecology (crab predation) compared to the environment (wave action) following Ravinet et al. (2016).

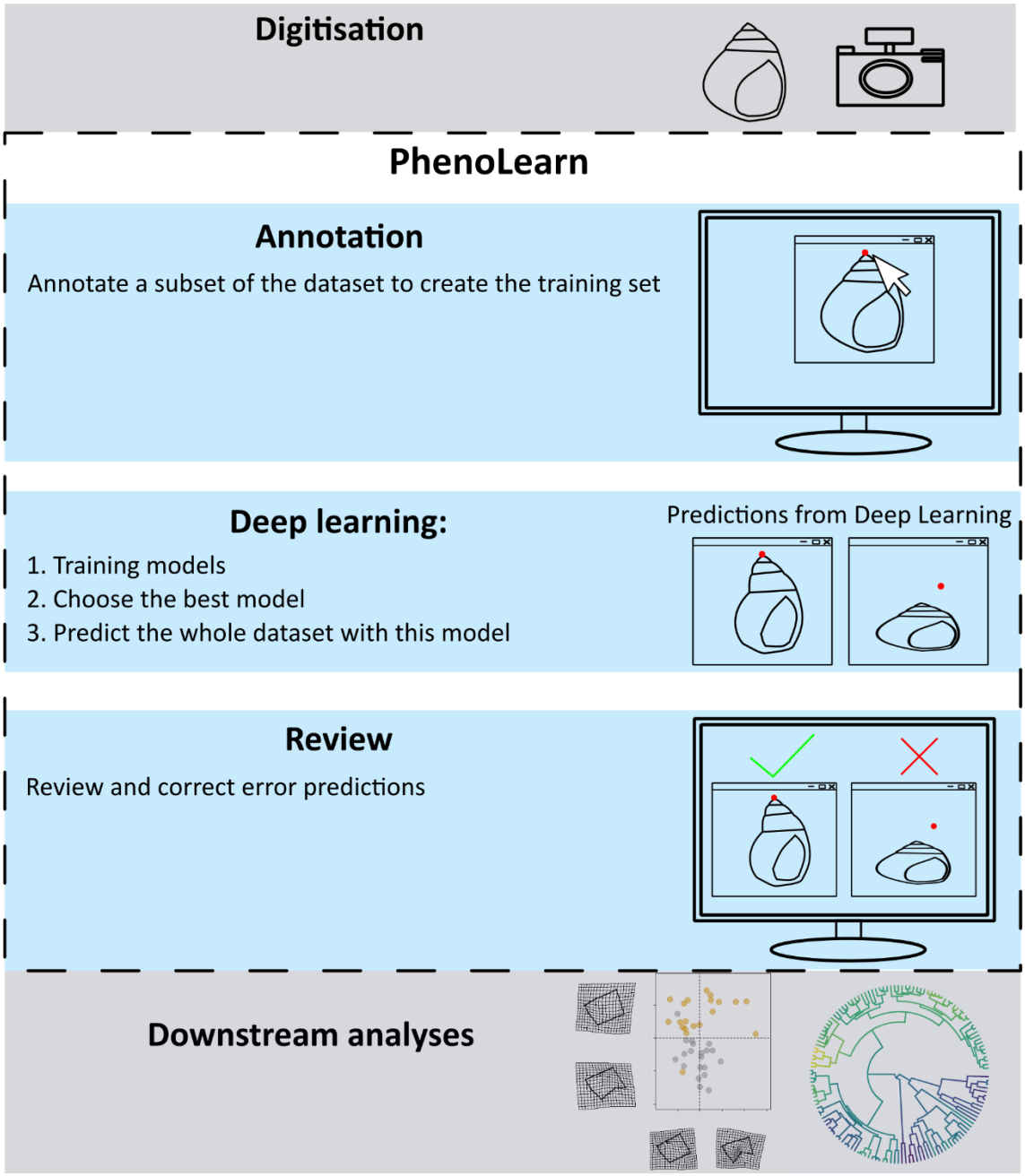


Figure 4.1. The workflow from digitisation to phenotypic analysis and major modules of PhenoLearn.

4.2 Software Description

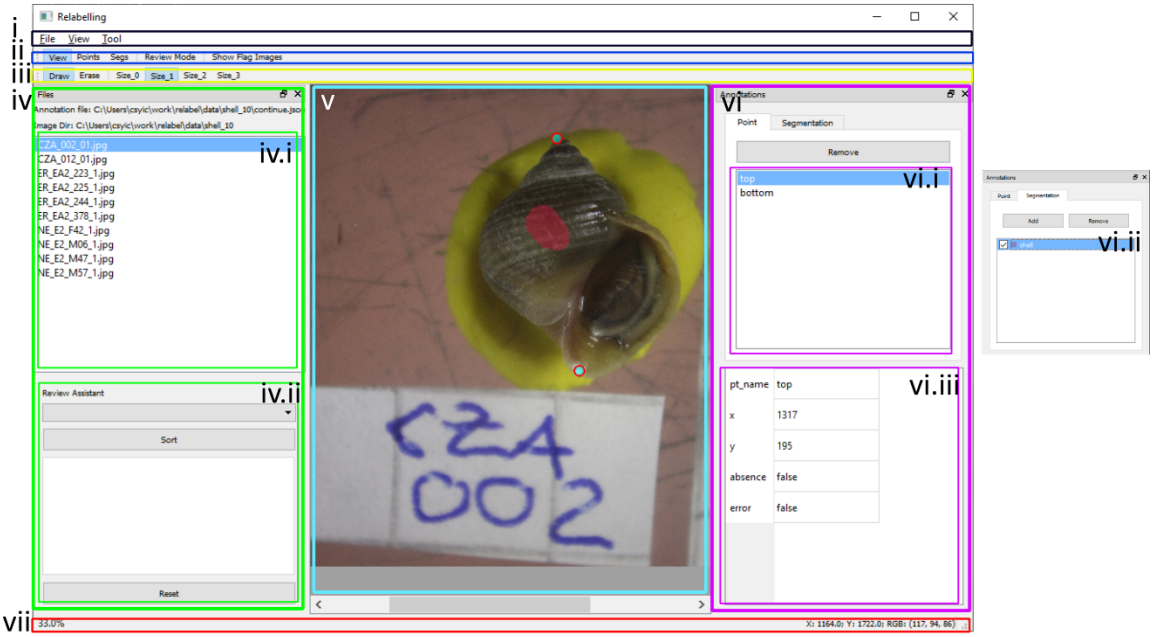
Below I describe the interface and main analytical components of a typical PhenoLearn workflow which covers annotating images, using deep learning to predict annotations and reviewing

predictions. PhenoLearn was written in Python 3.7 and tested on Mac OS and Windows 10. Deep learning methods were implemented using TensorFlow (Abadi et al. 2016), a machine and deep learning library. I provide a more detailed exploration of the software as the user manual in section 6.4.

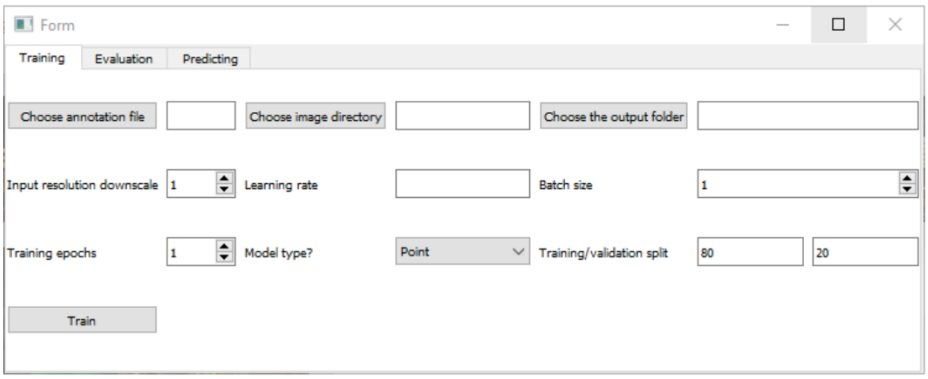
4.2.1 User interface

The main user interface for PhenoLearn is shown in Figure 4.2a. PhenoLearn has three main viewing panels: the file panel (Figure 4.2a.iv), the visualisation panel (Figure 4.2a.v) and the annotation panel (Figure 4.2a.vi). The file panel shows the data at the file level and each image in the dataset is listed in this panel. After the user selects a file, the image is displayed in the visualisation panel, and it allows users to interact with the image and place annotations. The visualisation panel also has a review mode (Figure 4.2c) that views multiple images increasing the efficiency of reviewing existed annotations. The review assistant can be used to filter images for reviewing (Figure 4.2a.iv.ii). The annotation panel shows annotation details and has two tabs (Figure 4.2a.vi.i and Figure 4.2a.vi.ii) for two types of annotations (points and segmentations), and each tab lists the names of existing annotations. The property editor (Figure 4.2a.vi.iii) below the annotation list shows properties of the selected annotation and allows users to change them as required. The menu (Figure 4.2a.i) has functions such as loading, saving annotations and zooming images. The mode bar (Figure 4.2a.ii) contains modes such as viewing and labelling images. The tool bar (Figure 4.2a.iii) shows tools for annotation (e.g. paint or erase segmented regions). The status bar (Figure 4.2a.vii), at the bottom of the window, shows the status such as the pixel coordinates and RGB value of the mouse location on the image. Functions that include deep learning are displayed in another interface, which is mainly used to read configurations for training (Figure 4.2b), evaluation and prediction.

(a)


i
ii
iii
iv
v
vi
vii
iv.i
iv.ii
vi.i
vi.ii
vi.iii
Relabelling
File View Tool
Review Assistant
Sort
Reset
Point
Segmentation
Remove
top
bottom
pt_name top
x 1317
y 195
absence false
error false

(b)


Form
Training
Evaluation
Predicting
Choose annotation file
Choose image directory
Choose the output folder
Input resolution downscale
Learning rate
Batch size
Training epochs
Model type?
Point
Training/validation split
Train

(c)

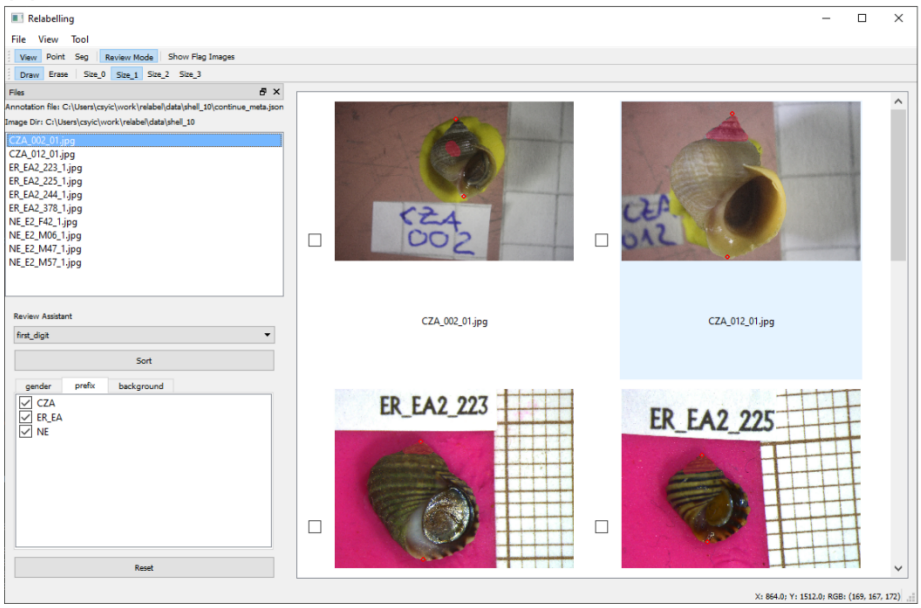

Relabelling
Review Assistant
Sort
Reset
CZA_002_01.jpg
CZA_012_01.jpg
ER_EA2_223
ER_EA2_225

Figure 4.2. The user interface of PhenoLearn. (a) The main interface consists of (i) Menu; (ii) Mode bar; (iii) Tool bar; (iv) The file panel displays image names in (iv.i) the file list; (iv.ii) The review assistant is used in reviewing images and is located in the bottom of the file panel; (v) The visualisation panel shows the selected image and its annotations; (vi) The annotation panel lists point names in (vi.i) the point tab. (vi.ii) The segmentation tab is the alternative tab of the annotation panel. Properties of the selected annotation are listed in (vi.iii) the property editor. (vii) The status bar shows information about the image. (b) The interface for configuring the training process. (c) The interface when the review mode is activated. The visualisation panel displays multiple images and their annotations. The annotation panel is hidden and the review assistant shows specimen characteristics that can be used to filter images.

4.2.2 Annotating digitised specimens

A high-quality training set is essential for deep learning. Although there is not a defined gold standard for the size of the training set, training sets of many well-known deep learning studies include thousands of images (Szegedy et al. 2014; He et al. 2016; Newell et al. 2016; Chen et al. 2018). Studies have shown that the size of the training set normally has a positive effect on the deep learning model accuracy (Joulin et al. 2016; Hestness et al. 2017; Sun et al. 2017). However, in Chapter 3, I was able to achieve high-quality segmentations on photos of bird specimens using about 200 training images. In any situation, the training set should be representative of the range of image types in the full data to maximise the quality of deep learning predictions. Random sampling is a common way to sample training images to avoid bias training sets.

PhenoLearn (www.github.com/EchanHe/PhenoLearn) provides tools to efficiently create and visually check training sets for the point annotation and segmentation of 2D images (See software manual and Supplementary Figure 6.3.1 for full instructions on the annotation process). The annotation process is started by selecting a folder containing the training images. The user can then view the images and add annotations (Supplementary Figure 6.3.2a shows an example of adding a point and Supplementary Figure 6.3.2b shows an example of adding segmentation areas). Each annotation (i.e. individual point or segmentation) is given a user-defined name that can be descriptive or simply a set of unique numeric identifies. Users can add any number of annotations and can also edit and delete points or segmentations for each image (Supplementary Figure 6.3.2c shows an example of deleting a point). The annotation file is saved as a JSON file

with image and annotation details stored using dictionary and array data types (see section 6.3.1 and Supplementary Figure 6.3.3). An annotation process can be continued at any time by loading a saved annotation file from an unfinished process.

The two deep learning networks used in PhenoLearn, Stacked Hourglass (Newell et al. 2016) and DeepLabV3+ (Chen et al. 2018), take input images under a uniform resolution. Image scaling and padding (i.e. add edges to an image to change its resolution) are commonly applied to make images into the same resolution (Simonyan and Zisserman 2014; He et al. 2016; Chen et al. 2017b). Therefore, images of a dataset that are going to be used in PhenoLearn should have a uniform resolution, either obtained directly from the digitisation or from unifying (e.g. image scaling) the resolution-varied images.

4.2.3 Applying deep learning models

Labelled images from the training set are used as the input to train deep learning models. I used two deep learning networks in PhenoLearn, (i) Stacked hourglass for predicting landmark points (Newell et al. 2016)s, and (ii) DeepLabV3+ for segmentations (Chen et al. 2018). These two networks performed well on predicting keypoints and segmentations on bird specimen photos in Chapter 2 and Chapter 3.

Hyperparameters define how a model is trained. A hyperparameter is a parameter that is set before the training (e.g. the learning rate), while parameters are referred to as parameters in the deep neural network which will be updated through the training process. Tuning hyperparameters and training different models can be time-consuming. Here, PhenoLearn only allows users to change some of the key hyperparameters such as the learning rate, input resolution, and training length (tuneable hyperparameters are listed in Supplementary Table 6.3.1). These hyperparameters can be entered in a form as shown in Figure 4.2b. Input and output options such as the annotation file (see annotation file in Supplementary Table 6.3.2) also need to be specific in order to start training (see software manual for the detail of setting up the training process). The training takes many iterations to optimise network parameters well. For each iteration, the network takes the input, generates the prediction, and calculates the loss function. The loss function shows the difference between the ground truth and predictions. Then,

gradient descent (Ruder 2016) using the gradient of the loss function reduces the loss and optimise network parameters. The loss is an important metric to measure how the network is trained and whether the training is converged (i.e. the point at which the degree of loss does not decrease further over long time periods).

During the training, users can view the loss of the model through Tensorboard (a model visualisation tool from Tensorflow; www.tensorflow.org/tensorboard) using the outputted log file (see log file in Supplementary Table 6.3.3). The result file (i.e. predicted annotation file), performance file (evaluation result using the metrics introduced later in this section) and checkpoint file (to restore the trained model in the predicting process) are generated once the training finishes (see the detail of these files in Supplementary Table 6.3.3). The result and performance files can be saved in CSV spreadsheet format, which can be used conveniently by other tools (e.g. R or MATLAB) for evaluations or analyses.

PhenoLearn also has build-in evaluation functions to evaluate results from different models. A trained model can be assessed by evaluating how predictions of the validation set are different from the expert labelled (ground truth) annotations. I used pixel distance and Percentage of Correct Keypoints (PCK) with a user-defined threshold in evaluating points prediction. PCK is the percentage of predictions that have pixel distances under a given threshold (Andriluka et al. 2014). The PCK is an intuitive metric, if a good threshold is selected. The intersection over union (IOU), precision and recall are used to evaluate the segmentation (see section 6.3.2 for detail of these metrics). Precision shows how many predicted areas are correct. Recall shows how many correct areas are predicted. IOU is a metric that takes account of both precision and recall. The mean IOU (mIOU) of all segmentation classes is considered the best metric for measuring the segmentation performance (Long et al. 2015; Chen et al. 2017b). After comparing the evaluation results of all trained models, the best model is then used to generate predictions for the whole dataset (see software manual for the detail of setting up the evaluation and predicting process. Supplementary Figure 6.3.4a and b show interfaces of evaluation and predicting. Supplementary Figure 6.3.5a shows an example of comparing results from different models).

The evaluation can also assess whether specimen characteristics (if provided, see metadata file in Supplementary Table 6.3.2) affect accuracy. The relation between specimen characteristics and the prediction accuracy helps users to understand datasets better. If a characteristic is numerical (e.g. body mass or length), the correlation is calculated between the characteristic and accuracy. For categorical characteristics (e.g. taxonomic ranks or ecological factors), boxplots are used to show differences among categories (an example is shown in Supplementary Figure 6.3.5b). These results show how different specimen characteristics affect the prediction accuracy, and users can later use these variables to help the review process (see section 4.2.4), for example, only review images with certain characteristics.

4.2.4 Reviewing the deep learning predictions

Manual review and correction of predictions are not commonly used to improve deep learning's accuracy. Investigators can accept the pure deep learning's results based on the speed-accuracy trade-off between reviewing huge datasets (more than tens of thousands of images) and perfect results. An extreme situation arises in many real-time tasks that require excessively high accuracy (e.g. self-driving and facial recognition), where manual correction cannot be applied quickly enough to meet the real-time requirement.

Some applications of machine learning to medical images require manual correction (Lustberg et al. 2018; Schlegl et al. 2018) but may also be useful in correcting predictions in a broader range of applications, including phenotyping digital natural history collections. Measurements should be as accurate as possible, because prediction errors will propagate to downstream analyses. Knowledge of the specimen characteristics can help to deal with the speed-accuracy trade-off by reviewing images that are likely to be incorrectly predicted based on these characteristics. For example, validation results of Chapter 2 and Chapter 3 have shown that the accuracies of some avian orders are lower than the average accuracy and would be priorities for error checking. I assume these orders will have low accuracies in the whole dataset (training sets are well selected to reflect the whole dataset). Images from these orders rather than all images are reviewed.

Users can review predictions in PhenoLearn using the prediction file and the dataset folder. Reviewing all predictions one by one is not very efficient and can be a tedious job. PhenoLearn

provides a review assistant in review mode with the aim to increase review efficiency (see software manual and Supplementary Figure 6.3.6 for detailed instructions on the review process). The review mode displays multiple thumbnails and their annotations in the visualisation panel as shown in Figure 4.2e. Users can rapidly review predictions using the review mode and note the images with incorrect predictions using a checkbox. Selecting images with the checkbox defines them as a subset for detailed checking and correction. The checked images can then be viewed and any incorrect predictions can then be corrected, or can be simply discarded. The review assistant also allows users to review images preferentially (e.g. images with higher chances of having incorrect annotations) by organizing specimen characteristics, such as sizes, if such features are known (Supplementary Figure 6.3.2d).

4.3 Application – morphology and ecotype in *Littorina*

4.3.1 Data and labels

I applied the pipeline introduced above to *Littorina* shell images to predict landmarks which are used to estimate the shell shape. I used the predictions from PhenoLearn to test the classic crab vs wave hypothesis on the effects of ecology on morphology (Butlin et al. 2014; Ravinet et al. 2016). The hypothesis predicts that the morphology of *Littorina* shells reflects ecological differences in habitat (crab-rich or wave-swept intertidal habitats) that determine the dominant selection pressure. Crab dominated habitats are expected to be associated with large, thick shells, with a relatively small aperture and wary behaviour, whereas wave-swept habitats with low crab predation are predicted to be associated with small, thin shells, a relatively large aperture and bold behaviour (Johannesson et al. 2010).

Shell images were provided courtesy of Prof. Roger Butlin and both labels and data collection protocols (outlined below) were provided by Zuzanna Zagrodzka (both Department of Animal and Plant Sciences, University of Sheffield). In each image, the shell was placed on a pink colour Play-Doh© background, next to a 5-mm graph paper, with a label of shell ID on the side. All shells were imaged under a Leica M80 (8:1 zoom) microscope using a Leica MC 170 HD camera. The digitised images have different pixel resolutions (1280 x 960, 1944 x 1458, 2074 x 1556, 2592x 1944, 5760 x 3840) and an example image is shown in Figure 4.3a.

As shown in Figure 4.3b, 15 morphometric landmarks (LMs) are placed on a shell following the configuration in Ravinet et al. (2016). The landmarks can be considered in two forms, independent and dependent. The landmarking rules are listed in

Supplementary Table 6.3.4. Dependent landmarks require reference lines that are extrapolated from independent landmarks. The rules for placing reference lines are listed in Supplementary Table 6.3.5. Based on the landmarking rules, landmarks can be manually placed in four steps. The first step is placing five independent landmarks (LM1, LM2, LM3, LM4 and LM13). The second step is drawing Line1 and Line4. LM10 and LM11 can then be placed. The third step is drawing Line2 and Line3. LM5 and LM12 can then be placed. The fourth step is drawing Line5, Line6, Line7 and Line8. LM6, LM7, LM8, LM9, LM 14, LM15 can then be placed. Inkscape was used to draw reference lines and tpsDig2 (Rohlf 2006) was used to place landmarks, and they are saved as TPS files.

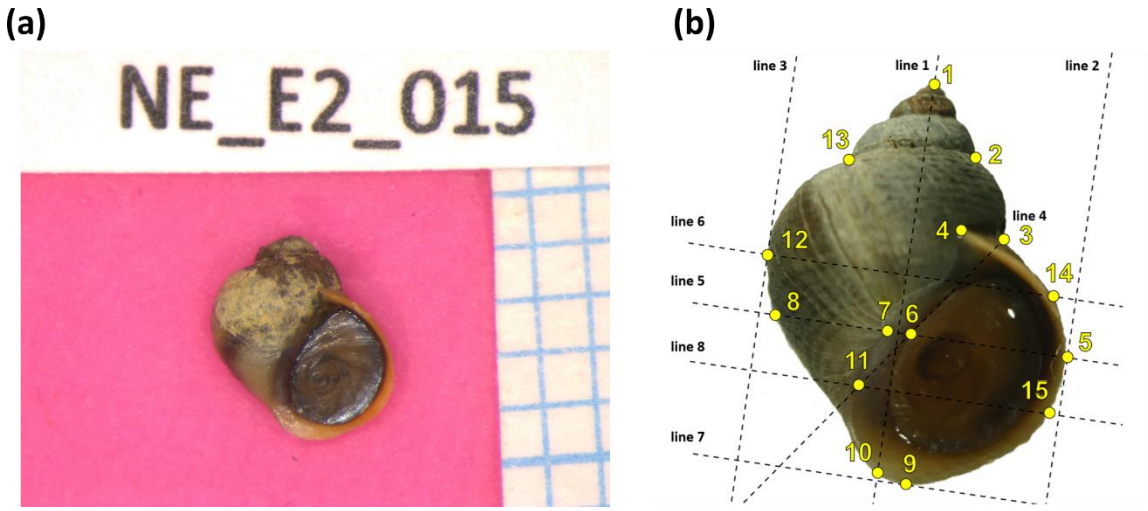


Figure 4.3. (a) An example of a digitised *Littorina* shell image. (b) The 15 landmarks and 8 reference lines are placed on a shell.

4.3.2 Method

A total of 681 labelled images were used in training. 544 images were used as training images, and the remaining 137 images were used as validation images. Images were first resized into a uniform resolution using bilinear image interpolation in OpenCV-python (a python computer vision library). The resolution of 2592 x 1944 was used as most of the images have this resolution.

Landmarks were scaled and converted into a CSV file that meets the input file standard in PhenoLearn and PhenoLearn was used to train the deep learning network.

After comparing pixel distances of the validation set for models with different configurations, the model with the learning rate of 0.01 (0.001 and 0.0001 were also tested), the batch size of four (one and two were also tested) and input resolution scale of 5 (7, 8 and 10 were also tested) generated the most accurate result. The resolution scaling gave an image resolution of 518 x 388 pixels (5 times smaller than 2592 x 1944 pixels). The data type was set to point, and the Stacked hourglass was therefore used as the deep learning network.

I then used the trained model to predict landmark positions on another dataset (188 *Littorina saxatilis* specimen images) that has manually placed landmarks and where the specimens belong to one of two ecotypes: crab (N=100) or wave (N=88). This can be used to evaluate whether deep learning can separate crab and wave specimens in morphospace, as is expected from previous studies on manual landmarks (Butlin et al. 2014; Ravinet et al. 2016). The ecotype describes the dominant selection pressure based on the location of collection. Specimens in the crab ecotype are likely to be subject to crab predation, whereas specimens in the wave ecotype are exposed to regular strong wave action (Johannesson et al. 2010). These alternative selection pressures are predicted to drive shell shape. A subset of the images from the dataset was labelled a second time by the same expert. I compared the landmarks between the first and second labelling to estimate the human variance in labelling *Littorina* shells. Among the re-labelled images, 20 crab and 20 wave images were re-labelled.

4.3.2.1 Inferring dependent landmarks

The raw deep learning predictions learn and predict each landmark independently, as a result, the placement of some landmarks may not meet the landmarking rules (e.g. the line passing LM5 and LM8 may not be parallel to the line passes LM12 and LM14). Shell outlines for the example data set have previously been estimated using thresholding methods to remove the pink background. I, therefore, implemented a landmarking generation algorithm to place dependent landmarks based on the independent landmarks and the shell outline using computational geometry (e.g. finding the tangent line and its intersection point to an outline; see Algorithm 1 in

section 6.3.5). Independent landmarks (LM1, LM2, LM3, LM4, LM10, LM11 and LM13) are used to calculate dependent landmarks and reference lines. Independent LMs that are on the shell outline (I refer as independent outline LMs, which are LM1, LM2, LM3, LM10 and LM13) are firstly optimized. The optimisation checks whether independent outline predictions are on the outline, and change locations of not-on-outline predictions to their closest points on the outline.

Dependent LMs can be split into semi- and fully-dependent landmarks. Semi-dependent landmarks (LM6 and LM7) use their deep learning predictions as prior positions. The final prediction is generated based on the prior position and the landmarking rules (e.g. the final prediction of LM7 is the nearest location from deep learning prediction LM7 to Line5). Dependent landmarks (LM5, LM8, LM9, LM12, LM14 and LM15) are calculated using independent landmarks and reference lines.

Stacked Hourglass (Newell et al. 2016) was used to predict independent and semi-dependent landmarks. The landmark generation algorithm was then applied to deep learning predictions to generate final landmark predictions. The assumptions for this algorithm are that deep learning predictions for the independent and semi-independent landmarks and the thresholding of the shell outline are sufficiently accurate to allow inference of the dependent landmarks.

The main drawback of this algorithm is that, if I follow the manual rules, it is impossible to automatically generate some landmarks (e.g. LM6, LM7 and LM11) and lines (e.g. Line1) that are related to the lip, operculum and aperture. This is because the manual landmarking configuration does not include any information of outlines of the lip, operculum and aperture. For example, Line1 is defined as passing LM1 and touching the inner margin of the aperture on the left and is therefore impossible to calculate using computational geometry without the operculum outline. Line1 is then used to define the LM10 on the outline of the shell. To solve this problem, they were generated in a different way. Specifically, I defined LM1 and LM10 as independent landmarks and calculate Line1 by connecting LM1 and LM10 in the algorithm. This does not guarantee that Line1 touches the operculum, but does allow geometry-based placement of other lines and landmarks. A future step of solving this drawback is to generate accurate operculum and lip outlines automatically. It is possible, for example, to use PhenoLearn to generate operculum and lip

segmentations if suitable training data are available. Here, I refer to the processed deep learning result or GeomDL as the outcome of applying the geometric landmark generation algorithm to the raw deep learning result.

Besides evaluating the processed deep learning result, I also assessed the causes of error in the placement of dependent landmarks. The dependent LM error of the processed deep learning result can be caused by inaccuracy of the independent LM predictions or by the placement of landmarks with the landmark generation algorithm. The latter would occur if the algorithm places dependent LMs differently to a manual implementation of the rules. This is most likely if the threshold outline differs from the human perception of the outline of the shell (e.g. LM12 can be placed differently even if Line1 is the same, as humans and machine learning may recognise the intersection between Line3 and the shell differently). The additional assessment gives four alternative sets of results to compare among themselves and with GeomDL. Alternative result 1 is the raw deep learning result (RawDL). Alternative result 2 is the processed deep learning result but with the predicted LM10 replaced with the ground truth LM10 (GeomDL_10). Alternative result 3 is the processed deep learning result but with the predicted LM1 and LM10 replaced with the ground truth LM1 and LM10 (GeomDL_1_10). Alternative result 4 is the result of applying the algorithm to the ground truth independent landmarks (I refer to this as the processed ground truth or GeomGT). The algorithm generates correct dependent LMs if both the independent LMs and the shell outline are correct. Similarly, if the manually placed independent landmarks are correctly placed then GeomGT has, in theory, correct LMs. The comparison of GeomGT to the raw ground truth result shows if the dependent landmark differences are due to placements of manual landmarks or to the geometric landmark generation algorithm. In addition, I compared GeomDL to GeomGT, which shows whether GeomDL is closer to GeomGT or to the raw ground truth.

4.3.2.2 Morphospace

After training and selecting the best model, I used the model outputs described above to generate landmark predictions, output as a CSV file. To evaluate GeomDL and alternative results, I calculated pixel distances (under the original image resolution of 2592 x 1944) of landmarks and

compared morphospaces to the ground truth. Although the pixel distance is a straightforward metric to measure prediction accuracies, it may not reflect accuracies closely on each individual shell as sizes of shells varied among images. To control the effect from the shell size, I used PCK with proportions of each shell's height (the distance between LM1 to LM10) as thresholds. I used 5% and 10% of the shell height (PCKht@0.05 and PCKht@0.1). This is similar to PCKh (h stands for the head) used in human pose estimation tasks (Andriluka et al. 2014; Newell et al. 2016), which uses a proportion of the head of each human as the threshold to measure the correct percentage.

TPS files were converted from CSV files, which can be used as inputs for geometric morphometric analyses (Gower 1975; Bookstein 1991). Landmarks were aligned using Generalized Procrustes Analysis (GPA). Principal component analysis (PCA) was then applied on aligned landmark coordinates to summarise the shape variation. GPA and PCA were applied to landmarks from the ground truth and predicted results (a total of 16,920 landmarks from six sets of landmarks). This ensures that all landmarks are arrayed on the same set of PC axes, which ensures the PCs and morphospaces of the six sets of landmarks (the ground truth, GeomDL and four alternative results) are directly comparable. I then visualised the morphospaces to assess how well the alternative landmarking methods (e.g. deep learning with the landmark generation algorithm) differentiate between crab and wave ecotypes. These steps were run using the R package Geomorph (Adams and Otárola-Castillo 2013).

4.3.3 Results

The pixel distances and correlations of PC1 to PC9 between the ground truth and re-labelled landmarks of a subset of the validation dataset (N=40; crab=20, wave=20) are shown in Supplementary Figure 6.3.7. These results of PC1-9 provide a baseline expectation of variation in landmarking with which to compare the machine learning results. The landmark with the largest pixel distance is LM15 (~30 pixels). The remaining landmarks have pixel distances less than 20 pixels as shown in Supplementary Figure 6.3.7a. The first 9 principal component (PC) axes explain more than 95% of the total variation in shell shape, and their correlations are shown in Supplementary Figure 6.3.7b. The majority of correlation coefficients are greater than 0.8. Only

PC2 has correlation coefficients below 0.8. These results suggest that variation in manual landmarking is minor.

I then used the best model to predict the validation dataset (N=188; crab=100, wave=88) and generated results such as GeomDL. Landmark accuracy, measured as pixel distance between landmark positions from GeomDL and the ground truth (manually labelled landmarks) varied from ~9-60 pixels. LM14 was the worst predicted landmarks for GeomDL, with an average of 59.8 pixels between the predicted and the ground truth position (4.9% of the average shell height which is 1,214 pixels measured from LM1 to LM10). LM2 was the best predicted LM scoring the pixel distance of 9.9 (0.8% of the average shell height). Dependent LMs generally had larger pixel distances than independent ones (Figure 4.4).

Among the independent LMs, those on the outline (LM1, LM2, LM3, LM10 and LM13) were predicted more accurately than those that are not on the outline (LM4 and LM11) as shown in Figure 4.4a. The optimisation of shell outline LMs improved accuracy for LM2 and LM13. The RawDL were more accurate than GeomDL in all dependent LMs besides LM15, which highly depends on LM11 (the worst LM among independent LMs). GeomDL_10 and GeomDL_1_10 did not improve dependent LM accuracy comparing to GeomDL as shown in Figure 4.4b. The difference between GeomDL and GeomGT is smaller than the difference between GeomDL and the raw ground truth result (grey and dark blue plots in Figure 4.4b). Dependent LMs (except for LM15) of GeomGT were closer to GeomDL than to the raw ground truth (light yellow and dark blue plots in Figure 4.4b).

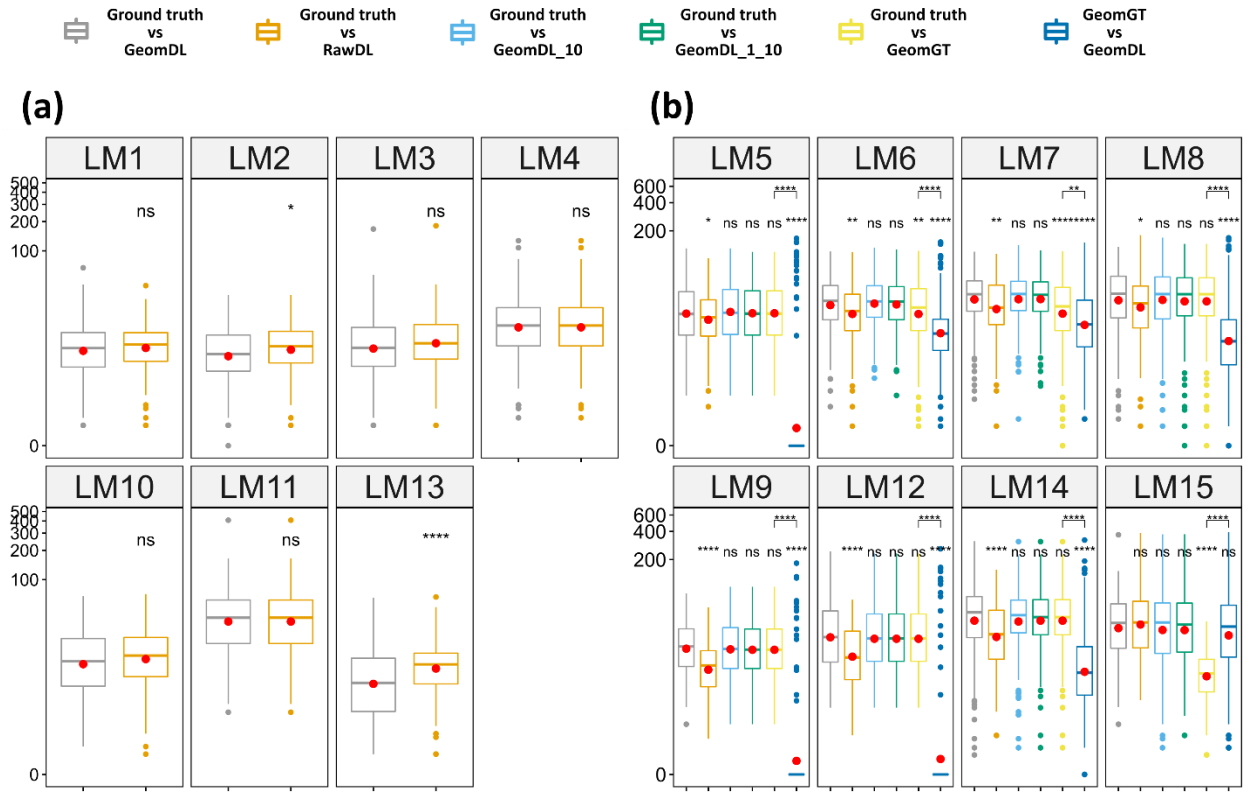


Figure 4.4. Boxplots of per-landmark pixel distance of different comparison groups. (a) Independent landmarks; (b) Dependent landmarks. Groups of pixel distance result: the ground truth vs GeomDL (grey); the ground truth vs RawDL (dark yellow); the ground truth vs GeomDL_10 (light blue); the ground truth vs GeomDL_1_10 (green); the ground truth vs GeomGT (light yellow) and GeomGT vs GeomDL (dark blue). Significance symbols above the brackets in (b) show the comparison between GeomGT vs GeomDL and the ground truth vs GeomGT. Rest of the symbols are comparison between the ground truth vs GeomDL (grey plots) and other groups (ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001).

PCK values were evaluated for three groups, (i) the ground truth vs GeomDL, (ii) the ground truth vs RawDL and (iii) GeomGT vs GeomDL (Supplementary Table 6.3.6). More than 98% of the predicted landmarks (N=2,820) were located with 10% of shell heights. PCKht@0.05 of the ground truth vs GeomDL had the lowest value (85.2) among the three groups. However, when comparing GeomDL to the GeomGT, the overall PCKht@0.05 was 93, as PCKht@0.05 of the most dependent LMs were improved greatly. The PCK result shows that the majority of the predictions

were located within 5% to 10% of their shell heights, suggesting that predictions were reliably accurate.

I generated principal components based on Procrustes aligned coordinates to quantify the morphospace. The first four PC axes explain 80% of the total variation, and PC points of these axes of the ground truth and GeomDL are shown in Figure 4.5. PC1 is positively related to the width of the operculum. PC2 is positively related to the length of the aperture and negatively related to the proportion of the operculum to the whole shell. PC3 mainly explains the distance from LM11 to the operculum. PC4 is positively related to the distance of the end of the suture (LM4) to the shell outline. PCs between crab and wave are distributed similarly in the morphospace in both the ground truth and GeomDL. In particular, there was a clear separation between crab and wave on PC2, where crab specimens had large PC2 values and wave specimens had small PC2 values.

Most of the distributions between crab and wave of the first eight PC axes (explain 93% of the variation) were similar among the ground truth and all tested results (see Supplementary Figure 6.3.8 and Supplementary Figure 6.3.9). The PC4 values between crab and wave of the six results were not significantly different from each other except for RawDL. The PC6 values between crab and wave were not significantly different except for the ground truth.

All PCs of GeomDL correlated better with GeomGT than with the raw ground truth as shown in Supplementary Figure 6.3.10. GeomGT and the ground truth were highly correlated in PC2 (R=0.99). GeomDL_10 and GeomDL_1_10 did not improve correlations with the ground truth much. Among GeomDL, RawDL, GeomDL_10 and GeomDL_1_10, RawDL had the best correlations with the ground truth in PC1-6.

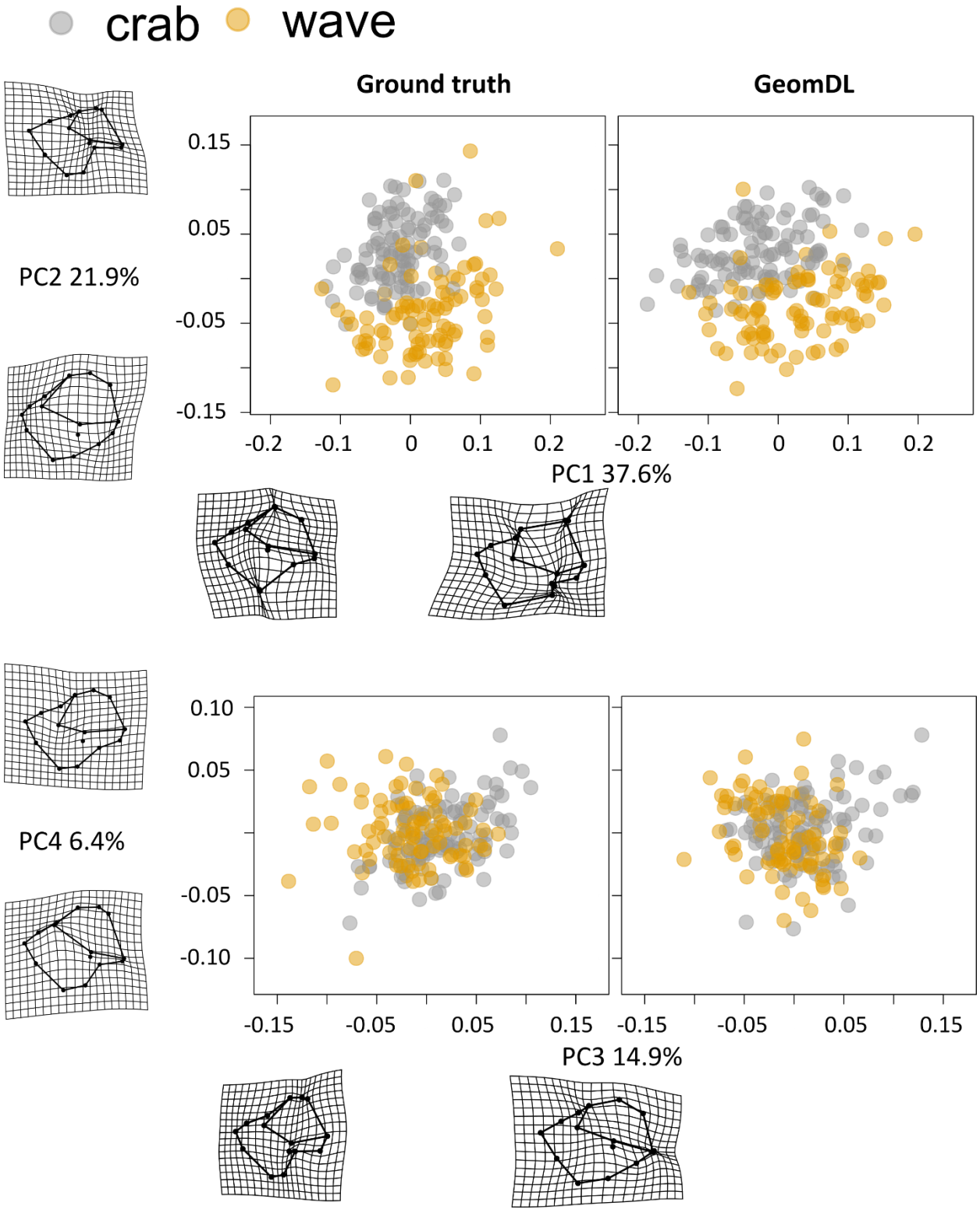


Figure 4.5. Distributions of PC1-2 and PC3-4 from the ground truth and GeomDL (N=188. 100 crab specimens and 88 wave specimens). Grey points are specimens with the crab ecotype. Yellow points are specimens with the wave ecotype.

4.3.4 *Littorina* discussion

The deep learning result (GeomDL or RawDL) is able to detect the difference between the two ecotypes along multiple axes and this is consistent with trends from the ground truth data (Figure 4.5, Supplementary Figure 6.3.8). Crab and wave specimens were split clearly along multiple PC axes (PC1, PC2, PC3, PC7) with a strong separation on PC2. Most of the crab ecotype shells had large PC2 values, which describes a longer spire and smaller aperture than small PC2 values associated with wave-ecotype shells. Studies using the same landmark setup have shown that crab ecotype shells tend to have a smaller aperture and higher spire than wave ecotype ones (Butlin et al. 2014; Ravinet et al. 2016). Morphospaces from these studies also separated two ecotypes well. Therefore, a well-trained deep learning landmark prediction can provide meaningful results in geometric morphometric analyses. Although the biological conclusions are robust to the landmarking method, the different approaches do yield differences in the position of specimens within the morphospace.

Deep learning predicts landmarks with distinct features well (e.g. the shell apex). The overall average pixel distance of GeomDL to the ground truth is 31.5 (2.5% of the average shell height). In general, independent outline LMs were predicted more accurately than independent LMs that are not on the outline. LM11 is the worst predicted independent LM, because it is not on the outline and it has fewer features than LM4 (the end of the suture is distinct in shells) for a deep learning network to learn and predict. The PCs of GeomDL are consistently positively correlated with PCs created from the ground truth. The high correlation in PC2 axis is especially important (PC2 has the highest correlation among PC1-8, R=0.88. See Supplementary Figure 6.3.10a), because it is the axis that most strongly separates the crab and wave (Figure 4.5).

Errors of GeomDL can be caused by inaccurate deep learning predicted independent LMs, which will generate inaccurate dependent LMs. To test whether it is the main cause, I tried to increase the accuracy of Line1 (a line passes LM1 and LM10). Line1 is the most important reference line used in the algorithm for generating dependent LMs. GeomDL_10 (using the ground truth LM10 instead of deep learning's) and GeomDL_1_10 (using the ground truth LM1 and LM10 instead of deep learning's) were compared to the ground truth. However, an accurate Line1 did not improve

the geometrical and morphological differences, as shown in the evaluation of pixel distances (Figure 4.4) and PC correlations (Supplementary Figure 6.3.10c and d).

The error can also be caused by the landmark generation algorithm. The landmark generation algorithm seems to introduce error in the dependent LM generation. Pixel distances of dependent LMs between GeomGT and the raw ground truth have shown that dependent LMs placed by human and the algorithm were geometrically different (light yellow plots in Figure 4.4b). Their pixel distances were similar to distances between GeomDL and ground truth (grey and light yellow plots in Figure 4.4b). The raw ground truth and GeomGT were highly correlated in PC2 (R=0.99) which primarily describes the length of the aperture, and is mainly defined by independent LMs (LM1, LM2, LM3 and LM13). Correlations of other PCs were not correlated as well as PC2, only PC7 and PC8 had R greater than 0.85 (Supplementary Figure 6.3.10e).

The reason that the algorithm generates dependent LMs differently from human maybe because outlines detected by humans differ from outlines from thresholding, which will cause intersection points to be placed differently. Outlines within a small range of pixels can all be considered correct. Supplementary Figure 6.3.11 shows an example of unprocessed and processed ground truth landmarks. Landmarks (except for the processed LM6 and LM7) and reference lines are correctly placed based on the rules (

Supplementary Table 6.3.4 and Supplementary Table 6.3.5). The LM6 and LM7 (both semi-dependent LMs) of the processed ground truth (GeomGT) are defined as the nearest points on Line5 to their raw positions (see Algorithm 1). In Supplementary Figure 6.3.11, they do not match the landmarking rule (LM6 and LM7 should be on the lip edges). Currently, this landmark placing algorithm is a reliable way to ensure LM6 and LM7 are placed on Line5. If the outline of the lip is provided, it is possible to use intersection points between Line5 and the lip outline as the LM6 and LM7 (LM6: right intersection, LM7: left intersection). In addition, variations seem to accumulate through the algorithm steps. LM8 has a larger pixel distance than LM5, and LM14 has a larger pixel distance than LM12 (see Figure 4.4). LM5 and LM12 are first generated dependent LMs that are used to calculate LM8 and LM14 (see Algorithm 1). PC correlations between GeomGT and the raw ground truth were not perfectly correlated (e.g. the R-value of PC1 is 0.74,

see Supplementary Figure 6.3.10e) suggesting that the geometric landmark generation algorithm introduces variances in placements of dependent LMs.

Since the landmark generation algorithm introduces variations on dependent LMs' positions. It is more sensible to compare dependent LMs and morphospaces of GeomDL to GeomGT than to the raw ground truth. The evaluation results showed that differences between GeomDL and GeomGT were generally smaller than differences between GeomDL and the raw ground truth. The average pixel distance of GeomDL vs GeomGT is 19.67 (GeomDL vs the ground truth is 31.51) and 93.0% of the predictions were located within 5% of the shell heights. Dependent LM pixel distances (except for LM15) of GeomDL vs GeomGT had significantly lower pixel distances than GeomDL vs the raw ground truth (grey and dark blue plots in Figure 4.4b). LM15 was generated based on LM11 and the shell outline. And LM11 was the worst predicted independent LM. In contrast, the prediction error is the main cause of the LM15 error. For PC axes correlations between GeomDL and GeomGT, of PC1-7 were above 0.7, among them PC1 and PC2 had higher than 0.88 (Supplementary Figure 6.3.10f), which explain almost 60% of the variation.

Another interesting thing is that RawDL had a more similar morphospace (based correlations of PC axes) and smaller dependent LM pixel distance to the ground truth. The raw landmark predictions do not follow the landmarking rule strictly (e.g. the line passes LM5 and LM8 may not be parallel to the line passes LM12 and LM14) but because the deep learning network learns to predict each landmark separately without constraint, the predictions are close to the results from manual labelling.

An important message raised in this example application is that the design of a dataset has great impacts on the deep learning performance. The landmarks used here for *Littorina* were not designed for a deep learning application, instead they were designed specifically for manual landmarking. As a result, there are difficulties in identifying some LMs or reference lines where the training data does not provide adequate information to train the deep learning algorithm. There are several approaches that could be considered in designing the training data. One solution may be to segment the operculum as well, so deep learning can be used to predict the operculum and generate more accurate reference lines for the dependent LMs. Alternatively, an

additional landmark could be added on the landmark to allow the placement of line 1. Although the additional landmark may not be needed for subsequent geometric morphometric analysis, it could improve the placement of dependent landmarks. Finally, deep learning results can be manually reviewed for accuracy and discarded or corrected as necessary. This can be done with the PhenoLearn review module. In general, emergent issues illustrate that ideally machine learning training sets should be designed from the outset with prediction in mind.

4.4 Conclusions

Together, the result shows the potential of deep learning on high-throughput phenotyping of digitised photos. PhenoLearn implements the pipeline of phenotyping digital images using deep learning with a user interface and result visualisation, including functions to (i) build training sets, (ii) train deep learning networks, (iii) use trained networks to predict annotations which are used to measure phenotypic traits and (iv) review the predictions. With these functions implemented, PhenoLearn can fill the gap between digitisation and biological analysis by implementing a high-throughput and accurate phenotyping pipeline. Along with automatic and high-throughput digitisation methods (Blagoderov et al. 2010; Hudson et al. 2015) and powerful software tools and packages for phenotypic analysis such as geometric morphometric analysis (Klingenberg 2011; Adams and Otárola-Castillo 2013) and colour and pattern analysis (Maia et al. 2013; Troscianko and Stevens 2015; Van Belleghem et al. 2018), the time for turning natural history collections to macro-scale phenotypic datasets will be greatly reduced (Figure 4.1).

Two aspects of the design of PhenoLearn make it suitable for users that are not experts in machine learning. The first is that PhenoLearn encapsulates deep learning networks and only allows users to tune some of the key hyperparameters which are normally tuned first in many studies. By only tuning these hyperparameters, PhenoLearn can achieve accurate trait measurements on digitised specimen photos and can generate biological meaningful results (see Chapter 2, Chapter 3 and the Application section) that are consistent with studies that use manually measured traits (Stoddard and Prum 2011; Ravinet et al. 2016).

Second, from a biological study perspective, I suggest that manually reviewing predictions or trying simple post-processing methods to increase measurement accuracy is more practical than

digging into the network optimisation to increase deep learning accuracy. Reviewing a large size set of predictions manually can be time-consuming. PhenoLearn provides annotation visualisation on checking and correcting predictions. The review mode, which displays multiple images, speeds up the review process. By adding the metadata information (e.g. taxonomical information), images with high chances of being incorrectly predicted can be reviewed first. Therefore PhenoLearn allows users themselves to balance the trade-off between measurement accuracy and time cost.

New features can be added to PhenoLearn easily due to the modular design. PhenoLearn is open-source, so people can contribute functions that fit their goals or analysis. With more and more functions added, PhenoLearn has the potential of becoming a powerful platform for high-throughput phenotyping. Improvements of new features can be made from aspects such as these:

- **Adding new types of annotations.** Placing bounding boxes can be added for detecting focal regions such as recognising specimens on the museum trays (Hudson et al. 2015). Deep learning networks used in object detection that identify bounding boxes such as R-CNN (Ren et al. 2015) can be used to predict bounding boxes.
- **Adding new and improving current evaluation metrics.** The PCK in PhenoLearn only supports a fixed value threshold. Dynamic thresholds (e.g. PCKht in the Application section) can be added as a way of selecting the threshold, and PCK can measure accuracies more intuitively for datasets that have size varied specimens (e.g. the *Littorina* dataset).
- **Adding post-processing methods.** Post-processing predictions can make them more accurate or fit closer phenotyping goals. For example, eroding segmentations is a common method to shrink the segmentation area (Haralick et al. 1987). It can be applied to predictions, if the goal is to generate conservative inferences on the focal areas. Adding build in post-processing functions in PhenoLearn can let users apply and evaluate post-processing in PhenoLearn without doing it on another platform (e.g. ImageJ).
- **Implementing a web application.** Web applications have been widely used in crowdsourcing labelling on digitised specimens (Chang and Alfaro 2016; Cooney et al.

2017). In addition, citizens can contribute to reviewing the predictions from deep learning (Keshavan et al. 2019).

Taken together PhenoLearn can fill the gap between the fast-growing automatic digitisation methods and computational analysis on phenotypes, introducing an accurate and high-throughput pipeline for measuring digital specimen images using deep learning.

Chapter 5 General Discussion

To achieve the main aim of building high-throughput pipelines of phenotyping or measuring digitised specimen datasets, I applied deep learning and other computational methods to place annotations on specimen images from two datasets (bird specimens and *Littorina* shells). Annotations have been placed accurately from both algorithm-based (e.g. whether predicted annotations match the expert-labelled annotations) and biological perspectives. The time to generate annotations for large-scale datasets (especially the bird images, of which there are more than 120,000 images) is reduced from processing in units by-months or years to processing by hours or days. Annotations were used to build and visualise bird plumage colour spaces (Chapter 2 and 3) as well as morphospaces of *Littorina* shells in Chapter 4.

5.1 Deep learning in phenotyping

Deep learning models generated reliable predictions (points and segmentations) for digitised specimen datasets for all three data chapters. My results suggest that deep learning could be a solution for a high-throughput pipeline for measuring phenotypic traits on specimen images in a range of applications. In all applications examined here, including keypoint placement and segmentation, deep learning performed well.

Keypoints were correctly placed on 95% of the validation images from Project Plumage (N=5,094 images) after checking by experts. The Percentage of Correct Keypoints (PCK) using 100 pixels as the threshold scored 100 for all five reflectance standards, suggesting all points were predicted correctly which were located inside standards circles. Because colours (values of selected pixels) from body regions were extracted using predicted body region points, a reasonable way of extracting pixels is important. Geometric shapes with fixed sizes around the points (e.g. squares and circles) fail to capture the size variance across body regions in the data set (e.g. the crown is normally smaller than the mantle). The preceding results of points from the Stacked Hourglass (Newell et al. 2016) are heatmaps (the value of each pixel in a heatmap is the likelihood of the pixel being the location of a point). Using heatmaps is a useful alternative method for extracting areas, as the sizes of the heatmaps vary among birds and body regions (see Chapter 2). Using

heatmap output instead of the final point outputs can be generalised to datasets that aim to locate multiple small regions (e.g. body regions). Chromatic information (i.e. hues) is often used in bird plumage colour analysis (Stoddard and Prum 2008, 2011). The hue information extracted from deep learning predictions of individual body regions were correlated well with hue information extracted from ground truth labels (all coefficients were larger than 0.93).

I applied keypoint placement to an alternative dataset of *Littorina* shells collected for geometric morphometric analysis. Here there is a need for higher precision in keypoint placement. Although landmark points on *Littorina* shell images were not checked by humans, PCKs using proportions (5% and 10%) of shell heights as thresholds (PCKht@0.05 and PCKht@0.1) were intuitive to evaluate predicted landmark accuracies. Six out of seven independent landmarks (landmarks predicted directly by deep learning) had PCKht@0.05 greater than 97, suggesting 97% of predictions were located within 5% height of the shell. The worst independent landmark (LM11) had 73.4 PCKht@0.05 and 96.8 PCKht@0.1, showing predictions of LM11 were not drastically worse than the rest landmarks. The biological conclusions drawn from geometric morphometric analysis of deep learning landmark points were consistent with those from manual landmarking.

Finally, I applied deep learning models to segment bird specimens from their background for the measurement of overall body colour. More than 95% of the segmented (segmented as the plumage area) pixels were correctly segmented (high precision) and more than 95% of plumage area pixels were segmented (high recall) using DeepLabv3+ (Chen et al. 2018). Importantly, DeepLabv3+ also outperformed the four classic methods tested in Chapter 3.

Taken together, the evaluation results (from human, geometric and colour perspectives) showed that deep learning produced reliably accurate point and segmentation predictions on both bird (Project Plumage) and *Littorina* shell images.

Besides deep learning models, I have used other computational approaches for measuring phenotypes in this thesis, which again reinforces the importance of utilising computational power to build high-throughput phenotyping pipelines. Customised metrics were designed to fit datasets better in the evaluation and quantifying phenotypes. For the Project Plumage dataset, the colour information correlation was used to evaluate colours measured by points, which is

more meaningful than pixel distances for colour measurements. Proportional colour diversity, the alpha shape, and the convex hull all have different biological meanings in describing the colour diversity of bird plumage colour (Chapter 3). In some circumstances, deep learning may not provide an end-to-end result (e.g. *Littorina* landmarks in Chapter 4). Therefore, other automatic methods may be required, such as the landmark placing algorithm used in Chapter 4 that generates a set of landmarks using deep learning predicted landmarks.

Although deep learning predictions are almost inevitably worse (i.e. the best possible outcome is a perfect match) than manual annotations by experts, they are nonetheless reliable predictions, often within the bounds of inter-human variation (Chapter 2). If predictions from deep learning are not consistently accurate enough and it is difficult to increase the accuracy using computational methods, some extra manual work (e.g. reviewing and editing) can be done to improve predictions, which still takes less time than placing annotations manually.

Not only did the evaluation results show that deep learning generates accurate results, but the three predicted datasets (body region heatmaps and whole-body segmentation from Project Plumage, and morphological landmarks on *Littorina* shells) in my data chapters were used to generate biological results which were consistent with previous studies. This lends further weight to the argument that deep learning is potentially a powerful and reliable tool for phenotypic analysis.

5.1.1 Bird plumage colour

In Chapter 2 I showed that the male birds are more colour-diverse than female birds for all and each individual body regions. This has previously been suggested by studies which found that male's plumage colour has evolved into wider colour spaces (Dale et al. 2015; Cooney et al. 2019). Colour space volumes measured by the convex hull or the alpha shape across more than 7,000 species showed that bird maximally occupied only about one-third of the total possible volume of the tetrahedral colour space. Stoddard and Prum's work (2011) found a similar result with colour volume only filling a fraction of the whole colour space. They suggest that this may be caused by limitations associated with the different colour producing mechanisms on birds. The colour diversity of the overall plumage colour was similar among closely related species, and is

moderately correlated with the bird phylogeny. Phylogenies visualised in Chapter 3 showed which clades have more diverse plumage colours (e.g. Psittaciformes and Coraciiformes) than others and largely confirm intuition based on observations in the human visible spectrum.

The plumage colour measured by deep learning included predictions for 8,509 bird species. This is by far the largest colour data set for birds using objectively measured colour (i.e. as opposed to scoring from illustrations or plates in field guides and handbooks). As such it can provide a plumage colour dataset that is useful to analyse evolutionary, ecological, and biodiversity questions on a macro-scale. In that sense, it is similar to other recent large scale data sets, such as the Mark My Bird (www.markmybird.org) bird beak dataset which measures beak shapes using morphological landmarks (2,028 species were used in the analysis and 8,896 species were labelled; Cooney et al. 2017), and the morphology data set of Pigot et al. (2020). Such data sets potentially open up new avenues for research and show the importance of having large-scale phenotypic datasets.

5.1.2 *Littorina* shell morphology

Both manual and deep learning landmarks revealed morphological differences between wave and crab ecotypes. Wave ecotype shells have larger apertures and lower spires than crab ecotypes (see Chapter 4). Previous studies have shown similar results between crab and wave ecotypes using manual landmarks (Butlin et al. 2014; Ravinet et al. 2016), again confirming the robustness of deep learning, even on data sets that were not designed with deep learning in mind.

5.2 Pipelines and applications

The major advantage of deep learning is the speed with which huge data sets can be generated. Here, predictions of points and segmentations of more than 120,000 images were generated in less than six days using one computer and one high-end consumer level but affordable graphics processing unit (NVIDIA GTX 1080Ti). Using deep learning has greatly increased the speed of measuring colour information on Project Plumage images (it was estimated to take years to finish the annotation process on Project Plumage). For many digitised specimen datasets where their phenotypic traits were measured by manual annotations (e.g. placing morphological landmarks)

in previous studies (Denton and Adams 2015; Zelditch et al. 2015; Maestri et al. 2017), deep learning may create faster measurements than manual annotation.

With the help of deep learning and the increasing computational power, biologists can try to build phenotypic datasets using measurements placed by deep learning models. Where possible the first step is to design annotations or measurements that are suitable for deep learning. There is not a universal rule or gold standard for how to design annotations for deep learning, but annotations should be placed by people with expert knowledge without difficulties. It is highly recommended to build high-quality (annotations are correctly placed) and manually labelled training sets, as deep learning models will learn what annotations are fed during the training. People should consider whether annotations are logically-workable for deep learning. Dependent landmarks of *Littorina* shells in Chapter 4, for example, cannot be predicted by deep learning models because these landmarks can only be calculated using pre-existing landmarks. Chapter 3 and 4 used low-quality datasets and achieved predictions almost as accurate as using original images suggesting that the digitisation set-up (e.g. lights, backgrounds and specimen placements) can be flexible. It is not necessary to spend a huge amount of time keeping everything under extremely high standards and consistency. The number of images labelled by experts does not need to be large at the beginning - my advice is to start from annotating a small number (e.g. tens or hundreds of specimens depends on the dataset) of images. After evaluating trained models, extra labelled images can be added, if the accuracy does not match the expectation.

Coding deep learning models and tuning model hyperparameters can be time-consuming. I introduced the software pipeline PhenoLearn (Chapter 4) that can be used to visualise the hyperparameter-tuning and allow biologists or others with limited deep learning expertise to train deep learning models. Training, evaluating models, and using models to predict annotations are all encapsulated in PhenoLearn. Users only need to control the user interface to produce predictions for large scale datasets. PhenoLearn also supports manual annotations of images, which creates training images. Sometimes, biologists want higher-quality measurements than the raw deep learning result. In PhenoLearn, manual review and editing can be achieved in a simple but effective and intuitive way. PhenoLearn has a review module to enable users to review

and edit predictions efficiently by displaying multiple images and annotations and prioritising images with high error-rate characteristics (if metadata or specimen information is provided).

Reviewing does not have to rely on experts, and crowdsourcing can now play an important role in reviewing as deep learning can generate mostly accurate annotations (Keshavan et al. 2019). Reviewing annotations is generally easier than placing them, as it requires less time and operations. Biologists can also decide to let citizens do either the review or both review and correction depending on the task difficulties. Based on the performance of the validation set, an overall correct prediction rate can be estimated. If, as I find here, deep learning generates high proportions of accurate predictions (e.g. 95% validation images are correct in Chapter 2), then correcting error predictions would not be a massive task.

Although I focus on specimen images (both from museum collections and the field), results from this thesis and previous studies have strengthened the case for the use of deep learning methods for a wide range of biodiversity datasets. Aside from classifying whether animals are presented or locating animals with bounding boxes using deep learning (Kellenberger et al. 2017, 2018; Norouzzadeh et al. 2018; Schneider et al. 2018), it is possible to identify keypoints and segment animals on images from camera traps or drones. Other studies have shown how deep learning can be used to provide extra annotations from automated scans of specimens in museum drawers (Mantle et al. 2012; Hudson et al. 2015). PhenoLearn can be a platform for biologists to carry out many of these manipulations. Because it is open-source, the functions can be supplemented and extended by other contributors to make it even more suitable for a greater range of biodiversity datasets.

5.3 Limitations

The automatic phenotyping pipeline introduced in PhenoLearn may not perfectly predict measurements previously designed for biodiversity datasets. Many measurements were designed to fit the way of measuring photos (e.g. manual annotation), which may not be suitable for deep learning networks to learn and predict. Landmarks of the *Littorina* shell predicted in Chapter 4 were originally designed by Ravinet et al. (2016) with the expectation that landmarking would be manual, that is, deep learning was not considered in the study design. It is impossible

to use deep learning to predict all landmarks in the *Littorina* data set, as some landmarks (dependent landmarks) can only be correctly placed based on other landmarks and outlines of shell parts. In addition, dependent landmarks are difficult for automatic methods to generate. Humans can place landmarks without explicitly highlighting some information (e.g. a human can recognise shell outlines without explicitly drawing them), while computational approaches may need this information as a guide. Sometimes, changes need to be applied to make measurements suitable for deep learning. For example, placing points on Project Plumage can be considered as a measurement alteration. The dataset and the way of measuring body regions were designed based on the measurements used by Cooney et.al (2019), in which they manually placed polygons, rather than points, to measure plumage colours in the bird body region. However, identifying multiple small areas or polygons that can be similar is difficult for deep learning. Therefore, points were used instead of polygons, since points can more easily be predicted by using well-developed pose estimation networks.

Methods used in measuring plumage colours introduced in Chapter 2 and 3 are not fully-automatic to match the annotation rules from Project Plumage. One main obstacle is that there is not an accurate and high-throughput method to detect occluded body regions. Pose estimation networks identify locations of occluded body regions (Andriluka et al. 2014; Newell et al. 2016; Wei et al. 2016), which is different from the goal of detecting and removing occluded points. Failure to detect occluded regions can lead to bias in analysing colours, especially for the rump, which is the region that has the highest chance of being occluded (16% of rumps were fully-occluded among the expert-labelled images). Some of the 'rumps' measured by deep learning predictions are wings, because wings can cover the entire rump in some specimens. Therefore, results from the rump should be cautiously treated or those 'rumps' should be detected and discarded using either manual or automatic methods.

The way of measuring the overall plumage area can also be improved. There are duplicated areas if using segmentations for all three views to measure plumage areas (e.g. areas like wings may appear in photos of different views). Although methods can be designed to automatically detect duplicate areas, re-designing how the overall plumage area is measured, for example, only using segmentations from two views (e.g. using only back and belly view), is a quick and simple solution.

The ideal deep learning pipeline should use high-quality images where possible within the constraints of the available GPU hardware. The Project Plumage images are originally in RAW format and are converted to JPG to reduce file sizes prior to uploading to projectplumage.org. For consistency with the images seen by users on project plumage, I also used JPG images for deep learning tasks. This has some advantages in that JPG is a common format and it is easy for tools like Python to read and write. Therefore, I used JPG images for training. However, JPG is a lossy format and image information might be compressed. The use of JPG images may therefore lower the deep learning performance (Dodge and Karam 2016). Due to the time limitation, I did not generate loss-less formats from raw images, however, compared to studies that assessed the effects of image quality (Dodge and Karam 2016), images used in this thesis appear to retain, at least to the human eye, high quality.

5.4 Future work

This section focuses on four main aspects of potential future work, (i) improving prediction accuracy; (ii) future analysis using annotations generated in this thesis; (iii) extending PhenoLearn; and (iv) applying deep learning to other types of natural history datasets.

(i) Improving prediction accuracy

Model prediction accuracy can be improved by tuning the deep learning architecture and hyperparameters but this can require a great deal of trial and error and expertise. In contrast, post-processing methods can be a simple but effective way to improve accuracy. For example, I suggest applying erosion techniques (Haralick et al. 1987) on the segmentations. Erosion works by shrinking the segmented areas. This means that segmentations are more likely to be placed inside the plumage area (i.e. high accuracy), which fits the Project Plumage segmentation goal better (i.e. segment only the plumage area). Dilation can be applied to expand segmented areas, which can be used for the contrary goal (i.e. the actual focal areas tend to be larger than the predicted areas).

Another area that may prove fruitful is to improve accuracy by combining predicted points and segmentation. Automatic checking can be used to flag abnormal predictions, if predicted body

region points (excluding points for identifying reflectance standards) were located outside the predicted segmentation. For a stricter rule, extraction areas (based on output heatmaps) can be used to replace points, as colours were extracted in these areas. Flagged images can have either error predicted points (Figure 5.1a, heatmaps if the strict rule is used), very conservative segmentations (Figure 5.1b) or both conditions. Flagged images can be either manually edited by humans or fixed by algorithms. A possible idea for the algorithm can be first detecting what causes the error. If it is a Figure 5.1a case, outlier points are moved inside the segmentation. While if it is a Figure 5.1b case, segmented areas can be increased till all points are covered.

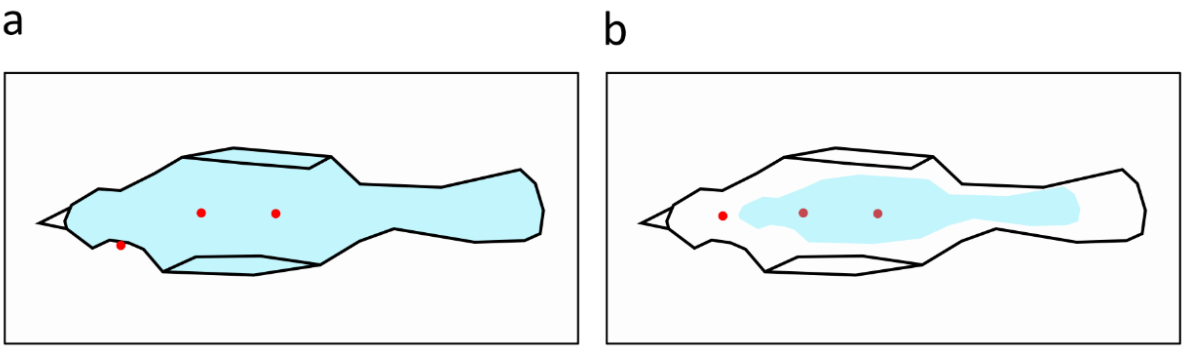


Figure 5.1. Examples of flagged images by checking the geometric relation between the points (red points) and the segmentation (areas in blue) using a drawing to simulate the belly view of a bird specimen. (a) the segmentation was correctly predicted but the neck point (the leftmost point) was placed outside the segmentation. (b) All points were correctly predicted while the segmentation only occupied a proportion of the bird, which makes the neck point outside of the conservatively predicted segmentation.

(ii) Future analysis with model predictions

In this thesis, I focused only on the colours directly. However, segmentations generated in Chapter 3 can also be used in analysing colour pattern to understand questions like mimicry (Stoddard and Stevens 2010; Van Belleghem et al. 2020), camouflage (Troscianko et al. 2016) and mate choice/attraction signalling traits (Ferns and Hinsley 2004; Pérez-Rodríguez et al. 2013; Marques et al. 2016). Colour patterns can be quantified using functions and models. As summarised in the work of Pérez-Rodríguez et al. (2017), patterns like barred (Gluckman and Cardoso 2009) and spotted (Stoddard and Stevens 2010), which exist on bird plumages (e.g. barred owl and European starling), can be quantified. Software tools and packages have been

implemented to quantify colour patterns (Troscianko and Stevens 2015) and colour pattern variations (Van Belleghem et al. 2018). These studies and tools have suggested it is worth considering the colour pattern analysis on the Project Plumage dataset.

Further explorations include comparing results between deep learning and crowdsourcing. In Chapter 2 and 3, I only evaluated the differences between deep learning results to annotations labelled by people with expert knowledge on bird anatomy. However, Project Plumage is a citizen science project and is generating large datasets itself. Since completing chapter 2, the Project Plumage workflows for placing points on body regions have been completed including >72,000 sets of landmark placements from >24,000 images of 8409 species. Workflows for segmentation are currently ~50% complete. When citizens have completed labelling Project Plumage images, it will be interesting to estimate annotation differences between deep learning and crowd-sourcing. The result will provide extra evidence on (i) whether deep learning is an ideal solution for phenotyping large scale biodiversity datasets, (ii) whether citizen science data is sufficiently reliable to replace, or supplement expert data as the training set for deep learning.

(iii) Extending PhenoLearn

Adding new functions to PhenoLearn, such as supporting new annotation types or developing a web-based version (Figure 5.2) have the potential to further expand the accessibility of deep learning methods to biologists. New annotations like bounding boxes can be added, so users can annotate bounding boxes and PhenoLearn can predict them using object detection neural networks such as faster R-CNN (Ren et al. 2015). With bounding boxes implemented, animals can be detected on images such as those from camera-traps. Implementing a web-based PhenoLearn can provide a platform for the public to label training images (if they can provide high-quality annotations) and review predictions.

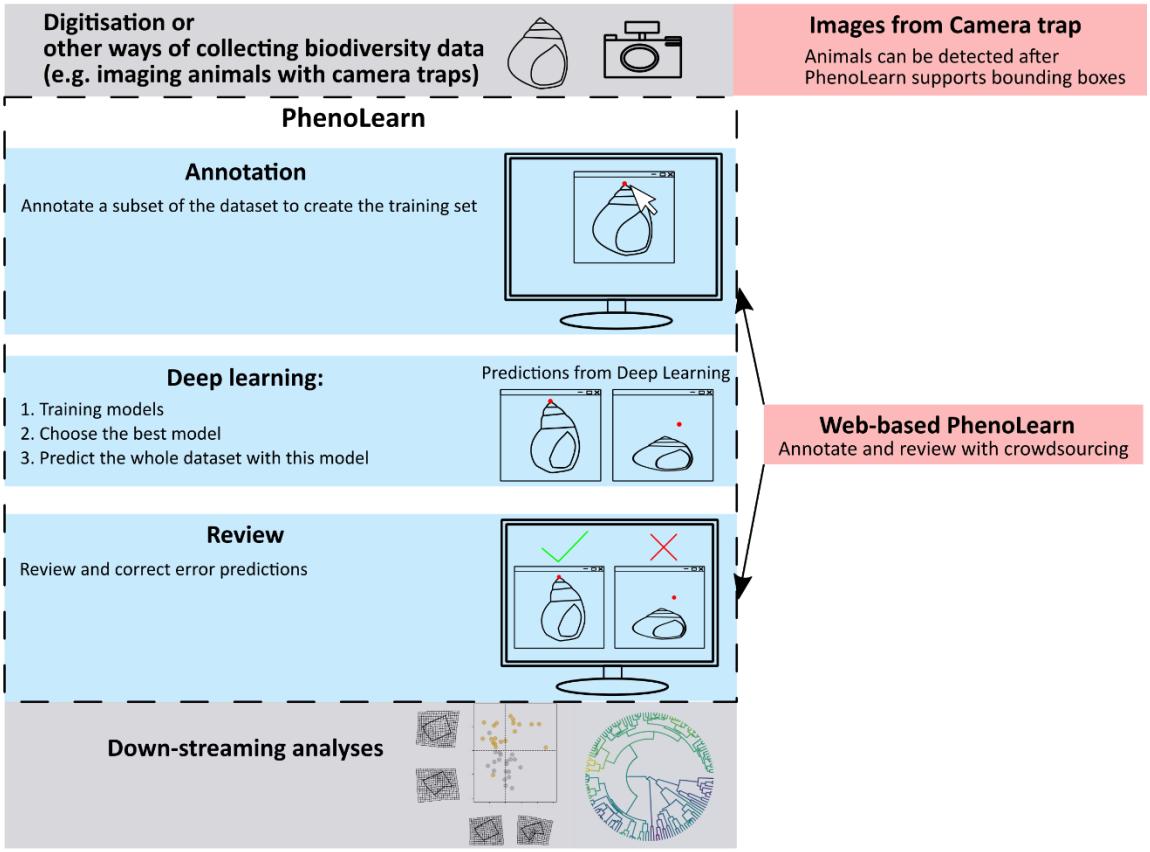


Figure 5.2. Possible new features (red rectangles) that extend PhenoLearn based on the pipeline (left) proposed in Chapter 4. With PhenoLearn supporting bounding boxes, tasks for detecting animals using camera trap images can be carried out on PhenoLearn. The public can participate to annotate and review images with web-based PhenoLearn.

(iv) Other types of natural history data set. Apply deep learning or machine learning to Mark My Bird 3D beaks.

Many studies use geometric morphometrics based on three-dimensional (3D) scans (Cooney et al. 2017; Buser et al. 2018; Giacomini et al. 2019; Felice et al. 2020). Analysing a specimen in 3D can capture both 2D and 3D (i.e. shape variations along the third axis) shape variations, while 2D images from multiple views (e.g. dorsal, ventral and lateral view) are required to capture these features.

A good example dataset is the Mark My Bird dataset (collected from bird collections at Natural History Museum, Tring), which placed 3D landmarks on 3D beak scans (more than 8,000 scans

and they are stored as 3D polygon meshes) to quantify beak shapes (Cooney et al. 2017). Landmarks and semi-landmarks were used to identify homologous structures like the beak tip and curves like tomial edges (see Figure 5.3 for the detail of landmarks). I previously explored automatical landmarking by thresholding 3D mesh properties. Properties like curvatures (Koenderink and van Doorn 1992; Rusinkiewicz 2004) and the shape diameter function (SDF, Shapira et al. 2008) which quantifies whether the space around the vertex is wide or narrow (e.g. the beak tip should have a small SDF value as the space around the tip is narrow) indicate distinct features on 3D data. Using these methods, only candidate points of the beak tip (low SDF values) and tomial edges (low curvature values) can be generated (Figure 5.4a). However, this result is far from the final landmarks (e.g. having too many candidate points) and it can be inaccurate on uncommon bird beaks (Figure 5.4b). Studies have explored classifying and segmenting 3D objects (Guo et al. 2015; Cherabier et al. 2016). A study has estimated 3D postures of *Drosophila* based on 2D images taken from cameras at different angles (Günel et al. 2019). Therefore, it is promising to apply deep learning methods on 3D beak scans, and may well be used in labelling datasets like Phenome10k (phenome10k.org, Goswami 2015), an online repository for 3D scans of biological and palaeontological specimens from the Natural History Museums.

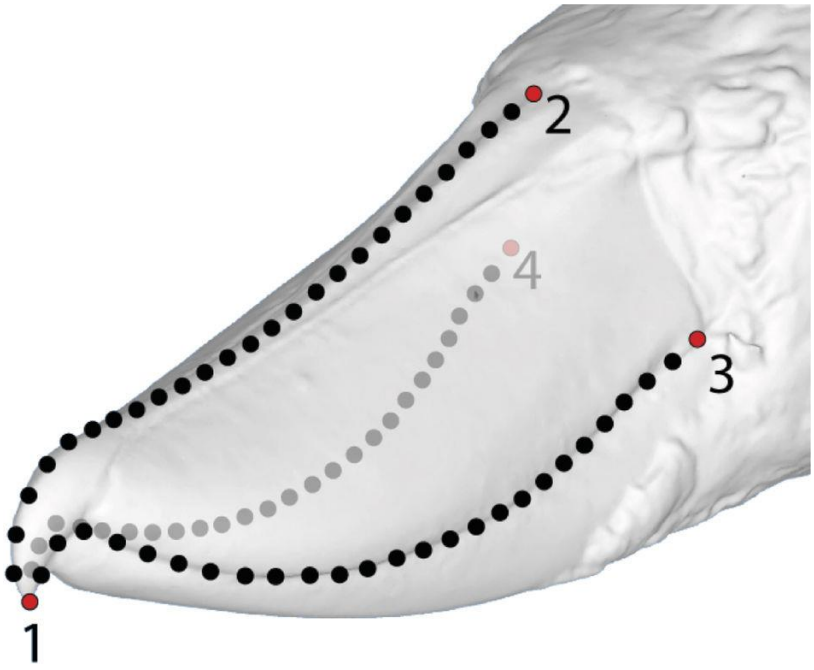


Figure 5.3. Landmarks and semi-landmarks that were placed on scans of the Mark My Bird 3D beak dataset. Four landmarks need to be placed to identify the tip (red point 1), posterior margin along the midline dorsal profile (red point 2), left and right tomial edges (red point 3 and 4). A total of 75 semi-landmarks were placed along the dorsal profile (black points between red point 1 and 2, N=25), left (black points between red point 1 and 3, N=25) and right (black points between red point 1 and 4, N=25) tomial edges.

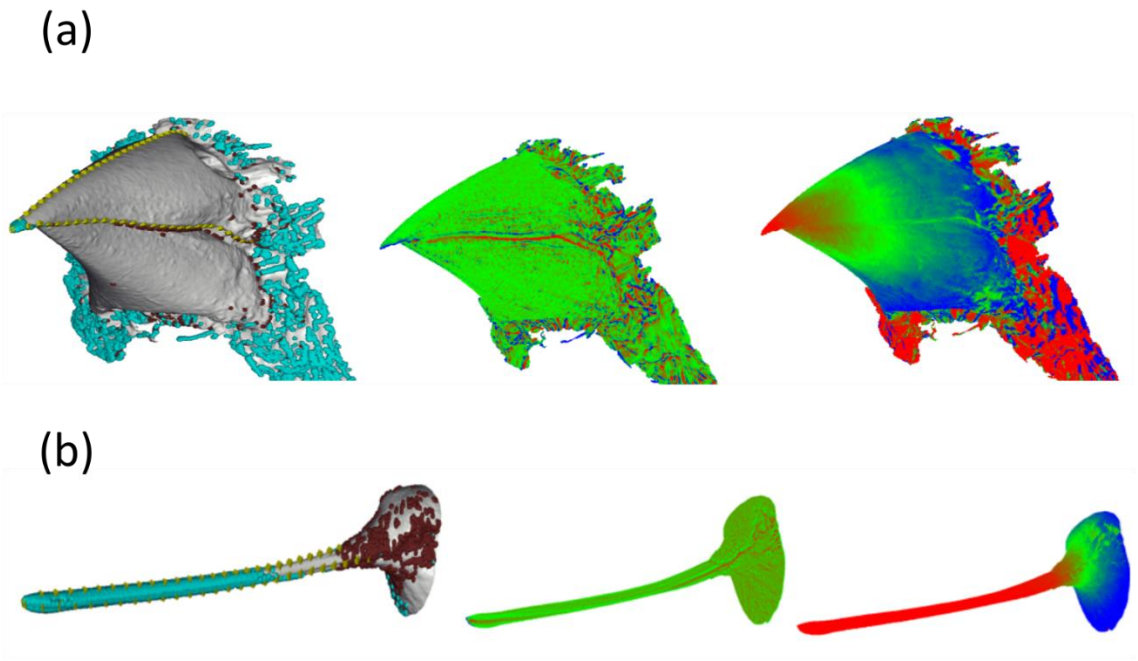


Figure 5.4. Example results of applying mesh properties on (a) a common beak and (b) an uncommon beak. Left: Candidate points of the beak tip (brown points) and tomial edges (cyan points) and the manually-labelled landmarks (yellow points). Centre: Mean curvature values (high: blue, medium: green, low: red). Right: SDF values (high: blue, medium: green, low: red).

5.5 Conclusion

In this thesis, I have explored the use of deep learning models and other computational algorithms in improving the speed of phenotypic measurements (e.g. placing annotations on images to measure phenotypic traits) while remaining reliably accurate compared to measurements from traditional methods (e.g. manual annotations). My results were encouraging, suggesting that deep learning has a great deal of potential for high throughput phenotyping of specimens. Two datasets of digitised specimen photos were assessed, a bird specimen dataset and a *Littorina* shell dataset. Points and segmentations needed to be placed on bird images to measure plumage colours from body regions and the overall bird. Deep learning placed 95% of points and more than 95% segmentations (based on correctly segmented pixels) correctly among 5,094 expert-labelled images. I then generated points and segmentations for all bird images (more than 120,000 images covering more than 8,000 bird species) in less than a week, which could take years for manual labelling. Bird plumage colour spaces across more than 7,000 bird

species were built, and I found that there is a moderately strong phylogenetic signal in bird plumage colour diversity (closely related species share similar colour diversity), as well as male birds tend to be more colour-diverse than female ones which are from the same species. Morphological landmarks were placed by deep learning and additional computational algorithms. Morphospaces from manual and deep learning landmarks were similar. PhenoLearn, a software tool which aims for setting up high-throughput phenotyping pipelines using deep learning models, was introduced in this thesis. Its functions which cover manual annotation, applying deep learning models and reviewing deep learning predictions are all implemented with user interfaces. In summary, the results from this thesis suggest deep learning can speed up the measurement process on digitised specimens (either from museum collections or live specimens) while providing accurate and reliable measurements.

Chapter 6 Appendix

6.1 Chapter 2 supplementary Material

6.1.1 Alpha shape

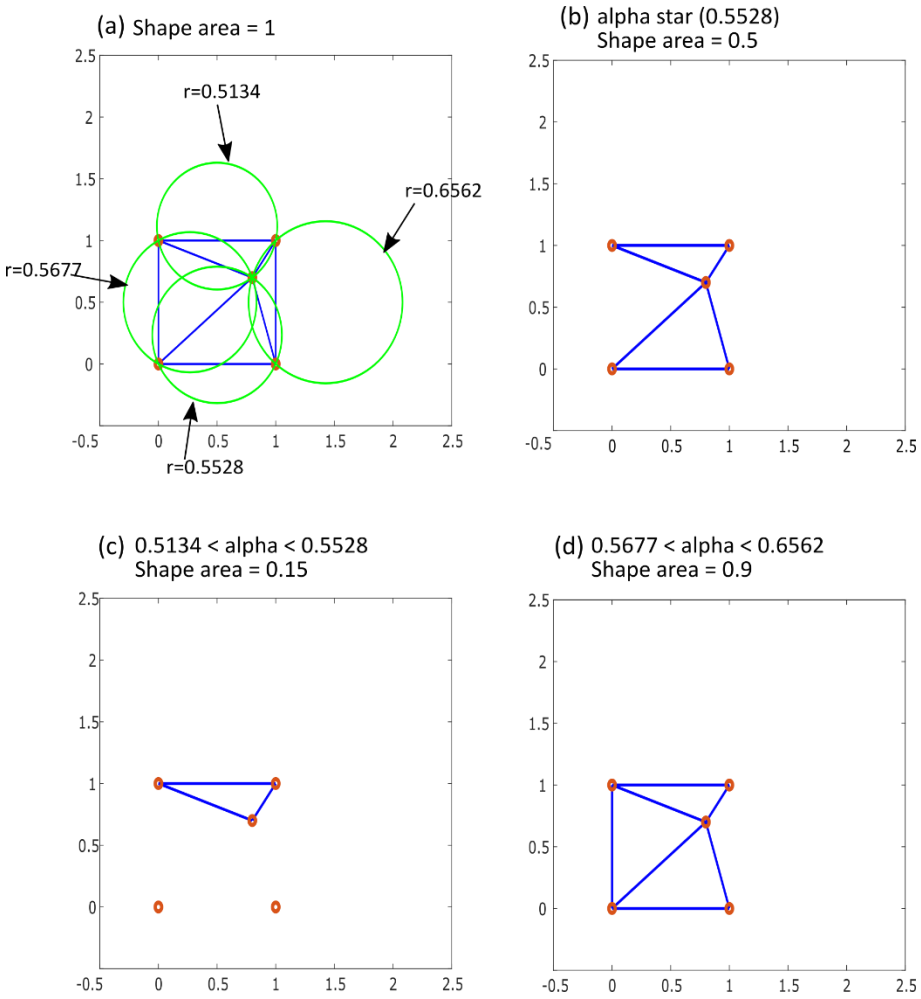
Here, I introduce the alpha shape definition in 2D space, which is easy to illustrate. The 3D alpha shape has the same definition as the 2D version after replacing triangles with tetrahedrons, circumcircles with circumspheres and areas with volumes (Edelsbrunner and Mücke 1994).

Before calculating the alpha shape, the Delaunay triangulation need to be first applied to the data points. The Delaunay triangulation of a set of points are triangles that use data points as vertices, while no points are inside circumcircles of any triangles as shown in Supplementary Figure 6.1.1a. The shape of the Delaunay triangulation is identical to the convex hull. An alpha shape with a given α value is the shape of triangles with circumcircle radii smaller than α (Edelsbrunner and Mücke 1994; Cholewo and Love 1999). α* is the α value that makes the area of the alpha shape smallest while containing all points (Supplementary Figure 6.1.1b). Alpha shapes with α values smaller than α* can not contain all data points (Supplementary Figure 6.1.1c). An α value larger than α* may increase the area (e.g. Supplementary Figure 6.1.1d).

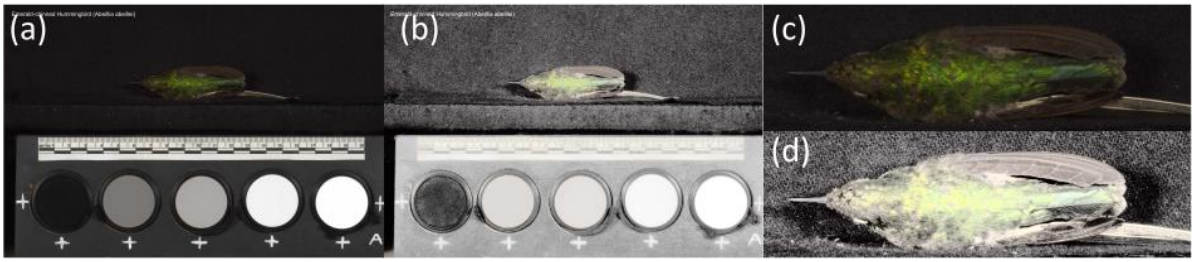
α* varies among different sets of data points so that, for example, α* for colour measures from the crown may not be the same as α* measured from the belly. Using respective α* for every alpha shape may make volume comparisons inaccurate and inconsistent. Supplementary Figure 6.1.4a shows an example of how alpha shapes (using its own α*) are not ideal for comparison (the example is in 2D, the area is the equivalent value for the volume in 3D). Dataset A has a greater convex hull area than Dataset B (Supplementary Figure 6.1.4a.1 and a.2), but a smaller alpha shape area (Supplementary Figure 6.1.4a.3 and a.4). One reason is that there are many points along the x-axis for Dataset A, forming many triangles with small circumcircle radii in the Delaunay triangulation (Cholewo and Love 1999). So the α* can be small and many triangles can be removed to create a very concave alpha shape (see section 6.1.1 and Supplementary Figure 6.1.1). On the contrary, Dataset B has sparse points across the x-axis, and most of the triangles have similar circumcircle radii. There is not an α value that can remove many triangles while

containing all points in the alpha shape. So the alpha shape of Dataset B (Supplementary Figure 6.1.4a.4) looks less concave than the one in Dataset A (Supplementary Figure 6.1.4a.3). The problem can be seen in Supplementary Figure 6.1.4c showing alpha shapes of flight feathers for both sexes combined, males-only, and females-only using α*. The female α* and alpha volumes are larger than the male volume or the total volume of males and females. The female alpha shape volume is very similar to its convex hull volume (Supplementary Figure 6.1.4b). This non-intuitive result illustrates the difficulty in comparing volumes based on different α*.

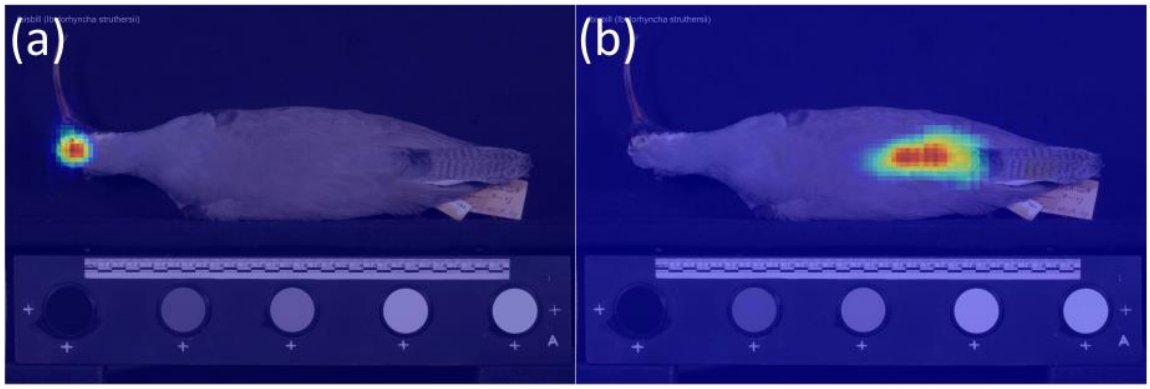
6.1.2 Supplementary Figures



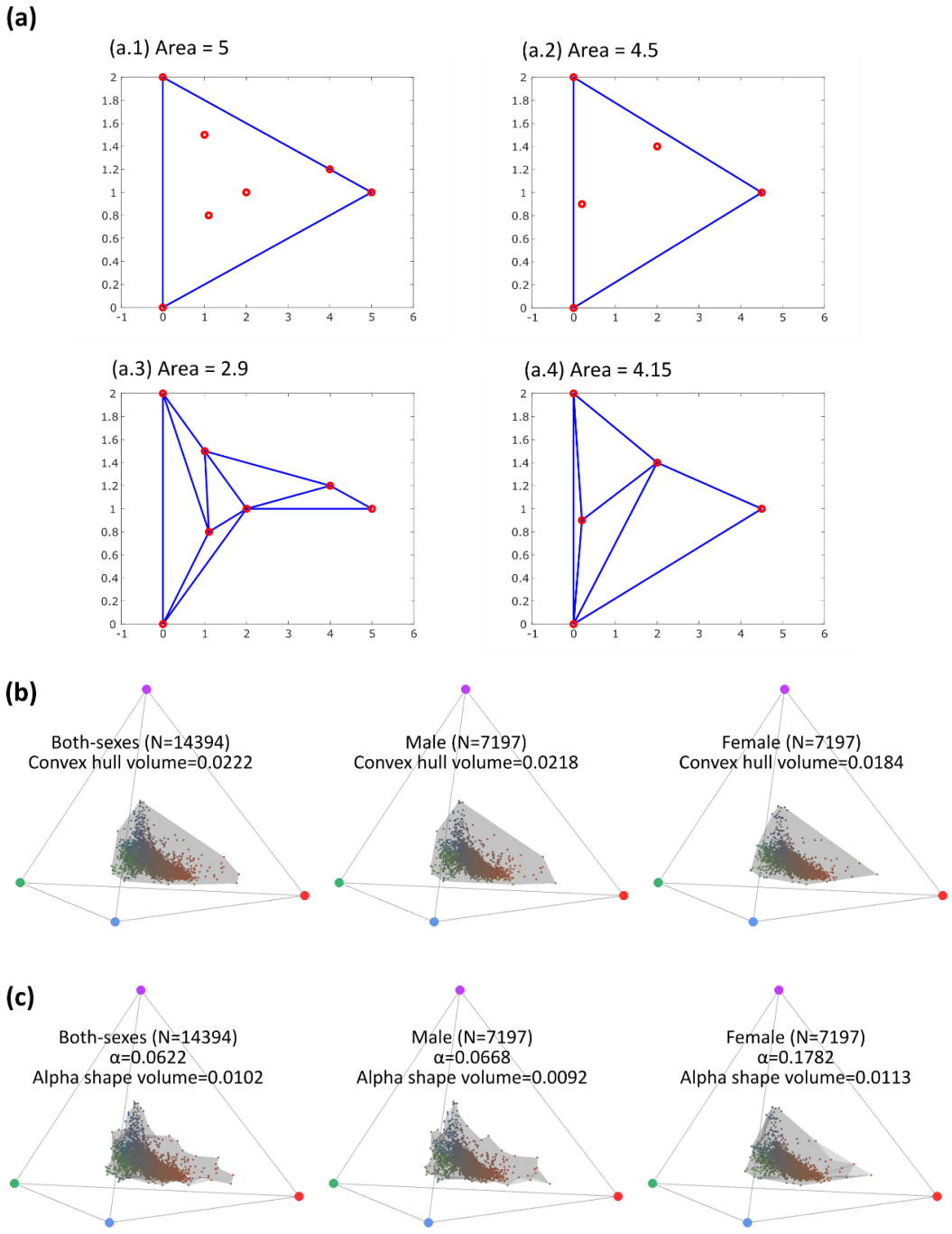
Supplementary Figure 6.1.1. An example of the relation between alpha shape areas and the α values for 5 data points (red points) in 2D. (a) The Delaunay triangulation of the data, which contains 4 triangles (blue triangles). Green circles are the circumcircles of the 4 triangles. The shape of the triangulation is identical to the convex hull or the alpha shape with an α value larger than the radius of the largest circumcircle which is 0.6562 in this example. (b) The alpha shape using α* (0. 5528). Triangles with circumcircle radii larger than 0. 5528 are removed, while all data points are included in the alpha shape (c) The alpha shape using an α value between 0.5134 to 0.5528 (smaller than α*). Only the triangle on the top is kept as it has a circumcircle radius of 0.5134. (d) The alpha shape using an α value between 0.5677 to 0.6562. Only the triangle on the right is removed because it has a circumcircle radius of 0.6562.



Supplementary Figure 6.1.2. Examples of pre-processed photos. (a) The original image; (b) applying histogram equalisation on the original image; (c) the specimen-only image after cropping; (d) the histogram equalised specimen-only image.

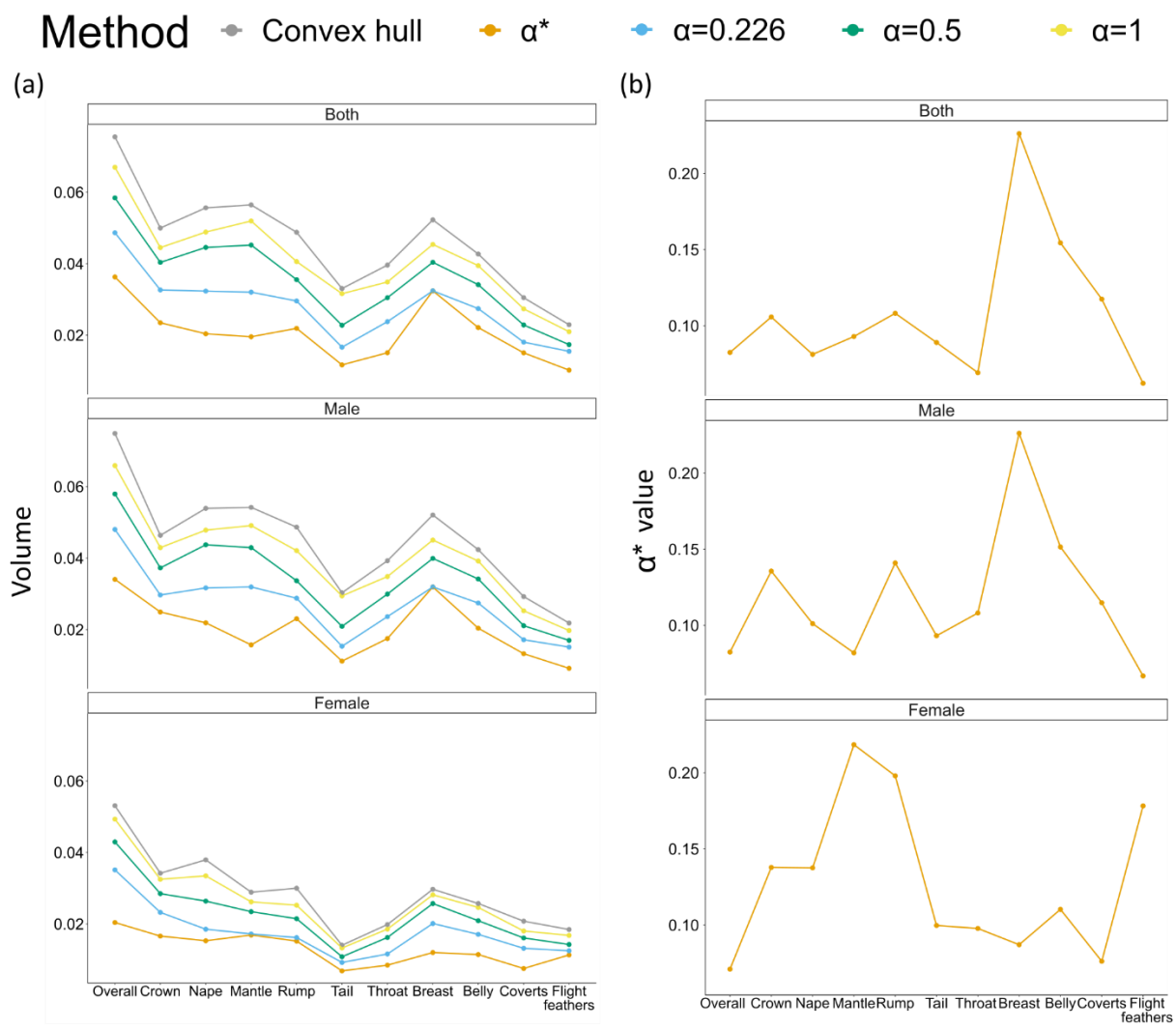


Supplementary Figure 6.1.3. Predicted heatmaps of two body regions on an example image. (a) The predicted heatmap of the crown has a smaller area than the one of the (b) rump which has an ellipse-like shape and can capture more area of the rump region than using a fixed-size area.

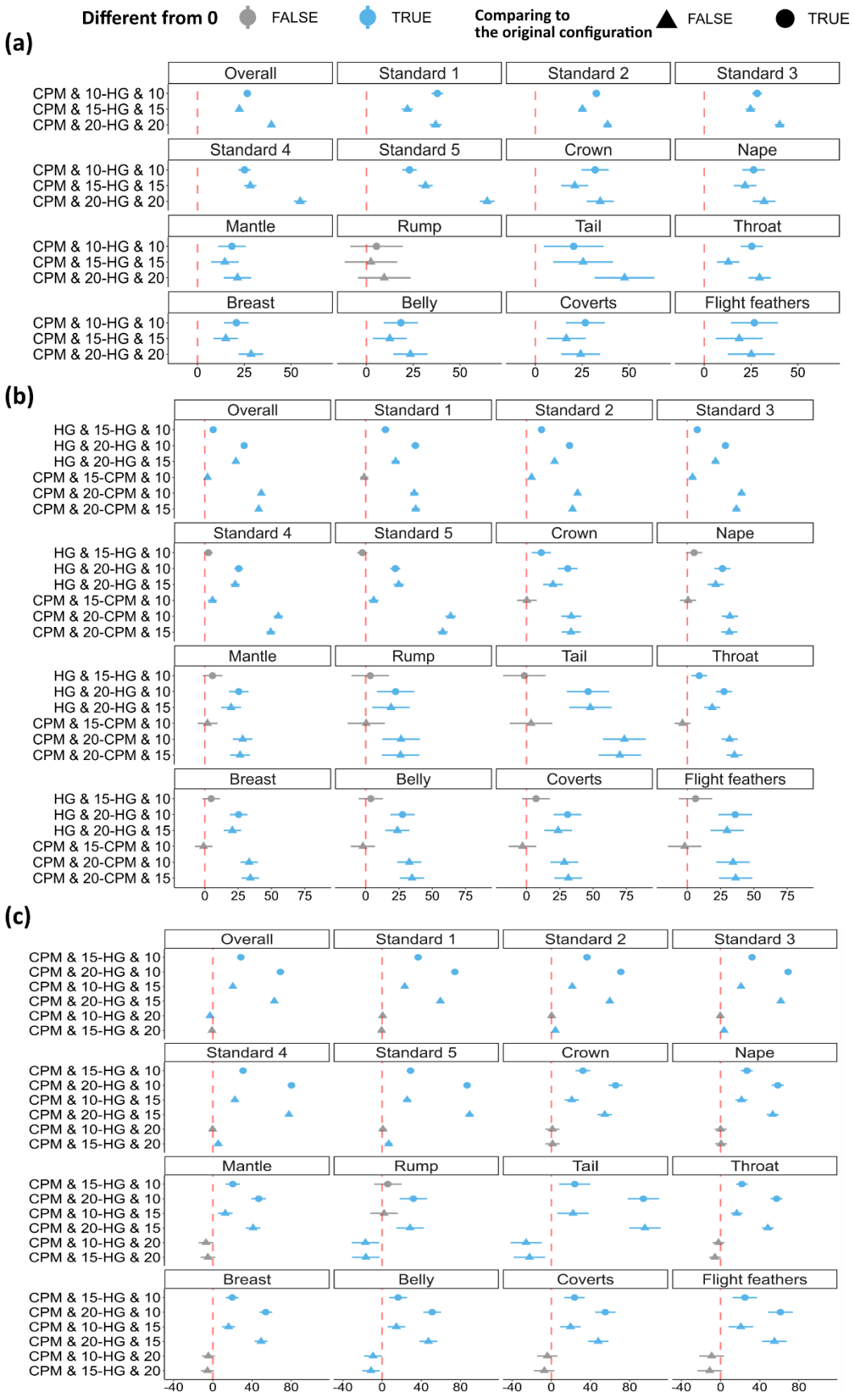


Supplementary Figure 6.1.4. (a.1) The convex hull of Dataset A, which has an area of 5; (a.2) The convex hull of Dataset B which has an area of 4.5; (a.3) The alpha shape using α* of Dataset A, which has an area of 2.9; (a.4) The alpha shape using α* of Dataset B, which has an area of 4.15. (b) Convex hulls of the flight

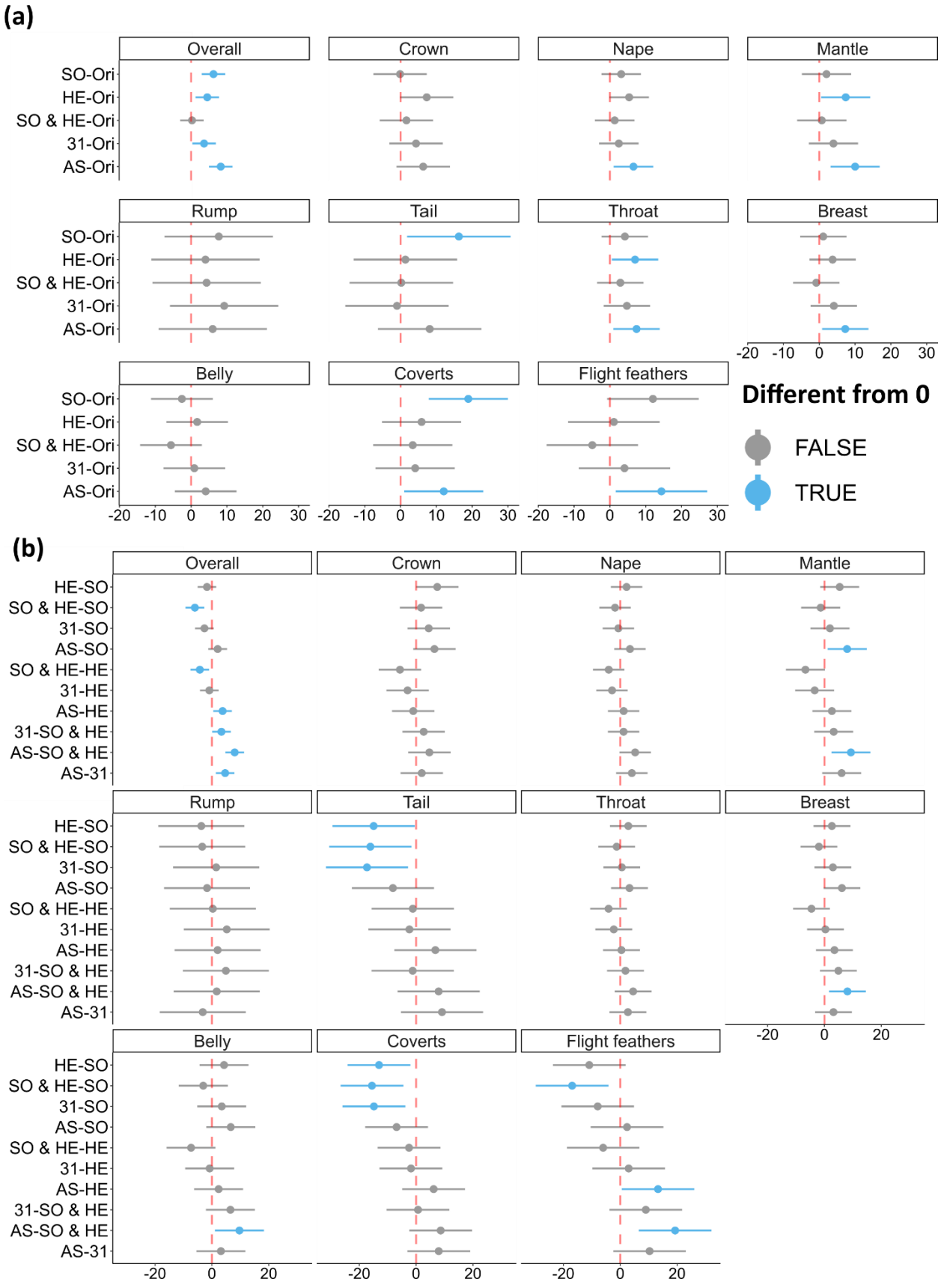
feathers for both sexes, male and female. (c) Alpha shapes using respective α* of the flight feathers for both sexes, male and female.



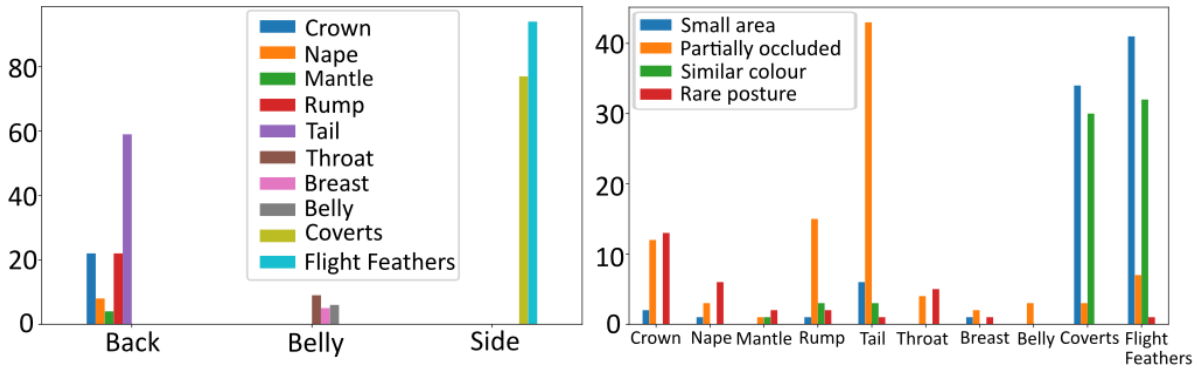
Supplementary Figure 6.1.5. (a) Volumes of overall and individual patches' colour points for both sexes, male and female using convex hull and alpha shape with four α values which are respective α*, α=0.226 (the largest α* among these patches), α=0.5 and α=1. (b) α* values of the all and individual body patches for both sexes, male and female.


Different from 0
FALSE
TRUE
Comparing to the original configuration
FALSE
TRUE
(a)
Overall
Standard 1
Standard 2
Standard 3
Standard 4
Standard 5
Crown
Nape
Mantle
Rump
Tail
Throat
Breast
Belly
Coverts
Flight feathers
CPM & 10-HG & 10
CPM & 15-HG & 15
CPM & 20-HG & 20
0
25
50
(b)
HG & 15-HG & 10
HG & 20-HG & 10
HG & 20-HG & 15
CPM & 15-CPM & 10
CPM & 20-CPM & 10
CPM & 20-CPM & 15
0
25
50
75
(c)
CPM & 15-HG & 10
CPM & 20-HG & 10
CPM & 10-HG & 15
CPM & 20-HG & 15
CPM & 10-HG & 20
CPM & 15-HG & 20
-40
0
40
80

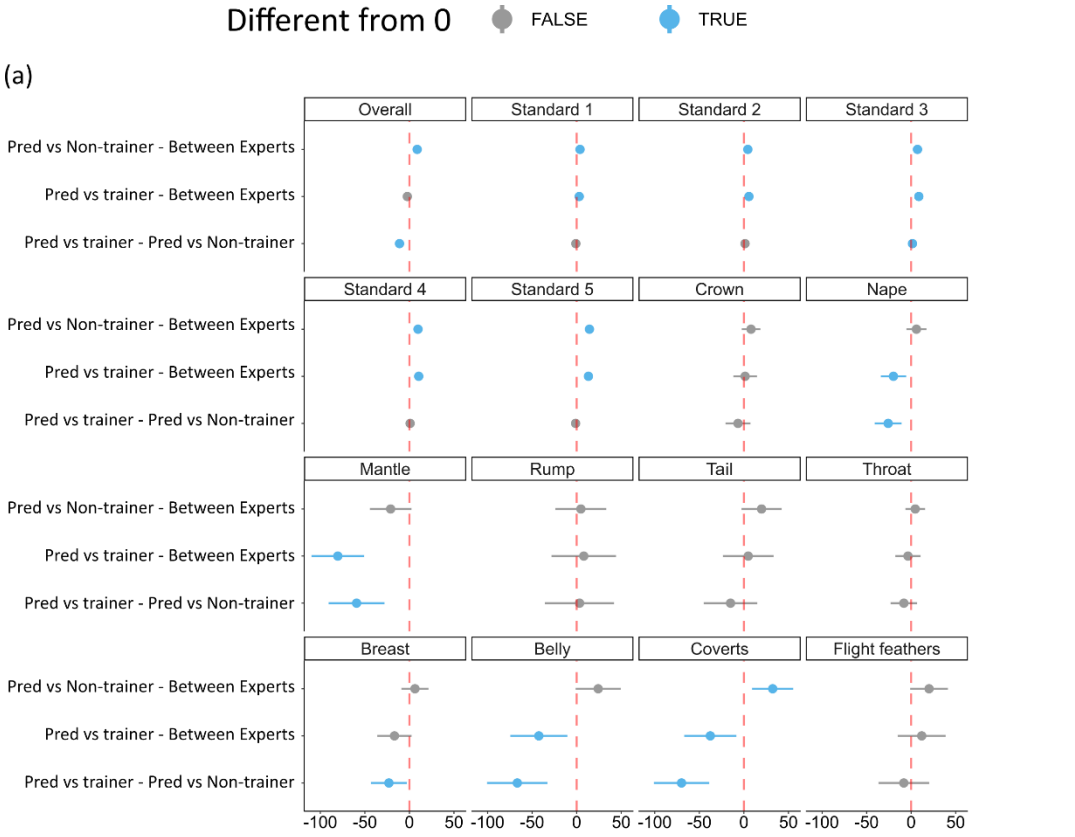
Supplementary Figure 6.1.6. Plots of Tukey's test (95% family-wise confidence level) on whether pixel distances differences of all and individual keypoints between architectures and resolutions are significantly different (blue: significance; grey: no significance) from 0 (red dotted lines). (a) Comparisons for configurations with different networks but the same resolution. (b) Comparisons for configurations with different resolutions but the same network. (c) Rest pair-wise Comparisons. Differences that include the original configuration (Stacked Hourglass with the resolution of 494 x 328 pixels) are plotted using the circle shape, while the rest of the differences are plotted using the triangle shape. (Abbreviations: HG: Stacked Hourglass; CPM: Convolutional pose machine; 10: Resolution of 494 x 328 pixels; 15: Resolution of 329 x 218 pixels; 20: Resolution of 247 x 164 pixels)

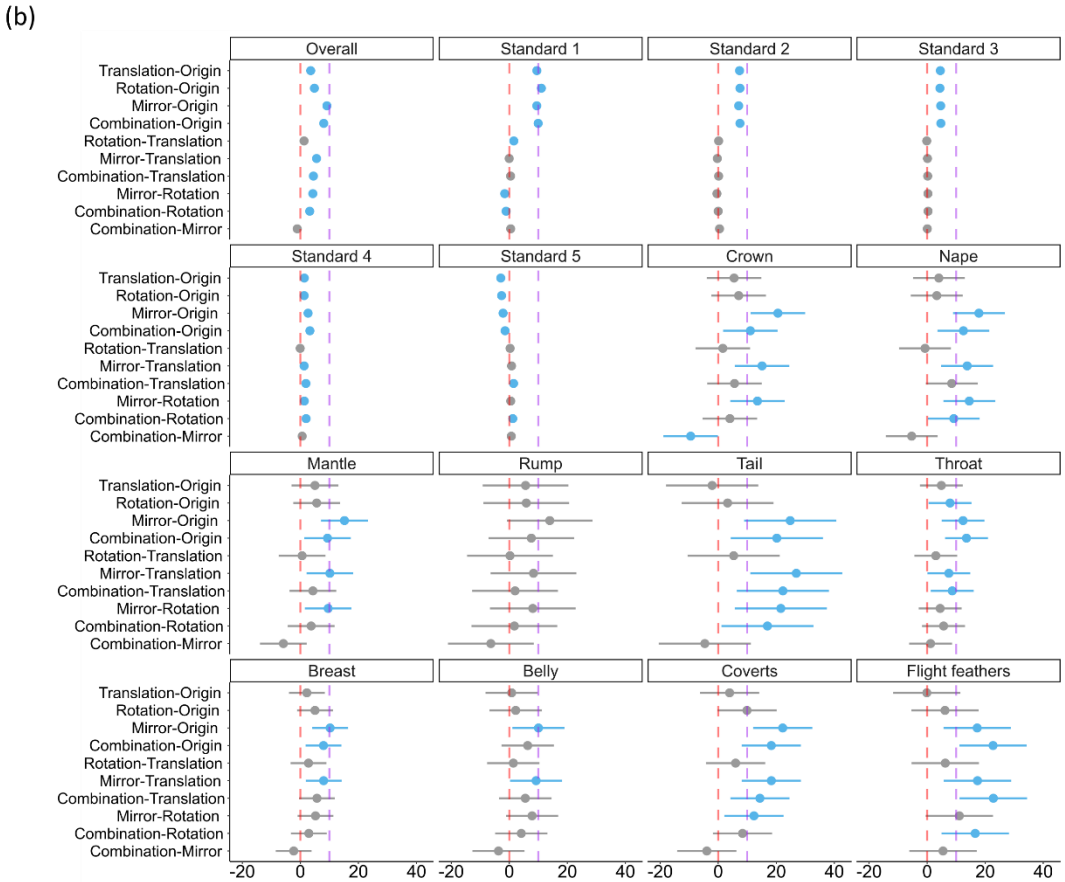

(a)
Overall
Crown
Nape
Mantle
Rump
Tail
Throat
Breast
Belly
Coverts
Flight feathers
SO-Ori
HE-Ori
SO & HE-Ori
31-Ori
AS-Ori
-20 -10 0 10 20 30
Different from 0
FALSE
TRUE
(b)
HE-SO
SO & HE-SO
31-SO
AS-SO
SO & HE-HE
31-HE
AS-HE
31-SO & HE
AS-SO & HE
AS-31
-20 0 20

Supplementary Figure 6.1.7. Plots of Tukey's test (95% family-wise confidence level) on whether pixel distances differences of all and each individual body region points between the original configuration (Stacked Hourglass, 494 x328 pixels and 15 epochs) and experimental manipulations are significantly different (blue: significance; grey: no significance) from 0 (red dotted lines). (a) Comparisons that include the original configuration. (b) Comparisons for that do not include the original configuration. (Abbreviations, Ori: Original configuration; SO: Specimen-only images; HE: Histogram equalised images; SO & HE: Histogram equalised specimen-only images; 31: Model trained for 31 epochs; AS: Image augmentation and model subsetting).

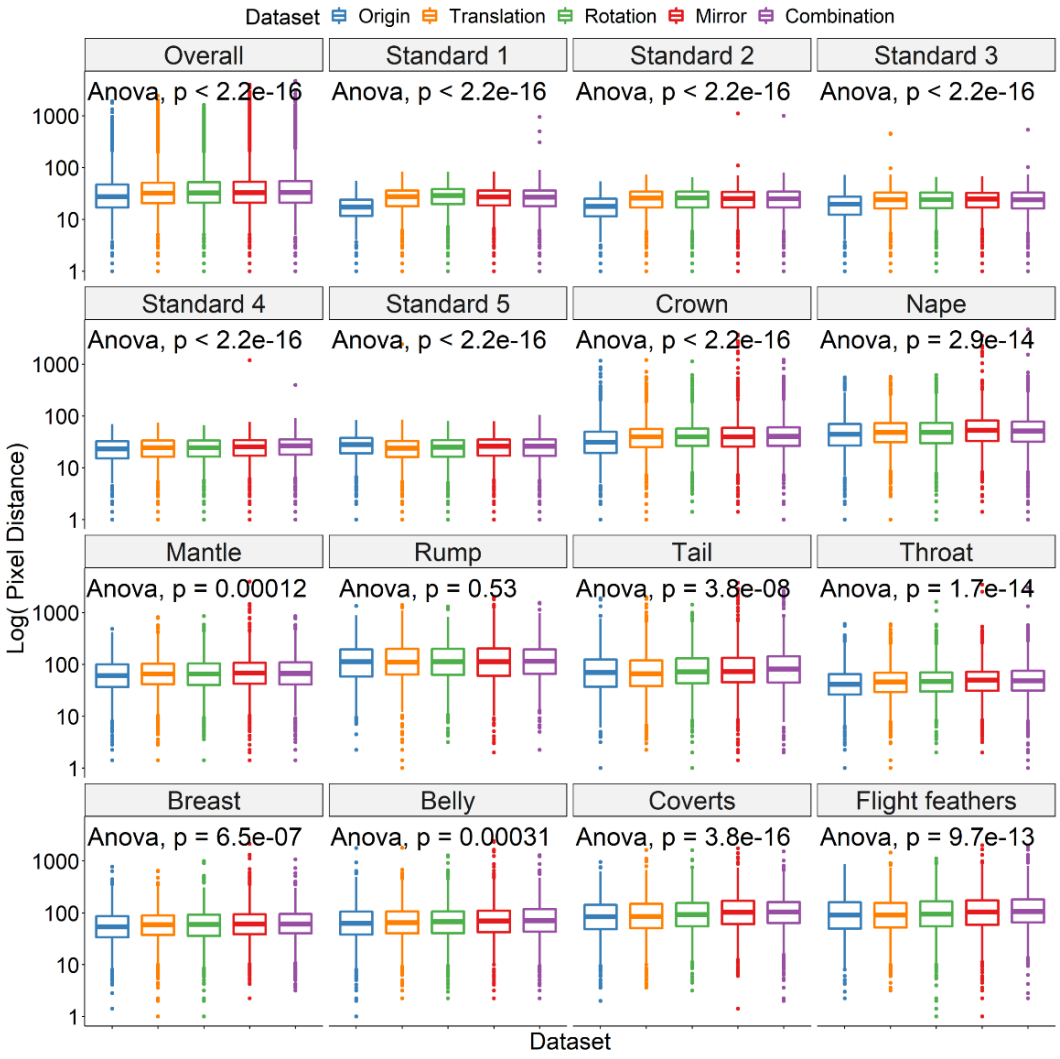


Supplementary Figure 6.1.8. Histograms of the expert evaluation (a) Error prediction counts (N=308) of each body region (N=308). (b) Feature counts of error predicts from each body regions.

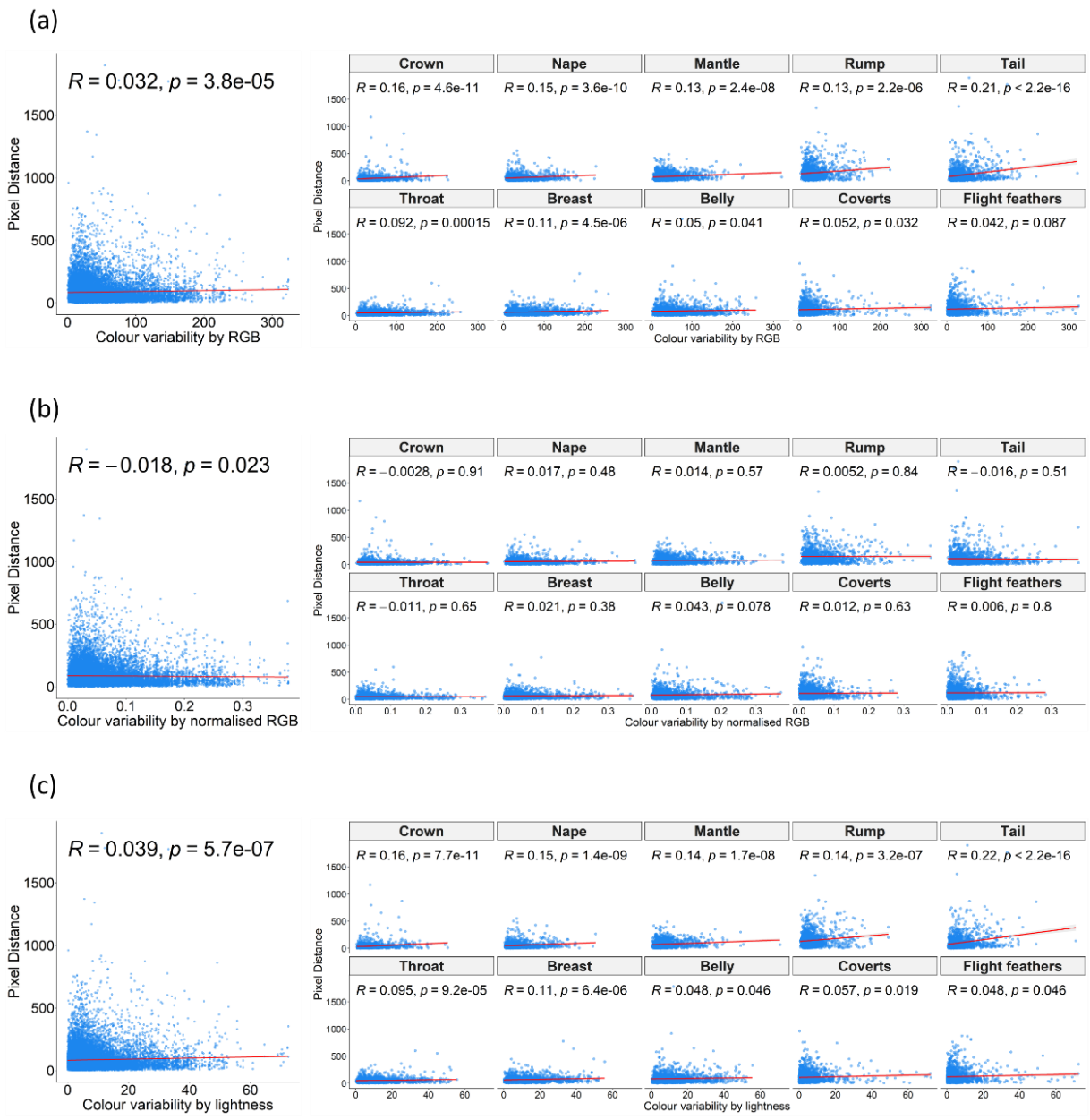

Different from 0
FALSE
TRUE
(a)
Overall
Standard 1
Standard 2
Standard 3
Standard 4
Standard 5
Crown
Nape
Mantle
Rump
Tail
Throat
Breast
Belly
Coverts
Flight feathers
Pred vs Non-trainer - Between Experts
Pred vs trainer - Between Experts
Pred vs trainer - Pred vs Non-trainer
-100
-50
0
50


(b)
Overall
Standard 1
Standard 2
Standard 3
Standard 4
Standard 5
Crown
Nape
Mantle
Rump
Tail
Throat
Breast
Belly
Coverts
Flight feathers
Translation-Origin
Rotation-Origin
Mirror-Origin
Combination-Origin
Rotation-Translation
Mirror-Translation
Combination-Translation
Mirror-Rotation
Combination-Rotation
Combination-Mirror
-20
0
20
40

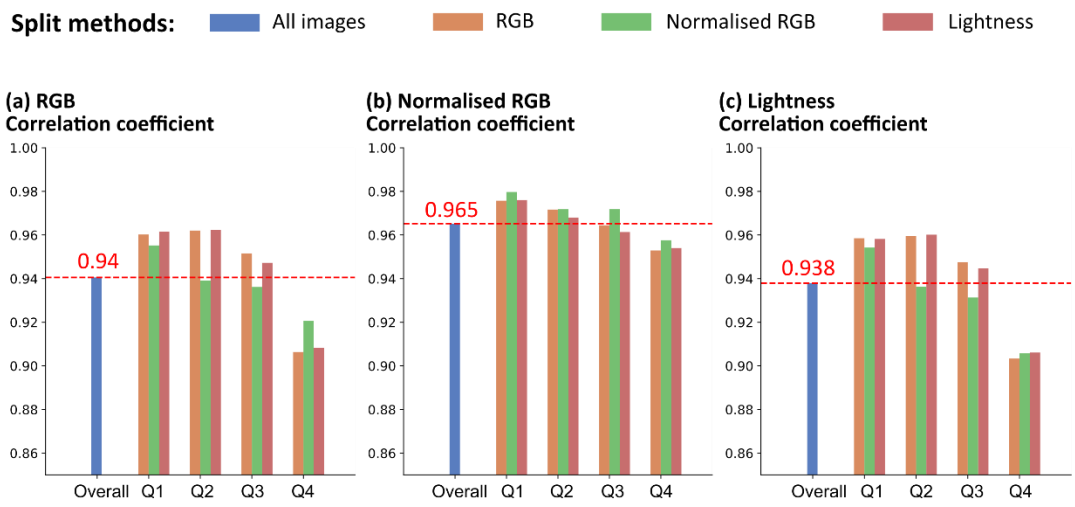
Supplementary Figure 6.1.9. (a) Plots of Tukey's test (95% family-wise confidence level) on whether pixel distances differences of all and individual keypoints across three groups from section 2.2.5.1 (predictions vs trainer, predictions vs non-trainer and between experts) are significantly different (blue: significance; grey: no significance) from 0 (red dotted lines). (b) Plots of Tukey's test (95% family-wise confidence level) on whether pixel distances differences of all and individual keypoints across the original and low-quality datasets are significantly different (blue: significance; grey: no significance) from 0 (red dotted lines). Purple dotted lines are pixel distance of 10.



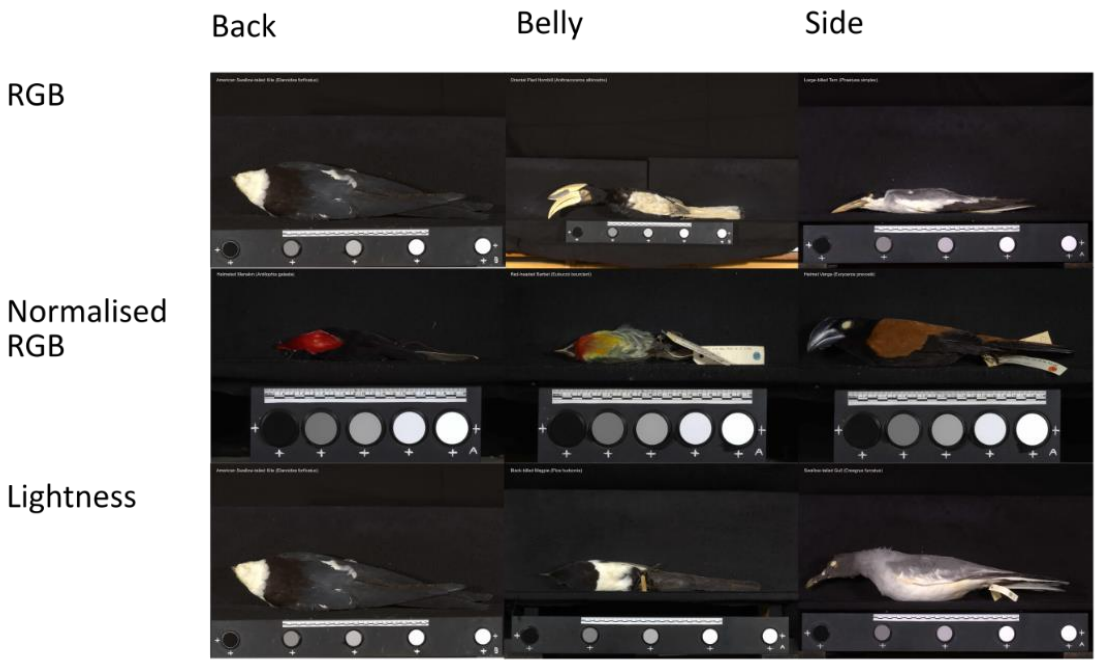
Supplementary Figure 6.1.10. Pixel distances of all and individual keypoints for the original and four low-quality datasets.



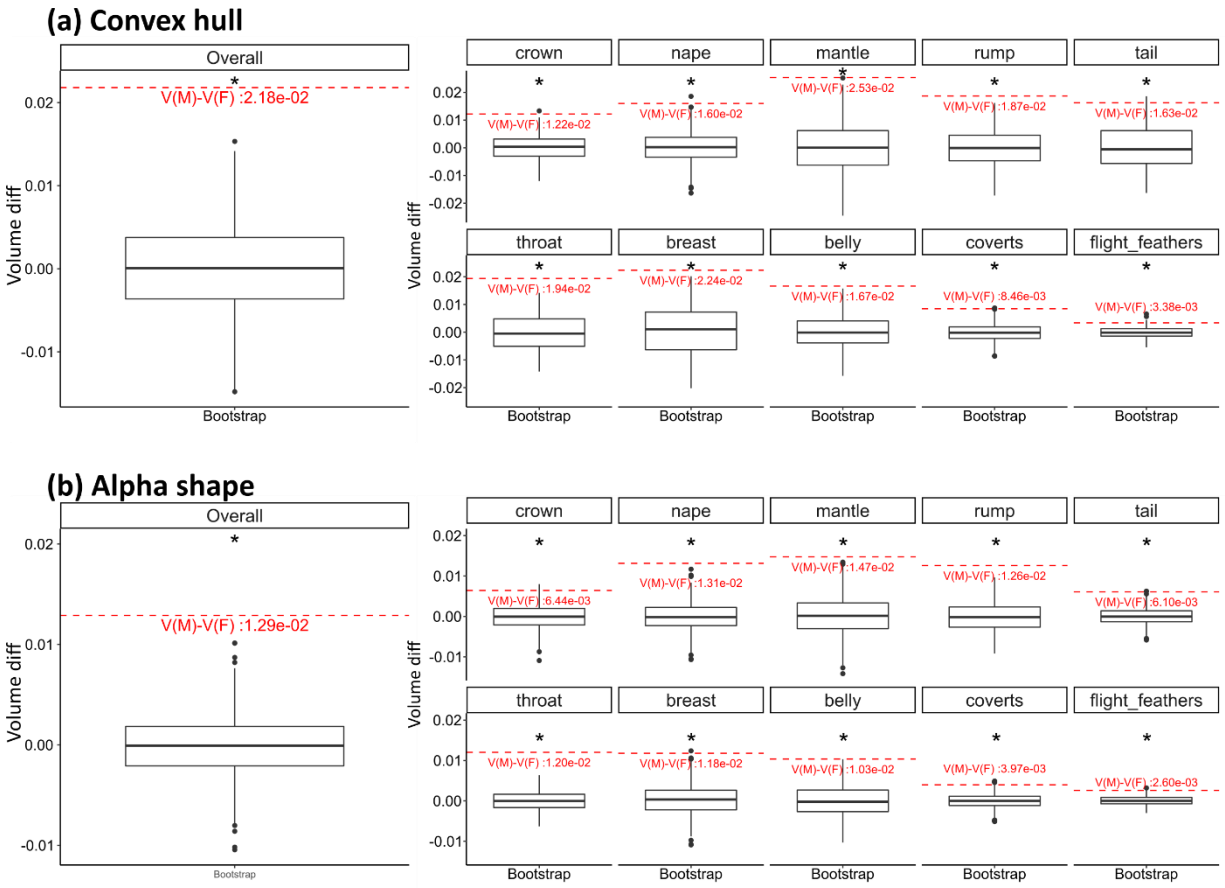
Supplementary Figure 6.1.11. Correlations between pixel distances and colour variabilities measured by (a) RGB, (b) normalised RGB and (c) lightness of all and individual body regions.



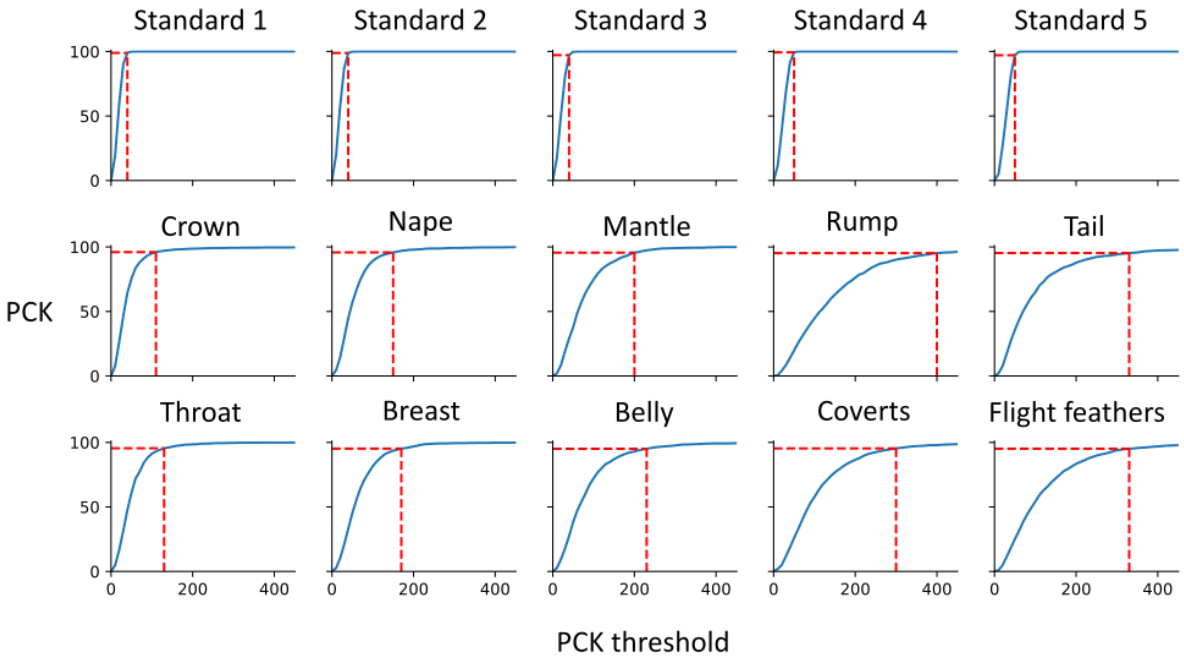
Supplementary Figure 6.1.12. RGB, normalised RGB and lightness correlation coefficients of the overall prediction and predictions from each quartile split by colour variabilities measured by (a) RGB, (b) normalised RGB and (c) lightness.



Supplementary Figure 6.1.13. Images with the highest colour variabilities (average pair-wise colour distances of RGB, normalised RGB and lightness) of each view (back, belly and side).



Supplementary Figure 6.1.14. Comparing (a) Convex hull and (b) alpha shape (α=0.2262) volume differences between male and female to 1000 volume differences between two groups of random-split data points for overall and individual body patches. The red dotted lines are the volume differences between male and female. * means the volume difference between male and female is significantly different from the difference between two random groups and NS means no significance.



Supplementary Figure 6.1.15. Plots of PCK against PCK thresholds (0 to 500 pixels) of 15 keypoints. Red dotted lines are where PCKs score more than 95.

6.1.3 Supplementary Tables

Supplementary Table 6.1.1. Volume differences between both sexes and male (B-M), both sexes and female (B-F) and female and male (M-F) from 5 volume calculation methods (Convex hull, alpha shape using α*, alpha shape using α=0.2262, alpha shape using α=0.5 and alpha shape using α=1). Black numbers are differences above zero and red numbers are differences.

	Convex hull			**α***		
	B-M	B-F	M-F	B-M	B-F	M-F
Overall	6.05E-04	2.24E-02	2.18E-02	2.25E-03	1.59E-02	1.37E-02
crown	3.62E-03	1.58E-02	1.22E-02	-1.43E-03	6.88E-03	8.31E-03
nape	1.72E-03	1.77E-02	1.60E-02	-1.49E-03	5.10E-03	6.59E-03
mantle	2.27E-03	2.76E-02	2.53E-02	3.85E-03	2.67E-03	-1.18E-03
rump	1.46E-04	1.88E-02	1.87E-02	-1.13E-03	6.70E-03	7.83E-03
tail	2.75E-03	1.90E-02	1.63E-02	5.22E-04	4.83E-03	4.31E-03
throat	3.60E-04	1.98E-02	1.94E-02	-2.39E-03	6.62E-03	9.01E-03
breast	2.17E-04	2.26E-02	2.24E-02	4.89E-04	2.04E-02	1.99E-02
belly	3.40E-04	1.70E-02	1.67E-02	1.76E-03	1.07E-02	8.94E-03
coverts	1.26E-03	9.72E-03	8.46E-03	1.82E-03	7.52E-03	5.71E-03
Flight feathers	1.10E-03	4.48E-03	3.38E-03	1.06E-03	-1.11E-03	-2.18E-03

	The Largest α*, α=0.2262			**α=0.5**		
	B-M	B-F	M-F	B-M	B-F	M-F
Overall	6.52E-04	1.35E-02	1.29E-02	4.88E-04	1.55E-02	1.50E-02
crown	2.98E-03	9.42E-03	6.44E-03	3.12E-03	1.19E-02	8.78E-03
nape	6.80E-04	1.38E-02	1.31E-02	8.39E-04	1.82E-02	1.73E-02
mantle	1.06E-04	1.48E-02	1.47E-02	2.29E-03	2.18E-02	1.95E-02
rump	8.01E-04	1.34E-02	1.26E-02	1.89E-03	1.41E-02	1.22E-02
tail	1.31E-03	7.41E-03	6.10E-03	1.86E-03	1.19E-02	1.01E-02
throat	1.67E-04	1.22E-02	1.20E-02	5.54E-04	1.43E-02	1.37E-02
breast	4.89E-04	1.23E-02	1.18E-02	4.86E-04	1.47E-02	1.42E-02
belly	1.03E-05	1.04E-02	1.03E-02	-2.30E-05	1.32E-02	1.33E-02
coverts	9.00E-04	4.87E-03	3.97E-03	1.77E-03	6.77E-03	5.00E-03
Flight feathers	3.87E-04	2.99E-03	2.60E-03	3.95E-04	3.11E-03	2.71E-03

	α=1		
	B-M	B-F	M-F
Overall	1.10E-03	1.77E-02	1.66E-02
crown	1.58E-03	1.20E-02	1.04E-02
nape	1.03E-03	1.54E-02	1.44E-02
mantle	2.86E-03	2.58E-02	2.29E-02
rump	-1.48E-03	1.54E-02	1.69E-02
tail	2.20E-03	1.83E-02	1.62E-02
throat	4.85E-05	1.63E-02	1.63E-02
breast	3.55E-04	1.73E-02	1.69E-02
belly	2.81E-04	1.48E-02	1.46E-02

coverts	2.11E-03	9.35E-03	7.24E-03
Flight feathers	1.24E-03	4.18E-03	2.93E-03

Supplementary Table 6.1.2. ANOVA results on pixel distances of overall and individual keypoints across tested input image resolutions (494 x 328, 329 x 218 and 247 x 164 pixels) and network architectures (Stacked hourglass and CPM).

	ANOVA
Overall	F=3529.6; df=5.0, 252864.0; p<0.01
Standard 1	F=1093.3; df=5.0, 30558.0; p<0.01
Standard 2	F=2263.7; df=5.0, 30558.0; p<0.01
Standard 3	F=1368.7; df=5.0, 30558.0; p<0.01
Standard 4	F=1242.7; df=5.0, 30558.0; p<0.01
Standard 5	F=1142.0; df=5.0, 30558.0; p<0.01
Crown	F=154.0; df=5.0, 10164.0; p<0.01
Nape	F=189.3; df=5.0, 10164.0; p<0.01
Mantle	F=79.9; df=5.0, 10176.0; p<0.01
Rump	F=13.2; df=5.0, 8526.0; p<0.01
Tail	F=81.2; df=5.0, 10062.0; p<0.01
Throat	F=174.4; df=5.0, 10182.0; p<0.01
Breast	F=136.2; df=5.0, 10182.0; p<0.01
Belly	F=66.8; df=5.0, 10182.0; p<0.01
Coverts	F=56.4; df=5.0, 10170.0; p<0.01
Feathers	F=49.7; df=5.0, 10182.0; p<0.01

Supplementary Table 6.1.3. ANOVA results on pixel distances of overall and individual keypoints across the original model (Stacked hourglass with the input resolution of 494 x 328 pixels, the training duration of 15 epochs and using the unmanipulated images and labels) and the five tested experimental manipulations (using specimen-only images, using histogram equalised images, using histogram equalised specimen-only images, trained for 31 epochs and using image augmentation and model subsetting. See section 2.2.4 for detail).

	ANOVA
Overall	F=16.1; df=5.0, 100044.0; p<0.01
Crown	F=3.0; df=5.0, 10164.0; p<0.01
Nape	F=3.2; df=5.0, 10164.0; p<0.01
Mantle	F=5.4; df=5.0, 10176.0; p<0.01
Rump	F=0.7; df=5.0, 8526.0; p=0.589
Tail	F=3.6; df=5.0, 10062.0; p<0.01
Throat	F=2.9; df=5.0, 10182.0; p=0.012
Breast	F=3.6; df=5.0, 10182.0; p<0.01
Belly	F=2.5; df=5.0, 10182.0; p=0.026
Coverts	F=6.4; df=5.0, 10170.0; p<0.01
Flight feathers	F=5.5; df=5.0, 10182.0; p<0.01

Supplementary Table 6.1.4. Numbers of error images and error rates per order among the 5,094 expert-labelled images. There are 234 incorrect predicted images from the final result.

Order	Error Images	Error Rate (%)
Galliformes (N=15)	11	73.3
Sphenisciformes (N=3)	2	66.7
Procellariiformes (N=3)	1	33.3
Pelecaniformes (N=6)	2	33.3
Ciconiiformes (N=9)	2	22.2
Pteroclidiformes (N=6)	1	16.7
Mesitornithiformes (N=6)	1	16.7
Bucerotiformes (N=42)	6	14.3
Gruiformes (N=111)	14	12.6
Caprimulgiformes (N=48)	5	10.4
Apodiformes (N=363)	34	9.4
Charadriiformes (N=246)	18	7.3
Strigiformes (N=66)	4	6.1
Accipitriformes (N=195)	11	5.6
Trogoniformes (N=18)	1	5.6
Coraciiformes (N=96)	4	4.2
Piciformes (N=198)	7	3.5
Columbiformes (N=120)	4	3.3
Passeriformes (N=3405)	104	3.0
Cuculiformes (N=66)	2	3.0
Leptosomiformes (N=3)	0	0.0
Falconiformes (N=33)	0	0.0
Musophagiformes (N=15)	0	0.0
Opisthocomiformes (N=3)	0	0.0
Otidiformes (N=6)	0	0.0
Eurypygiformes (N=3)	0	0.0
Coliiformes (N=6)	0	0.0

Supplementary Table 6.1.5. ANOVA results on pixel distances of overall and individual keypoints across three comparison groups (predictions vs trainer, predictions vs non-trainer and between experts). Data size: 300 images.

	ANOVA
Overall	F=26.2; df=2.0, 14910.0; p<0.01
Standard 1	F=33.0; df=2.0, 1797.0; p<0.01
Standard 2	F=79.3; df=2.0, 1797.0; p<0.01
Standard 3	F=199.2; df=2.0, 1797.0; p<0.01
Standard 4	F=226.2; df=2.0, 1797.0; p<0.01
Standard 5	F=307.0; df=2.0, 1797.0; p<0.01
Crown	F=1.7; df=2.0, 597.0; p=0.190
Nape	F=8.3; df=2.0, 597.0; p<0.01
Mantle	F=20.4; df=2.0, 597.0; p<0.01
Rump	F=0.2; df=2.0, 517.0; p=0.850
Tail	F=2.2; df=2.0, 592.0; p=0.110
Throat	F=0.9; df=2.0, 597.0; p=0.400
Breast	F=3.6; df=2.0, 597.0; p=0.029
Belly	F=10.6; df=2.0, 597.0; p<0.01
Coverts	F=14.5; df=2.0, 597.0; p<0.01
Flight feathers	F=2.5; df=2.0, 597.0; p=0.079

Supplementary Table 6.1.6. ANOVA results on pixel distances of overall and individual keypoints across the original and four tested low-quality datasets.

	ANOVA
Overall	F=87.2; df=4.0, 210720.0; p<0.01
Standard 1	F=532.7; df=4.0, 25465.0; p<0.01
Standard 2	F=260.0; df=4.0, 25465.0; p<0.01
Standard 3	F=146.1; df=4.0, 25465.0; p<0.01
Standard 4	F=33.5; df=4.0, 25465.0; p<0.01
Standard 5	F=18.4; df=4.0, 25465.0; p<0.01
Crown	F=9.9; df=4.0, 8470.0; p<0.01
Nape	F=10.1; df=4.0, 8470.0; p<0.01
Mantle	F=7.3; df=4.0, 8480.0; p<0.01
Rump	F=1.7; df=4.0, 7105.0; p=0.145
Tail	F=9.1; df=4.0, 8385.0; p<0.01
Throat	F=8.4; df=4.0, 8485.0; p<0.01
Breast	F=6.7; df=4.0, 8485.0; p<0.01
Belly	F=3.3; df=4.0, 8485.0; p=0.011
Coverts	F=12.6; df=4.0, 8475.0; p<0.01
Flight feathers	F=11.8; df=4.0, 8485.0; p<0.01

Supplementary Table 6.1.7. Colour statistics for both sexes, male and female across patches. Alpha shapes were all calculated using α=0.2262

Patch	Sex	Convex hull Volume	% colour space	Alpha shape volume	% colour space
Crown	Both	5.00E-02	23.1	3.27E-02	15.1
	Male	4.64E-02	21.4	2.97E-02	13.7
	Female	3.42E-02	15.8	2.32E-02	10.7
Nape	Both	5.57E-02	25.7	3.23E-02	14.9
	Male	5.39E-02	24.9	3.17E-02	14.6
	Female	3.80E-02	17.5	1.85E-02	8.6
Mantle	Both	5.65E-02	26.1	3.20E-02	14.8
	Male	5.42E-02	25	3.19E-02	14.8
	Female	2.89E-02	13.3	1.72E-02	7.9
Rump	Both	4.88E-02	22.5	2.96E-02	13.7
	Male	4.87E-02	22.5	2.88E-02	13.3
	Female	3.00E-02	13.9	1.62E-02	7.5
Tail	Both	3.31E-02	15.3	1.66E-02	7.7
	Male	3.03E-02	14	1.53E-02	7.1
	Female	1.41E-02	6.5	9.24E-03	4.3
Throat	Both	3.96E-02	18.3	2.38E-02	11
	Male	3.93E-02	18.1	2.36E-02	10.9
	Female	1.98E-02	9.2	1.16E-02	5.4
Breast	Both	5.23E-02	24.2	3.24E-02	15
	Male	5.21E-02	24.1	3.19E-02	14.7
	Female	2.97E-02	13.7	2.01E-02	9.3
Belly	Both	4.27E-02	19.7	2.74E-02	12.7
	Male	4.24E-02	19.6	2.74E-02	12.7
	Female	2.57E-02	11.9	1.71E-02	7.9
Coverts	Both	3.05E-02	14.1	1.81E-02	8.3
	Male	2.92E-02	13.5	1.72E-02	7.9
	Female	2.08E-02	9.6	1.32E-02	6.1
Flight Feathers	Both	2.29E-02	10.6	1.55E-02	7.2
	Male	2.18E-02	10.1	1.16E-02	7
	Female	1.84E-02	8.5	2.36E-02	5.8

Patch	Sex	Mean colour span	Colour span Variance	Mean hue disparity	hue disparity Variance
Crown	Both	1.05E-01	6.39E-03	5.51E-01	3.59E-01
	Male	1.18E-01	8.23E-03	6.49E-01	4.40E-01
	Female	9.10E-02	4.32E-03	4.45E-01	2.53E-01

Nape	Both	1.04E-01	6.16E-03	5.83E-01	4.20E-01
	Male	1.16E-01	7.84E-03	6.85E-01	5.11E-01
	Female	9.07E-02	4.24E-03	4.75E-01	3.00E-01
Mantle	Both	9.31E-02	4.19E-03	4.94E-01	3.18E-01
	Male	1.01E-01	5.08E-03	5.76E-01	4.01E-01
	Female	8.42E-02	3.18E-03	4.06E-01	2.15E-01
Rump	Both	9.76E-02	4.79E-03	4.82E-01	3.04E-01
	Male	1.06E-01	5.89E-03	5.52E-01	3.83E-01
	Female	8.90E-02	3.58E-03	4.08E-01	2.11E-01
Tail	Both	8.21E-02	3.18E-03	4.93E-01	3.73E-01
	Male	8.55E-02	3.54E-03	5.48E-01	4.29E-01
	Female	7.81E-02	2.78E-03	4.35E-01	3.07E-01
Throat	Both	9.17E-02	4.87E-03	3.90E-01	1.81E-01
	Male	1.01E-01	6.13E-03	4.64E-01	2.55E-01
	Female	8.13E-02	3.48E-03	3.11E-01	9.25E-02
Breast	Both	1.07E-01	5.92E-03	4.49E-01	2.25E-01
	Male	1.18E-01	7.38E-03	5.21E-01	2.94E-01
	Female	9.54E-02	4.30E-03	3.74E-01	1.44E-01
Belly	Both	9.36E-02	4.78E-03	3.77E-01	1.51E-01
	Male	1.02E-01	5.71E-03	4.22E-01	1.92E-01
	Female	8.52E-02	3.77E-03	3.30E-01	1.05E-01
Coverts	Both	8.94E-02	3.66E-03	4.72E-01	3.17E-01
	Male	9.45E-02	4.13E-03	5.33E-01	3.82E-01
	Female	8.36E-02	3.11E-03	4.09E-01	2.43E-01
Flight Feathers	Both	8.64E-02	3.33E-03	4.61E-01	3.21E-01
	Male	8.13E-02	3.48E-03	3.11E-01	9.25E-02
	Female	1.01E-01	6.13E-03	4.64E-01	2.55E-01

6.2 Chapter 3 supplementary Material

6.2.1 Supplementary Figures

(a) Ground Truth

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

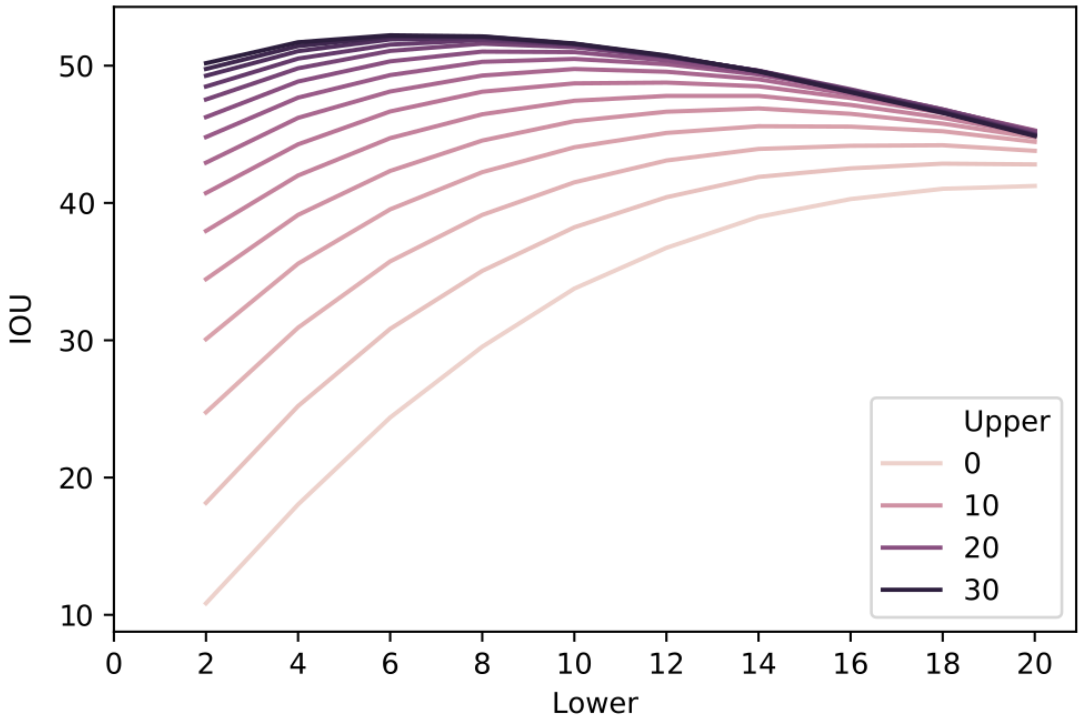
(b) Eroded segmentation
Precision=100%, Recall<100%

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	1	1	0	0	0
0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

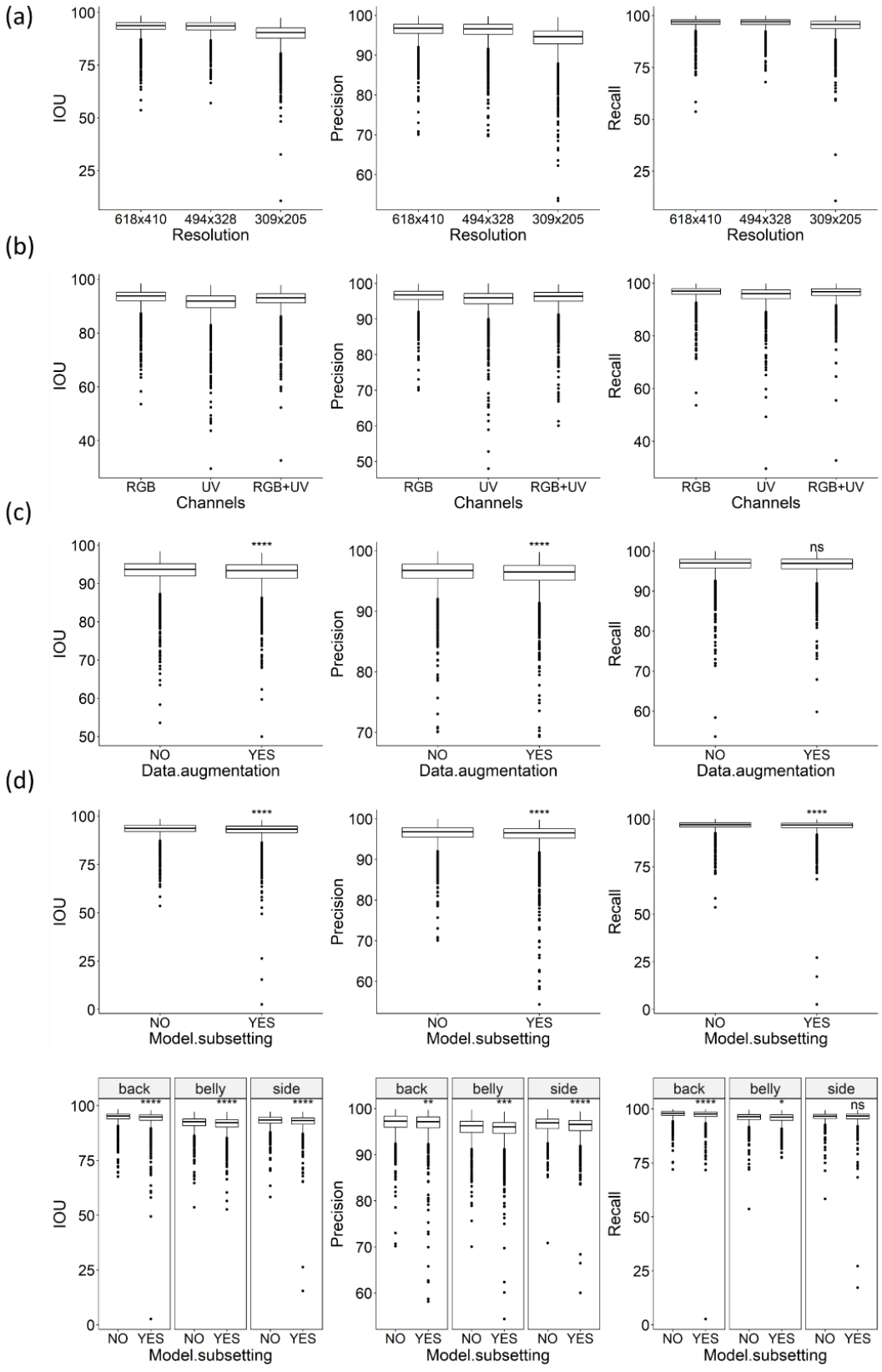
(c) Dilated segmentation
Precision<100%, Recall=100%

0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0

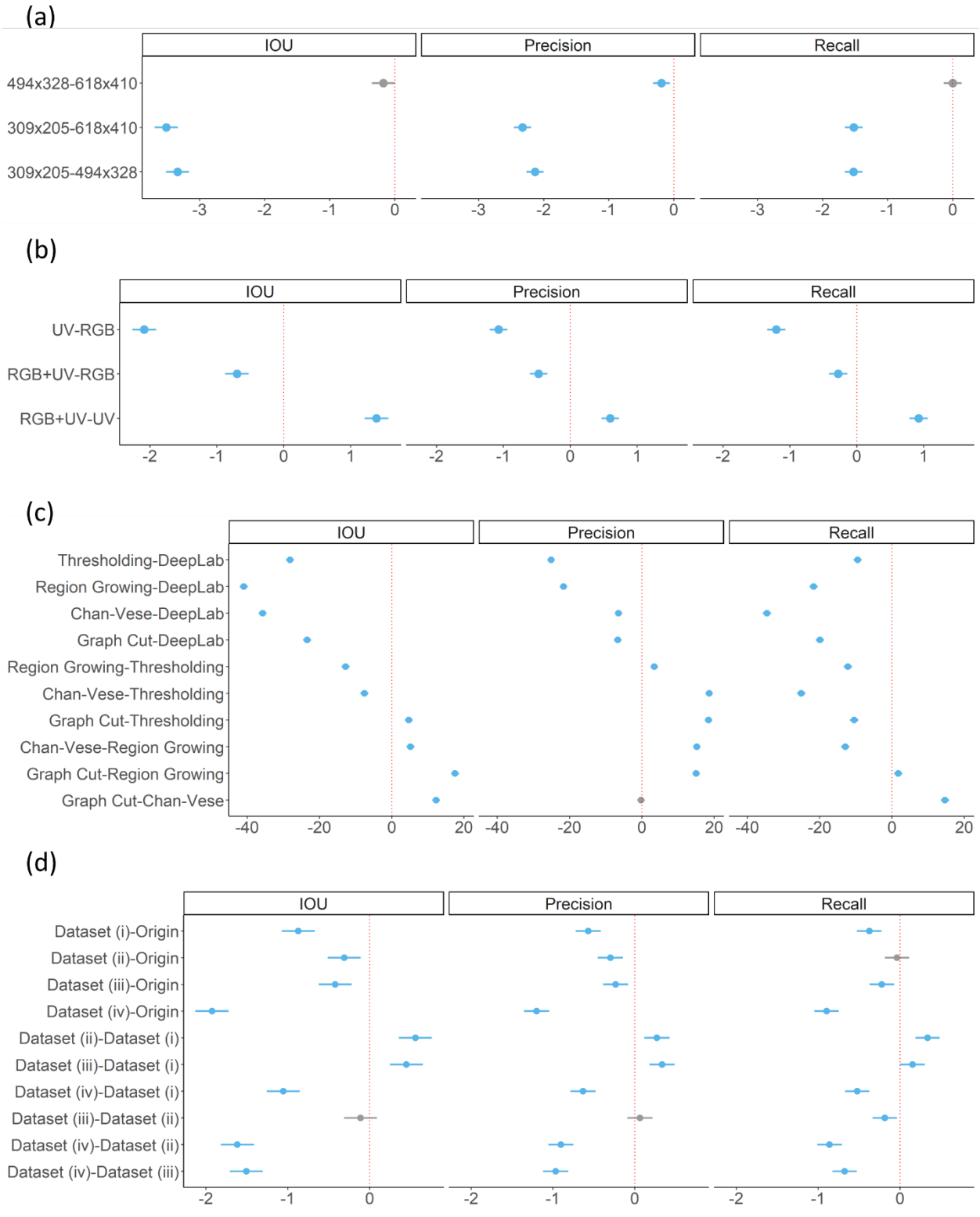
Supplementary Figure 6.2.1. Example segmentations and corresponding heatmaps and grid cells can be seen as pixels. Heatmaps have grid cells of 1 for plumage area and grid cells of 0 for non-plumage area. The example in (a) is defined as the ground truth segmentation of an images with plumage area segmented as yellow cells and non-plumage area as white cells. The example in (b) is an eroded segmentation (grey cells) based on (a). The example in (c) is a dilated segmentation (both green and yellow cells) based on (a).



Supplementary Figure 6.2.2. The IOU (y-axis) of region growing using different upper (colours) and lower boundaries (x-axis).

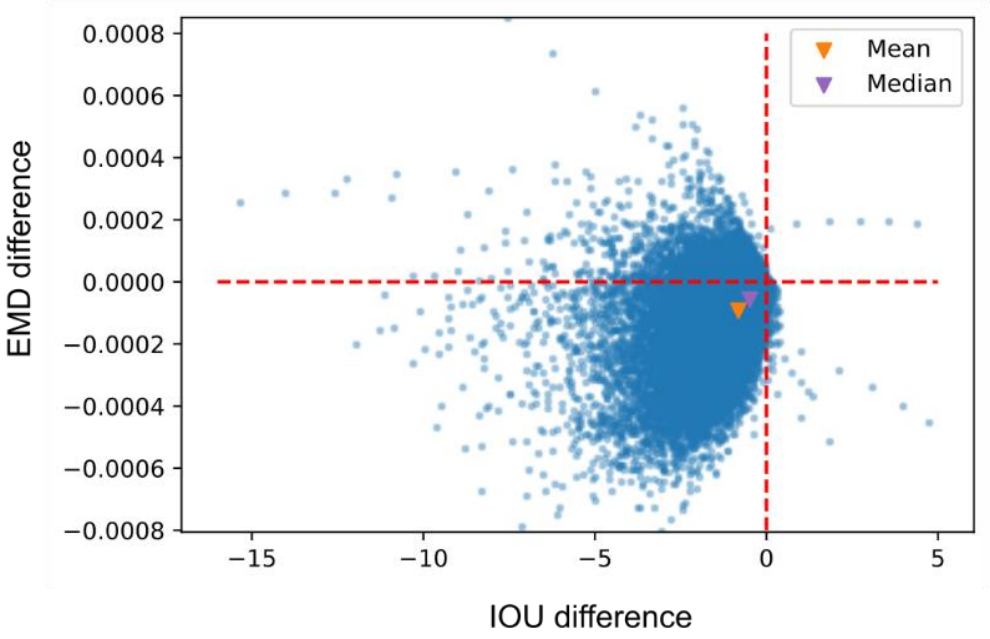

(a)
IOU
Precision
Recall
618x410
494x328
309x205
Resolution
(b)
RGB
UV
RGB+UV
Channels
(c)
NO
YES
Data.augmentation
****
ns
(d)
Model.subsetting
back
belly
side
**
***
*

Supplementary Figure 6.2.3. The performances (IOU, precision and recall) of predictions (N=5,094) from all experimental runs and the original model. (a) Input resolutions. (b) Input channels. (c) Image augmentation. (d) Subsetting models. The original model uses one DeepLabv3+ model to train non-augmented dataset for all views with 618 x 410 pixels as the input resolution, RGB as the input channels. Significant symbols are t-test results between the original model (the left most column) and (c) and (d) experimental runs (ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001).

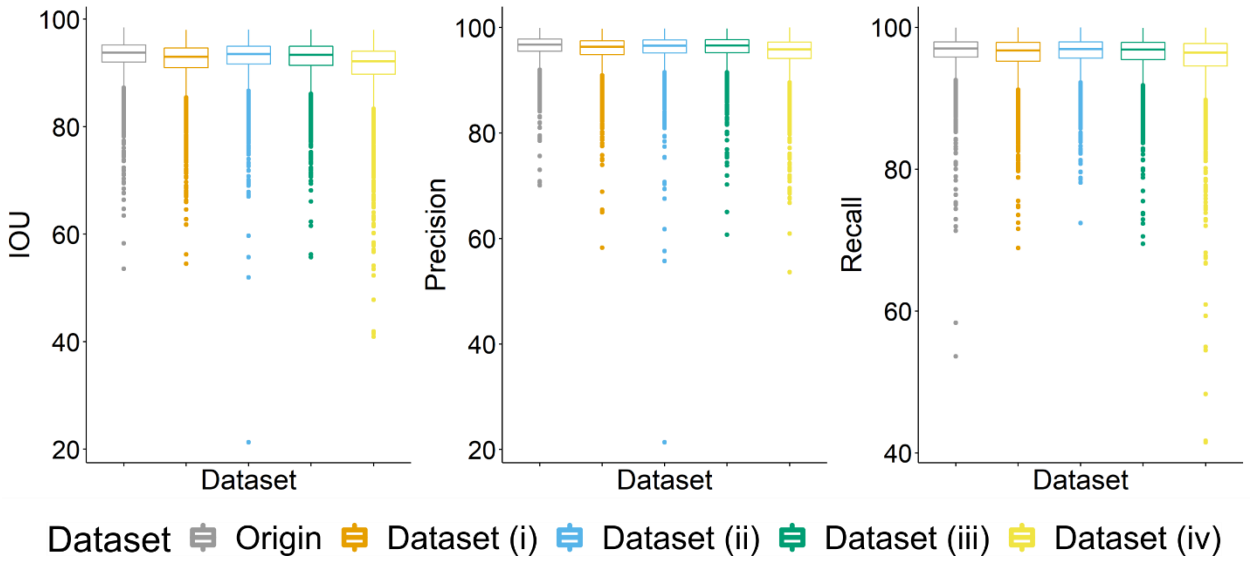


Supplementary Figure 6.2.4. Plots of Tukey's test (95% family-wise confidence level) on whether metric (IOU, precision and recall) differences between experimental runs are significantly different (blue:

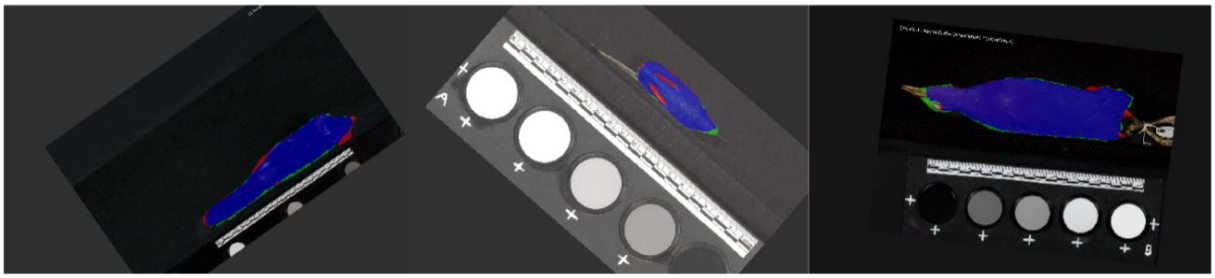
significance; grey: no significance) from 0 (red dotted lines). Experimental runs are (a) input resolutions; (b) input channels; (c) DeepLabv3+ and classic methods; and (d) low-quality datasets.



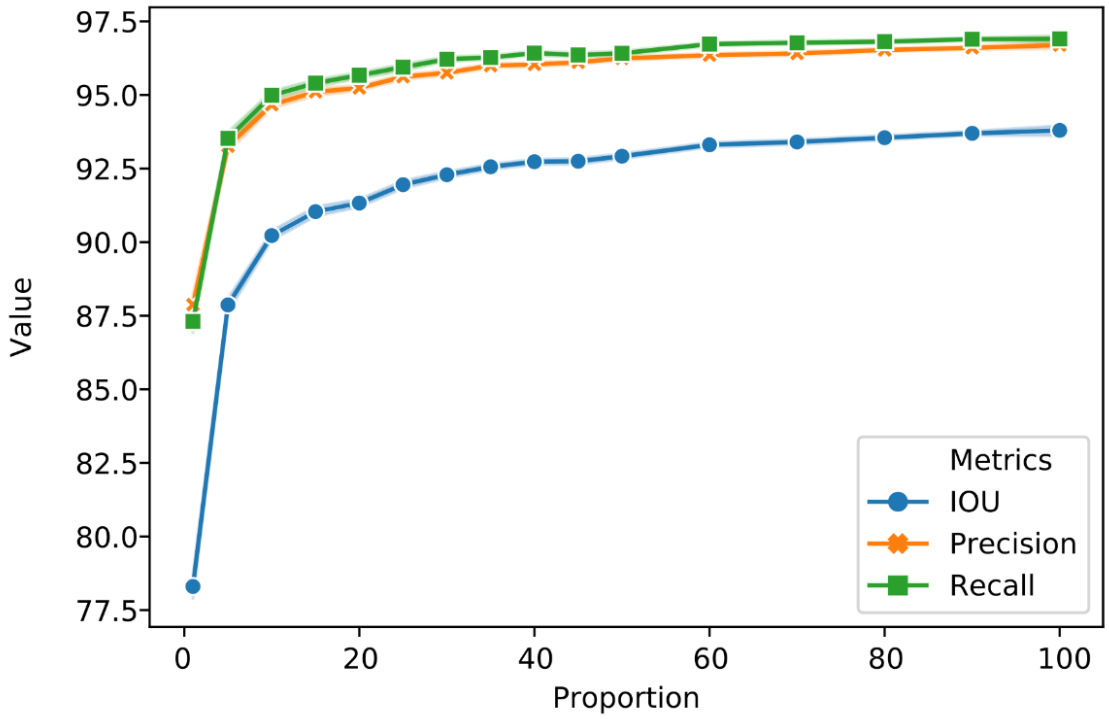
Supplementary Figure 6.2.5. The IOU difference (x-axis) and EMD difference (y-axis) of every eroded ground truth segmentation subtracting its corresponding dilated version (N=50,940).



Supplementary Figure 6.2.6. The performances (IOU, precision and recall) of predictions (N=5,094) using the original dataset and low-quality datasets. Dataset (i) rotated (angles between -45 to 45), translated (-500 to 500 pixels on x and y axes) and scaled (scale ratio from 0.8 to 1.2) images; Dataset (ii) horizontal flip 50% images randomly; Dataset (iii) images with random contrast and brightness; Dataset (iv) the combination of (i), (ii) and (iii).



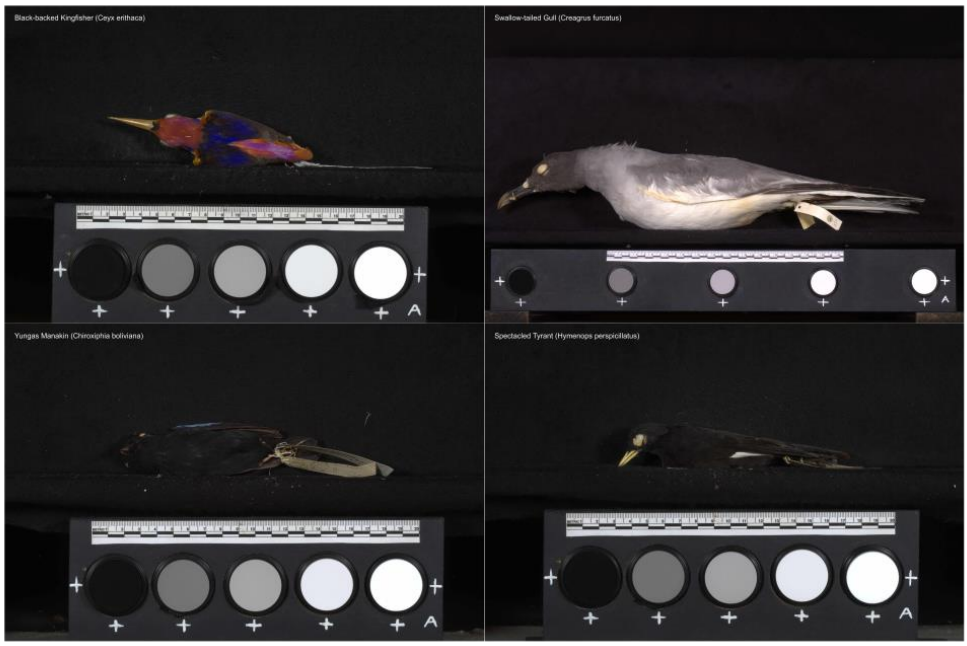
Supplementary Figure 6.2.7. Examples of predictions using Dataset (iv) of the low-quality datasets.



Supplementary Figure 6.2.8. The performances (IOU, precision and recall) of the same validation set (N=1,018) using 15 proportions (1%, every 5% from 5% to 50% and every 10% from 50% to 90%) of the original training set.

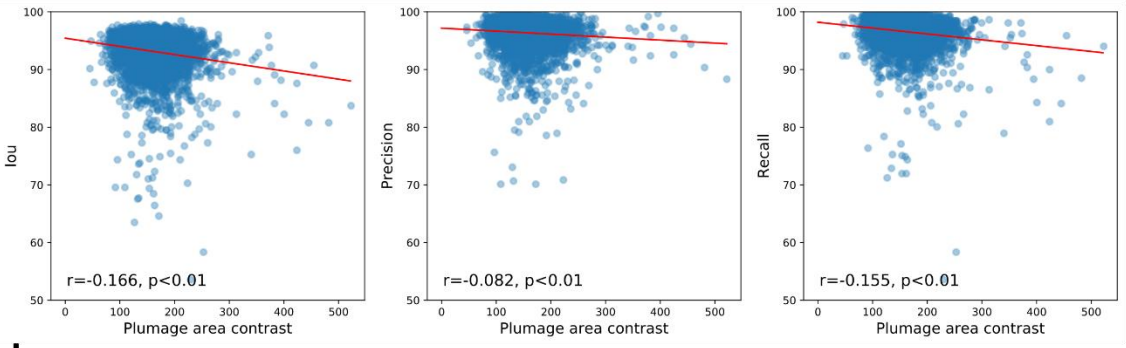
(a) Plumage area contrast (b) RGB variability

(c) Normalised RGB variability (d) Lightness variability

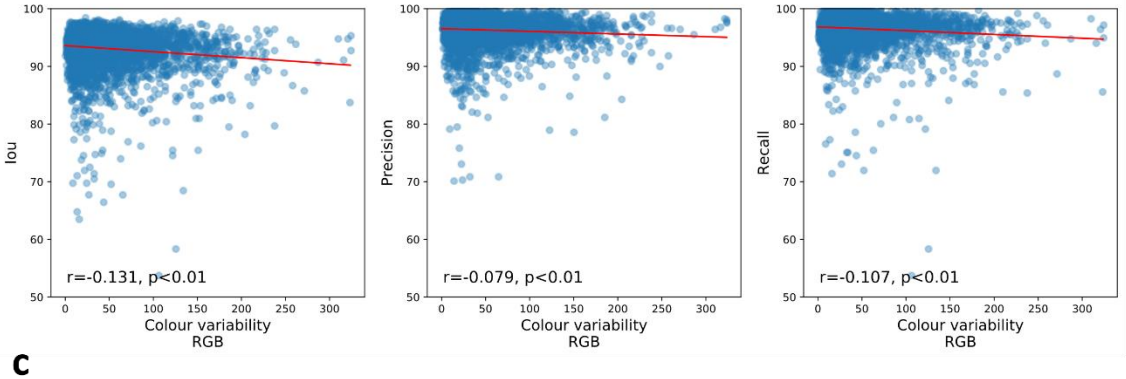


Supplementary Figure 6.2.9. Photos with the maximum (top) and minimum (bottom) (a) plumage area contrast, (b) Pair-wise RGB distance, (c) Pair-wise normalised RGB distance and (d) Pair-wise lightness distance.

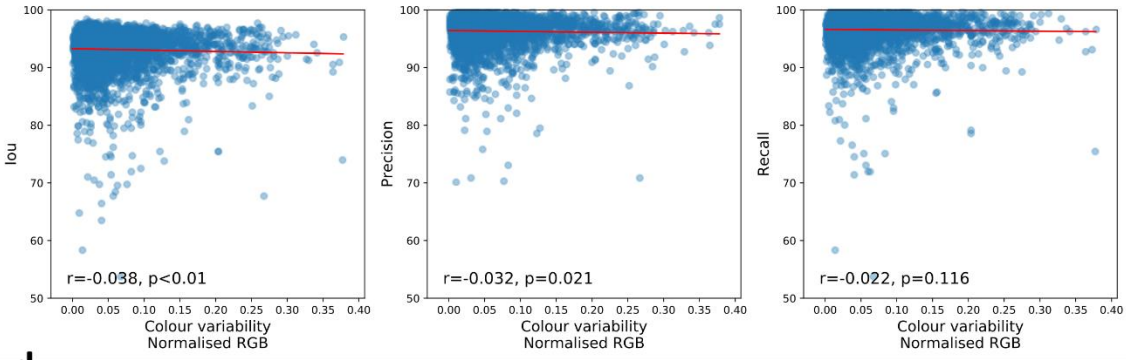
a


Iou
Precision
Recall
r=-0.166, p<0.01
r=-0.082, p<0.01
r=-0.155, p<0.01
Plumage area contrast

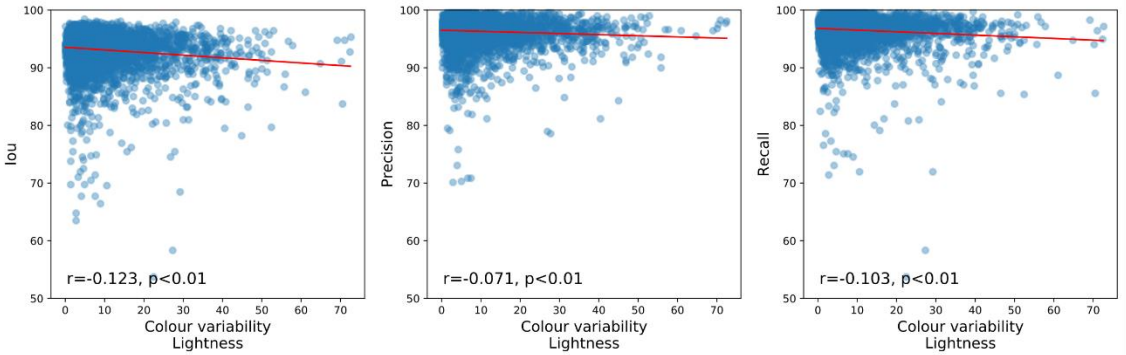
b


Iou
Precision
Recall
r=-0.131, p<0.01
r=-0.079, p<0.01
r=-0.107, p<0.01
Colour variability
RGB

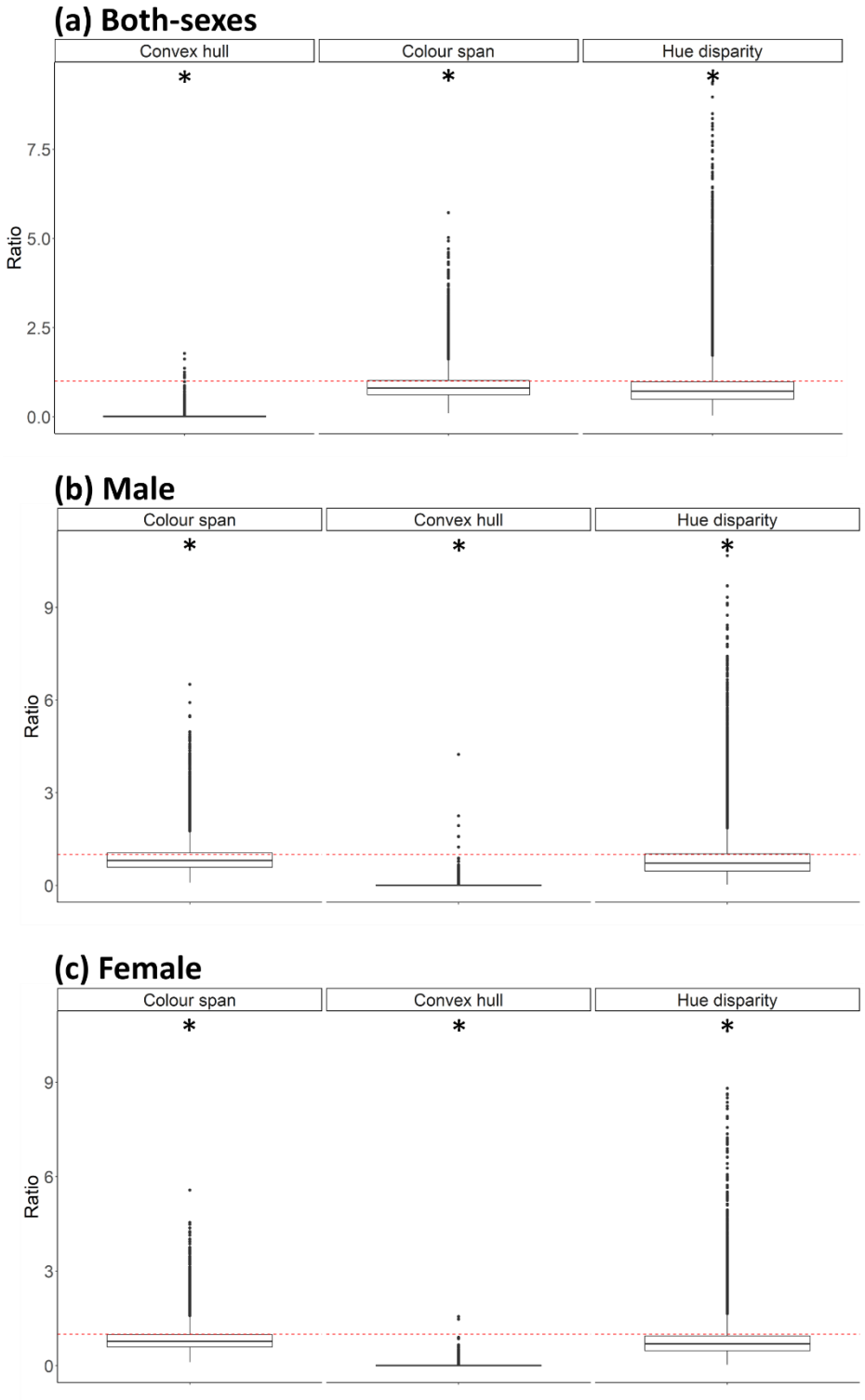
c


Iou
Precision
Recall
r=-0.038, p<0.01
r=-0.032, p=0.021
r=-0.022, p=0.116
Colour variability
Normalised RGB

d

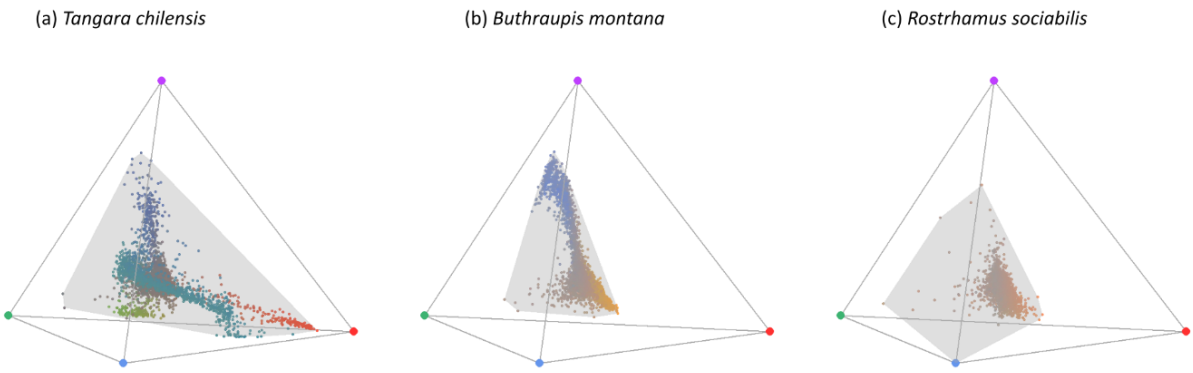

Iou
Precision
Recall
r=-0.123, p<0.01
r=-0.071, p<0.01
r=-0.103, p<0.01
Colour variability
Lightness

Supplementary Figure 6.2.10. Scatter plots of performances (IOU, precision and recall) of predictions (N=5,094) and images' colour properties (a) Plumage area contrast; (b) Pair-wise RGB distance; (c) Pair-wise normalized RGB distance; (d) Pair-wise lightness distance. The r-value and p-value are shown in each plot.



Supplementary Figure 6.2.11. The ratios of the convex hull volume, colour span and hue disparity from the patch data to ones from the segmentation data for (a) both sexes, (b) male and (c) female. Ratios were

tested whether they are different from 1 (red dotted lines) using t-test, significance are shown as *, non-significances are shown as ns.



Supplementary Figure 6.2.12. Colour points and their convex hull (Grey) in the tetrahedral colour space of three species, (a) Tangara chilensis species (N=3,600), (b) Buthraupis montana (N=3,600) and (c) Rostrhamus sociabilis (N=3,600).

6.2.2 Supplementary Tables

Supplementary Table 6.2.1. T-test results of IOU, precision and recall between all-views model and subsetting models.

	IOU	PRECISION	RECALL
OVERALL (N=5094)	Mean difference=0.6 t(10186)=7.98, p<0.01	Mean difference=0.4 t(10186)=6.75, p<0.01	Mean difference=0.3 t(10186)=5.19, p<0.01
BACK (N=1698)	Mean difference=0.7 t(10186)=5.80, p<0.01	Mean difference=0.3 t(10186)=2.89, p<0.01	Mean difference=0.5 t(10186)=5.53, p<0.01
BELLY (N=1698)	Mean difference=0.5 t(10186)=4.22, p<0.01	Mean difference=0.4 t(10186)=3.68, p<0.01	Mean difference=0.2 t(10186)=2.44, p=0.015
SIDE (N=1698)	Mean difference=0.5 t(10186)=4.36, p<0.01	Mean difference=0.4 t(10186)=5.96, p<0.01	Mean difference=0.1 t(10186)=1.42, p=0.15

Supplementary Table 6.2.2. IOU, precision and recall of predictions from the best DeepLabV3+ model per bird order.

ORDER	IOU	PRECISION	RECALL
Columbiformes (N=120)	95.2	97.8	97.2
Leptosomiformes (N=3)	94.7	97.2	97.4
Procellariiformes (N=3)	94.6	97.3	97.2
Strigiformes (N=66)	94.6	96.6	97.9
Charadriiformes (N=246)	94.3	96.7	97.5
Gruiformes (N=111)	94.3	97.2	97.0
Mesitornithiformes (N=6)	94.3	96.6	97.6
Caprimulgiformes (N=48)	94.2	96.7	97.3
Falconiformes (N=33)	94.2	96.5	97.6
Trogoniformes (N=18)	93.6	96.5	96.9
Piciformes (N=198)	93.6	96.4	96.9
Accipitriformes (N=195)	93.5	96.8	96.5
Pteroclidiformes (N=6)	93.4	96.9	96.2
Passeriformes (N=3408)	93.3	96.4	96.6
Cuculiformes (N=66)	93.3	96.4	96.7
Musophagiformes (N=15)	93.1	96.7	96.2
Coraciiformes (N=96)	93.0	96.3	96.5
Sphenisciformes (N=3)	92.9	94.7	98.0
Opisthocomiformes (N=3)	92.0	95.8	95.9
Eurypygiformes (N=3)	91.5	96.0	95.1
Pelecaniformes (N=6)	91.4	93.2	97.9
Otidiformes (N=6)	91.4	95.8	95.2
Coliiformes (N=6)	90.8	95.2	95.3
Apodiformes (N=363)	90.3	94.8	95.0
Bucerotiformes (N=42)	89.1	94.0	94.4
Ciconiiformes (N=9)	85.3	88.2	96.2

Galliformes (N=15)	84.6	94.9	88.8

Supplementary Table 6.2.3. Top 200 convex hull volume species

Species name	Colour volume	% Avian colour space	Family	Order
Pitta ussheri	7.73E-02	35.7	Pittidae	Passeriformes
Tangara chilensis	7.28E-02	33.6	Thraupidae	Passeriformes
Pitta granatina	6.02E-02	27.8	Pittidae	Passeriformes
Cyanerpes cyaneus	5.89E-02	27.2	Thraupidae	Passeriformes
Aethopyga gouldiae	5.44E-02	25.1	Nectariniidae	Passeriformes
Vini australis	5.38E-02	24.8	Psittacidae	Psittaciformes
Alisterus chloropterus	5.28E-02	24.4	Psittacidae	Psittaciformes
Pitta megarhyncha	5.26E-02	24.3	Pittidae	Passeriformes
Aglaiocercus coelestis	5.16E-02	23.8	Trochilidae	Apodiformes
Rostrhamus sociabilis	5.03E-02	23.2	Accipitridae	Accipitriformes
Nectarinia hunteri	5.02E-02	23.2	Nectariniidae	Passeriformes
Pitta gurneyi	4.86E-02	22.4	Pittidae	Passeriformes
Tanysiptera galatea	4.83E-02	22.3	Alcedinidae	Coraciiformes
Pyrilia pyrilia	4.71E-02	21.8	Psittacidae	Psittaciformes
Malurus splendens	4.71E-02	21.8	Maluridae	Passeriformes
Pitta superba	4.69E-02	21.6	Pittidae	Passeriformes
Charmosyna papou	4.62E-02	21.3	Psittacidae	Psittaciformes
Alisterus amboinensis	4.58E-02	21.1	Psittacidae	Psittaciformes
Tangara fastuosa	4.53E-02	20.9	Thraupidae	Passeriformes
Chlorophonia pyrrhophrys	4.46E-02	20.6	Thraupidae	Passeriformes
Nectarinia regia	4.43E-02	20.5	Nectariniidae	Passeriformes
Vini kuhlii	4.40E-02	20.3	Psittacidae	Psittaciformes
Prosopeia splendens	4.37E-02	20.2	Psittacidae	Psittaciformes
Platycercus eximius	4.36E-02	20.1	Psittacidae	Psittaciformes
Heliodoxa jacula	4.25E-02	19.6	Trochilidae	Apodiformes
Forpus modestus	4.17E-02	19.2	Psittacidae	Psittaciformes
Phigys solitarius	4.15E-02	19.2	Psittacidae	Psittaciformes
Pteroglossus viridis	4.12E-02	19	Ramphastidae	Piciformes
Campylopterus hemileucurus	4.10E-02	18.9	Trochilidae	Apodiformes
Psittacula longicauda	4.09E-02	18.9	Psittacidae	Psittaciformes
Nectarinia loveridgei	4.05E-02	18.7	Nectariniidae	Passeriformes
Parotia helenae	4.04E-02	18.7	Paradisaeidae	Passeriformes
Nectarinia chloropygia	4.00E-02	18.5	Nectariniidae	Passeriformes
Aglaiocercus kingi	3.96E-02	18.3	Trochilidae	Apodiformes
Pitta sordida	3.95E-02	18.2	Pittidae	Passeriformes
Psephotus varius	3.92E-02	18.1	Psittacidae	Psittaciformes
Dryocopus javensis	3.87E-02	17.9	Picidae	Piciformes

Pitta baudii	3.87E-02	17.9	Pittidae	Passeriformes
Passerina ciris	3.85E-02	17.8	Emberizidae	Passeriformes
Astrapia nigra	3.83E-02	17.7	Paradisaeidae	Passeriformes
Pitta steerii	3.82E-02	17.7	Pittidae	Passeriformes
Erythrura pealii	3.82E-02	17.7	Estrildidae	Passeriformes
Anisognathus notabilis	3.82E-02	17.7	Thraupidae	Passeriformes
Trichoglossus haematodus	3.80E-02	17.6	Psittacidae	Psittaciformes
Nectarinia fuelleborni	3.79E-02	17.5	Nectariniidae	Passeriformes
Forpus coelestis	3.79E-02	17.5	Psittacidae	Psittaciformes
Malurus amabilis	3.78E-02	17.4	Maluridae	Passeriformes
Lorius lory	3.75E-02	17.3	Psittacidae	Psittaciformes
Pitta moluccensis	3.73E-02	17.2	Pittidae	Passeriformes
Megalaima rafflesii	3.70E-02	17.1	Ramphastidae	Piciformes
Syrmaticus soemmerringii	3.69E-02	17	Phasianidae	Galliformes
Ceyx erithaca	3.66E-02	16.9	Alcedinidae	Coraciiformes
Pitta arcuata	3.64E-02	16.8	Pittidae	Passeriformes
Prosopeia tabuensis	3.63E-02	16.8	Psittacidae	Psittaciformes
Parotia wahnesi	3.58E-02	16.5	Paradisaeidae	Passeriformes
Loddigesia mirabilis	3.58E-02	16.5	Trochilidae	Apodiformes
Touit purpuratus	3.56E-02	16.5	Psittacidae	Psittaciformes
Anthreptes metallicus	3.56E-02	16.4	Nectariniidae	Passeriformes
Pteridophora alberti	3.56E-02	16.4	Paradisaeidae	Passeriformes
Nectarinia sperata	3.54E-02	16.4	Nectariniidae	Passeriformes
Aethopyga siparaja	3.52E-02	16.3	Nectariniidae	Passeriformes
Nectarinia afra	3.51E-02	16.2	Nectariniidae	Passeriformes
Alcedo meninting	3.47E-02	16	Alcedinidae	Coraciiformes
Tangara cyanocephala	3.45E-02	15.9	Thraupidae	Passeriformes
Ceyx lecontei	3.42E-02	15.8	Alcedinidae	Coraciiformes
Neophema splendida	3.41E-02	15.8	Psittacidae	Psittaciformes
Nectarinia notata	3.39E-02	15.6	Nectariniidae	Passeriformes
Musophaga rossae	3.35E-02	15.5	Musophagidae	Musophagiformes
Lorius hypoinochrous	3.35E-02	15.5	Psittacidae	Psittaciformes
Actophilornis africanus	3.34E-02	15.4	Jacanidae	Charadriiformes
Goura scheepmakeri	3.33E-02	15.4	Columbidae	Columbiformes
Bycanistes bucinator	3.32E-02	15.3	Bucerotidae	Bucerotiformes
Charmosyna josefinae	3.30E-02	15.3	Psittacidae	Psittaciformes
Pyrrhura rhodocephala	3.30E-02	15.2	Psittacidae	Psittaciformes
Nectarinia superba	3.25E-02	15	Nectariniidae	Passeriformes
Nectarinia mediocris	3.24E-02	15	Nectariniidae	Passeriformes
Coracias caudatus	3.23E-02	14.9	Coraciidae	Coraciiformes
Pitta maxima	3.22E-02	14.9	Pittidae	Passeriformes
Alcedo quadribrachys	3.22E-02	14.9	Alcedinidae	Coraciiformes
Purpureicephalus spurius	3.22E-02	14.9	Psittacidae	Psittaciformes
Nectarinia moreaui	3.21E-02	14.8	Nectariniidae	Passeriformes

Ara chloropterus	3.20E-02	14.8	Psittacidae	Psittaciformes
Tangara velia	3.20E-02	14.8	Thraupidae	Passeriformes
Euphonia finschi	3.18E-02	14.7	Thraupidae	Passeriformes
Nectarinia johannae	3.18E-02	14.7	Nectariniidae	Passeriformes
Phaenicophaeus cumingi	3.16E-02	14.6	Cuculidae	Cuculiformes
Chloropsis hardwickii	3.15E-02	14.5	Chloropseidae	Passeriformes
Tangara seledon	3.14E-02	14.5	Thraupidae	Passeriformes
Pyrilia caica	3.11E-02	14.4	Psittacidae	Psittaciformes
Tangara callophrys	3.08E-02	14.2	Thraupidae	Passeriformes
Picus puniceus	3.08E-02	14.2	Picidae	Piciformes
Pitta iris	3.07E-02	14.2	Pittidae	Passeriformes
Atthis ellioti	3.07E-02	14.2	Trochilidae	Apodiformes
Chlorophonia callophrys	3.07E-02	14.2	Thraupidae	Passeriformes
Campephilus principalis	3.07E-02	14.2	Picidae	Piciformes
Eunymphicus cornutus	3.06E-02	14.1	Psittacidae	Psittaciformes
Ptiloris magnificus	3.04E-02	14	Paradisaeidae	Passeriformes
Irena puella	3.03E-02	14	Irenidae	Passeriformes
Aethopyga mystacalis	3.02E-02	14	Nectariniidae	Passeriformes
Pyrilia barrabandi	3.01E-02	13.9	Psittacidae	Psittaciformes
Nectarinia tacazze	3.01E-02	13.9	Nectariniidae	Passeriformes
Nectarinia calcostetha	3.01E-02	13.9	Nectariniidae	Passeriformes
Trichoglossus ornatus	2.99E-02	13.8	Psittacidae	Psittaciformes
Musophaga violacea	2.97E-02	13.7	Musophagidae	Musophagiformes
Psophia crepitans	2.95E-02	13.6	Psophiidae	Gruiformes
Rhinoplax vigil	2.94E-02	13.6	Bucerotidae	Bucerotiformes
Manucodia keraudrenii	2.93E-02	13.5	Paradisaeidae	Passeriformes
Cicinnurus respublica	2.93E-02	13.5	Paradisaeidae	Passeriformes
Malurus cyanocephalus	2.93E-02	13.5	Maluridae	Passeriformes
Loriculus galgulus	2.92E-02	13.5	Psittacidae	Psittaciformes
Hemicircus concretus	2.92E-02	13.5	Picidae	Piciformes
Charmosyna margarethae	2.90E-02	13.4	Psittacidae	Psittaciformes
Nectarinia preussi	2.89E-02	13.4	Nectariniidae	Passeriformes
Syrmaticus reevesii	2.89E-02	13.3	Phasianidae	Galliformes
Pyrilia pulchra	2.88E-02	13.3	Psittacidae	Psittaciformes
Ramphomicron microrhynchum	2.88E-02	13.3	Trochilidae	Apodiformes
Lepidothrix isidorei	2.88E-02	13.3	Pipridae	Passeriformes
Ramphocelus flammigerus	2.86E-02	13.2	Thraupidae	Passeriformes
Aethopyga ignicauda	2.86E-02	13.2	Nectariniidae	Passeriformes
Coracias benghalensis	2.86E-02	13.2	Coraciidae	Coraciiformes
Pitta angolensis	2.86E-02	13.2	Pittidae	Passeriformes
Agapornis pullarius	2.86E-02	13.2	Psittacidae	Psittaciformes
Nisaetus alboniger	2.84E-02	13.1	Accipitridae	Accipitriformes
Hylocharis sapphirina	2.84E-02	13.1	Trochilidae	Apodiformes

Campephilus robustus	2.84E-02	13.1	Picidae	Piciformes
Cyclopsitta diophthalma	2.84E-02	13.1	Psittacidae	Psittaciformes
Nectarinia violacea	2.84E-02	13.1	Nectariniidae	Passeriformes
Anthreptes platurus	2.84E-02	13.1	Nectariniidae	Passeriformes
Chloropsis aurifrons	2.84E-02	13.1	Chloropseidae	Passeriformes
Prioniturus flavicans	2.84E-02	13.1	Psittacidae	Psittaciformes
Tauraco ruspolii	2.83E-02	13.1	Musophagidae	Musophagiformes
Grus virgo	2.83E-02	13.1	Gruidae	Gruiformes
Nectarinia stuhlmanni	2.82E-02	13	Nectariniidae	Passeriformes
Pionopsitta pileata	2.80E-02	12.9	Psittacidae	Psittaciformes
Agapornis fischeri	2.79E-02	12.9	Psittacidae	Psittaciformes
Psarisomus dalhousiae	2.79E-02	12.9	Eurylaimidae	Passeriformes
Fregata magnificens	2.79E-02	12.9	Fregatidae	Suliformes
Pitta nympha	2.79E-02	12.9	Pittidae	Passeriformes
Elvira chionura	2.79E-02	12.9	Trochilidae	Apodiformes
Chrysuronia oenone	2.78E-02	12.9	Trochilidae	Apodiformes
Nectarinia purpureiventris	2.78E-02	12.8	Nectariniidae	Passeriformes
Chrysolophus amherstiae	2.77E-02	12.8	Phasianidae	Galliformes
Psophia viridis	2.76E-02	12.8	Psophiidae	Gruiformes
Colibri serrirostris	2.75E-02	12.7	Trochilidae	Apodiformes
Cyanerpes caeruleus	2.75E-02	12.7	Thraupidae	Passeriformes
Momotus aequatorialis	2.75E-02	12.7	Momotidae	Coraciiformes
Myophonus horsfieldii	2.75E-02	12.7	Turdidae	Passeriformes
Chiroxiphia boliviana	2.74E-02	12.7	Pipridae	Passeriformes
Halcyon pileata	2.74E-02	12.6	Alcedinidae	Coraciiformes
Cyanerpes nitidus	2.73E-02	12.6	Thraupidae	Passeriformes
Platycercus icterotis	2.72E-02	12.6	Psittacidae	Psittaciformes
Parotia lawesii	2.72E-02	12.6	Paradisaeidae	Passeriformes
Anthreptes neglectus	2.72E-02	12.5	Nectariniidae	Passeriformes
Pavo cristatus	2.72E-02	12.5	Phasianidae	Galliformes
Ceyx lepidus	2.71E-02	12.5	Alcedinidae	Coraciiformes
Klais guimeti	2.71E-02	12.5	Trochilidae	Apodiformes
Damophila julie	2.71E-02	12.5	Trochilidae	Apodiformes
Eclectus roratus	2.71E-02	12.5	Psittacidae	Psittaciformes
Manucodia jobiensis	2.70E-02	12.5	Paradisaeidae	Passeriformes
Anthreptes aurantium	2.69E-02	12.4	Nectariniidae	Passeriformes
Aethopyga saturata	2.69E-02	12.4	Nectariniidae	Passeriformes
Niltava sundara	2.69E-02	12.4	Muscicapidae	Passeriformes
Pitta reichenowi	2.68E-02	12.4	Pittidae	Passeriformes
Campephilus rubricollis	2.66E-02	12.3	Picidae	Piciformes
Merops muelleri	2.65E-02	12.3	Meropidae	Coraciiformes
Psittacula cyanocephala	2.65E-02	12.3	Psittacidae	Psittaciformes
Niltava grandis	2.65E-02	12.3	Muscicapidae	Passeriformes
Charmosyna placentis	2.65E-02	12.2	Psittacidae	Psittaciformes

Alcedo cristata	2.64E-02	12.2	Alcedinidae	Coraciiformes
Pitta elegans	2.64E-02	12.2	Pittidae	Passeriformes
Ramphocelus passerinii	2.64E-02	12.2	Thraupidae	Passeriformes
Nectarinia minulla	2.64E-02	12.2	Nectariniidae	Passeriformes
Astrapia mayeri	2.63E-02	12.2	Paradisaeidae	Passeriformes
Lepidothrix coronata	2.62E-02	12.1	Pipridae	Passeriformes
Thalurania furcata	2.60E-02	12	Trochilidae	Apodiformes
Brotogeris cyanoptera	2.60E-02	12	Psittacidae	Psittaciformes
Ploceus nelicourvi	2.60E-02	12	Ploceidae	Passeriformes
Nectarinia bocagii	2.58E-02	11.9	Nectariniidae	Passeriformes
Lophorina superba	2.58E-02	11.9	Paradisaeidae	Passeriformes
Nectarinia pembae	2.58E-02	11.9	Nectariniidae	Passeriformes
Alisterus scapularis	2.58E-02	11.9	Psittacidae	Psittaciformes
Campochaera sloetii	2.57E-02	11.9	Campephagidae	Passeriformes
Eriocnemis luciani	2.56E-02	11.8	Trochilidae	Apodiformes
Chiroxiphia caudata	2.56E-02	11.8	Pipridae	Passeriformes
Charmosyna pulchella	2.55E-02	11.8	Psittacidae	Psittaciformes
Astrapia splendidissima	2.55E-02	11.8	Paradisaeidae	Passeriformes
Psittacula calthropae	2.55E-02	11.8	Psittacidae	Psittaciformes
Ara macao	2.54E-02	11.8	Psittacidae	Psittaciformes
Manucodia comrii	2.53E-02	11.7	Paradisaeidae	Passeriformes
Psophia leucoptera	2.52E-02	11.6	Psophiidae	Gruiformes
Nectarinia coccinigaster	2.52E-02	11.6	Nectariniidae	Passeriformes
Colibri coruscans	2.51E-02	11.6	Trochilidae	Apodiformes
Laniocera rufescens	2.51E-02	11.6	Cotingidae	Passeriformes
Barnardius zonarius	2.50E-02	11.6	Psittacidae	Psittaciformes
Epimachus meyeri	2.50E-02	11.6	Paradisaeidae	Passeriformes
Micropsitta bruijnii	2.50E-02	11.5	Psittacidae	Psittaciformes
Platycercus elegans	2.48E-02	11.5	Psittacidae	Psittaciformes
Tauraco schuetti	2.48E-02	11.5	Musophagidae	Musophagiformes
Aethopyga nipalensis	2.47E-02	11.4	Nectariniidae	Passeriformes
Ptiloris intercedens	2.47E-02	11.4	Paradisaeidae	Passeriformes

Supplementary Table 6.2.4. Top 200 convex hull volume male species

Species name	Colour volume	% Avian colour space	Family	Order
Tangara chilensis	3.09E+01	30.9	Thraupidae	Passeriformes
Cyanerpes cyaneus	2.66E+01	26.6	Thraupidae	Passeriformes

Aethopyga gouldiae	2.51E+01	25.1	Nectariniidae	Passeriformes
Vini australis	2.27E+01	22.7	Psittacidae	Psittaciformes
Pitta ussheri	2.22E+01	22.2	Pittidae	Passeriformes
Alisterus chloropterus	2.21E+01	22.1	Psittacidae	Psittaciformes
Aglaiocercus coelestis	2.21E+01	22.1	Trochilidae	Apodiformes
Pitta granatina	2.15E+01	21.5	Pittidae	Passeriformes
Pitta megarhyncha	2.09E+01	20.9	Pittidae	Passeriformes
Pyrilia pyrilia	2.06E+01	20.6	Psittacidae	Psittaciformes
Nectarinia regia	2.05E+01	20.5	Nectariniidae	Passeriformes
Prosopeia splendens	2.00E+01	20	Psittacidae	Psittaciformes
Pitta superba	1.97E+01	19.7	Pittidae	Passeriformes
Pteroglossus viridis	1.87E+01	18.7	Ramphastidae	Piciformes
Malurus splendens	1.86E+01	18.6	Maluridae	Passeriformes
Nectarinia loveridgei	1.83E+01	18.3	Nectariniidae	Passeriformes
Psephotus varius	1.81E+01	18.1	Psittacidae	Psittaciformes
Dryocopus javensis	1.78E+01	17.8	Picidae	Piciformes
Psittacula longicauda	1.77E+01	17.7	Psittacidae	Psittaciformes
Aglaiocercus kingi	1.76E+01	17.6	Trochilidae	Apodiformes
Vini kuhlii	1.75E+01	17.5	Psittacidae	Psittaciformes
Pitta baudii	1.72E+01	17.2	Pittidae	Passeriformes
Charmosyna papou	1.66E+01	16.6	Psittacidae	Psittaciformes
Platycercus eximius	1.66E+01	16.6	Psittacidae	Psittaciformes
Lorius lory	1.65E+01	16.5	Psittacidae	Psittaciformes
Astrapia nigra	1.64E+01	16.4	Paradisaeidae	Passeriformes
Heliodoxa jacula	1.64E+01	16.4	Trochilidae	Apodiformes
Phigys solitarius	1.63E+01	16.3	Psittacidae	Psittaciformes
Megalaima rafflesii	1.61E+01	16.1	Ramphastidae	Piciformes
Anisognathus notabilis	1.59E+01	15.9	Thraupidae	Passeriformes
Pitta arcuata	1.59E+01	15.9	Pittidae	Passeriformes
Parotia helenae	1.58E+01	15.8	Paradisaeidae	Passeriformes
Forpus modestus	1.56E+01	15.6	Psittacidae	Psittaciformes
Pteridophora alberti	1.55E+01	15.5	Paradisaeidae	Passeriformes
Tangara cyanocephala	1.54E+01	15.4	Thraupidae	Passeriformes
Nectarinia afra	1.52E+01	15.2	Nectariniidae	Passeriformes
Nectarinia sperata	1.51E+01	15.1	Nectariniidae	Passeriformes
Alisterus amboinensis	1.50E+01	15	Psittacidae	Psittaciformes
Trichoglossus haematodus	1.48E+01	14.8	Psittacidae	Psittaciformes
Anthreptes metallicus	1.48E+01	14.8	Nectariniidae	Passeriformes
Pitta gurneyi	1.47E+01	14.7	Pittidae	Passeriformes
Nectarinia fuelleborni	1.47E+01	14.7	Nectariniidae	Passeriformes
Alcedo meninting	1.47E+01	14.7	Alcedinidae	Coraciiformes
Pitta moluccensis	1.45E+01	14.5	Pittidae	Passeriformes
Nectarinia hunteri	1.44E+01	14.4	Nectariniidae	Passeriformes
Purpureicephalus spurius	1.44E+01	14.4	Psittacidae	Psittaciformes
Tangara seledon	1.43E+01	14.3	Thraupidae	Passeriformes

Chloropsis hardwickii	1.42E+01	14.2	Chloropseidae	Passeriformes
Neophema splendida	1.41E+01	14.1	Psittacidae	Psittaciformes
Parotia wahnesi	1.39E+01	13.9	Paradisaeidae	Passeriformes
Campylopterus hemileucurus	1.39E+01	13.9	Trochilidae	Apodiformes
Nectarinia superba	1.39E+01	13.9	Nectariniidae	Passeriformes
Chlorophonia pyrrhophrys	1.38E+01	13.8	Thraupidae	Passeriformes
Malurus amabilis	1.38E+01	13.8	Maluridae	Passeriformes
Nectarinia mediocris	1.38E+01	13.8	Nectariniidae	Passeriformes
Lorius hypoinochrous	1.37E+01	13.7	Psittacidae	Psittaciformes
Pyrilia barrabandi	1.35E+01	13.5	Psittacidae	Psittaciformes
Pitta maxima	1.34E+01	13.4	Pittidae	Passeriformes
Nectarinia notata	1.34E+01	13.4	Nectariniidae	Passeriformes
Tanysiptera galatea	1.33E+01	13.3	Alcedinidae	Coraciiformes
Irena puella	1.33E+01	13.3	Irenidae	Passeriformes
Loriculus galgulus	1.33E+01	13.3	Psittacidae	Psittaciformes
Nectarinia tacazze	1.32E+01	13.2	Nectariniidae	Passeriformes
Aethopyga mystacalis	1.32E+01	13.2	Nectariniidae	Passeriformes
Cicinnurus respublica	1.32E+01	13.2	Paradisaeidae	Passeriformes
Nectarinia violacea	1.31E+01	13.1	Nectariniidae	Passeriformes
Forpus coelestis	1.31E+01	13.1	Psittacidae	Psittaciformes
Nectarinia johannae	1.31E+01	13.1	Nectariniidae	Passeriformes
Prioniturus flavicans	1.31E+01	13.1	Psittacidae	Psittaciformes
Nectarinia stuhlmanni	1.30E+01	13	Nectariniidae	Passeriformes
Trichoglossus ornatus	1.29E+01	12.9	Psittacidae	Psittaciformes
Actophilornis africanus	1.29E+01	12.9	Jacanidae	Charadriiformes
Tangara velia	1.28E+01	12.8	Thraupidae	Passeriformes
Ara chloropterus	1.28E+01	12.8	Psittacidae	Psittaciformes
Nectarinia preussi	1.28E+01	12.8	Nectariniidae	Passeriformes
Aethopyga ignicauda	1.27E+01	12.7	Nectariniidae	Passeriformes
Anthreptes platurus	1.27E+01	12.7	Nectariniidae	Passeriformes
Aethopyga siparaja	1.27E+01	12.7	Nectariniidae	Passeriformes
Pyrilia pulchra	1.27E+01	12.7	Psittacidae	Psittaciformes
Manucodia keraudrenii	1.25E+01	12.5	Paradisaeidae	Passeriformes
Elvira chionura	1.24E+01	12.4	Trochilidae	Apodiformes
Loddigesia mirabilis	1.23E+01	12.3	Trochilidae	Apodiformes
Pitta sordida	1.22E+01	12.2	Pittidae	Passeriformes
Alcedo quadribrachys	1.22E+01	12.2	Alcedinidae	Coraciiformes
Touit purpuratus	1.21E+01	12.1	Psittacidae	Psittaciformes
Cyclopsitta diophthalma	1.21E+01	12.1	Psittacidae	Psittaciformes
Eunymphicus cornutus	1.19E+01	11.9	Psittacidae	Psittaciformes
Nectarinia purpureiventris	1.18E+01	11.8	Nectariniidae	Passeriformes
Prosopeia tabuensis	1.18E+01	11.8	Psittacidae	Psittaciformes
Pitta steerii	1.17E+01	11.7	Pittidae	Passeriformes
Aethopyga saturata	1.17E+01	11.7	Nectariniidae	Passeriformes
Picus puniceus	1.17E+01	11.7	Picidae	Piciformes

Chlorophonia callophrys	1.16E+01	11.6	Thraupidae	Passeriformes
Nectarinia calcostetha	1.16E+01	11.6	Nectariniidae	Passeriformes
Syrmaticus reevesii	1.16E+01	11.6	Phasianidae	Galliformes
Ramphocelus flammigerus	1.16E+01	11.6	Thraupidae	Passeriformes
Chloropsis aurifrons	1.15E+01	11.5	Chloropseidae	Passeriformes
Charmosyna placentis	1.15E+01	11.5	Psittacidae	Psittaciformes
Pitta angolensis	1.15E+01	11.5	Pittidae	Passeriformes
Nectarinia moreaui	1.14E+01	11.4	Nectariniidae	Passeriformes
Anthreptes neglectus	1.14E+01	11.4	Nectariniidae	Passeriformes
Ptiloris magnificus	1.14E+01	11.4	Paradisaeidae	Passeriformes
Passerina ciris	1.14E+01	11.4	Emberizidae	Passeriformes
Touit huetii	1.14E+01	11.4	Psittacidae	Psittaciformes
Ploceus nelicourvi	1.13E+01	11.3	Ploceidae	Passeriformes
Ramphomicron microrhynchum	1.13E+01	11.3	Trochilidae	Apodiformes
Tauraco ruspolii	1.13E+01	11.3	Musophagidae	Musophagiformes
Pitta elegans	1.13E+01	11.3	Pittidae	Passeriformes
Nectarinia talatala	1.12E+01	11.2	Nectariniidae	Passeriformes
Nectarinia pembae	1.12E+01	11.2	Nectariniidae	Passeriformes
Aethopyga nipalensis	1.11E+01	11.1	Nectariniidae	Passeriformes
Myophonus horsfieldii	1.11E+01	11.1	Turdidae	Passeriformes
Psittacula cyanocephala	1.11E+01	11.1	Psittacidae	Psittaciformes
Campochaera sloetii	1.10E+01	11	Campephagidae	Passeriformes
Hylocharis sapphirina	1.10E+01	11	Trochilidae	Apodiformes
Micropsitta bruijnii	1.10E+01	11	Psittacidae	Psittaciformes
Pionopsitta pileata	1.10E+01	11	Psittacidae	Psittaciformes
Nisaetus alboniger	1.10E+01	11	Accipitridae	Accipitriformes
Coeligena helianthea	1.09E+01	10.9	Trochilidae	Apodiformes
Epimachus meyeri	1.09E+01	10.9	Paradisaeidae	Passeriformes
Charmosyna josefinae	1.09E+01	10.9	Psittacidae	Psittaciformes
Nectarinia minulla	1.09E+01	10.9	Nectariniidae	Passeriformes
Alcedo vintsioides	1.08E+01	10.8	Alcedinidae	Coraciiformes
Charmosyna margarethae	1.08E+01	10.8	Psittacidae	Psittaciformes
Parotia lawesii	1.08E+01	10.8	Paradisaeidae	Passeriformes
Ceyx lepidus	1.07E+01	10.7	Alcedinidae	Coraciiformes
Campephilus imperialis	1.07E+01	10.7	Picidae	Piciformes
Astrapia mayeri	1.07E+01	10.7	Paradisaeidae	Passeriformes
Nectarinia shelleyi	1.07E+01	10.7	Nectariniidae	Passeriformes
Nectarinia jugularis	1.07E+01	10.7	Nectariniidae	Passeriformes
Manucodia comrii	1.06E+01	10.6	Paradisaeidae	Passeriformes
Pitta reichenowi	1.06E+01	10.6	Pittidae	Passeriformes
Todus angustirostris	1.06E+01	10.6	Todidae	Coraciiformes
Nectarinia mariquensis	1.06E+01	10.6	Nectariniidae	Passeriformes

Lepidothrix isidorei	1.06E+01	10.6	Pipridae	Passeriformes
Erythrura gouldiae	1.06E+01	10.6	Estrildidae	Passeriformes
Chrysolophus amherstiae	1.06E+01	10.6	Phasianidae	Galliformes
Lepidothrix coronata	1.05E+01	10.5	Pipridae	Passeriformes
Ramphocelus passerinii	1.05E+01	10.5	Thraupidae	Passeriformes
Anisognathus igniventris	1.05E+01	10.5	Thraupidae	Passeriformes
Pavo muticus	1.05E+01	10.5	Phasianidae	Galliformes
Astrapia splendidissima	1.05E+01	10.5	Paradisaeidae	Passeriformes
Pavo cristatus	1.05E+01	10.5	Phasianidae	Galliformes
Tachyphonus coronatus	1.05E+01	10.5	Thraupidae	Passeriformes
Masius chrysopterus	1.04E+01	10.4	Pipridae	Passeriformes
Neophema pulchella	1.04E+01	10.4	Psittacidae	Psittaciformes
Musophaga violacea	1.03E+01	10.3	Musophagidae	Musophagiformes
Oreopsittacus arfaki	1.03E+01	10.3	Psittacidae	Psittaciformes
Anthreptes aurantium	1.03E+01	10.3	Nectariniidae	Passeriformes
Megalaima oorti	1.03E+01	10.3	Ramphastidae	Piciformes
Halcyon pileata	1.03E+01	10.3	Alcedinidae	Coraciiformes
Halcyon leucocephala	1.03E+01	10.3	Alcedinidae	Coraciiformes
Nectarinia ludovicensis	1.02E+01	10.2	Nectariniidae	Passeriformes
Nectarinia minima	1.02E+01	10.2	Nectariniidae	Passeriformes
Charmosyna wilhelminae	1.02E+01	10.2	Psittacidae	Psittaciformes
Nectarinia coccinigaster	1.02E+01	10.2	Nectariniidae	Passeriformes
Eumomota superciliosa	1.02E+01	10.2	Momotidae	Coraciiformes
Myzomela rosenbergii	1.02E+01	10.2	Meliphagidae	Passeriformes
Thalurania furcata	1.02E+01	10.2	Trochilidae	Apodiformes
Anthreptes rhodolaemus	1.01E+01	10.1	Nectariniidae	Passeriformes
Alcedo cristata	1.01E+01	10.1	Alcedinidae	Coraciiformes
Charmosyna pulchella	1.01E+01	10.1	Psittacidae	Psittaciformes
Pyrilia haematotis	1.01E+01	10.1	Psittacidae	Psittaciformes
Ara ararauna	1.01E+01	10.1	Psittacidae	Psittaciformes
Agapornis pullarius	1.01E+01	10.1	Psittacidae	Psittaciformes
Agapornis fischeri	1.01E+01	10.1	Psittacidae	Psittaciformes
Eupodotis melanogaster	1.01E+01	10.1	Otididae	Otidiformes
Ceyx erithaca	1.01E+01	10.1	Alcedinidae	Coraciiformes
Merops muelleri	1.01E+01	10.1	Meropidae	Coraciiformes
Hemicircus concretus	1.00E+01	10	Picidae	Piciformes
Aethopyga christinae	1.00E+01	10	Nectariniidae	Passeriformes
Eclectus roratus	1.00E+01	10	Psittacidae	Psittaciformes
Psittinus cyanurus	1.00E+01	10	Psittacidae	Psittaciformes
Lophorina superba	1.00E+01	10	Paradisaeidae	Passeriformes
Seleucidis melanoleucus	1.00E+01	10	Paradisaeidae	Passeriformes
Agapornis roseicollis	9.90E+00	9.9	Psittacidae	Psittaciformes
Ptilorrhoa castanonota	9.90E+00	9.9	Eupetidae	Passeriformes
Niltava vivida	9.90E+00	9.9	Muscicapidae	Passeriformes

Niltava davidi	9.90E+00	9.9	Muscicapidae	Passeriformes
Aceros waldeni	9.90E+00	9.9	Bucerotidae	Bucerotiformes
Chiroxiphia boliviana	9.80E+00	9.8	Pipridae	Passeriformes
Platycercus icterotis	9.80E+00	9.8	Psittacidae	Psittaciformes
Lorius chlorocercus	9.80E+00	9.8	Psittacidae	Psittaciformes
Nectarinia senegalensis	9.80E+00	9.8	Nectariniidae	Passeriformes
Nectarinia johnstoni	9.80E+00	9.8	Nectariniidae	Passeriformes
Ephippiorhynchus asiaticus	9.80E+00	9.8	Ciconiidae	Ciconiiformes
Erythrura regia	9.70E+00	9.7	Estrildidae	Passeriformes
Lamprolaima rhami	9.70E+00	9.7	Trochilidae	Apodiformes
Niltava grandis	9.70E+00	9.7	Muscicapidae	Passeriformes
Ara macao	9.70E+00	9.7	Psittacidae	Psittaciformes
Cnemophilus loriae	9.70E+00	9.7	Cnemophilidae	Passeriformes
Ploceus velatus	9.60E+00	9.6	Ploceidae	Passeriformes
Pyrilia caica	9.60E+00	9.6	Psittacidae	Psittaciformes
Ceyx lecontei	9.60E+00	9.6	Alcedinidae	Coraciiformes
Nectarinia zeylonica	9.60E+00	9.6	Nectariniidae	Passeriformes
Parotia carolae	9.60E+00	9.6	Paradisaeidae	Passeriformes
Calyptomena hosii	9.60E+00	9.6	Eurylaimidae	Passeriformes
Monarcha chrysomela	9.60E+00	9.6	Monarchidae	Passeriformes
Cyanocorax yucatanicus	9.60E+00	9.6	Corvidae	Passeriformes
Nectarinia habessinica	9.50E+00	9.5	Nectariniidae	Passeriformes

Supplementary Table 6.2.5. Top 200 convex hull volume female species

Species name	Colour volume	% Avian colour space	Family	Order
Tangara chilensis	6.36E-02	29.4	Thraupidae	Passeriformes
Pitta ussheri	6.05E-02	27.9	Pittidae	Passeriformes
Pitta granatina	5.06E-02	23.4	Pittidae	Passeriformes
Charmosyna papou	3.89E-02	18	Psittacidae	Psittaciformes
Phigys solitarius	3.75E-02	17.3	Psittacidae	Psittaciformes
Vini australis	3.71E-02	17.1	Psittacidae	Psittaciformes
Pitta megarhyncha	3.55E-02	16.4	Pittidae	Passeriformes
Pitta sordida	3.45E-02	15.9	Pittidae	Passeriformes
Pitta superba	3.38E-02	15.6	Pittidae	Passeriformes
Goura scheepmakeri	3.24E-02	15	Columbidae	Columbiformes
Touit purpuratus	3.24E-02	14.9	Psittacidae	Psittaciformes
Prosopeia tabuensis	3.20E-02	14.8	Psittacidae	Psittaciformes
Alisterus amboinensis	3.20E-02	14.8	Psittacidae	Psittaciformes
Pitta moluccensis	3.19E-02	14.7	Pittidae	Passeriformes

Musophaga rossae	3.16E-02	14.6	Musophagidae	Musophagiformes
Pitta steerii	3.11E-02	14.4	Pittidae	Passeriformes
Pitta iris	3.07E-02	14.2	Pittidae	Passeriformes
Tanysiptera galatea	3.07E-02	14.2	Alcedinidae	Coraciiformes
Syrmaticus soemmerringii	3.01E-02	13.9	Phasianidae	Galliformes
Ceyx erithaca	2.99E-02	13.8	Alcedinidae	Coraciiformes
Trichoglossus haematodus	2.92E-02	13.5	Psittacidae	Psittaciformes
Vini kuhlii	2.91E-02	13.4	Psittacidae	Psittaciformes
Bycanistes bucinator	2.87E-02	13.3	Bucerotidae	Bucerotiformes
Charmosyna josefinae	2.84E-02	13.1	Psittacidae	Psittaciformes
Anisognathus notabilis	2.79E-02	12.9	Thraupidae	Passeriformes
Megalaima rafflesii	2.77E-02	12.8	Ramphastidae	Piciformes
Platycercus eximius	2.77E-02	12.8	Psittacidae	Psittaciformes
Alisterus chloropterus	2.70E-02	12.5	Psittacidae	Psittaciformes
Tangara velia	2.60E-02	12	Thraupidae	Passeriformes
Musophaga violacea	2.59E-02	12	Musophagidae	Musophagiformes
Pyrilia caica	2.57E-02	11.9	Psittacidae	Psittaciformes
Psophia crepitans	2.53E-02	11.7	Psophiidae	Gruiformes
Erythrura pealii	2.53E-02	11.7	Estrildidae	Passeriformes
Lorius hypoinochrous	2.53E-02	11.7	Psittacidae	Psittaciformes
Rostrhamus sociabilis	2.52E-02	11.6	Accipitridae	Accipitriformes
Coracias benghalensis	2.48E-02	11.5	Coraciidae	Coraciiformes
Charmosyna margarethae	2.47E-02	11.4	Psittacidae	Psittaciformes
Chlorophonia pyrrhophrys	2.46E-02	11.4	Thraupidae	Passeriformes
Eunymphicus cornutus	2.42E-02	11.2	Psittacidae	Psittaciformes
Brotogeris cyanoptera	2.41E-02	11.1	Psittacidae	Psittaciformes
Fregata magnificens	2.38E-02	11	Fregatidae	Suliformes
Ceyx lecontei	2.35E-02	10.8	Alcedinidae	Coraciiformes
Manucodia jobiensis	2.34E-02	10.8	Paradisaeidae	Passeriformes
Pyrilia pyrilia	2.31E-02	10.7	Psittacidae	Psittaciformes
Vini stepheni	2.31E-02	10.7	Psittacidae	Psittaciformes
Psarisomus dalhousiae	2.30E-02	10.6	Eurylaimidae	Passeriformes
Alcedo quadribrachys	2.28E-02	10.6	Alcedinidae	Coraciiformes
Pitta arcuata	2.27E-02	10.5	Pittidae	Passeriformes
Laniocera rufescens	2.25E-02	10.4	Cotingidae	Passeriformes
Rhinoplax vigil	2.21E-02	10.2	Bucerotidae	Bucerotiformes
Coracias caudatus	2.16E-02	10	Coraciidae	Coraciiformes
Charmosyna pulchella	2.16E-02	10	Psittacidae	Psittaciformes
Pogonocichla stellata	2.14E-02	9.9	Muscicapidae	Passeriformes
Platycercus elegans	2.14E-02	9.9	Psittacidae	Psittaciformes
Pitta nympha	2.14E-02	9.9	Pittidae	Passeriformes
Pyrilia pulchra	2.13E-02	9.8	Psittacidae	Psittaciformes
Coracias temminckii	2.10E-02	9.7	Coraciidae	Coraciiformes

Psophia leucoptera	2.09E-02	9.6	Psophiidae	Gruiformes
Phaenicophaeus cumingi	2.08E-02	9.6	Cuculidae	Cuculiformes
Cyanocorax sanblasianus	2.08E-02	9.6	Corvidae	Passeriformes
Agapornis fischeri	2.07E-02	9.6	Psittacidae	Psittaciformes
Lybius dubius	2.07E-02	9.6	Ramphastidae	Piciformes
Momotus aequatorialis	2.06E-02	9.5	Momotidae	Coraciiformes
Cephalopterus ornatus	2.03E-02	9.4	Cotingidae	Passeriformes
Lorius lory	2.03E-02	9.4	Psittacidae	Psittaciformes
Colibri coruscans	2.02E-02	9.3	Trochilidae	Apodiformes
Ramphastos vitellinus	1.99E-02	9.2	Ramphastidae	Piciformes
Campephilus principalis	1.96E-02	9.1	Picidae	Piciformes
Pitta reichenowi	1.92E-02	8.9	Pittidae	Passeriformes
Eos cyanogenia	1.89E-02	8.7	Psittacidae	Psittaciformes
Pitta maxima	1.89E-02	8.7	Pittidae	Passeriformes
Pitta angolensis	1.88E-02	8.7	Pittidae	Passeriformes
Micropsitta keiensis	1.88E-02	8.7	Psittacidae	Psittaciformes
Centropus goliath	1.87E-02	8.7	Cuculidae	Cuculiformes
Metopidius indicus	1.87E-02	8.6	Jacanidae	Charadriiformes
Megalaima asiatica	1.86E-02	8.6	Ramphastidae	Piciformes
Tauraco schuetti	1.85E-02	8.6	Musophagidae	Musophagiformes
Phalacrocorax africanus	1.84E-02	8.5	Phalacrocoracidae	Suliformes
Pyrrhura egregia	1.84E-02	8.5	Psittacidae	Psittaciformes
Phoeniculus purpureus	1.82E-02	8.4	Phoeniculidae	Bucerotiformes
Centropus anselli	1.81E-02	8.4	Cuculidae	Cuculiformes
Buthraupis eximia	1.81E-02	8.3	Thraupidae	Passeriformes
Prosopeia splendens	1.80E-02	8.3	Psittacidae	Psittaciformes
Aratinga solstitialis	1.80E-02	8.3	Psittacidae	Psittaciformes
Merops bulocki	1.80E-02	8.3	Meropidae	Coraciiformes
Trichoglossus ornatus	1.80E-02	8.3	Psittacidae	Psittaciformes
Pyrrhura rhodocephala	1.80E-02	8.3	Psittacidae	Psittaciformes
Lathamus discolor	1.79E-02	8.3	Psittacidae	Psittaciformes
Klais guimeti	1.78E-02	8.2	Trochilidae	Apodiformes
Chloropsis aurifrons	1.78E-02	8.2	Chloropseidae	Passeriformes
Eos histrio	1.77E-02	8.2	Psittacidae	Psittaciformes
Ara macao	1.76E-02	8.2	Psittacidae	Psittaciformes
Psophia viridis	1.76E-02	8.1	Psophiidae	Gruiformes
Chlorophonia callophrys	1.76E-02	8.1	Thraupidae	Passeriformes
Turnix pyrrhothorax	1.76E-02	8.1	Turnicidae	Charadriiformes
Agapornis pullarius	1.76E-02	8.1	Psittacidae	Psittaciformes
Neophema chrysostoma	1.76E-02	8.1	Psittacidae	Psittaciformes
Psittaculirostris edwardsii	1.75E-02	8.1	Psittacidae	Psittaciformes
Balearica regulorum	1.74E-02	8.1	Gruidae	Gruiformes
Ploceus melanogaster	1.74E-02	8	Ploceidae	Passeriformes

Lamprotornis purpureiceps	1.74E-02	8	Sturnidae	Passeriformes
Brotogeris chrysoptera	1.73E-02	8	Psittacidae	Psittaciformes
Malurus cyanocephalus	1.73E-02	8	Maluridae	Passeriformes
Brotogeris tirica	1.72E-02	7.9	Psittacidae	Psittaciformes
Ara chloropterus	1.71E-02	7.9	Psittacidae	Psittaciformes
Campephilus robustus	1.71E-02	7.9	Picidae	Piciformes
Campephilus rubricollis	1.71E-02	7.9	Picidae	Piciformes
Eurystomus glaucurus	1.70E-02	7.9	Coraciidae	Coraciiformes
Eriocnemis luciani	1.70E-02	7.9	Trochilidae	Apodiformes
Syrmaticus reevesii	1.70E-02	7.9	Phasianidae	Galliformes
Alcedo semitorquata	1.70E-02	7.8	Alcedinidae	Coraciiformes
Amazona finschi	1.68E-02	7.7	Psittacidae	Psittaciformes
Megalaima henricii	1.67E-02	7.7	Ramphastidae	Piciformes
Aratinga auricapillus	1.66E-02	7.7	Psittacidae	Psittaciformes
Halcyon pileata	1.66E-02	7.7	Alcedinidae	Coraciiformes
Leptoptilos javanicus	1.66E-02	7.7	Ciconiidae	Ciconiiformes
Ceyx lepidus	1.65E-02	7.6	Alcedinidae	Coraciiformes
Psittacula calthropae	1.65E-02	7.6	Psittacidae	Psittaciformes
Ducula concinna	1.65E-02	7.6	Columbidae	Columbiformes
Merops muelleri	1.65E-02	7.6	Meropidae	Coraciiformes
Grus virgo	1.64E-02	7.6	Gruidae	Gruiformes
Eos reticulata	1.63E-02	7.5	Psittacidae	Psittaciformes
Tangara mexicana	1.62E-02	7.5	Thraupidae	Passeriformes
Coracias cyanogaster	1.62E-02	7.5	Coraciidae	Coraciiformes
Pogoniulus bilineatus	1.61E-02	7.5	Ramphastidae	Piciformes
Malimbus scutatus	1.61E-02	7.4	Ploceidae	Passeriformes
Malurus amabilis	1.61E-02	7.4	Maluridae	Passeriformes
Ramphastos dicolorus	1.61E-02	7.4	Ramphastidae	Piciformes
Erythrura regia	1.60E-02	7.4	Estrildidae	Passeriformes
Myophonus melanurus	1.60E-02	7.4	Turdidae	Passeriformes
Botaurus stellaris	1.60E-02	7.4	Ardeidae	Pelecaniformes
Eulampis holosericeus	1.60E-02	7.4	Trochilidae	Apodiformes
Malimbus rubricollis	1.59E-02	7.3	Ploceidae	Passeriformes
Tanysiptera riedelii	1.59E-02	7.3	Alcedinidae	Coraciiformes
Actophilornis africanus	1.58E-02	7.3	Jacanidae	Charadriiformes
Cyanoramphus unicolor	1.57E-02	7.3	Psittacidae	Psittaciformes
Glossopsitta concinna	1.57E-02	7.2	Psittacidae	Psittaciformes
Phalacrocorax urile	1.56E-02	7.2	Phalacrocoracidae	Suliformes
Anastomus lamelligerus	1.55E-02	7.2	Ciconiidae	Ciconiiformes
Lybius bidentatus	1.55E-02	7.1	Ramphastidae	Piciformes
Buccanodon duchaillui	1.54E-02	7.1	Ramphastidae	Piciformes
Tangara cyanocephala	1.54E-02	7.1	Thraupidae	Passeriformes
Loriculus philippensis	1.53E-02	7.1	Psittacidae	Psittaciformes

Merops gularis	1.53E-02	7.1	Meropidae	Coraciiformes
Tanygnathus megalorynchos	1.53E-02	7.1	Psittacidae	Psittaciformes
Dryocopus schulzi	1.52E-02	7	Picidae	Piciformes
Cyanocorax luxuosus	1.52E-02	7	Corvidae	Passeriformes
Aquila verreauxii	1.52E-02	7	Accipitridae	Accipitriformes
Micropsitta pusio	1.51E-02	7	Psittacidae	Psittaciformes
Butorides striata	1.51E-02	7	Ardeidae	Pelecaniformes
Florisuga mellivora	1.51E-02	7	Trochilidae	Apodiformes
Chloropsis palawanensis	1.51E-02	7	Chloropseidae	Passeriformes
Amazona viridigenalis	1.51E-02	7	Psittacidae	Psittaciformes
Tauraco macrorhynchus	1.50E-02	7	Musophagidae	Musophagiformes
Baryphthengus martii	1.50E-02	6.9	Momotidae	Coraciiformes
Psittaculirostris desmarestii	1.50E-02	6.9	Psittacidae	Psittaciformes
Merops variegatus	1.50E-02	6.9	Meropidae	Coraciiformes
Bostrychia carunculata	1.50E-02	6.9	Threskiornithidae	Pelecaniformes
Pyrrhura picta	1.49E-02	6.9	Psittacidae	Psittaciformes
Centropus cupreicaudus	1.49E-02	6.9	Cuculidae	Cuculiformes
Chloropsis kinabaluensis	1.49E-02	6.9	Chloropseidae	Passeriformes
Ardeotis arabs	1.49E-02	6.9	Otididae	Otidiformes
Aprosmictus jonquillaceus	1.49E-02	6.9	Psittacidae	Psittaciformes
Boissonneaua jardini	1.48E-02	6.8	Trochilidae	Apodiformes
Nectarinia oritis	1.48E-02	6.8	Nectariniidae	Passeriformes
Agapornis roseicollis	1.48E-02	6.8	Psittacidae	Psittaciformes
Centropus violaceus	1.48E-02	6.8	Cuculidae	Cuculiformes
Oreopsittacus arfaki	1.48E-02	6.8	Psittacidae	Psittaciformes
Eudocimus ruber	1.47E-02	6.8	Threskiornithidae	Pelecaniformes
Galerida modesta	1.47E-02	6.8	Alaudidae	Passeriformes
Todiramphus lazuli	1.47E-02	6.8	Alcedinidae	Coraciiformes
Dicrurus aeneus	1.46E-02	6.7	Dicruridae	Passeriformes
Chalcopsitta sintillata	1.46E-02	6.7	Psittacidae	Psittaciformes
Alectroenas pulcherrima	1.45E-02	6.7	Columbidae	Columbiformes
Tauraco schalowi	1.45E-02	6.7	Musophagidae	Musophagiformes
Buthraupis montana	1.45E-02	6.7	Thraupidae	Passeriformes
Amazona oratrix	1.45E-02	6.7	Psittacidae	Psittaciformes
Urocissa caerulea	1.45E-02	6.7	Corvidae	Passeriformes
Anisognathus igniventris	1.44E-02	6.7	Thraupidae	Passeriformes
Pionites melanocephalus	1.44E-02	6.7	Psittacidae	Psittaciformes
Halcyon malimbica	1.44E-02	6.7	Alcedinidae	Coraciiformes
Centropus monachus	1.44E-02	6.6	Cuculidae	Cuculiformes

Melanerpes chrysogenys	1.44E-02	6.6	Picidae	Piciformes
Dryocopus hodgei	1.43E-02	6.6	Picidae	Piciformes
Momotus mexicanus	1.43E-02	6.6	Momotidae	Coraciiformes
Pteroglossus torquatus	1.43E-02	6.6	Ramphastidae	Piciformes
Centropus sinensis	1.43E-02	6.6	Cuculidae	Cuculiformes
Ceyx melanurus	1.42E-02	6.6	Alcedinidae	Coraciiformes
Manucodia comrii	1.42E-02	6.5	Paradisaeidae	Passeriformes
Tachuris rubrigastra	1.42E-02	6.5	Tyrannidae	Passeriformes
Psittacula longicauda	1.41E-02	6.5	Psittacidae	Psittaciformes
Megalaima nuchalis	1.40E-02	6.5	Ramphastidae	Piciformes
Dicrurus paradiseus	1.39E-02	6.4	Dicruridae	Passeriformes
Tangara seledon	1.38E-02	6.4	Thraupidae	Passeriformes
Amazona autumnalis	1.38E-02	6.4	Psittacidae	Psittaciformes
Lamprotornis purpureus	1.38E-02	6.4	Sturnidae	Passeriformes
Jacana jacana	1.38E-02	6.4	Jacanidae	Charadriiformes
Pteruthius rufiventer	1.38E-02	6.4	Timaliidae	Passeriformes
Aglaiocercus coelestis	1.37E-02	6.3	Trochilidae	Apodiformes
Chlorostilbon maugaeus	1.37E-02	6.3	Trochilidae	Apodiformes

Supplementary Table 6.2.6. Top 200 proportional colour diversity species

Species name	Proportional colour diversity	Family	Order
Phigys solitarius	1.79E-01	Psittacidae	Psittaciformes
Charmosyna papou	1.76E-01	Psittacidae	Psittaciformes
Pitta ussheri	1.61E-01	Pittidae	Passeriformes
Cyanerpes caeruleus	1.60E-01	Thraupidae	Passeriformes
Tangara callophrys	1.59E-01	Thraupidae	Passeriformes
Alcedo quadribrachys	1.59E-01	Alcedinidae	Coraciiformes
Anisognathus notabilis	1.57E-01	Thraupidae	Passeriformes
Alisterus amboinensis	1.57E-01	Psittacidae	Psittaciformes
Buthraupis montana	1.55E-01	Thraupidae	Passeriformes
Vini kuhlii	1.54E-01	Psittacidae	Psittaciformes
Pitta granatina	1.54E-01	Pittidae	Passeriformes
Charmosyna josefinae	1.53E-01	Psittacidae	Psittaciformes
Cyanerpes cyaneus	1.51E-01	Thraupidae	Passeriformes
Charmosyna margarethae	1.50E-01	Psittacidae	Psittaciformes
Lorius lory	1.48E-01	Psittacidae	Psittaciformes
Malurus splendens	1.48E-01	Maluridae	Passeriformes
Ara macao	1.46E-01	Psittacidae	Psittaciformes
Charmosyna pulchella	1.43E-01	Psittacidae	Psittaciformes
Actenoides bougainvillei	1.42E-01	Alcedinidae	Coraciiformes
Lorius hypoinochrous	1.37E-01	Psittacidae	Psittaciformes
Cyanerpes lucidus	1.36E-01	Thraupidae	Passeriformes

Pitta arcuata	1.36E-01	Pittidae	Passeriformes
Platycercus elegans	1.36E-01	Psittacidae	Psittaciformes
Vini australis	1.35E-01	Psittacidae	Psittaciformes
Ceyx lecontei	1.35E-01	Alcedinidae	Coraciiformes
Vini stepheni	1.33E-01	Psittacidae	Psittaciformes
Alcedo leucogaster	1.33E-01	Alcedinidae	Coraciiformes
Chlorochrysa nitidissima	1.32E-01	Thraupidae	Passeriformes
Ceyx erithaca	1.31E-01	Alcedinidae	Coraciiformes
Eclectus roratus	1.30E-01	Psittacidae	Psittaciformes
Pipra filicauda	1.30E-01	Pipridae	Passeriformes
Ceyx pictus	1.29E-01	Alcedinidae	Coraciiformes
Tanysiptera sylvia	1.28E-01	Alcedinidae	Coraciiformes
Anisognathus igniventris	1.27E-01	Thraupidae	Passeriformes
Prosopeia splendens	1.27E-01	Psittacidae	Psittaciformes
Alisterus chloropterus	1.27E-01	Psittacidae	Psittaciformes
Lorius chlorocercus	1.27E-01	Psittacidae	Psittaciformes
Ara chloropterus	1.26E-01	Psittacidae	Psittaciformes
Loriculus amabilis	1.25E-01	Psittacidae	Psittaciformes
Icterus croconotus	1.25E-01	Icteridae	Passeriformes
Lorius albidinucha	1.24E-01	Psittacidae	Psittaciformes
Alisterus scapularis	1.24E-01	Psittacidae	Psittaciformes
Ara ararauna	1.23E-01	Psittacidae	Psittaciformes
Icterus auratus	1.23E-01	Icteridae	Passeriformes
Trichoglossus haematodus	1.22E-01	Psittacidae	Psittaciformes
Halcyon pileata	1.22E-01	Alcedinidae	Coraciiformes
Niltava vivida	1.22E-01	Muscicapidae	Passeriformes
Platycercus eximius	1.22E-01	Psittacidae	Psittaciformes
Alcedo azurea	1.21E-01	Alcedinidae	Coraciiformes
Eos histrio	1.20E-01	Psittacidae	Psittaciformes
Vini peruviana	1.20E-01	Psittacidae	Psittaciformes
Urocissa ornata	1.19E-01	Corvidae	Passeriformes
Alcedo vintsioides	1.19E-01	Alcedinidae	Coraciiformes
Ceyx lepidus	1.19E-01	Alcedinidae	Coraciiformes
Alcedo cristata	1.19E-01	Alcedinidae	Coraciiformes
Pitta baudii	1.18E-01	Pittidae	Passeriformes
Todiramphus leucopygius	1.17E-01	Alcedinidae	Coraciiformes
Loriculus sclateri	1.17E-01	Psittacidae	Psittaciformes
Pericrocotus flammeus	1.16E-01	Campephagidae	Passeriformes
Alcedo meninting	1.16E-01	Alcedinidae	Coraciiformes
Pericrocotus brevirostris	1.15E-01	Campephagidae	Passeriformes
Passerina ciris	1.15E-01	Emberizidae	Passeriformes
Loriculus philippensis	1.14E-01	Psittacidae	Psittaciformes
Icterus jamacaii	1.14E-01	Icteridae	Passeriformes
Psittaculirostris edwardsii	1.14E-01	Psittacidae	Psittaciformes
Malurus pulcherrimus	1.14E-01	Maluridae	Passeriformes

Prosopeia tabuensis	1.14E-01	Psittacidae	Psittaciformes
Tanysiptera ellioti	1.14E-01	Alcedinidae	Coraciiformes
Tanysiptera carolinae	1.13E-01	Alcedinidae	Coraciiformes
Trichoglossus ornatus	1.13E-01	Psittacidae	Psittaciformes
Icterus pectoralis	1.13E-01	Icteridae	Passeriformes
Piranga leucoptera	1.13E-01	Cardinalidae	Passeriformes
Telophorus dohertyi	1.13E-01	Malaconotidae	Passeriformes
Icterus icterus	1.12E-01	Icteridae	Passeriformes
Pitta dohertyi	1.12E-01	Pittidae	Passeriformes
Touit purpuratus	1.11E-01	Psittacidae	Psittaciformes
Icterus maculialatus	1.11E-01	Icteridae	Passeriformes
Cosmopsarus regius	1.11E-01	Sturnidae	Passeriformes
Lorius garrulus	1.11E-01	Psittacidae	Psittaciformes
Pericrocotus igneus	1.10E-01	Campephagidae	Passeriformes
Eos cyanogenia	1.10E-01	Psittacidae	Psittaciformes
Piranga rubriceps	1.10E-01	Cardinalidae	Passeriformes
Coracias temminckii	1.10E-01	Coraciidae	Coraciiformes
Lorius domicella	1.10E-01	Psittacidae	Psittaciformes
Pachycephala aurea	1.09E-01	Pachycephalidae	Passeriformes
Cotinga cotinga	1.09E-01	Cotingidae	Passeriformes
Niltava davidi	1.08E-01	Muscicapidae	Passeriformes
Loriculus exilis	1.08E-01	Psittacidae	Psittaciformes
Erythrura regia	1.08E-01	Estrildidae	Passeriformes
Niltava sundara	1.07E-01	Muscicapidae	Passeriformes
Pionites leucogaster	1.07E-01	Psittacidae	Psittaciformes
Baryphthengus martii	1.07E-01	Momotidae	Coraciiformes
Pyrilia pyrilia	1.06E-01	Psittacidae	Psittaciformes
Malurus cyanocephalus	1.06E-01	Maluridae	Passeriformes
Icterus graceannae	1.06E-01	Icteridae	Passeriformes
Pericrocotus solaris	1.06E-01	Campephagidae	Passeriformes
Pipra fasciicauda	1.06E-01	Pipridae	Passeriformes
Aratinga solstitialis	1.06E-01	Psittacidae	Psittaciformes
Piranga olivacea	1.05E-01	Cardinalidae	Passeriformes
Ramphastos dicolorus	1.05E-01	Ramphastidae	Piciformes
Todiramphus farquhari	1.05E-01	Alcedinidae	Coraciiformes
Tangara gyrola	1.05E-01	Thraupidae	Passeriformes
Pericrocotus ethologus	1.05E-01	Campephagidae	Passeriformes
Tangara seledon	1.04E-01	Thraupidae	Passeriformes
Tangara arthus	1.04E-01	Thraupidae	Passeriformes
Todiramphus winchelli	1.04E-01	Alcedinidae	Coraciiformes
Campochaera sloetii	1.04E-01	Campephagidae	Passeriformes
Icterus gularis	1.04E-01	Icteridae	Passeriformes
Halcyon cyanoventris	1.04E-01	Alcedinidae	Coraciiformes
Telophorus viridis	1.04E-01	Malaconotidae	Passeriformes
Aprosmictus erythropterus	1.04E-01	Psittacidae	Psittaciformes

Touit huetii	1.03E-01	Psittacidae	Psittaciformes
Antilophia galeata	1.03E-01	Pipridae	Passeriformes
Merops muelleri	1.03E-01	Meropidae	Coraciiformes
Chlorophanes spiza	1.03E-01	Thraupidae	Passeriformes
Cyornis superbus	1.03E-01	Muscicapidae	Passeriformes
Paroaria dominicana	1.02E-01	Emberizidae	Passeriformes
Cyanerpes nitidus	1.02E-01	Thraupidae	Passeriformes
Telophorus quadricolor	1.02E-01	Malaconotidae	Passeriformes
Pyrilia caica	1.02E-01	Psittacidae	Psittaciformes
Tangara velia	1.02E-01	Thraupidae	Passeriformes
Ficedula sapphira	1.02E-01	Muscicapidae	Passeriformes
Icterus auricapillus	1.02E-01	Icteridae	Passeriformes
Passerina rositae	1.01E-01	Emberizidae	Passeriformes
Cyclopsitta gulielmitertii	1.01E-01	Psittacidae	Psittaciformes
Neopsittacus pullicauda	1.01E-01	Psittacidae	Psittaciformes
Myzomela rosenbergii	1.01E-01	Meliphagidae	Passeriformes
Ceyx fallax	1.01E-01	Alcedinidae	Coraciiformes
Pyrocephalus rubinus	1.01E-01	Tyrannidae	Passeriformes
Niltava sumatrana	1.00E-01	Muscicapidae	Passeriformes
Loriculus galgulus	1.00E-01	Psittacidae	Psittaciformes
Merops breweri	1.00E-01	Meropidae	Coraciiformes
Tanysiptera galatea	1.00E-01	Alcedinidae	Coraciiformes
Icterus wagleri	1.00E-01	Icteridae	Passeriformes
Eubucco bourcierii	1.00E-01	Ramphastidae	Piciformes
Ramphocelus nigrogularis	1.00E-01	Thraupidae	Passeriformes
Euphonia musica	9.99E-02	Thraupidae	Passeriformes
Ramphocelus bresilius	9.98E-02	Thraupidae	Passeriformes
Euphonia cyanocephala	9.97E-02	Thraupidae	Passeriformes
Tangara chilensis	9.95E-02	Thraupidae	Passeriformes
Ramphocelus flammigerus	9.95E-02	Thraupidae	Passeriformes
Pyrilia pulchra	9.94E-02	Psittacidae	Psittaciformes
Aratinga jandaya	9.91E-02	Psittacidae	Psittaciformes
Euphonia elegantissima	9.90E-02	Thraupidae	Passeriformes
Amblyramphus holosericeus	9.90E-02	Icteridae	Passeriformes
Icterus mesomelas	9.90E-02	Icteridae	Passeriformes
Aratinga erythrogenys	9.86E-02	Psittacidae	Psittaciformes
Ploceus insignis	9.86E-02	Ploceidae	Passeriformes
Chrysothlypis salmoni	9.83E-02	Thraupidae	Passeriformes
Pipraeidea melanonota	9.83E-02	Thraupidae	Passeriformes
Semnornis ramphastinus	9.81E-02	Ramphastidae	Piciformes
Pitta erythrogaster	9.80E-02	Pittidae	Passeriformes
Calyptomena whiteheadi	9.79E-02	Eurylaimidae	Passeriformes
Haematoderus militaris	9.79E-02	Cotingidae	Passeriformes
Platycercus icterotis	9.79E-02	Psittacidae	Psittaciformes
Neophema pulchella	9.78E-02	Psittacidae	Psittaciformes

Coracias cyanogaster	9.77E-02	Coraciidae	Coraciiformes
Alcedo semitorquata	9.72E-02	Alcedinidae	Coraciiformes
Icterus pustulatus	9.70E-02	Icteridae	Passeriformes
Ramphocelus passerinii	9.70E-02	Thraupidae	Passeriformes
Cyclopsitta diophthalma	9.69E-02	Psittacidae	Psittaciformes
Mino anais	9.68E-02	Sturnidae	Passeriformes
Icterus galbula	9.66E-02	Icteridae	Passeriformes
Ploceus dorsomaculatus	9.64E-02	Ploceidae	Passeriformes
Ara glaucogularis	9.64E-02	Psittacidae	Psittaciformes
Cyanolyca pulchra	9.62E-02	Corvidae	Passeriformes
Ploceus bicolor	9.61E-02	Ploceidae	Passeriformes
Calyptomena hosii	9.60E-02	Eurylaimidae	Passeriformes
Tangara fastuosa	9.60E-02	Thraupidae	Passeriformes
Melanerpes flavifrons	9.59E-02	Picidae	Piciformes
Polytelis swainsonii	9.59E-02	Psittacidae	Psittaciformes
Piranga erythrocephala	9.58E-02	Cardinalidae	Passeriformes
Garrulus lidthi	9.57E-02	Corvidae	Passeriformes
Pyrilia barrabandi	9.56E-02	Psittacidae	Psittaciformes
Euphonia saturata	9.56E-02	Thraupidae	Passeriformes
Anaplectes rubriceps	9.55E-02	Ploceidae	Passeriformes
Agapornis fischeri	9.54E-02	Psittacidae	Psittaciformes
Iridosornis rufivertex	9.54E-02	Thraupidae	Passeriformes
Campephaga lobata	9.53E-02	Campephagidae	Passeriformes
Aglaiocercus kingi	9.51E-02	Trochilidae	Apodiformes
Cyornis turcosus	9.51E-02	Muscicapidae	Passeriformes
Aprosmictus jonquillaceus	9.51E-02	Psittacidae	Psittaciformes
Euphonia laniirostris	9.50E-02	Thraupidae	Passeriformes
Ramphocelus sanguinolentus	9.50E-02	Thraupidae	Passeriformes
Halcyon leucocephala	9.48E-02	Alcedinidae	Coraciiformes
Pitta moluccensis	9.46E-02	Pittidae	Passeriformes
Foudia madagascariensis	9.46E-02	Ploceidae	Passeriformes
Malaconotus cruentus	9.46E-02	Malaconotidae	Passeriformes
Pachycare flavogriseum	9.44E-02	Pachycephalidae	Passeriformes
Loriculus stigmatus	9.42E-02	Psittacidae	Psittaciformes
Aethopyga mystacalis	9.41E-02	Nectariniidae	Passeriformes
Delothraupis castaneoventris	9.41E-02	Thraupidae	Passeriformes
Icterus oberi	9.40E-02	Icteridae	Passeriformes
Jacamerops aureus	9.40E-02	Galbulidae	Piciformes
Charmosyna amabilis	9.39E-02	Psittacidae	Psittaciformes
Passerina leclancherii	9.38E-02	Emberizidae	Passeriformes
Meropogon forsteni	9.37E-02	Meropidae	Coraciiformes
Hirundo nigrorufa	9.34E-02	Hirundinidae	Passeriformes
Eos reticulata	9.34E-02	Psittacidae	Psittaciformes
Tangara desmaresti	9.33E-02	Thraupidae	Passeriformes

Supplementary Table 6.2.7. Top 200 proportional colour diversity male species

Species name	Proportional colour diversity	Family	Order
Cyanerpes cyaneus	1.95E-01	Thraupidae	Passeriformes
Cyanerpes caeruleus	1.88E-01	Thraupidae	Passeriformes
Phigys solitarius	1.84E-01	Psittacidae	Psittaciformes
Vini kuhlii	1.68E-01	Psittacidae	Psittaciformes
Pitta ussheri	1.64E-01	Pittidae	Passeriformes
Anisognathus notabilis	1.63E-01	Thraupidae	Passeriformes
Actenoides bougainvillei	1.62E-01	Alcedinidae	Coraciiformes
Charmosyna papou	1.61E-01	Psittacidae	Psittaciformes
Pitta granatina	1.59E-01	Pittidae	Passeriformes
Malurus pulcherrimus	1.58E-01	Maluridae	Passeriformes
Charmosyna josefinae	1.58E-01	Psittacidae	Psittaciformes
Alisterus chloropterus	1.57E-01	Psittacidae	Psittaciformes
Alisterus amboinensis	1.56E-01	Psittacidae	Psittaciformes
Charmosyna pulchella	1.54E-01	Psittacidae	Psittaciformes
Lorius lory	1.54E-01	Psittacidae	Psittaciformes
Pitta arcuata	1.52E-01	Pittidae	Passeriformes
Buthraupis montana	1.49E-01	Thraupidae	Passeriformes
Malurus splendens	1.48E-01	Maluridae	Passeriformes
Charmosyna margarethae	1.46E-01	Psittacidae	Psittaciformes
Alcedo quadribrachys	1.44E-01	Alcedinidae	Coraciiformes
Niltava vivida	1.44E-01	Muscicapidae	Passeriformes
Aprosmictus erythropterus	1.44E-01	Psittacidae	Psittaciformes
Niltava sumatrana	1.44E-01	Muscicapidae	Passeriformes
Ceyx erithaca	1.43E-01	Alcedinidae	Coraciiformes
Platycercus eximius	1.42E-01	Psittacidae	Psittaciformes
Niltava sundara	1.41E-01	Muscicapidae	Passeriformes
Vini australis	1.39E-01	Psittacidae	Psittaciformes
Pipra filicauda	1.39E-01	Pipridae	Passeriformes
Vini stepheni	1.38E-01	Psittacidae	Psittaciformes
Alcedo leucogaster	1.37E-01	Alcedinidae	Coraciiformes
Niltava davidi	1.35E-01	Muscicapidae	Passeriformes
Lorius hypoinochrous	1.34E-01	Psittacidae	Psittaciformes
Passerina ciris	1.33E-01	Emberizidae	Passeriformes
Alcedo cristata	1.32E-01	Alcedinidae	Coraciiformes
Tanysiptera sylvia	1.32E-01	Alcedinidae	Coraciiformes
Chlorochrysa nitidissima	1.32E-01	Thraupidae	Passeriformes
Loriculus galgulus	1.30E-01	Psittacidae	Psittaciformes
Ara chloropterus	1.29E-01	Psittacidae	Psittaciformes
Anisognathus igniventris	1.29E-01	Thraupidae	Passeriformes

Lorius chlorocercus	1.29E-01	Psittacidae	Psittaciformes
Irena puella	1.29E-01	Irenidae	Passeriformes
Ara ararauna	1.28E-01	Psittacidae	Psittaciformes
Ceyx lepidus	1.28E-01	Alcedinidae	Coraciiformes
Ara macao	1.28E-01	Psittacidae	Psittaciformes
Prosopeia splendens	1.28E-01	Psittacidae	Psittaciformes
Euphonia cyanocephala	1.27E-01	Thraupidae	Passeriformes
Loriculus sclateri	1.27E-01	Psittacidae	Psittaciformes
Cyornis superbus	1.26E-01	Muscicapidae	Passeriformes
Loriculus amabilis	1.26E-01	Psittacidae	Psittaciformes
Icterus croconotus	1.25E-01	Icteridae	Passeriformes
Alisterus scapularis	1.25E-01	Psittacidae	Psittaciformes
Euphonia laniirostris	1.25E-01	Thraupidae	Passeriformes
Vini peruviana	1.25E-01	Psittacidae	Psittaciformes
Euphonia saturata	1.24E-01	Thraupidae	Passeriformes
Alcedo meninting	1.24E-01	Alcedinidae	Coraciiformes
Eos cyanogenia	1.24E-01	Psittacidae	Psittaciformes
Ceyx pictus	1.24E-01	Alcedinidae	Coraciiformes
Lorius albidinucha	1.24E-01	Psittacidae	Psittaciformes
Todiramphus leucopygius	1.23E-01	Alcedinidae	Coraciiformes
Platycercus icterotis	1.23E-01	Psittacidae	Psittaciformes
Euphonia violacea	1.22E-01	Thraupidae	Passeriformes
Euphonia elegantissima	1.22E-01	Thraupidae	Passeriformes
Ficedula sapphira	1.22E-01	Muscicapidae	Passeriformes
Pitta baudii	1.22E-01	Pittidae	Passeriformes
Erythrura regia	1.22E-01	Estrildidae	Passeriformes
Cotinga cotinga	1.21E-01	Cotingidae	Passeriformes
Halcyon pileata	1.20E-01	Alcedinidae	Coraciiformes
Loriculus philippensis	1.20E-01	Psittacidae	Psittaciformes
Urocissa ornata	1.20E-01	Corvidae	Passeriformes
Psittaculirostris edwardsii	1.19E-01	Psittacidae	Psittaciformes
Alcedo azurea	1.19E-01	Alcedinidae	Coraciiformes
Trichoglossus haematodus	1.19E-01	Psittacidae	Psittaciformes
Platycercus elegans	1.19E-01	Psittacidae	Psittaciformes
Cyanerpes lucidus	1.19E-01	Thraupidae	Passeriformes
Ceyx lecontei	1.17E-01	Alcedinidae	Coraciiformes
Passerina leclancherii	1.17E-01	Emberizidae	Passeriformes
Telophorus dohertyi	1.17E-01	Malaconotidae	Passeriformes
Pachycephala aurea	1.16E-01	Pachycephalidae	Passeriformes
Lacedo pulchella	1.16E-01	Alcedinidae	Coraciiformes
Tangara seledon	1.16E-01	Thraupidae	Passeriformes
Euphonia musica	1.15E-01	Thraupidae	Passeriformes
Alcedo vintsioides	1.15E-01	Alcedinidae	Coraciiformes
Icterus jamacaii	1.14E-01	Icteridae	Passeriformes
Nectarinia regia	1.14E-01	Nectariniidae	Passeriformes

Euphonia concinna	1.14E-01	Thraupidae	Passeriformes
Tangara velia	1.14E-01	Thraupidae	Passeriformes
Aethopyga gouldiae	1.14E-01	Nectariniidae	Passeriformes
Pipra fasciicauda	1.14E-01	Pipridae	Passeriformes
Cosmopsarus regius	1.14E-01	Sturnidae	Passeriformes
Tanysiptera carolinae	1.13E-01	Alcedinidae	Coraciiformes
Eubucco bourcierii	1.13E-01	Ramphastidae	Piciformes
Icterus pectoralis	1.13E-01	Icteridae	Passeriformes
Tangara chilensis	1.13E-01	Thraupidae	Passeriformes
Pipraeidea melanonota	1.12E-01	Thraupidae	Passeriformes
Calyptomena hosii	1.12E-01	Eurylaimidae	Passeriformes
Icterus icterus	1.12E-01	Icteridae	Passeriformes
Muscicapella hodgsoni	1.11E-01	Muscicapidae	Passeriformes
Sericulus chrysocephalus	1.11E-01	Ptilonorhynchidae	Passeriformes
Icterus maculialatus	1.11E-01	Icteridae	Passeriformes
Campochaera sloetii	1.11E-01	Campephagidae	Passeriformes
Lorius garrulus	1.11E-01	Psittacidae	Psittaciformes
Chlorophonia cyanea	1.10E-01	Thraupidae	Passeriformes
Pyrilia pyrilia	1.10E-01	Psittacidae	Psittaciformes
Calyptomena whiteheadi	1.10E-01	Eurylaimidae	Passeriformes
Euphonia rufiventris	1.10E-01	Thraupidae	Passeriformes
Touit purpuratus	1.10E-01	Psittacidae	Psittaciformes
Charmosyna placentis	1.10E-01	Psittacidae	Psittaciformes
Euphonia anneae	1.10E-01	Thraupidae	Passeriformes
Lorius domicella	1.10E-01	Psittacidae	Psittaciformes
Icterus bullockii	1.10E-01	Icteridae	Passeriformes
Cyclopsitta gulielmitertii	1.09E-01	Psittacidae	Psittaciformes
Telophorus quadricolor	1.09E-01	Malaconotidae	Passeriformes
Telophorus viridis	1.09E-01	Malaconotidae	Passeriformes
Icterus cucullatus	1.09E-01	Icteridae	Passeriformes
Ramphocelus nigrogularis	1.09E-01	Thraupidae	Passeriformes
Cyclopsitta diophthalma	1.09E-01	Psittacidae	Psittaciformes
Tangara arthus	1.09E-01	Thraupidae	Passeriformes
Prosopeia tabuensis	1.09E-01	Psittacidae	Psittaciformes
Antilophia galeata	1.09E-01	Pipridae	Passeriformes
Aglaiocercus kingi	1.09E-01	Trochilidae	Apodiformes
Neophema splendida	1.08E-01	Psittacidae	Psittaciformes
Loriculus stigmatus	1.08E-01	Psittacidae	Psittaciformes
Loriculus exilis	1.08E-01	Psittacidae	Psittaciformes
Halcyon leucocephala	1.08E-01	Alcedinidae	Coraciiformes
Eos histrio	1.08E-01	Psittacidae	Psittaciformes
Tangara gyrola	1.08E-01	Thraupidae	Passeriformes
Ramphastos dicolorus	1.08E-01	Ramphastidae	Piciformes
Tarsiger hyperythrus	1.08E-01	Muscicapidae	Passeriformes
Malimbus racheliae	1.08E-01	Ploceidae	Passeriformes

Trichoglossus ornatus	1.07E-01	Psittacidae	Psittaciformes
Mycerobas affinis	1.07E-01	Fringillidae	Passeriformes
Aglaiocercus coelestis	1.07E-01	Trochilidae	Apodiformes
Icterus graceannae	1.07E-01	Icteridae	Passeriformes
Trogon elegans	1.07E-01	Trogonidae	Trogoniformes
Merops muelleri	1.07E-01	Meropidae	Coraciiformes
Euphonia xanthogaster	1.07E-01	Thraupidae	Passeriformes
Chrysothlypis chrysomelas	1.07E-01	Thraupidae	Passeriformes
Campephaga lobata	1.07E-01	Campephagidae	Passeriformes
Pipra erythrocephala	1.06E-01	Pipridae	Passeriformes
Phoenicircus nigricollis	1.06E-01	Cotingidae	Passeriformes
Aratinga solstitialis	1.06E-01	Psittacidae	Psittaciformes
Xanthopsar flavus	1.06E-01	Icteridae	Passeriformes
Chlorophonia pyrrhophrys	1.06E-01	Thraupidae	Passeriformes
Sialia mexicana	1.06E-01	Turdidae	Passeriformes
Polytelis swainsonii	1.06E-01	Psittacidae	Psittaciformes
Malurus amabilis	1.05E-01	Maluridae	Passeriformes
Piranga rubriceps	1.05E-01	Cardinalidae	Passeriformes
Icterus wagleri	1.05E-01	Icteridae	Passeriformes
Icterus galbula	1.05E-01	Icteridae	Passeriformes
Monticola cinclorhynchus	1.05E-01	Muscicapidae	Passeriformes
Piranga ludoviciana	1.05E-01	Cardinalidae	Passeriformes
Cyornis turcosus	1.05E-01	Muscicapidae	Passeriformes
Icterus mesomelas	1.05E-01	Icteridae	Passeriformes
Cyanolanius madagascarinus	1.05E-01	Vangidae	Passeriformes
Myzomela rosenbergii	1.05E-01	Meliphagidae	Passeriformes
Euplectes gierowii	1.05E-01	Ploceidae	Passeriformes
Cyornis caerulatus	1.04E-01	Muscicapidae	Passeriformes
Actenoides concretus	1.04E-01	Alcedinidae	Coraciiformes
Semnornis ramphastinus	1.04E-01	Ramphastidae	Piciformes
Tangara cyanocephala	1.04E-01	Thraupidae	Passeriformes
Pipra chloromeros	1.04E-01	Pipridae	Passeriformes
Aethopyga ignicauda	1.04E-01	Nectariniidae	Passeriformes
Tangara lavinia	1.04E-01	Thraupidae	Passeriformes
Pyrilia pulchra	1.03E-01	Psittacidae	Psittaciformes
Baryphthengus martii	1.03E-01	Momotidae	Coraciiformes
Touit huetii	1.03E-01	Psittacidae	Psittaciformes
Pyrilia barrabandi	1.03E-01	Psittacidae	Psittaciformes
Halcyon cyanoventris	1.03E-01	Alcedinidae	Coraciiformes
Aprosmictus jonquillaceus	1.02E-01	Psittacidae	Psittaciformes
Paroaria dominicana	1.02E-01	Emberizidae	Passeriformes
Aethopyga nipalensis	1.02E-01	Nectariniidae	Passeriformes
Spermophaga ruficapilla	1.02E-01	Estrildidae	Passeriformes
Ploceus dorsomaculatus	1.02E-01	Ploceidae	Passeriformes
Cyanerpes nitidus	1.02E-01	Thraupidae	Passeriformes

Pachycare flavogriseum	1.02E-01	Pachycephalidae	Passeriformes
Merops breweri	1.02E-01	Meropidae	Coraciiformes
Aethopyga mystacalis	1.02E-01	Nectariniidae	Passeriformes
Psittacula roseata	1.01E-01	Psittacidae	Psittaciformes
Tangara cyanoventris	1.01E-01	Thraupidae	Passeriformes
Icterus gularis	1.01E-01	Icteridae	Passeriformes
Icterus auricapillus	1.01E-01	Icteridae	Passeriformes
Euphonia trinitatis	1.01E-01	Thraupidae	Passeriformes
Mino anais	1.01E-01	Sturnidae	Passeriformes
Spermophaga haematina	1.00E-01	Estrildidae	Passeriformes
Melanerpes flavifrons	1.00E-01	Picidae	Piciformes
Ramphocelus sanguinolentus	1.00E-01	Thraupidae	Passeriformes
Pericrocotus brevirostris	1.00E-01	Campephagidae	Passeriformes
Pericrocotus flammeus	1.00E-01	Campephagidae	Passeriformes
Thraupis bonariensis	9.98E-02	Thraupidae	Passeriformes
Lanio aurantius	9.98E-02	Thraupidae	Passeriformes
Aratinga jandaya	9.98E-02	Psittacidae	Psittaciformes
Pipra mentalis	9.96E-02	Pipridae	Passeriformes
Euphonia finschi	9.95E-02	Thraupidae	Passeriformes
Platysteira concreta	9.94E-02	Platysteiridae	Passeriformes
Phoenicurus frontalis	9.94E-02	Muscicapidae	Passeriformes
Psittacula cyanocephala	9.93E-02	Psittacidae	Psittaciformes
Chlorophonia callophrys	9.92E-02	Thraupidae	Passeriformes
Chlorochrysa calliparaea	9.90E-02	Thraupidae	Passeriformes
Prionochilus thoracicus	9.88E-02	Dicaeidae	Passeriformes
Pitta erythrogaster	9.86E-02	Pittidae	Passeriformes

Supplementary Table 6.2.8. Top 200 proportional colour diversity female species

Species name	proportional colour diversity	BLFamilyLatin	IOCOrder
Charmosyna papou	1.80E-01	Psittacidae	Psittaciformes
Phigys solitarius	1.68E-01	Psittacidae	Psittaciformes
Ara macao	1.60E-01	Psittacidae	Psittaciformes
Alcedo quadribrachys	1.57E-01	Alcedinidae	Coraciiformes
Buthraupis montana	1.56E-01	Thraupidae	Passeriformes
Pitta ussheri	1.55E-01	Pittidae	Passeriformes
Charmosyna margarethae	1.53E-01	Psittacidae	Psittaciformes
Alisterus amboinensis	1.50E-01	Psittacidae	Psittaciformes
Pitta granatina	1.49E-01	Pittidae	Passeriformes
Charmosyna josefinae	1.46E-01	Psittacidae	Psittaciformes
Anisognathus notabilis	1.45E-01	Thraupidae	Passeriformes
Lorius hypoinochrous	1.40E-01	Psittacidae	Psittaciformes
Ceyx lecontei	1.37E-01	Alcedinidae	Coraciiformes

Platycercus elegans	1.33E-01	Psittacidae	Psittaciformes
Vini stepheni	1.32E-01	Psittacidae	Psittaciformes
Charmosyna pulchella	1.32E-01	Psittacidae	Psittaciformes
Vini australis	1.30E-01	Psittacidae	Psittaciformes
Vini kuhlii	1.28E-01	Psittacidae	Psittaciformes
Alcedo leucogaster	1.27E-01	Alcedinidae	Coraciiformes
Halcyon pileata	1.24E-01	Alcedinidae	Coraciiformes
Lorius lory	1.24E-01	Psittacidae	Psittaciformes
Trichoglossus haematodus	1.22E-01	Psittacidae	Psittaciformes
Alcedo azurea	1.22E-01	Alcedinidae	Coraciiformes
Eos histrio	1.21E-01	Psittacidae	Psittaciformes
Pitta arcuata	1.21E-01	Pittidae	Passeriformes
Ara ararauna	1.19E-01	Psittacidae	Psittaciformes
Loriculus amabilis	1.18E-01	Psittacidae	Psittaciformes
Ceyx lepidus	1.18E-01	Alcedinidae	Coraciiformes
Anisognathus igniventris	1.17E-01	Thraupidae	Passeriformes
Coracias temminckii	1.16E-01	Coraciidae	Coraciiformes
Ara chloropterus	1.16E-01	Psittacidae	Psittaciformes
Prosopeia tabuensis	1.16E-01	Psittacidae	Psittaciformes
Vini peruviana	1.14E-01	Psittacidae	Psittaciformes
Lorius chlorocercus	1.14E-01	Psittacidae	Psittaciformes
Alcedo vintsioides	1.14E-01	Alcedinidae	Coraciiformes
Tanysiptera ellioti	1.14E-01	Alcedinidae	Coraciiformes
Prosopeia splendens	1.12E-01	Psittacidae	Psittaciformes
Ceyx pictus	1.12E-01	Alcedinidae	Coraciiformes
Pitta moluccensis	1.12E-01	Pittidae	Passeriformes
Todiramphus leucopygius	1.11E-01	Alcedinidae	Coraciiformes
Baryphthengus martii	1.09E-01	Momotidae	Coraciiformes
Touit purpuratus	1.09E-01	Psittacidae	Psittaciformes
Urocissa ornata	1.09E-01	Corvidae	Passeriformes
Ceyx erithaca	1.09E-01	Alcedinidae	Coraciiformes
Alcedo meninting	1.08E-01	Alcedinidae	Coraciiformes
Halcyon cyanoventris	1.07E-01	Alcedinidae	Coraciiformes
Icterus jamacaii	1.07E-01	Icteridae	Passeriformes
Eos cyanogenia	1.07E-01	Psittacidae	Psittaciformes
Cosmopsarus regius	1.07E-01	Sturnidae	Passeriformes
Tanysiptera galatea	1.06E-01	Alcedinidae	Coraciiformes
Actenoides bougainvillei	1.06E-01	Alcedinidae	Coraciiformes
Icterus gularis	1.06E-01	Icteridae	Passeriformes
Loriculus philippensis	1.06E-01	Psittacidae	Psittaciformes
Todiramphus winchelli	1.06E-01	Alcedinidae	Coraciiformes
Alcedo cristata	1.06E-01	Alcedinidae	Coraciiformes
Icterus graceannae	1.05E-01	Icteridae	Passeriformes
Telophorus dohertyi	1.05E-01	Malaconotidae	Passeriformes
Todiramphus farquhari	1.05E-01	Alcedinidae	Coraciiformes

Icterus icterus	1.05E-01	Icteridae	Passeriformes
Platycercus eximius	1.04E-01	Psittacidae	Psittaciformes
Anisognathus somptuosus	1.03E-01	Thraupidae	Passeriformes
Alcedo semitorquata	1.03E-01	Alcedinidae	Coraciiformes
Psittaculirostris edwardsii	1.02E-01	Psittacidae	Psittaciformes
Piranga rubriceps	1.02E-01	Cardinalidae	Passeriformes
Pyrilia caica	1.01E-01	Psittacidae	Psittaciformes
Loriculus sclateri	1.01E-01	Psittacidae	Psittaciformes
Neopsittacus pullicauda	1.01E-01	Psittacidae	Psittaciformes
Malurus cyanocephalus	1.01E-01	Maluridae	Passeriformes
Aratinga erythrogenys	1.00E-01	Psittacidae	Psittaciformes
Trichoglossus ornatus	1.00E-01	Psittacidae	Psittaciformes
Icterus laudabilis	1.00E-01	Icteridae	Passeriformes
Pachycephala aurea	1.00E-01	Pachycephalidae	Passeriformes
Ramphastos dicolorus	9.97E-02	Ramphastidae	Piciformes
Aratinga solstitialis	9.93E-02	Psittacidae	Psittaciformes
Tangara arthus	9.86E-02	Thraupidae	Passeriformes
Telophorus viridis	9.84E-02	Malaconotidae	Passeriformes
Icterus auricapillus	9.82E-02	Icteridae	Passeriformes
Aratinga auricapillus	9.79E-02	Psittacidae	Psittaciformes
Pericrocotus igneus	9.76E-02	Campephagidae	Passeriformes
Delothraupis castaneoventris	9.74E-02	Thraupidae	Passeriformes
Campochaera sloetii	9.73E-02	Campephagidae	Passeriformes
Tangara gyrola	9.72E-02	Thraupidae	Passeriformes
Merops breweri	9.69E-02	Meropidae	Coraciiformes
Ploceus dorsomaculatus	9.67E-02	Ploceidae	Passeriformes
Eclectus roratus	9.64E-02	Psittacidae	Psittaciformes
Merops muelleri	9.62E-02	Meropidae	Coraciiformes
Ploceus bicolor	9.59E-02	Ploceidae	Passeriformes
Alisterus chloropterus	9.54E-02	Psittacidae	Psittaciformes
Icterus mesomelas	9.53E-02	Icteridae	Passeriformes
Erythrura regia	9.50E-02	Estrildidae	Passeriformes
Pyrilia pulchra	9.47E-02	Psittacidae	Psittaciformes
Chlorochrysa phoenicotis	9.45E-02	Thraupidae	Passeriformes
Parula pitiayumi	9.44E-02	Parulidae	Passeriformes
Coracias cyanogaster	9.44E-02	Coraciidae	Coraciiformes
Malaconotus cruentus	9.43E-02	Malaconotidae	Passeriformes
Agapornis fischeri	9.35E-02	Psittacidae	Psittaciformes
Psarisomus dalhousiae	9.35E-02	Eurylaimidae	Passeriformes
Ramphocelus flammigerus	9.33E-02	Thraupidae	Passeriformes
Ara militaris	9.31E-02	Psittacidae	Psittaciformes
Aprosmictus jonquillaceus	9.31E-02	Psittacidae	Psittaciformes
Alcedo cyanopectus	9.31E-02	Alcedinidae	Coraciiformes
Pyrilia pyrilia	9.31E-02	Psittacidae	Psittaciformes
Anisognathus lacrymosus	9.28E-02	Thraupidae	Passeriformes

Amblyramphus holosericeus	9.27E-02	Icteridae	Passeriformes
Tangara velia	9.26E-02	Thraupidae	Passeriformes
Mino anais	9.22E-02	Sturnidae	Passeriformes
Pitta erythrogaster	9.21E-02	Pittidae	Passeriformes
Jacamerops aureus	9.21E-02	Galbulidae	Piciformes
Tangara chilensis	9.20E-02	Thraupidae	Passeriformes
Melanerpes flavifrons	9.20E-02	Picidae	Piciformes
Icterus wagleri	9.18E-02	Icteridae	Passeriformes
Tanysiptera sylvia	9.16E-02	Alcedinidae	Coraciiformes
Pionites melanocephalus	9.12E-02	Psittacidae	Psittaciformes
Todiramphus diops	9.10E-02	Alcedinidae	Coraciiformes
Aratinga jandaya	9.02E-02	Psittacidae	Psittaciformes
Icterus chrysater	9.02E-02	Icteridae	Passeriformes
Eos squamata	9.02E-02	Psittacidae	Psittaciformes
Tangara seledon	9.02E-02	Thraupidae	Passeriformes
Ramphocelus sanguinolentus	9.02E-02	Thraupidae	Passeriformes
Buthraupis eximia	9.01E-02	Thraupidae	Passeriformes
Psittaculirostris desmarestii	9.00E-02	Psittacidae	Psittaciformes
Pityriasis gymnocephala	8.99E-02	Pityriaseidae	Passeriformes
Myioborus ornatus	8.97E-02	Parulidae	Passeriformes
Pyrrhura picta	8.93E-02	Psittacidae	Psittaciformes
Charmosyna amabilis	8.91E-02	Psittacidae	Psittaciformes
Megalaima rafflesii	8.91E-02	Ramphastidae	Piciformes
Pitta steerii	8.90E-02	Pittidae	Passeriformes
Malimbus malimbicus	8.88E-02	Ploceidae	Passeriformes
Telophorus quadricolor	8.85E-02	Malaconotidae	Passeriformes
Icterus nigrogularis	8.81E-02	Icteridae	Passeriformes
Merops oreobates	8.81E-02	Meropidae	Coraciiformes
Coracias caudatus	8.81E-02	Coraciidae	Coraciiformes
Gymnomystax mexicanus	8.80E-02	Icteridae	Passeriformes
Pyrilia haematotis	8.73E-02	Psittacidae	Psittaciformes
Eurystomus glaucurus	8.73E-02	Coraciidae	Coraciiformes
Ceyx melanurus	8.72E-02	Alcedinidae	Coraciiformes
Pyrrhura molinae	8.71E-02	Psittacidae	Psittaciformes
Todus multicolor	8.70E-02	Todidae	Coraciiformes
Lybius dubius	8.70E-02	Ramphastidae	Piciformes
Pyrrhura leucotis	8.70E-02	Psittacidae	Psittaciformes
Tangara desmaresti	8.70E-02	Thraupidae	Passeriformes
Nyctyornis amictus	8.70E-02	Meropidae	Coraciiformes
Alisterus scapularis	8.69E-02	Psittacidae	Psittaciformes
Loriculus stigmatus	8.68E-02	Psittacidae	Psittaciformes
Priotelus temnurus	8.68E-02	Trogonidae	Trogoniformes
Semnornis ramphastinus	8.67E-02	Ramphastidae	Piciformes
Pyrilia barrabandi	8.65E-02	Psittacidae	Psittaciformes
Coracias spatulatus	8.65E-02	Coraciidae	Coraciiformes

Cyanicterus cyanicterus	8.62E-02	Thraupidae	Passeriformes
Merops bulocki	8.61E-02	Meropidae	Coraciiformes
Malimbus scutatus	8.59E-02	Ploceidae	Passeriformes
Hyliota violacea	8.59E-02	Sylviidae	Passeriformes
Cymbirhynchus macrorhynchos	8.58E-02	Eurylaimidae	Passeriformes
Pitta angolensis	8.57E-02	Pittidae	Passeriformes
Brotogeris pyrrhoptera	8.55E-02	Psittacidae	Psittaciformes
Ptilinopus pulchellus	8.52E-02	Columbidae	Columbiformes
Eos reticulata	8.50E-02	Psittacidae	Psittaciformes
Nandayus nenday	8.50E-02	Psittacidae	Psittaciformes
Lybius rolleti	8.50E-02	Ramphastidae	Piciformes
Eumomota superciliosa	8.49E-02	Momotidae	Coraciiformes
Pitta megarhyncha	8.47E-02	Pittidae	Passeriformes
Ploceus insignis	8.46E-02	Ploceidae	Passeriformes
Musophaga rossae	8.45E-02	Musophagidae	Musophagiformes
Oreopsittacus arfaki	8.44E-02	Psittacidae	Psittaciformes
Dryocopus schulzi	8.44E-02	Picidae	Piciformes
Ramphastos vitellinus	8.43E-02	Ramphastidae	Piciformes
Heliothryx barroti	8.41E-02	Trochilidae	Apodiformes
Dinopium benghalense	8.41E-02	Picidae	Piciformes
Icterus prosthemelas	8.41E-02	Icteridae	Passeriformes
Ceyx rufidorsa	8.40E-02	Alcedinidae	Coraciiformes
Empidornis semipartitus	8.39E-02	Muscicapidae	Passeriformes
Loriculus beryllinus	8.38E-02	Psittacidae	Psittaciformes
Eunymphicus cornutus	8.38E-02	Psittacidae	Psittaciformes
Coracias benghalensis	8.37E-02	Coraciidae	Coraciiformes
Campephilus rubricollis	8.37E-02	Picidae	Piciformes
Centropus rectunguis	8.36E-02	Cuculidae	Cuculiformes
Lybius bidentatus	8.34E-02	Ramphastidae	Piciformes
Merops leschenaulti	8.34E-02	Meropidae	Coraciiformes
Hypopyrrhus pyrohypogaster	8.33E-02	Icteridae	Passeriformes
Pionus menstruus	8.33E-02	Psittacidae	Psittaciformes
Garrulus lidthi	8.32E-02	Corvidae	Passeriformes
Pachycare flavogriseum	8.30E-02	Pachycephalidae	Passeriformes
Alcedo pusilla	8.29E-02	Alcedinidae	Coraciiformes
Touit dilectissimus	8.29E-02	Psittacidae	Psittaciformes
Todiramphus lazuli	8.29E-02	Alcedinidae	Coraciiformes
Agapornis lilianae	8.27E-02	Psittacidae	Psittaciformes
Cossypha dichroa	8.26E-02	Muscicapidae	Passeriformes
Pseudeos fuscata	8.25E-02	Psittacidae	Psittaciformes
Neopsittacus musschenbroekii	8.24E-02	Psittacidae	Psittaciformes
Aratinga pertinax	8.24E-02	Psittacidae	Psittaciformes
Psitteuteles iris	8.23E-02	Psittacidae	Psittaciformes
Tanysiptera riedelii	8.23E-02	Alcedinidae	Coraciiformes
Todus todus	8.23E-02	Todidae	Coraciiformes

Pyrrhura calliptera	8.21E-02	Psittacidae	Psittaciformes
Cyclopsitta gulielmitertii	8.20E-02	Psittacidae	Psittaciformes
Aratinga nana	8.19E-02	Psittacidae	Psittaciformes
Amazona ventralis	8.18E-02	Psittacidae	Psittaciformes
Halcyon coromanda	8.17E-02	Alcedinidae	Coraciiformes
Parula gutturalis	8.16E-02	Parulidae	Passeriformes
Ptilorrhoa castanonota	8.16E-02	Eupetidae	Passeriformes

6.3 Chapter 4 supplementary material

6.3.1 Data

A diagram of the data class is shown in Supplementary Figure 6.3.3a. An object of this data class should have at least one image and use the image name as the index. One image can have multiple annotations (points or segmentations) and specimen characteristics. Annotation names are used to index annotations. The absence is used to define whether the part that is supposed to be labelled is missing (e.g. a broken snail shell does not have the shell tip). X and y coordinates are stored as the point location. Contours are used to define one segmentation, and a contour is defined as a list of x and y coordinates that represent contour vertices (see Supplementary Figure 6.3.3b). The area inside a contour should be segmented. While the intersected area between contours should remain non-segmented.

6.3.2 Metrics

Pixel distance of point i between ground truth and predicted points is defined as the Euclidean distance between two points, $d = \sqrt{(x_i' - x_i)^2 + (y_i' - y_i)^2}$, where x_i' and y_i' are the x and y coordinates of predicted point i, x_i and y_i are the x and y coordinates of the ground truth point i.

IOU of class i is defined as

$$IOU_i = \frac{p_{ii}}{p_{ii} + p_{ij} + p_{ji}}$$

precision of class i is defined as

$$Precision_i = \frac{p_{ii}}{p_{ii} + p_{ji}}$$

recall of class i is defined as

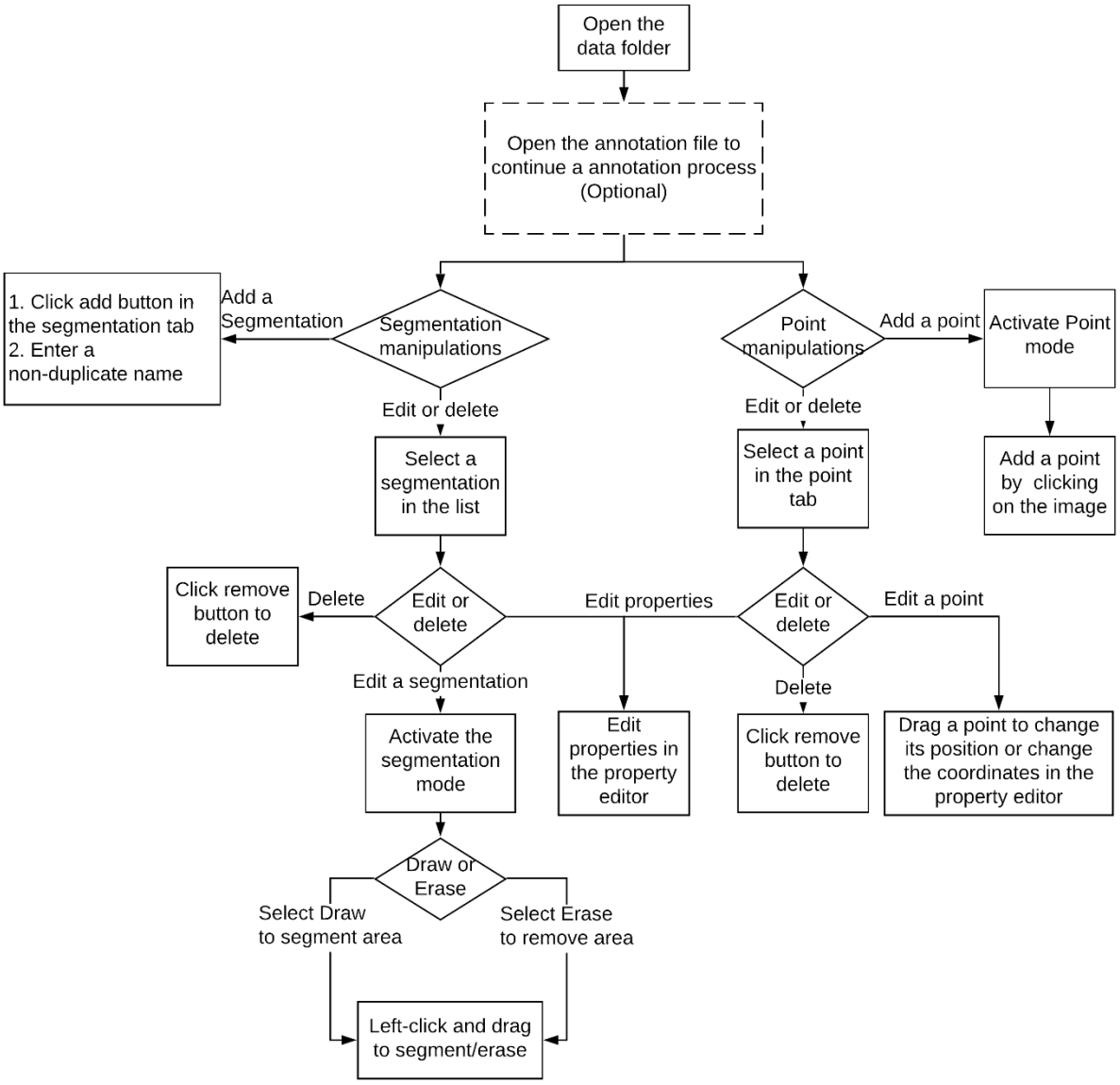
$$Recall_i = \frac{p_{ii}}{p_{ii} + p_{ij}}$$

and mIOU is defined as

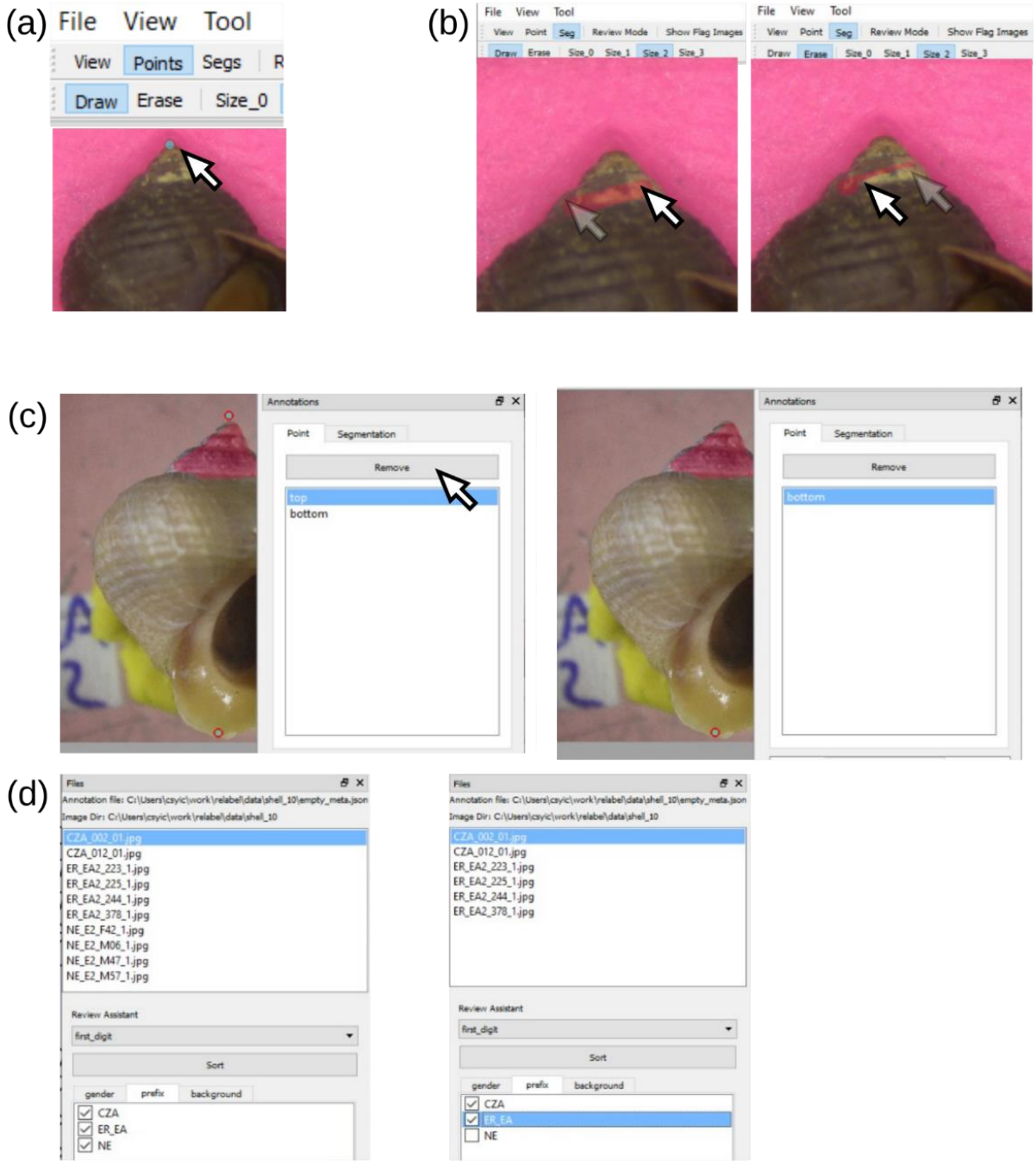
$$mIOU = \frac{\sum_{i=1}^{n} \frac{p_{ii}}{p_{ii} + p_{ij} + p_{ji}}}{n}$$

Where n is the number of the class of the segmentation (there are two classes for segmenting a specimen from a photo, specimen and background), i is one of the segmentation class, p_{ii} are pixels of class i and classified as class i (true positive); p_{ij} are pixels of class i but classified as other classes (false negative); and p_{ji} are pixels of others classes classified as class i (false positive).

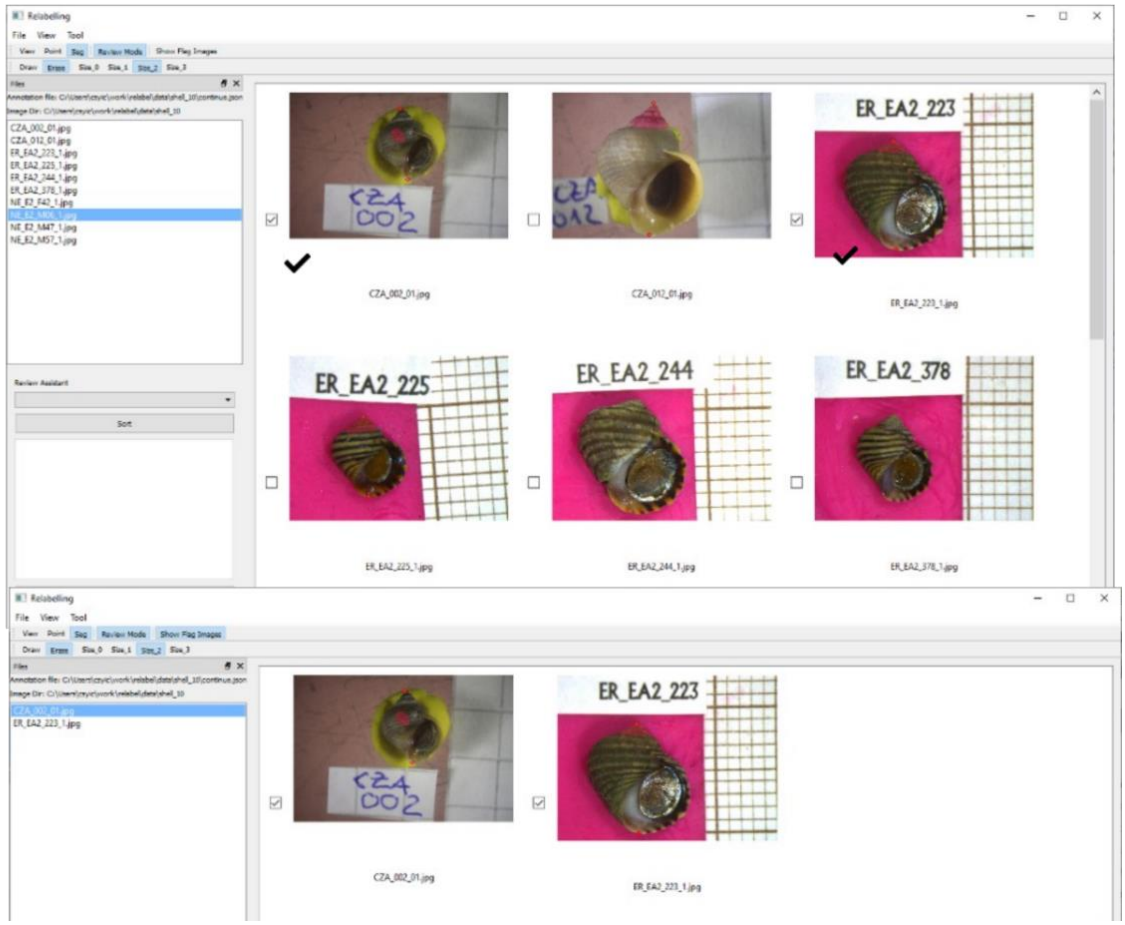
6.3.3 Supplementary Figures



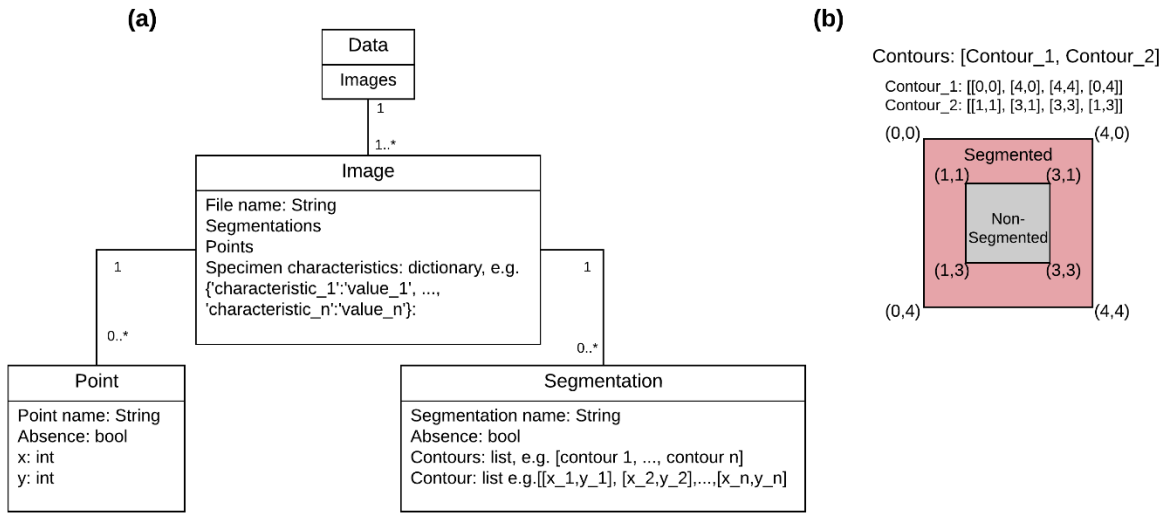
Supplementary Figure 6.3.1. The flow chart of the annotation process. It explains steps from opening the folder to handling annotations (points and segmentations).


(a)
File View Tool
View Points Segs R
Draw Erase Size_0
(b)
File View Tool
View Point Seg Review Mode Show Flag Images
Draw Erase Size_0 Size_1 Size_2 Size_3
(c)
Annotations
Point Segmentation
Remove
top
bottom
(d)
Files
Review Assistant
first_digit
Sort
gender prefix background
CZA
ER_EA
NE

(e)

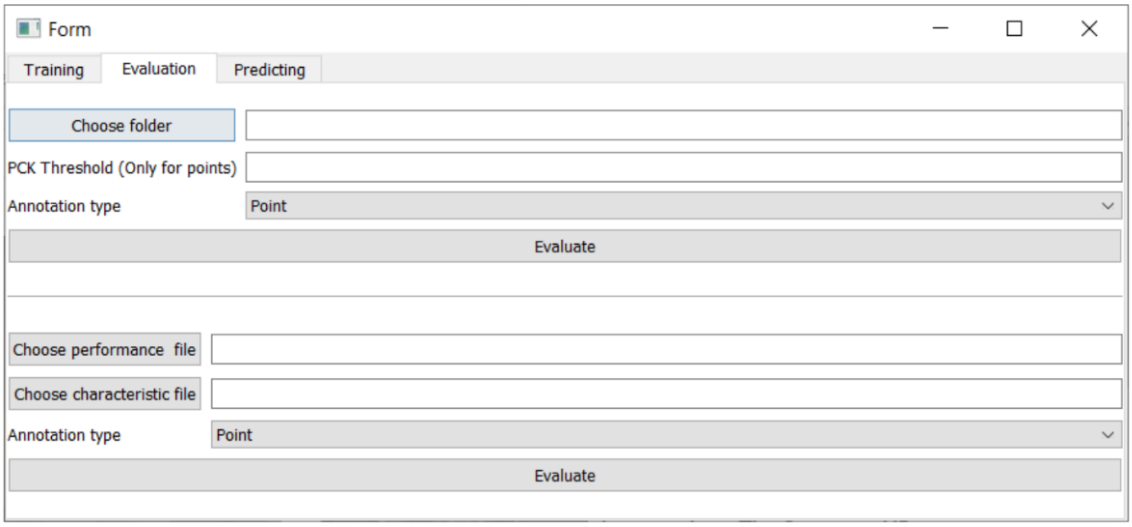


Supplementary Figure 6.3.2. Examples of (a) placing a point on top of the snail when point mode is activated; (b) Left: clicking and dragging to paint the segmentation areas when segmentation mode is activated and draw is selected. Right: removing segmentation areas when erase is selected; (c) Deleting a point; (d) Using review assistant to remove images with the prefix of 'NE'; (e) top: ticking two images in the review mode. bottom: displaying ticked images when Show Flag Images is activated.

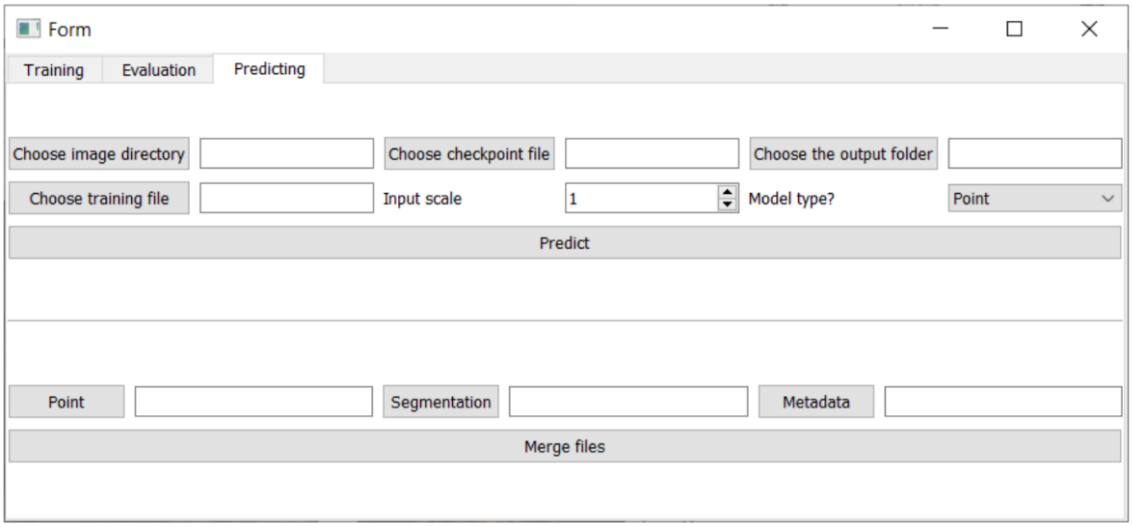


Supplementary Figure 6.3.3. (a) A diagram of the data class. One data class has at least one image, and one image can have multiple points or segmentations. (b) An example of how contours define a segmentation. The intersected square between contour_1 and contour_2 has not been segmented (grey). Rest areas are inside contour_1 and are not intersected areas, so they are segmented (pink).

(a)

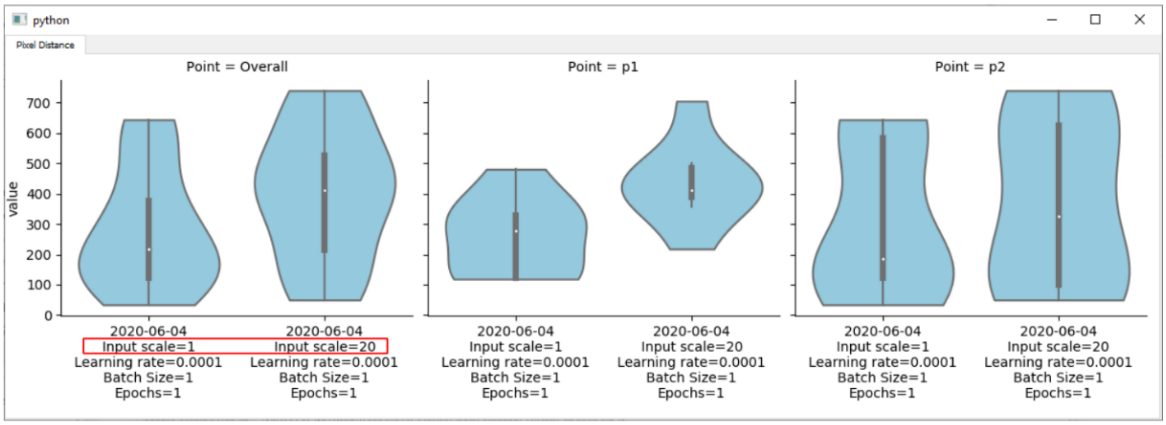


(b)

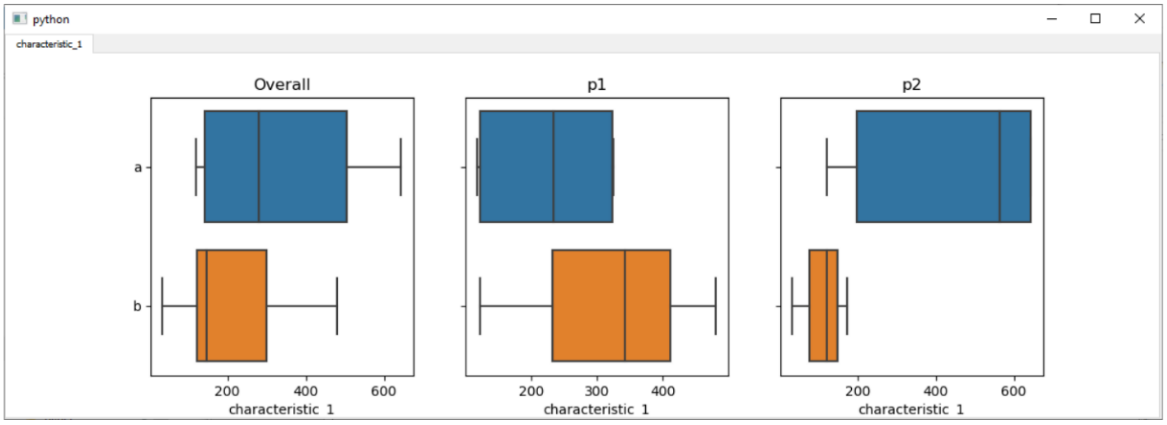


Supplementary Figure 6.3.4. User interfaces of the deep learning tool. (a) The evaluation sub-module; (b) The predicting sub-module.

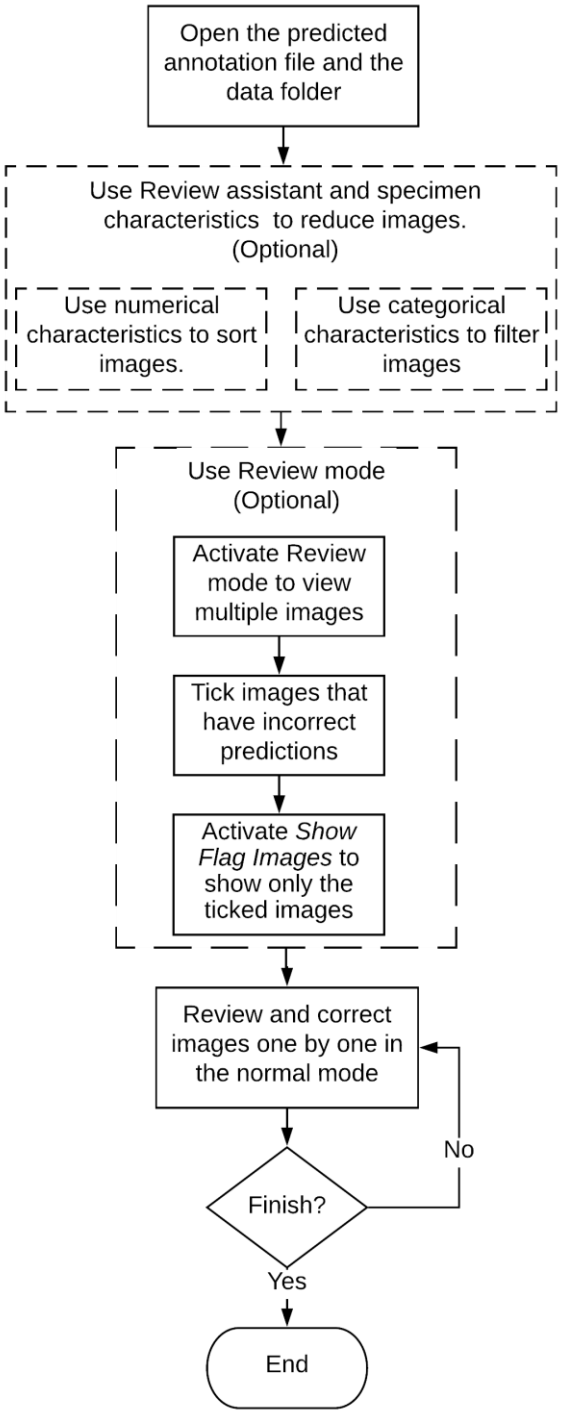
(a)



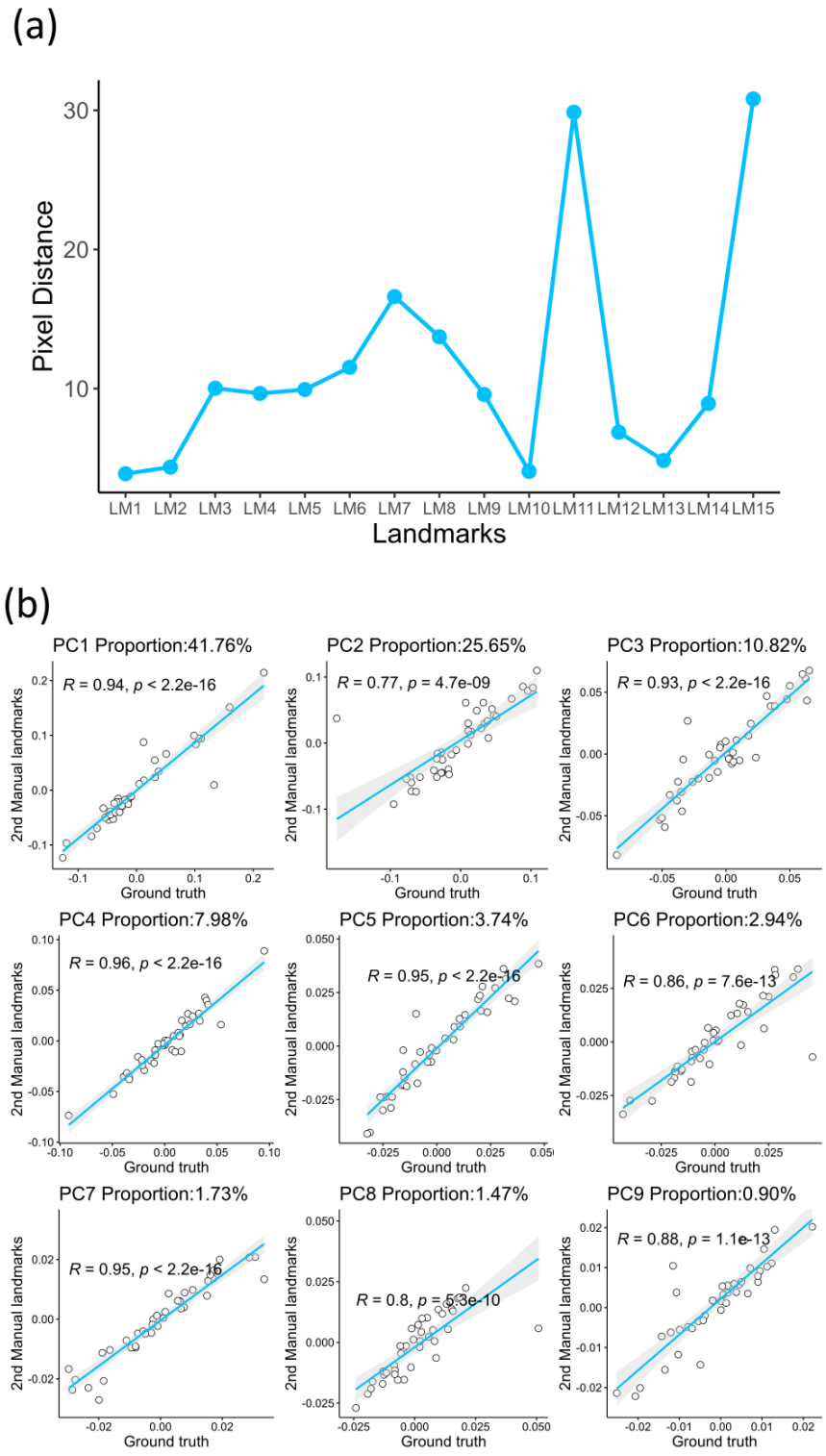
(b)



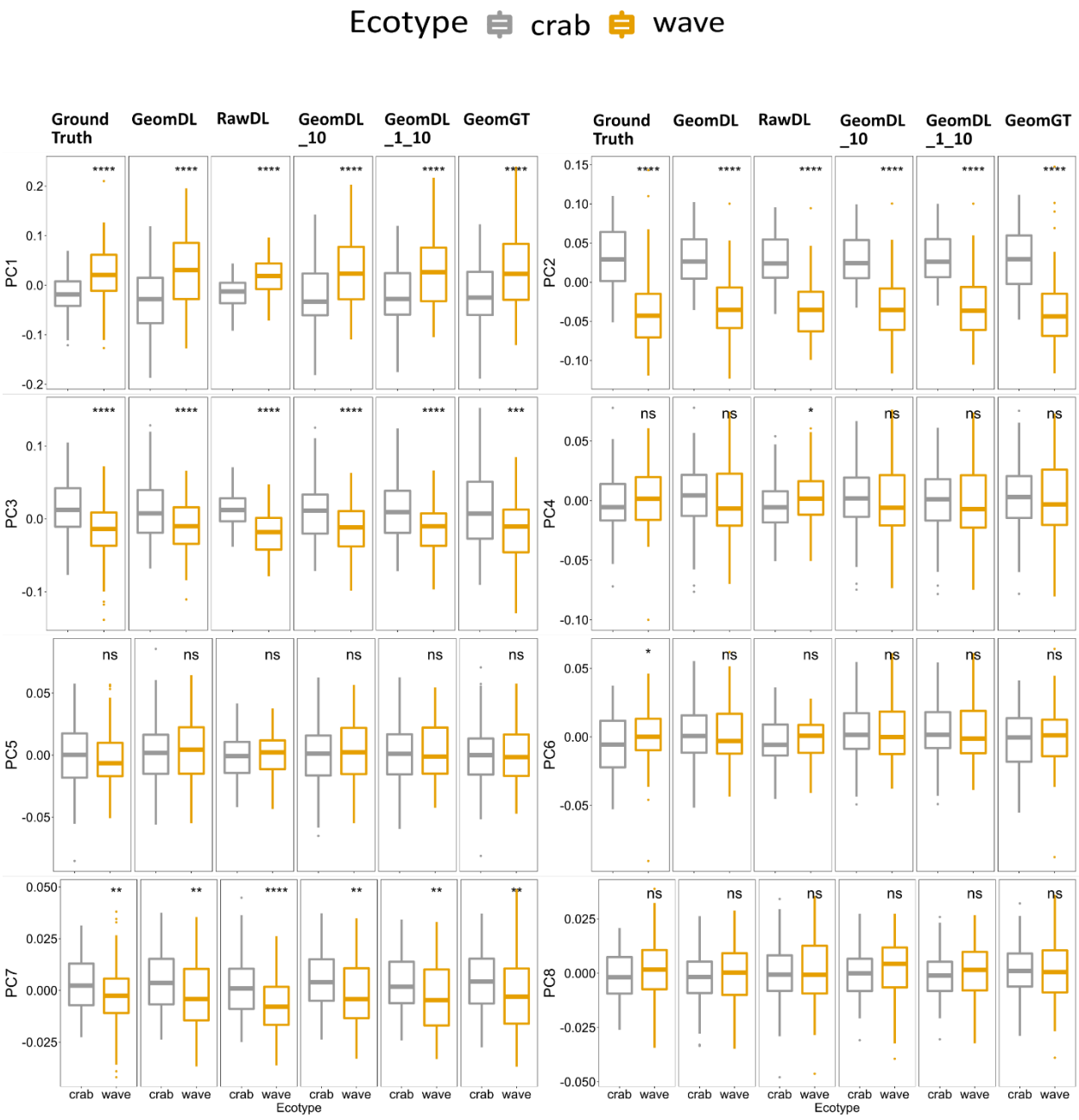
Supplementary Figure 6.3.5. Examples of evaluating results. (a) Comparing the pixel distances of all and each individual points (p1 and p2) from two configurations (only different in input resolution scale. Left uses the scale of 1 and right uses the scale of 20). And pixel distances from using the scale of 1 is smaller, suggesting it is the better configuration. (b) Relations between pixel distances of all and each individual points (p1 and p2) from two categories (blue plots are category a and yellow plots are category b). The result shows that images from category b were predicted more accurate than those from category a.



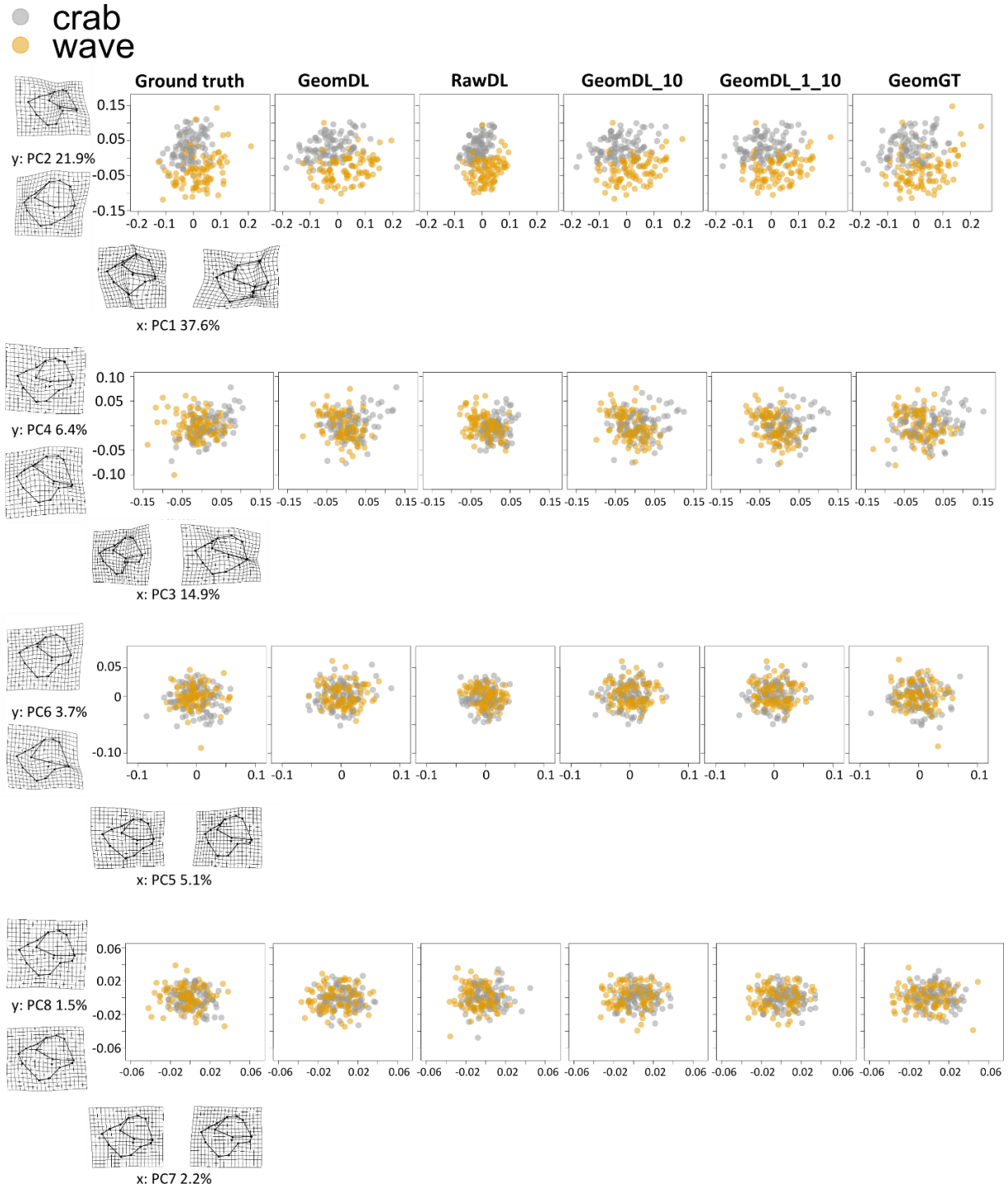
Supplementary Figure 6.3.6. The flow chart of the review process. It shows how to use review mode and the review assistant to increase the reviewing efficient.



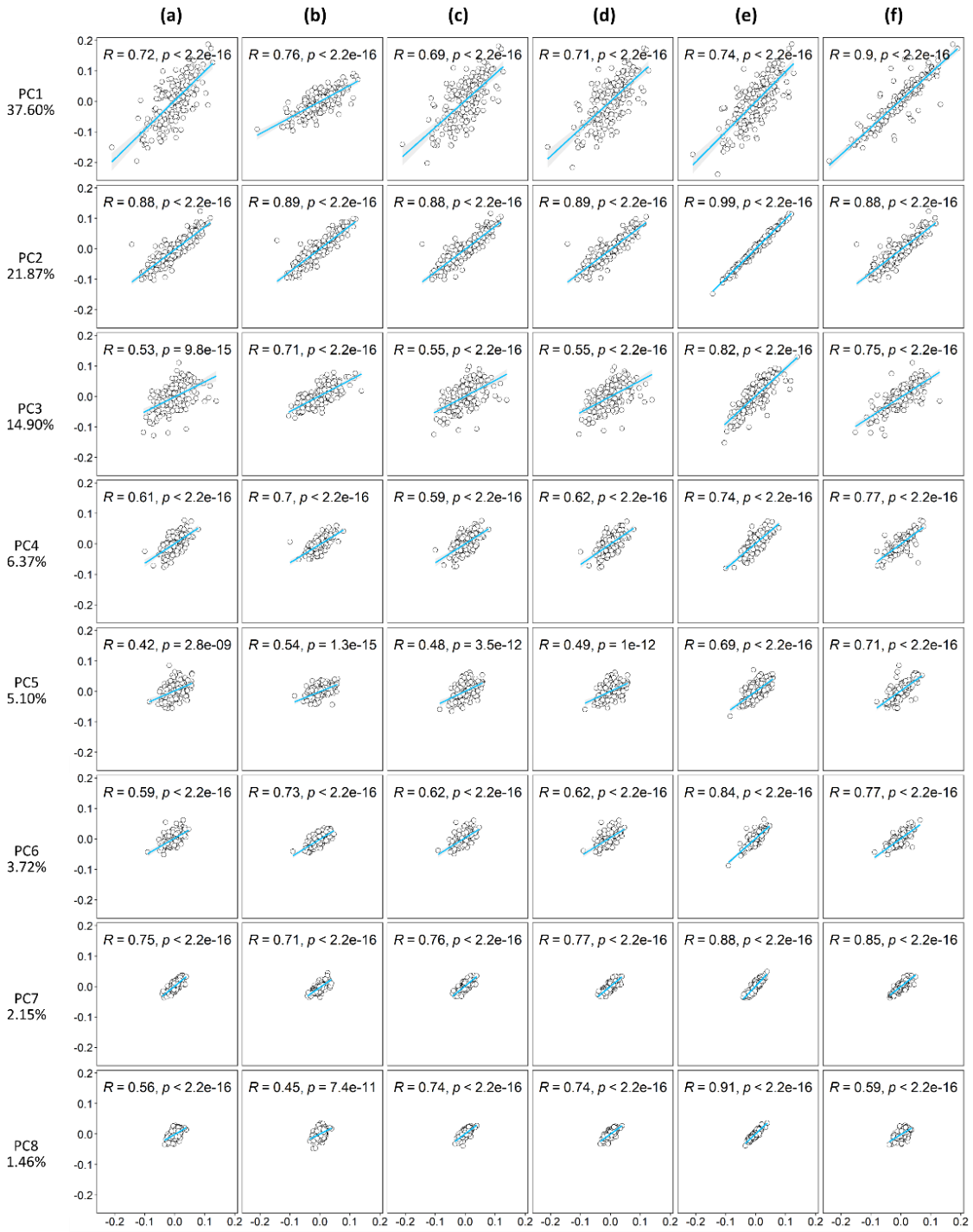
Supplementary Figure 6.3.7. Comparisons between the ground truth and the manually re-labelled landmarks (40 images). (a) The average pixel distances for individual landmarks. (b) Correlations of PC1-9 axes.



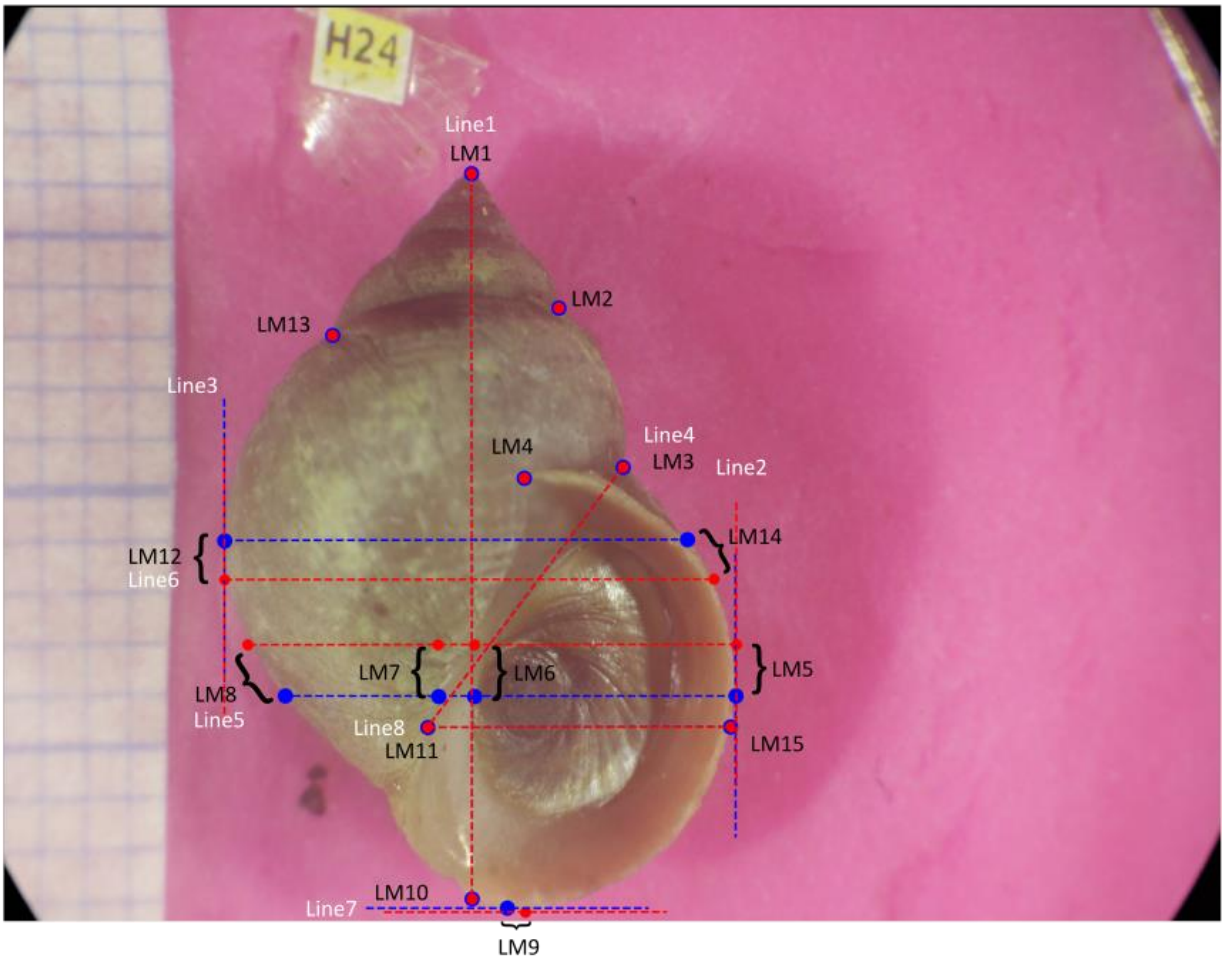
Supplementary Figure 6.3.8. Boxplots of crab (grey, N=100) and wave (yellow, N=88) PC1-8 from the ground truth and all tested results. Significant symbols are t-test results between the crab ecotype and the wave ecotype (ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001).



Supplementary Figure 6.3.9. Distributions of PC1-2, PC3-4, PC5-6 and PC7-8 from the ground truth and all tested results (N=188). Grey points are crab ecotype (N=100); yellow points are wave ecotype (N=88).



Supplementary Figure 6.3.10. Correlations of PC1-8 between two results. The comparison groups are listed column-wise and they are: (a) The ground truth vs GeomDL; (b) The ground truth vs RawDL; (c) The ground truth vs GeomDL_10; (d) The ground truth vs GeomDL_1_10; (e) The ground truth vs GeomGT and (f) GeomGT vs GeomDL.



Supplementary Figure 6.3.11. An example of ground truth landmarks and reference lines (blue) and processed ground truth (GeomGT) landmarks and reference lines (red). Independent landmarks have the same coordinates. Dependent landmarks are placed in different locations.

6.3.4 Supplementary Tables

Supplementary Table 6.3.1. Hyperparameters that are tuneable in the training process.

Hyperparameters	
Input resolution downscale	Due to the limitation of the graphics processing unit memory, deep learning networks can not take images with excessively large resolutions. Images sometimes needed to be downscaled. The resolution fed into the network is $\frac{Width}{scale} \times \frac{Height}{scale}$. A scale of 1 means the original resolution.
Learning rate	The learning rate controls the degree of the optimization of network parameters. A small learning rate can lead to a slow learning process, while a large learning rate may not find a good optimal in the parameter space as it updates parameters too much. It is important for users to look at the loss changes in the log file and to decide suitable learning rates for their datasets.
Batch size	Batch size stands for how many images are fed into the network for each training step. Batch sizes smaller than 32 are commonly used in Stacked Hourglass and DeepLab (Newell et al. 2016; Chen et al. 2017b). I used batch sizes of 4 for both networks on Project Plumage (See Chapter 2 and Chapter 3).
Training epochs	An epoch is defined as one pass of the full training set for the network. Training epochs defined the length of the training more straightforward than training steps. $\text{Training steps} = \frac{\text{Epochs} \times \text{Training set size}}{\text{Batch size}}$.There is no an one-for-all epoch number. I recommend to start with large training epochs and see when the loss converges.
Training/validation set split	The split between the training set and validation set. A number (between 1 and 100) is entered to defined the proportion of the training set and the proportion of the validation set (the sum of them should be 100). Common choices are 80/20, 70/30 and 60/40.

Supplementary Table 6.3.2. Files used as inputs in PhenoLearn.

Files	
Annotation file	An annotation file provides image names and annotation information. The file can either be a JSON file from PhenoLearn or a CSV spreadsheet. The structure of the CSV spreadsheet. Point: Columns: file, <name of the point 1>_x, <name of the point 1>_y, …, <name of the point n>_x, <name of the point n>_y Segmentation: Columns: file, <name of the segmentation 1>, …,<name of the segmentation n>
Metadata file	A metadata file is a CSV spread sheet that contains image names and specimen characteristics. Columns: file, < characteristic 1>, …, < characteristic n>

Supplementary Table 6.3.3. Files generated in the training and predicting process.

Files	
Validation result file	A validation result saves predicted annotations of validation images. One CSV and one JSON prediction file is generated after the training process. Both CSV and JSON files have the same structures as the ones from the annotation file in Supplementary Table 6.3.2. File location: *<output directory>/result_<date>_<configurations>.json* *<output directory>/result_<date>_<configurations>.csv*
Performance file	A performance file is a file that quantifies the performance of the model on the validation set and is generated after the training process. It is a spreadsheet that contains evaluation metrics for every validation images. Point: Columns: file, <name of the point 1>, …, <name of the point n> Each row contains the name of the file and the pixel distances of corresponding points. Segmentation: Columns: file, iou_<name of the segmentation 1>, precision_<name of the segmentation 1>, recall_<name of the segmentation 1>, …, iou_<name of the segmentation n>, precision_<name of the segmentation n>, recall_<name of the segmentation n> Each row A spreadsheet with names of validation images and evaluation results of their annotations (i.e. pixel distances for point predictions; IOU, precision and recall for segmentation predictions). The performance file is used in the evaluation module to visualise the performance of this configuration. File location: *<output directory>/performance_<date>_<configurations>.csv*
Log file	Tensorflow logs the training metrics (e.g. the training loss) during the training process in this file. It can be visualized using Tensorboard (See manual for detail). File location: *<output directory>/log/<date>_<configurations>/*
Checkpoint file	Tensorflow saves the parameters of a trained network in this file. The checkpoint file can be loaded when predicting the whole dataset. File location: *<output directory>/checkpoint/<date>_<configurations>*

Prediction result file	A prediction file has predicted annotations for images used in the prediction. One CSV and one JSON prediction file is generated after prediction. Both CSV and JSON files have the same structures as the ones from the annotation file in Supplementary Table 6.3.2. File location: *<output directory>/pred_<date>_<configurations>.json* *<output directory>/ pred_<date>_<configurations>.csv*

Supplementary Table 6.3.4. Landmarks placed on *Littorina* shells

Landmarks	Definitions
LM1	The apex of the shell
LM2	The upper suture of the penultimate whorl (right)
LM3	The lower suture of the penultimate whorl (right)
LM4	The end of the suture
LM5	The point of intersection between Line2 and the contour of the shell (right)
LM6	The point of intersection between Line5 and edge of the lip (right)
LM7	The point of intersection between Line5 and edge of the lip (left)
LM8	The point of intersection between Line5 and the contour of the shell (left)
LM9	The point of intersection between Line7 and the contour of the shell (bottom)
LM10	The point of intersection between Line1 and the contour of the shell (bottom)
LM11	The point of intersection between Line4 and the external edge of the lip
LM12	The point of intersection between Line3 and the contour of the shell (left)
LM13	The upper suture of the penultimate whorl (left)
LM14	The point of intersection between Line6 and the contour of the shell (right)
LM15	The point of intersection between Line8 and the contour of the shell (right)

Supplementary Table 6.3.5. Reference lines placed on *Littorina* shells

Lines	Definitions
Line1	The line starts at the apex of the shell (LM1) and is tangent to the inner margin of the aperture on the left.
Line2	The line which is parallel to Line1 (parallel line) and is tangent to the shell (right)
Line3	The line which is parallel to Line1 (parallel line) and is tangent to the shell (left)
Line4	The line starts at the terminal whorl meets the edge of the outer lip of the aperture (LM3) and is the tangent line of the left operculum.
Line5	The line which passes LM 5 and is perpendicular to Line1
Line6	The line which passes LM 12 and is perpendicular to Line1.
Line7	The line which is perpendicular to Line1 and is tangent to the shell (bottom).
Line8	The line which passes LM 11 and is perpendicular to Line1.

Supplementary Table 6.3.6. PCKht@0.1 and PCKht@0.05 of all and each individual landmarks for results of (i) the ground truth vs GeomDL, (ii) the ground truth vs RawDL and (iii) GeomGT vs GeomDL.

	Ground truth vs GeomDL		**Ground truth vs RawDL**		**GeomGT vs GeomDL**	
	PCKht@0.05	**PCKht@0.1**	**PCKht@0.05**	**PCKht@0.1**	**PCKht@0.05**	**PCKht@0.1**
Overall	85.2	98.3	90.0	98.7	93.0	98.5
LM1	99.5	100	100	100	99.5	100
LM2	100	100	100	100	100	100
LM3	98.9	99.5	99.5	99.5	98.9	99.5
LM4	97.9	99.5	97.9	99.5	97.9	99.5
LM5	87.8	99.5	92	100	93.1	98.4
LM6	81.4	99.5	87.2	98.9	93.6	98.9
LM7	75	98.9	82.4	97.9	89.9	98.9
LM8	69.7	96.8	79.3	95.7	93.1	97.3
LM9	92	99.5	100	100	96.3	98.9
LM10	98.9	100	98.9	100	98.9	100
LM11	73.4	96.8	73.4	96.8	73.4	96.8
LM12	77.1	97.3	96.3	100	94.1	96.8
LM13	99.5	100	100	100	99.5	100
LM14	58	91	78.2	98.4	93.6	96.3
LM15	68.6	96.3	64.4	93.1	72.9	96.3

6.3.5 Algorithms

Algorithm 1. Generating dependent landmarks from independent landmarks on *Littorina* shell images

```
### Algorithm ###
## Independent landmarks optimisation
# Iterate through independent LM predictions that are on the shell outline
# Move them to the outline
for LM in [LM1, LM2, LM3, LM10, LM13]:
    LM = Find_nearest_location_on_outline(shell_outline, LM)

## Dependent landmarks and reference lines
# Calculate Line1 using LM1 and LM10
Line1 = Calculate_line(LM1, LM10)

# Line1 and shell outline are used to calculate LM5, LM12, Line2 and Line3
Line2, LM5 = Find_parallel_line_is_tangent_to_outline(Line1, shell_outline, Position = 'left')
Line3, LM12 = Find_parallel_line_is_tangent_to_outline(Line1, shell_outline, Position =
'right')

# Line1, LM5, LM12 and LM11 are used to calculate Line5, Line6 and Line8
# Then Line5, Line6, Line8 and the shell outline are used to calculated LM8, LM14, LM15, Line7
and LM9
Line5 = Find_perpendicular_line_passes_a_point(Line1, LM5)
Line6 = Find_perpendicular_line_passes_a_point(Line1, LM12)
Line8 = Find_perpendicular_line_passes_a_point(Line1, LM11)
LM8 = Find_interstion_between_line_and_outline(shell_outline, Line5)
LM14 = Find_interstion_between_line_and_outline(shell_outline, Line6)
LM15 = Find_interstion_between_line_and_outline(shell_outline, Line8)
Line7, LM9 = Find_parallel_line_is_tangent_to_outline(Line5, shell_outline, Position = 'bot-
tom')

## Semi-dependent landmarks
# LM6 and LM7 are moved to Line5
LM6 = Find_nearest_location_on_Line(Line5, LM6)
LM7 = Find_nearest_location_on_Line(Line5, LM7)

### Functions ###
Find_nearest_location_on_outline(Outline, Point)
# Returns:
##   a point that is on Outline and has the shortest distance to Point

Calculate_line(Point1, Point2)
# Returns:
##   a line that passes Point1 and Point2

Find_parallel_line_is_tangent_to_outline(Line, Outline, Position=['top','bot-
tom','right','left'])
# Returns:
##   A line that is parallel to Line and tangent to Outline. Position specifies which tangent
line to return
##   The intersection point between the returned line and Outline

Find_perpendicular_line_passes_a_point(Line, Point)
# Returns:
##   A line that is perpendicular to Line and passes Point

Find_interstion_between_line_and_outline(Outline, Line)
# Returns:
##   a point that is on both Line and Outline.

Find_nearest_location_on_Line(Line, Point)
# Returns:
##   a point that is on Line and has the shortest distance to Point
```

6.4 Chapter 4 software manual

6.4.1 Environments

Python 3.7

Packages:

PyQt5 5.15.0

numpy 1.18

pandas 1.0

opencv-python 4.2.0

tensorflow 1.15

tensorflow-gpu 1.15 (if GPU available)

6.4.2 Annotation

To start a new project, use the ***Menu->File->open image directory*** to open the directory of training set images. All images (supported formats: JPG, PNG, TIF) in the folder are listed in the file panel. Images can be selected in the file panel, and the selected image will be displayed in the visualisation panel (Figure 4.2a.v). Users can examine and zoom images (hold Ctrl and scroll mouse wheel, or click zoom functions in *Menu->View*) in the visualisation panel. The status bar shows the coordinates and pixel value of the mouse on the image. The default mode is view mode (mode selection is located in the mode bar, see Figure 4.2a.ii), users can only view images under this mode.

To add points on an image, point mode needs to be activated first. Points then can be added by left-clicking on the image (Supplementary Figure 6.3.2a). A point name is entered in a pop-up dialogue box, the name should not be duplicated to other point names in the current image. Existed points are listed in the point tab (Figure 4.2a.iv.i). When a point is selected, its properties (e.g. point name and coordinates) are shown in the property editor (Figure 4.2a.vi.iii). Users can edit point in the property editor, or simply click a point in the image and drag it to change its

location. A point can be removed by selecting it and click the *remove* button in the point tab as shown in Supplementary Figure 6.3.2c.

A segmentation can be added using the *add* button in the segmentation tab (Figure 4.2a.vi.ii). A segmentation I referred here is a segmentation class rather than a connected segmented region, therefore a segmentation can have multiple unconnected segmented regions. One colour is allocated to one segmentation for displaying its segmented regions. The palette has eight colours, if one image has more than eight segmentation, segmentations may share colours. Users need to type a name in a pop-up dialogue box to finish adding a new segmentation, the segmentation is then listed in the segmentation tab along with other existed segmentations in the current image. A newly added segmentation does not have any segmented regions. To segment images, users first need to activated *mode bar->segmentation* and *tool bar->Draw*. Similar to paint programs, regions can be segmented by clicking and dragging the mouse on the image (Supplementary Figure 6.3.2b). The size of the painter can be selected in the *tool bar* (see Figure 4.2a.iii). Regions can be removed following the same instructions as segmenting but with *tool bar->Erase* activated. The *Auto fill* function solves the inefficiency of segmenting a large area by drawing. Users can first draw a closed contour of the focal region, and click *Auto fill* (located in the tool bar) to segment the whole region by filling areas inside the contour. A segmentation can be removed using the *Remove* button in the segmentation tab. Supplementary Figure 6.3.1 shows a flow chart of steps of placing points and segmenting images using PhenoLearn.

Many projects apply the same list of annotations to every image, entering annotation names for all images is trivial and time-consuming. The Quick Label mode allows users to label images without naming them for every image. Quick label can be activated when all required annotations have been all placed on the current image. Annotations of the current image are saved as the annotation templated and are shown in in the annotation panel as the instruction. After Quick Label is activated, instead of placing and naming a point, a point will be named following the template. Segmentations in the template will be automatically added for all images, and users then can add segmented regions to them manually.

The annotations are saved as a JSON file, which contains image information (e.g. names) and annotations (see section 6.3.1 and Supplementary Figure 6.3.3 for detail). PhenoLearn can read a saved annotation file and load annotations. Therefore users can save an incomplete annotation process, and continue the annotation process by opening the previously saved file and the image directory.

CSV is a file format supported by many software tools used in scientific analyses (e.g. R, Python, MATLAB). PhenoLearn also supports CSV import and export, which increases the compatibility of annotations from PhenoLearn. When importing a CSV file, columns needed to be assigned into the file, point, segmentation, property or unwanted columns. Annotations are then read into PhenoLearn after assigning columns. When exporting annotations to a CSV file. A column named *file* is used to save file names. To avoid duplicate or ambiguous column names, prefixes are added (*pt_* is added to point names, *seg_* is added to segmentation names and *prop_* is added to specimen characteristics). For example, x and y coordinates of point LM1 are stored in columns named *pt_LM1_x* and *pt_LM1_y*. Values of segmentation S1 are stored in a column named *seg_S1*. The species information is stored in a column named *prop_species*.

6.4.3 Deep Learning

The deep learning module provides three form-like sub-modules that help users to set up model training, evaluation, and predicting. The training module enables users to set up network hyperparameters (e.g. input resolution scale), I/O (annotation file, input and output directory) and the network (i.e. for points or segmentations) for the training (Figure 4.2b). Hyperparameter tuning can be very complex because of many hyperparameters are involved in training. To simplify it, PhenoLearn only allows users to change some of the key hyperparameters (see Supplementary Table 6.3.1 for detail of these hyperparameters).

A training process generates a log file at the beginning and constantly updates during the training. When training finishes, three more files are generated which are result, performance, and checkpoint (See Supplementary Table 6.3.3 for detail of these files). Performance and result files are saved as CSV files, users can import these files easily using their preferable tools such as R, Python or MATLAB.

The training normally takes a few days. A progress bar is displayed during the training, and the software should remain open until the training finishes. Users can monitor the training process using the Tensorboard (www.tensorflow.org/tensorboard), a model visualisation tool from Tensorflow (Abadi et al. 2016). By typing ***tensorboard --logdir=<location of the log file> --host localhost*** in the terminal, and entering http://localhost:6006 in a web browser, training metrics such as the relation between the training loss and the training time can be visualized in the web browser.

The evaluation module compares performances from different trained configurations by visualising performances across configurations (see Supplementary Figure 6.3.5b). For point predictions, plots of overall and each individual point pixel distance and PCK with a user-defined threshold are displayed. Mean and per-segmentation class IOU, precision and recall are plotted for segmentation predictions. Users can then select the best configuration based on evaluation metrics.

If there is a metadata file that provides specimen characteristics (see Supplementary Table 6.3.2 for the structure of the metadata file), it can be used in the evaluation module to visualise and quantify relations between prediction accuracies and characteristics. Characteristics that affect performance greatly can be useful in reviewing predictions of the whole dataset (See section 4.2.4 review). A scatter plot of performance against characteristic values is displayed for numerical characteristics (e.g. body mass or length). For categorical characteristics (e.g. taxonomic ranks or ecological factors), performances of different categories are displayed in boxplots. Examples are shown in Supplementary Figure 6.3.5c.

The prediction module reads the image folder, checkpoint file, output folder, the annotation file used in training, input resolution scale and the network type (i.e. for points or segmentations) as shown in Supplementary Figure 6.3.4b. The checkpoint file, annotation file, input resolution scale and the network type are used to create the deep learning network and restore its parameters for the prediction. Annotations of images inside the image folder are predicted. During the prediction, a progress bar shows the predicting progress. When it completes, the predicted annotations are saved as a CSV and a JSON file in the output folder (See Supplementary Table

6.3.3 for detail of the prediction file). Prediction files can be opened directly in PhenoLearn. The merge function can add specimen characteristics from the metadata file into the prediction file. Also, it can merge point and segmentation files together if a project has predictions for both annotations.

6.4.4 Review

Opening the predicted annotation file and the image folder begins the review process. Review can be done by scrolling through image files to check that annotations are incorrect, as shown in the Supplementary Figure 6.3.6. Incorrect annotations can then be edited. Reviewing all images one by one can be time-consuming, especially since deep learning models normally produce high-quality predictions. A way to optimise the trade-off between result precision and time is to review images with a priori expected higher error rates. Users can reduce the number of images to be reviewed using the relation between accuracy and the specimen characteristics. If there are specimen characteristics in the annotation file, they will be shown in the review assistant (Figure 4.2a.iv.ii). Numerical characteristics (e.g. body length) can be used to sort images in the file panel. Categorical characteristics (e.g. taxonomic ranks) can be used to reduce images by unticking unwanted categories in the review assistant as shown in Supplementary Figure 6.3.2d.

Review mode in mode bar can be activated to increase the efficiency of the review process. The Review mode displays multiple thumbnails and annotations in the visualisation panel as shown in Figure 4.2c. Users can tick images with incorrect annotations to flag them. *Mode Bar->Show Flag Images* removes unselected images (Supplementary Figure 6.3.2e), users can then deactivate Review mode and correct the flagged images (i.e. images with incorrect labels).

Reference

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. CoRR abs/1603.0.

Adams, D. C., and E. Otárola-Castillo. 2013. Geomorph: An r package for the collection and analysis of geometric morphometric shape data. Methods Ecol. Evol. 4:393–399.

Adams, D., J. L. Rohlf, and D. Slice. 2013. A field comes of age: geometric morphometrics in the 21 st century. Hystrix, Ital. J. Mammal. 24:7–14.

Adams, R., and L. Bischof. 1994. Seeded Region Growing. IEEE Trans. Pattern Anal. Mach. Intell. 16:641–647.

Al-Qizwini, M., I. Barjasteh, H. Al-Qassab, and H. Radha. 2017. Deep learning algorithm for autonomous driving using GoogLeNet. Pp. 89–96 *in* IEEE Intelligent Vehicles Symposium, Proceedings.

Aljabar, P., R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert. 2009. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. Neuroimage 46:726–738. Elsevier Inc.

Andriluka, M., L. Pishchulin, P. Gehler, and B. Schiele. 2014. 2D human pose estimation: New benchmark and state of the art analysis. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 3686–3693.

Araus, J. L., and J. E. Cairns. 2014. Field high-throughput phenotyping: The new crop breeding frontier. Trends Plant Sci. 19:52–61.

Ariño, A. 2010. Approaches to estimating the universe of natural history collections data.

Biodivers. Informatics 7:81–92.

Baiker, M., J. Milles, J. Dijkstra, T. D. Henning, A. W. Weber, I. Que, E. L. Kaijzel, C. W. G. M. Lwik, J. H. C. Reiber, and B. P. F. Lelieveldt. 2010. Atlas-based whole-body segmentation of mice from low-contrast Micro-CT data. Med. Image Anal. 14:723–737. Elsevier B.V.

Balafar, M. A., A. R. Ramli, M. I. Saripan, and S. Mashohor. 2010. Review of brain MRI image segmentation methods. Artif. Intell. Rev. 33:261–274.

Basu, S., S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani. 2015. DeepSat - A learning framework for satellite imagery. GIS Proc. ACM Int. Symp. Adv. Geogr. Inf. Syst. 03-06-Nove.

Benton, M. J. 2015. Exploring macroevolution using modern and fossil data. Proc. R. Soc. B Biol. Sci. 282:20150569.

Berzins, V. 1984. Accuracy of Laplacian edge detectors. Comput. Vision, Graph. Image Process. 27:195–210.

Bilder, R. M., F. Sabb, T. D. Cannon, E. D. London, J. D. Jentsch, S. Parker, R. a Poldrack, C. Evans, and N. B. Freimer. 2009. Phenomics: The systematic study of phenotypes on a genome- wide scale. Neuroscience 164:30–42.

Bishop, C. M. 2007. Pattern Recognition and Machine Learning_old_version.

Blagoderov, V., I. J. Kitching, L. Livermore, T. J. Simonsen, and V. S. Smith. 2012. No specimen left behind: Industrial scale digitization of natural history collections. Zookeys 209:133–146.

Blagoderov, V., I. Kitching, T. Simonsen, and V. Smith. 2010. Report on trial of SatScan tray scanner system by SmartDrive Ltd. Nat. Preced., doi: 10.1038/npre.2010.4486.1.

Bookstein, F. L. 1991. Morphometric tools for landmark data: geometry and biology.

Boykov, Y. Y., and M.-P. Jolly. 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. Pp. 105–112 *in* Proceedings eighth IEEE international conference on computer vision. ICCV 2001.

Bradski, G. 2000. The OpenCV Library. Dr. Dobb's J. Softw. Tools.

Bradski, G., and A. Kaehler. 2008. Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc."

Bulat, A., and G. Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). , doi: 10.1109/ICCV.2017.116.

Buser, T. J., B. L. Sidlauskas, and A. P. Summers. 2018. 2D or Not 2D? Testing the Utility of 2D Vs. 3D Landmark Data in Geometric Morphometrics of the Sculpin Subfamily Oligocottinae (Pisces; Cottoidea). Anat. Rec. 301:806–818.

Butlin, R. K., M. Saura, G. Charrier, B. Jackson, C. André, A. Caballero, J. A. Coyne, J. Galindo, J. W. Grahame, J. Hollander, and others. 2014. Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. Evolution (N. Y). 68:935–949. Wiley Online Library.

Cabezas, M., A. Oliver, X. Lladó, J. Freixenet, and M. Bach Cuadra. 2011. A review of atlas-based segmentation for magnetic resonance brain images. Comput. Methods Programs Biomed. 104:e158–e177. Elsevier Ireland Ltd.

Chan, T. F., and L. A. Vese. 2001. Active contours without edges. IEEE Trans. Image Process. 10:266–277.

Chang, J., and M. E. Alfaro. 2016. Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. Methods Ecol. Evol. 7:472–482.

Chang, Y. L., and X. Li. 1994. Adaptive Image Region-Growing. IEEE Trans. Image Process. 3:868–872.

Chen, J.-C., V. M. Patel, and R. Chellappa. 2016. Unconstrained face verification using deep cnn features. Pp. 1–9 *in* 2016 IEEE winter conference on applications of computer vision (WACV).

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv Prepr. arXiv1412.7062.

Chen, L.-C., G. Papandreou, F. Schroff, and H. Adam. 2017a. Rethinking atrous convolution for semantic image segmentation. arXiv Prepr. arXiv1706.05587.

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. Pp. 801–818 *in* Proceedings of the European conference on computer vision (ECCV).

Chen, L. C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2017b. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40:834–848.

Cherabier, I., C. Hane, M. R. Oswald, and M. Pollefeys. 2016. Multi-label semantic 3D reconstruction using voxel blocks. Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016 601–610.

Cholewo, T. J., and S. Love. 1999. Gamut boundary determination using alpha-shapes. Final Progr. Proc. - IS T/SID Color Imaging Conf. 200–204.

Conde-Padín, P., A. Caballero, and E. Rolán-Alvarez. 2009. Relative role of genetic determination and plastic response during ontogeny for shell-shape traits subjected to diversifying selection. Evolution (N. Y). 63:1356–1363.

Cooney, C. R., J. A. Bright, E. J. R. Capp, A. M. Chira, E. C. Hughes, C. J. A. Moody, L. O. Nouri, Z. K. Varley, and G. H. Thomas. 2017. Mega-evolutionary dynamics of the adaptive radiation of birds. Nature 542:344–347. Nature Publishing Group.

Cooney, C. R., Z. K. Varley, L. O. Nouri, C. J. A. Moody, M. D. Jardine, and G. H. Thomas. 2019. Sexual selection predicts the rate and direction of colour divergence in a large avian radiation. Nat. Commun. 10. Springer US.

Cootes, T., E. Baldock, and J. Graham. 2000. An introduction to active shape models. Image Process. Anal. 223–248.

Cootes, T. F., G. J. Edwards, and C. J. Taylor. 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23:681–685.

Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and

B. Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. Pp. 3213–3223 *in* Proceedings of the IEEE conference on computer vision and pattern recognition.

Cuthill, I. C., J. C. Partridge, A. T. D. Bennett, S. C. Church, N. S. Hart, and S. Hunt. 2000. Ultraviolet vision in birds. Pp. 159–214 *in* Advances in the Study of Behavior. Elsevier.

Dale, J., C. J. Dey, K. Delhey, B. Kempenaers, and M. Valcu. 2015. The effects of life history and sexual selection on male and female plumage colouration. Nature 527:367–370. Nature Publishing Group.

Dalrymple, R. L., D. J. Kemp, H. Flores-Moreno, S. W. Laffan, T. E. White, F. A. Hemmings, M. L. Tindall, and A. T. Moles. 2015. Birds, butterflies and flowers in the tropics are not more colourful than those at higher latitudes. Glob. Ecol. Biogeogr. 24:1424–1432.

Davies, T. G., I. A. Rahman, S. Lautenschlager, J. A. Cunningham, R. J. Asher, P. M. Barrett, K. T. Bates, S. Bengtson, R. B. J. Benson, D. M. Boyer, J. Braga, J. A. Bright, L. P. A. M. Claessens, P. G. Cox, X. P. Dong, A. R. Evans, P. L. Falkingham, M. Friedman, R. J. Garwood, A. Goswami, J. R. Hutchinson, N. S. Jeffery, Z. Johanson, R. Lebrun, C. Martínez-Pérez, J. Marugán-Lobón, P. M. O'Higgins, B. Metscher, M. Orliac, T. B. Rowe, M. Rücklin, M. R. Sánchez-Villagra, N. H. Shubin, S. Y. Smith, J. M. Starck, C. Stringer, A. P. Summers, M. D. Sutton, S. A. Walsh, V. Weisbecker, L. M. Witmer, S. Wroe, Z. Yin, E. J. Rayfield, and P. C. J. Donoghue. 2017. Open data and digital morphology. Proc. R. Soc. B Biol. Sci. 284.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. P. *in* CVPR09.

Denton, J. S. S., and D. C. Adams. 2015. A new phylogenetic test for comparing multiple high-dimensional evolutionary rates suggests interplay of evolutionary rates and modularity in lanternfishes (Myctophiformes; Myctophidae). Evolution (N. Y). 69:2425–2440.

Dodge, S., and L. Karam. 2016. Understanding how image quality affects deep neural networks. Pp. 1–6 *in* 2016 eighth international conference on quality of multimedia experience (QoMEX).

Dunn, P. O., J. K. Armenta, and L. A. Whittingham. 2015. Natural and sexual selection act on different axes of variation in avian plumage color. Sci. Adv. 1.

Dutta, A., and A. Zisserman. 2019. The VIA annotation software for images, audio and video. MM 2019 - Proc. 27th ACM Int. Conf. Multimed. 2276–2279.

Ebey Honeycutt, C., R. Plotnick, and F. Kenig. 2014. Breaking free from the matrix: Segmentation of fossil images. Palaeontol. Electron. 1–18.

Edelsbrunner, H., D. Kirkpatrick, and R. Seidel. 1983. On the shape of a set of points in the plane. IEEE Trans. Inf. theory 29:551–559. IEEE.

Edelsbrunner, H., and E. P. Mücke. 1994. Three-dimensional alpha shapes. ACM Trans. Graph. 13:43–72. ACM New York, NY, USA.

Endler, J. A., and P. W. Mielke. 2005. Comparing entire colour patterns as birds see them. Biol. J. Linn. Soc. 86:405–431.

Everingham, M., S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2015. The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. 111:98–136. Springer.

Ewen, J. G., P. Surai, R. Stradi, A. P. Møller, B. Vittorio, R. Griffiths, and D. P. Armstrong. 2006. Carotenoids, colour and conservation in an endangered passerine, the hihi or stitchbird (Notiomystis cincta). Anim. Conserv. 9:229–235.

Fan, J., D. K. Y. Yau, A. K. Elmagarmid, and W. G. Aref. 2001. Automatic image segmentation by integrating color-edge extraction and seeded region growing. IEEE Trans. Image Process. 10:1454–1466.

Farabet, C., C. Couprie, L. Najman, and Y. LeCun. 2012. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35:1915–1929. IEEE.

Felice, R. N., and A. Goswami. 2017. Developmental origins of mosaic evolution in the avian cranium. Proc. Natl. Acad. Sci. 115:201716437.

Felice, R. N., A. Watanabe, A. R. Cuff, M. Hanson, B. A. S. Bhullar, E. R. Rayfield, L. M. Witmer, M. A. Norell, and A. Goswami. 2020. Decelerated dinosaur skull evolution with the origin of birds. PLoS Biol. 18:e3000801.

Ferns, P. N., and S. A. Hinsley. 2004. Immaculate tits: Head plumage pattern as an indicator of quality in birds. Anim. Behav. 67:261–272.

Flemons, P., and P. Berents. 2012. Image based digitisation of entomology collections: Leveraging volunteers to increase digitization capacity. Zookeys 209:203–217.

Garnett, S. T., G. B. Ainsworth, and K. K. Zander. 2018. Are we choosing the right flagships? The bird species and traits australians find most attractive. PLoS One 13:1–17.

Gehan, M. A., N. Fahlgren, A. Abbasi, J. C. Berry, S. T. Callen, L. Chavez, A. N. Doust, M. J. Feldman, K. B. Gilbert, J. G. Hodge, J. S. Hoyer, A. Lin, S. Liu, C. Lizárraga, A. Lorence, M. Miller, E. Platon, M. Tessman, and T. Sax. 2017. PlantCV v2: Image analysis software for high-throughput plant phenotyping. PeerJ 2017:1–23.

Giacomini, G., D. Scaravelli, A. Herrel, A. Veneziano, D. Russo, R. P. Brown, and C. Meloro. 2019. 3D Photogrammetry of Bat Skulls: Perspectives for Macro-evolutionary Analyses. Evol. Biol. 46:249–259. Springer US.

Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Pp. 580–587 *in* Proceedings of the IEEE conference on computer vision and pattern recognition.

Gluckman, T. L., and G. C. Cardoso. 2009. A method to quantify the regularity of barred plumage patterns. Behav. Ecol. Sociobiol. 63:1837–1844.

Goldsmith, T. H. 1990. Optimization, Constraint, and History in the Evolution of Eyes. Q. Rev. Biol. 65:281–322. University of Chicago Press.

Gomes, A. C. R., M. D. Sorenson, and G. C. Cardoso. 2016. Speciation is associated with changing ornamentation rather than stronger sexual selection. Evolution (N. Y). 70:2823–2838.

Goswami, A. 2015. Phenome10K: a free online repository for 3-D scans of biological and

palaeontological specimens.

Gower, J. C. 1975. Generalized procrustes analysis. Psychometrika 40:33–51. Springer.

Gruson, H. 2020. Estimation of colour volumes as concave hypervolumes using α-shapes. Methods Ecol. Evol., doi: 10.1111/2041-210x.13398.

Günel, S., H. Rhodin, D. Morales, J. Campagnolo, and P. Fua. 2019. DeepFly3D : A deep learning-based approach for 3D limb and appendage tracking in tethered , adult Drosophila. 1–20.

Guo, K., D. Zou, and X. Chen. 2015. 3D mesh labeling via deep convolutional neural networks. ACM Trans. Graph. 35.

Haralick, R. M., S. R. Sternberg, and X. Zhuang. 1987. Image analysis using mathematical morphology. IEEE Trans. Pattern Anal. Mach. Intell. 532–550. IEEE.

Hariharan, B., P. Arbeláez, R. Girshick, and J. Malik. 2014. Simultaneous detection and segmentation. Pp. 297–312 *in* European Conference on Computer Vision.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. 770–778.

Heimann, T., and H. P. Meinzer. 2009. Statistical shape models for 3D medical image segmentation: A review. Med. Image Anal. 13:543–563. Elsevier B.V.

Henries, D. G., and R. Tashakkori. 2012. Extraction of leaves from herbarium images. IEEE Int. Conf. Electro Inf. Technol. 1–6. IEEE.

Hestness, J., S. Narang, N. Ardalani, G. F. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou. 2017. Deep Learning Scaling is Predictable, Empirically. CoRR abs/1712.0.

Hill, G. E., G. E. Hill, K. J. McGraw, J. Kevin, and others. 2006a. Bird coloration: function and evolution. Harvard University Press.

Hill, G. E., G. E. Hill, K. J. McGraw, J. Kevin, and others. 2006b. Bird coloration: mechanisms and measurements. Harvard University Press.

Hinton, G., N. Srivastava, and K. Swersky. 2012. Neural networks for machine learning lecture 6a

overview of mini-batch gradient descent. Cited on 14:8.

Hollander, J., and R. K. Butlin. 2010. The adaptive value of phenotypic plasticity in two ecotypes of a marine gastropod. BMC Evol. Biol. 10.

Holovachov, O., A. Zatushevsky, and I. Shydlovsky. 2014. Whole-Drawer Imaging of Entomological Collections: Benefits, Limitations and Alternative Applications. J. Conserv. Museum Stud. 12:1–13.

Houle, D., D. R. Govindaraju, and S. Omholt. 2010. Phenomics: the next challenge. Nat. Rev. Genet. 11:855–866. Nature Publishing Group.

Hudson, L. N., V. Blagoderov, A. Heaton, P. Holtzhausen, L. Livermore, B. W. Price, S. Van Der Walt, and V. S. Smith. 2015. Inselect: Automating the digitization of natural history collections. PLoS One 10:1–15.

Hussein, B. R., O. A. Malik, W.-H. Ong, and J. W. F. Slik. 2020. Semantic Segmentation of Herbarium Specimens Using Deep Learning Techniques. Pp. 321–330 *in* Computational Science and Technology. Springer.

Insafutdinov, E., L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. 2016. DeeperCut: {A} Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. CoRR abs/1605.0.

Ioffe, S., and C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd Int. Conf. Mach. Learn. ICML 2015 1:448–456.

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, and others. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science (80-.). 346:1320 LP – 1331. American Association for the Advancement of Science.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers. 2012. The global diversity of birds in space and time. Nature 491:444–448.

Johannesson, K., M. Panova, P. Kemppainen, C. André, E. Rolan-Alvarez, and R. K. Butlin. 2010. Repeated evolution of reproductive isolation in a marine snail: Unveiling mechanisms of

speciation. Philos. Trans. R. Soc. B Biol. Sci. 365:1735–1747.

Johnson, J. T., M. S. Hansen, I. Wu, L. J. Healy, C. R. Johnson, G. M. Jones, M. R. Capecchi, and C. Keller. 2006. Virtual histology of transgenic mouse embryos for high-throughput phenotyping. PLoS Genet. 2:471–477.

Johnson, S., and M. Everingham. 2011. Learning effective human pose estimation from inaccurate annotation. Pp. 1465–1472 *in* CVPR 2011.

Joulin, A., L. Van Der Maaten, A. Jabri, and N. Vasilache. 2016. Learning visual features from large weakly supervised data. Pp. 67–84 *in* European Conference on Computer Vision.

Kamar, E., S. Hacker, and E. Horvitz. 2012. Combining Human and Machine Intelligence in Large-scale Crowdsourcing. Pp. 467–474 *in* Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.

Karanth, K. U. 1995. Estimating tiger Panthera tigris populations from camera-trap data using capture-recapture models. Biol. Conserv. 71:333–338.

Kass, M., A. Witkin, and D. Terzopoulos. 1988. Snakes: Active contour models. Int. J. Comput. Vis. 1:321–331.

Kellenberger, B., D. Marcos, and D. Tuia. 2018. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. Remote Sens. Environ. 216:139–153. Elsevier.

Kellenberger, B., M. Volpi, and D. Tuia. 2017. Fast animal detection in UAV images using convolutional neural networks. Int. Geosci. Remote Sens. Symp. 2017-July:866–869.

Keshavan, A., J. D. Yeatman, and A. Rokem. 2019. Combining citizen science and deep learning to amplify expertise in neuroimaging. Front. Neuroinform. 13:1–13.

Khorrami, P., J. Wang, and T. Huang. 2012. Multiple Animal Species Detection Using Robust Principal Component Analysis and Large Displacement Optical Flow. Homepages.Inf.Ed.Ac.Uk.

Kim, J., J. Kwon Lee, and K. Mu Lee. 2016. Accurate Image Super-Resolution Using Very Deep Convolutional Networks.

Kingma, D. P., and J. Ba. 2014. Adam: {A} Method for Stochastic Optimization. CoRR abs/1412.6.

Klingenberg, C. P. 2011. MorphoJ: An integrated software package for geometric morphometrics. Mol. Ecol. Resour. 11:353–357.

Klingenberg, C. P., and N. A. Gidaszewski. 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Syst. Biol. 59:245–261.

Koenderink, J. J., and A. J. van Doorn. 1992. Surface shape and curvature scales. Image Vis. Comput. 10:557–564.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 1–9.

Kumar, Y. H. S., N. Manohar, and H. K. Chethan. 2015. Animal Classification System: A Block Based Approach. Procedia Comput. Sci. 45:336–343.

Kuzminsky, S. C., and M. S. Gardiner. 2012. Three-dimensional laser scanning: Potential uses for museum conservation and scientific research. J. Archaeol. Sci. 39:2744–2751. Elsevier Ltd.

Lazić, M. M., M. A. Carretero, J. Crnobrnja-Isailović, and A. Kaliontzopoulou. 2015. Effects of Environmental Disturbance on Phenotypic Variation: An Integrated Assessment of Canalization, Developmental Stability, Modularity, and Allometry in Lizard Head Shape. Am. Nat. 185:44–58.

LeCun, Y., B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. 1990. Handwritten digit recognition with a back-propagation network. Pp. 396–404 *in* Advances in neural information processing systems.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86:2278–2323.

Lee, S. H., C. S. Chan, P. Wilkin, and P. Remagnino. 2015. Deep-plant: Plant identification with

convolutional neural networks. Pp. 452–456 *in* 2015 IEEE International Conference on Image Processing (ICIP).

Li, X., H. Chen, X. Qi, Q. Dou, C. W. Fu, and P. A. Heng. 2018. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. IEEE Trans. Med. Imaging 37:2663–2674. IEEE.

Li, Y., J. Sun, C. K. Tang, and H. Y. Shum. 2004. Lazy snapping. ACM SIGGRAPH 2004 Pap. SIGGRAPH 2004 303–308.

Lin, T. Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. Microsoft COCO: Common objects in context. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 8693 LNCS:740–755.

Long, J., E. Shelhamer, and T. Darrell. 2015. Fully convolutional networks for semantic segmentation. Pp. 3431–3440 *in* Proceedings of the IEEE conference on computer vision and pattern recognition.

Loshchilov, I., and F. Hutter. 2016. {SGDR:} Stochastic Gradient Descent with Restarts. CoRR abs/1608.0.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. Pp. 1150–1157 *in* Proceedings of the seventh IEEE international conference on computer vision.

Lussier, Y. A., and Y. Liu. 2007. Computational Approaches to Phenotyping: High-Throughput Phenomics. Proc. Am. Thorac. Soc. 4:18–25.

Lustberg, T., J. van Soest, M. Gooding, D. Peressutti, P. Aljabar, J. van der Stoep, W. van Elmpt, and A. Dekker. 2018. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother. Oncol. 126:312–317. The Authors.

M. Pagel. 1999. Inferring historical patterns of biological evolution. Nature 401:877–884.

Maestri, R., L. R. Monteiro, R. Fornel, N. S. Upham, B. D. Patterson, and T. R. O. de Freitas. 2017. The ecology of a continental evolutionary radiation: Is the radiation of sigmodontine rodents adaptive? Evolution (N. Y). 71:610–632.

Maia, R., C. M. Eliason, P. P. Bitton, S. M. Doucet, and M. D. Shawkey. 2013. pavo: An R package for the analysis, visualization and organization of spectral data. Methods Ecol. Evol. 4:906–913.

Maia, R., H. Gruson, J. A. Endler, and T. E. White. 2019. pavo 2: New tools for the spectral and spatial analysis of colour in r. Methods Ecol. Evol. 10:1097–1107.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. MacKay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753. Nature Publishing Group.

Mantle, B. L., J. la Salle, and N. Fisher. 2012. Whole-drawer imaging for digital management and curation of a large entomological collection. Zookeys 209:147–163.

Marques, C. I. J., H. R. Batalha, and G. C. Cardoso. 2016. Signalling with a cryptic trait: The regularity of barred plumage in common waxbills. R. Soc. Open Sci. 3.

Mathis, A., P. Mamidanna, T. Abe, K. M. Cury, V. N. Murthy, M. W. Mathis, and M. Bethge. 2018a. Markerless tracking of user-defined features with deep learning. arXiv Prepr. arXiv1804.03142.

Mathis, A., P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. 2018b. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. 21. Springer US.

MATLAB. 2018. 9.7.0.1190202 (R2019b). The MathWorks Inc., Natick, Massachusetts.

Meijering, E. 2012. Cell segmentation: 50 Years down the road [life Sciences]. IEEE Signal Process. Mag. 29:140–145. IEEE.

Meyer, F., and S. Beucher. 1990. Morphological segmentation. J. Vis. Commun. Image Represent. 1:21–46. Elsevier.

Mharib, A. M., A. R. Ramli, S. Mashohor, and R. B. Mahmood. 2012. Survey on liver CT image segmentation methods. Artif. Intell. Rev. 37:83–95.

Milioto, A., P. Lottes, and C. Stachniss. 2018. Real-Time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. Proc. - IEEE Int. Conf. Robot. Autom. 2229–2235.

Miller, E. T., G. M. Leighton, B. G. Freeman, A. C. Lees, and R. A. Ligon. 2019. Ecological and geographical overlap drive plumage evolution and mimicry in woodpeckers. Nat. Commun. 10. Springer US.

Minervini, M., M. M. Abdelsamea, and S. A. Tsaftaris. 2014. Image-based plant phenotyping with incremental learning and active contours. Ecol. Inform. 23:35–48. Elsevier B.V.

Monteiro, L. R., J. A. F. Diniz-Filho, S. F. Dos Reis, and E. D. Araújo. 2002. Geometric estimates of heritability in biological shape. Evolution (N. Y). 56:563–572.

Nelson, G., D. Paul, G. Riccardi, and A. R. Mast. 2012. Five task clusters that enable efficient and effective digitization of biological collections. Zookeys 209:19–45.

Newell, A., K. Yang, and J. Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. Eur. Conf. Comput. Vis. 483–499.

Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. 115:E5716--E5725. National Acad Sciences.

NVIDIA. 2017. NVIDIA CUDA C, Programming Guide.

Otsu, N. 1979. Threshold Selection Method From Gray-Level Histograms. IEEE Trans Syst Man Cybern SMC-9:62–66.

Park, D. S., I. Breckheimer, A. C. Williams, E. Law, A. M. Ellison, and C. C. Davis. 2019. Herbarium specimens reveal substantial and unexpected variation in phenological sensitivity across the eastern United States. Philos. Trans. R. Soc. B Biol. Sci. 374.

Park, T., A. R. Evans, S. J. Gallagher, and E. M. G. Fitzgerald. 2017. Low-frequency hearing preceded the evolution of giant body size and filter feeding in baleen whales. Proc. R. Soc. B Biol. Sci. 284:9–11.

Pearson, K. D., G. Nelson, M. F. J. Aronson, P. Bonnet, L. Brenskelle, C. C. Davis, E. G. Denny, E. R. Ellwood, H. Goëau, J. Mason Heberling, A. Joly, T. Lorieul, S. J. Mazer, E. K. Meineke, B. J. Stucky, P. Sweeney, A. E. White, and P. S. Soltis. 2020. Machine learning using digitized herbarium specimens to advance phenological research. Bioscience 70:610–620.

Pereira, T. D., D. E. Aldarondo, L. Willmore, M. Kislin, S. S. H. S.-H. Wang, M. Murthy, and J. W. Shaevitz. 2019. Fast animal pose estimation using deep neural networks. Nat. Methods 16:117. Nature Publishing Group.

Pérez-Rodríguez, L., R. Jovani, and F. Mougeot. 2013. Fractal geometry of a complex plumage trait reveals bird's quality. Proc. R. Soc. B Biol. Sci. 280.

Pérez-Rodríguez, L., R. Jovani, and M. Stevens. 2017. Shape matters: Animal colour patterns as signals of individual quality. Proc. R. Soc. B Biol. Sci. 284.

Perez, L., and J. Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv Prepr. arXiv1712.04621.

Pigot, A. L., C. Sheard, E. T. Miller, T. P. Bregman, B. G. Freeman, U. Roll, N. Seddon, C. H. Trisos, B. C. Weeks, and J. A. Tobias. 2020. Macroevolutionary convergence connects morphological form to ecological function in birds. Nat. Ecol. Evol. 4:230–239. Springer US.

Pinheiro, P. O., and R. Collobert. 2014. Recurrent convolutional neural networks for scene labeling. 31st Int. Conf. Mach. Learn. ICML 2014 1:151–159.

Pishchulin, L., E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V Gehler, and B. Schiele. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. Pp. 4929–4937 *in* Proceedings of the IEEE conference on computer vision and pattern recognition.

Pohl, K. M., J. Fisher, W. E. L. Grimson, R. Kikinis, and W. M. Wells. 2006. A Bayesian model for joint segmentation and registration. Neuroimage 31:228–239.

Porto, A., and K. L. Voje. 2020. ML-morph: A fast, accurate and general approach for automated detection and landmarking of biological structures in images. Methods Ecol. Evol. 11:500–512.

Potena, C., D. Nardi, and A. Pretto. 2016. Fast and accurate crop and weed identification with summarized train sets for precision agriculture. Pp. 105–121 *in* International Conference on Intelligent Autonomous Systems.

Pound, M. P., J. A. Atkinson, A. J. Townsend, M. H. Wilson, M. Griffiths, A. S. Jackson, A. Bulat, G. Tzimiropoulos, D. M. Wells, E. H. Murchie, T. P. Pridmore, and A. P. French. 2017. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. Gigascience 6:1–10.

Price, J. J., and M. D. Eaton. 2014. Reconstructing the evolution of sexual dichromatism: Current color diversity does not reflect past rates of male and female change. Evolution (N. Y). 68:2026–2037.

Prum, R. O., J. S. Berv, A. Dornburg, D. J. Field, J. P. Townsend, E. M. Lemmon, and A. R. Lemmon. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

Ranjan, R., V. M. Patel, and R. Chellappa. 2017. HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. IEEE Trans. Pattern Anal. Mach. Intell. XX:1–16.

Rasband, W. S., and others. 1997. ImageJ. Bethesda, MD.

Rathi, D., S. Jain, and S. Indu. 2017. Underwater fish species classification using convolutional neural network and deep learning. Pp. 1–6 *in* 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR).

Ravinet, M., A. Westram, K. Johannesson, R. Butlin, C. André, and M. Panova. 2016. Shared and nonshared genomic divergence in parallel ecotypes of L ittorina saxatilis at a local scale. Mol. Ecol. 25:287–305. Wiley Online Library.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: Unified, real-time object detection. Pp. 779–788 *in* Proceedings of the IEEE conference on computer vision and pattern recognition.

Ren, S., K. He, R. Girshick, and J. Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Nips 91–99.

Renoult, J. P., A. Kelber, and H. M. Schaefer. 2017. Colour spaces in ecology and evolutionary biology. Biol. Rev. 92:292–315.

Revell, L. J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.

Roach, J. C., G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas. 2010. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science (80-.). 328:636 LP – 639.

Rohlf, F. J. 2006. tpsDig, version 2.10. http//life. bio. sunysb. edu/morph/index. html. Department of Ecology and Evolution, State University of New York at Stony Brook.

Ronneberger, O., P. Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. CoRR abs/1505.0.

Rother, C., V. Kolmogorov, and A. Blake. 2004. GrabCut - Interactive foreground extraction using iterated graph cuts. ACM SIGGRAPH 2004 Pap. SIGGRAPH 2004 309–314.

Rubner, Y., C. Tomasi, and L. J. Guibas. 2000. Earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. 40:99–121.

Ruder, S. 2016. An overview of gradient descent optimization algorithms. arXiv Prepr. arXiv1609.04747.

Rusinkiewicz, S. 2004. Estimating Curvature and Their Derivatives on Triangle Meshes. Symp. 3D Data Process. Vis. Transm. Sept. 2004 1–8.

Saito, S., T. Li, and H. Li. 2016. Real-Time Facial Segmentation and Performance Capture from RGB Input. Pp. 244–261 *in* B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. Computer Vision -- ECCV 2016. Springer International Publishing, Cham.

Scharr, H., M. Minervini, A. P. French, C. Klukas, D. M. Kramer, X. Liu, I. Luengo, J. M. Pape, G. Polder, D. Vukadinovic, X. Yin, and S. A. Tsaftaris. 2016. Leaf segmentation in plant phenotyping: a collation study. Mach. Vis. Appl. 27:585–606. Springer Berlin Heidelberg.

Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9:676–682.

Schlegl, T., S. M. Waldstein, H. Bogunovic, F. Endstraßer, A. Sadeghipour, A. M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth. 2018. Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning. Ophthalmology 125:549–558. American Academy of Ophthalmology.

Schneider, S., G. W. Taylor, and S. Kremer. 2018. Deep learning object detection methods for ecological camera trap data. Proc. - 2018 15th Conf. Comput. Robot Vision, CRV 2018 321–328. IEEE.

Schork, N. J. 1997. Genetics of complex disease: Approaches, problems, and solutions. P. *in* American Journal of Respiratory and Critical Care Medicine.

Sezgin, M., and B. Sankur. 2004. Survey over image thresholding techniques and quantitative performance evaluation. J. Electron. Imaging 13:146–165. SPIE.

Shapira, L., A. Shamir, and D. Cohen-Or. 2008. Consistent mesh partitioning and skeletonisation using the shape diameter function. Vis. Comput. 24:249–259.

Siam, M., S. Elkerdawy, M. Jagersand, and S. Yogamani. 2018. Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC 2018-March:1–8.

Simonyan, K., and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv Prepr. arXiv1409.1556 1–14.

Slagsvold, T., S. Dale, and A. Kruszewicz. 1995. Predation favours cryptic coloration in breeding male pied flycatchers. Anim. Behav. 50:1109–1121. Elsevier.

Soltis, P. S., G. Nelson, A. Zare, and E. K. Meineke. 2020. Plants meet machines: Prospects in machine learning for plant biology. Appl. Plant Sci. 8:1–6.

Stevens, M., C. A. Párraga, I. C. Cuthill, J. C. Partridge, and T. S. Troscianko. 2007. Using digital photography to study animal coloration. Biol. J. Linn. Soc. 90:211–237. Oxford University Press.

Stevens, M., J. Troscianko, J. K. Wilson-Aggarwal, and C. N. Spottiswoode. 2017. Improvement of individual camouflage through background choice in ground-nesting birds. Nat. Ecol. Evol. 1:1325–1333. Springer US.

Stoddard, M. C., and R. O. Prum. 2008. Evolution of avian plumage color in a tetrahedral color space: A phylogenetic analysis of new world buntings. Am. Nat. 171:755–776.

Stoddard, M. C., and R. O. Prum. 2011. How colorful are birds? Evolution of the avian plumage color gamut. Behav. Ecol. 22:1042–1052.

Stoddard, M. C., and M. Stevens. 2010. Pattern mimicry of host eggs by the common cuckoo, as seen through a bird's eye. Proc. R. Soc. B Biol. Sci. 277:1387–1393.

Sun, C., A. Shrivastava, S. Singh, and A. Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. Proc. IEEE Int. Conf. Comput. Vis. 2017-Octob:843–852.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2014. Going Deeper with Convolutions. arXiv:1409.4842, doi: 10.1109/CVPR.2015.7298594.

Treml, M., J. Arjona-medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich, B. Nessler, and S. Hochreiter. 2016. Speeding up Semantic Segmentation for Autonomous Driving. NIPS 2016 Work. MLITS 1–7.

Troscianko, J., and M. Stevens. 2015. Image calibration and analysis toolbox - a free software suite for objectively measuring reflectance, colour and pattern. Methods Ecol. Evol. 6:1320–1331.

Troscianko, J., J. Wilson-Aggarwal, M. Stevens, and C. N. Spottiswoode. 2016. Camouflage predicts survival in ground-nesting birds. Sci. Rep. 6:1–8. Nature Publishing Group.

Tuset, V. M., M. Farré, A. Lombarte, F. Bordes, R. Wienerroither, and P. Olivar. 2014. A comparative study of morphospace occupation of mesopelagic fish assemblages from the Canary Islands (North-eastern Atlantic). Ichthyol. Res. 61:152–158.

Tzutalin. 2015. LabelImg.

Unger, J., D. Merhof, and S. Renner. 2016. Computer vision applied to herbarium specimens of German trees: Testing the future utility of the millions of herbarium specimen images for automated identification. BMC Evol. Biol. 16:1–7. BMC Evolutionary Biology.

Van Belleghem, S. M., P. A. Alicea Roman, H. C. Gutierrez, B. A. Counterman, and R. Papa. 2020. Perfect mimicry between Heliconius butterflies is constrained by genetics and development. bioRxiv 2020.01.10.902494.

Van Belleghem, S. M., R. Papa, H. Ortiz-Zuazaga, F. Hendrickx, C. D. Jiggins, W. Owen McMillan, and B. A. Counterman. 2018. patternize: An R package for quantifying colour pattern variation. Methods Ecol. Evol. 9:390–398.

van den Oever, J. P., and M. Gofferjé. 2012. “From pilot to production”: Large scale digitisation project at naturalis biodiversity center. Zookeys 209:87–92.

van Vliet, L. J., I. T. Young, and G. L. Beckers. 1989. A nonlinear laplace operator as edge detector in noisy images. Comput. Vision, Graph. Image Process. 45:167–195.

Vincent, L., and P. Soille. 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 583–598. IEEE.

Wegge, P., C. P. Pokheral, and S. R. Jnawali. 2004. Effects of trapping effort and trap shyness on estimates of tiger abundance from camera trap studies. Anim. Conserv. 7:251–256.

Wei, S.-E., V. Ramakrishna, T. Kanade, and Y. Sheikh. 2016. Convolutional pose machines. Pp. 4724–4732 *in* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Whitwell, J. L. 2009. Voxel-based morphometry: An automated technique for assessing structural changes in the brain. J. Neurosci. 29:9661–9664.

Wickham, H. 2011. Ggplot2. Wiley Interdiscip. Rev. Comput. Stat. 3:180–185.

Willink, B., A. García-Rodríguez, F. Bolaños, and H. Pröhl. 2014. The interplay between multiple predators and prey colour divergence. Biol. J. Linn. Soc. 113:580–589.

Xing, F., and L. Yang. 2016. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. IEEE Rev. Biomed. Eng. 9:234–263. IEEE.

Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T. Y. Lam. 2017. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol. Evol. 8:28–36.

Zelditch, M. L., J. Li, L. A. P. Tran, and D. L. Swiderski. 2015. Relationships of diversity, disparity, and their evolutionary rates in squirrels (Sciuridae). Evolution (N. Y). 69:1284–1300.

Zhang, G., C. Li, Q. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, A. Antunes, M. J. Greenwold, R. W. Meredith, and others. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science (80-.). 346:1311–1320. American Association for the Advancement of Science.