Anatomical Landmark Localisation and Uncertainty Estimation

University of Sheffield

Lawrence Schöbs

Supervisor: Prof. Haiping Lu

This dissertation is submitted for the degree of
Doctor of Philosophy

in the

Department of Computer Science

September, 2023

Declaration

All sentences or passages quoted in this thesis from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this thesis have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in the degree examination as a whole.

Name: Lawrence Schöbs

Signature: Lawrence Schöbs

Date: September 1, 2023

Acknowledgements

First and foremost, I extend my deepest gratitude to Georgina Grace Lea for her unconditional, unwavering love and support. Her kindness, selflessness, and stoicism have served as a continual source of inspiration, teaching me the values of compassion, hard work and patience. With you, Georgina, the past years have not only been a chapter in my academic journey but the most fulfilling, meaningful, and beautiful era of my life.

I am incredibly grateful to Prof. Haiping Lu, whose guidance and wisdom have been invaluable. He has not merely been an advisor but a mentor in the truest sense, nurturing not just my research skills but also my mental resilience. Thank you for your boundless patience and encouragement. Equally, I express my heartfelt appreciation to Dr. Joab Winkler for his guidance, and to Dr. Andrew J. Swift and Dr. Samer Alabed for their valued collaborations.

Special thanks are extended to my labmates and friends: Dr. Chao Han, Rea Nkhumise, Pawel Pukowski, Xianyuan Liu, Thomas Baldwin-McDonald, Dr. Shuo Zhou, Sina Tabakhi, Dr. Juan José Giraldo, Peizhen Bai, Mohammod N.I. Suvon, Dr. Prasun C. Tripathi, Dr. Chunchao Ma, and so many more. These extraordinary individuals have transformed our lab into more than a workplace; they have made it a community, and coming to the lab has been a daily joy because of them.

With warm regard, I extend my appreciation to the remarkable individuals who have shaped my character and enriched my life beyond measure. To Matthew Horton, Brogan James, Balraj Johal, Anjali Manoj, Kinjal Meswani, James Ribbens, and Owyn Welch: you are cherished.

Finally, to my parents, sister and family: your faith in me and wholehearted support for every step I take are the foundation upon which all of my achievements are built. Your love and belief in me have been ceaseless, and for that, I am infinitely grateful.

Abstract

Machine learning promises transformative applications in medical image analysis. However, the black-box nature of Deep Neural Networks and data sensitivity issues hinders their clinical deployment. Addressing these challenges necessitates the development of lightweight models suitable for local deployment, accompanied by improved methods for uncertainty estimation of model predictions. Such uncertainty estimation methods could flag potentially erroneous predictions for a human-in-the-loop to review. In this thesis, we tackle these challenges, specifically focusing on the task of landmark localisation, a supervised task that involves identifying precise coordinates of anatomical structures within medical images.

Our first approach introduces PHD-Net, a lightweight, patch-based landmark localisation model that estimates prediction uncertainty heuristically. We experimentally show our approach performs exceptionally given its size and scales well with model capacity, offering an alternative perspective to landmark localisation with a unique uncertainty estimation property. Building on this foundational concept of uncertainty, we broaden its applicability to a wider range of landmark localisation models through the introduction of the Frequentist-inspired *Quantile Binning* framework. Our approach is general, applicable to any regression problem. Recognising the limitations of relying solely on localisation accuracy to holistically evaluate our models, we introduce evaluation metrics specifically designed for assessing binning-based uncertainty measures, enabling better model uncertainty estimation benchmarking. In our final work, we present the first application of Gaussian Processes to anatomical landmark localisation, achieving genuine Bayesian uncertainty.

Underpinning the impact of our research is a commitment to open-source accessibility. All our tools and innovations are made publicly available on Github within the low-code/no-code framework of *MediMarker*, or the *PyKale* library.

Contents

List of Figures

List of Tables

Symbols and Notations

Numbers and Arrays

a	A scalar (integer or real)
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$\boldsymbol{A}$	A tensor
A	A constant
$\mathcal{A}$	A Space
$\boldsymbol{I}$	Identity matrix with dimensionality implied by context

Sets, Probability and Spaces

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{E}$	Expectation
$\mathbb{P}$	Probability

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
$A_{i,j}$	Element i, j of matrix $\boldsymbol{A}$
$\mathbf{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\mathbf{A}_{:,i}$	Column i of matrix $\mathbf{A}$
$\boldsymbol{A}_{i,j,k}$	Element (i, j, k) of a 3 -D tensor $\boldsymbol{A}$
$\boldsymbol{A}_{:,:,i}$	2-D slice of a 3 -D tensor

Datasets

L_i	A landmark object, indexed by i
$\mathbf{a}$	A data vector, where $a^{(i)}$ refers to a_i
$\mathbf{A}$	An $m \times n$ data matrix of m vectors, where $\mathbf{a}^{(i)}$ refers to $\mathbf{A}_{i,:}$
$\boldsymbol{A}$	A $k \times m \times \ldots \times n$ tensor of k data tensors/matrices, where $\mathbf{A}^{(i)}$ refers to $\boldsymbol{A}_{i,:,\ldots,:}$
$\mathbf{x}$	An input vector
$\mathbf{y}$	A label/output vector
$\tilde{\mathbf{c}}$	A coordinate label vector
$\hat{\mathbf{c}}$	A coordinate prediction vector
$\hat{u}$	A scalar representing a predicted continuous uncertainty value

Operations and Functions

$k(\cdot, \cdot)$	Kernel function, e.g., linear, Gaussian, polynomial
$\mathrm{O}(\cdot)$	Computational complexity
$\cdot$	Dot product operation
$\odot$	The element-wise multiplication operation
$\mathcal{L}$	Loss function l l
$\mathbf{x}^T$	Transpose of $\mathbf{x}$
$\circ$	A convolution operation
$\|\|A\|\|_F$	Frobenius norm of matrix A
$\|\mathbb{S}\|$	Size of set $\mathbb{S}$

Chapter 1

Introduction

1.1 Motivation

Deep Learning has become the status quo in medical image analysis research, promising a future of high-performing, low-cost solutions for healthcare. However, the transition from research to real-world applications has been slower than anticipated. The incredible ability for neural networks to extract complex, informative features that enable their ground-breaking performance comes at the cost of interpretability. Currently, Deep Neural Networks (DNNs) are black boxes; a series of simple arithmetic operations representing increasingly abstract concepts, quickly dissolving into computational static to a human observer.

Stakeholders in the healthcare sector are reluctant to rely on models whose predictions cannot be interpreted, a natural response considering the potentially life-threatening consequences an incorrect prediction may have. A component of model interpretability is the ability to recognise and accurately quantify when the model is unsure of its answer. Therefore, a model that offers a reliable uncertainty score beside its prediction is an attractive proposition, providing a more nuanced recommendation compared to a binary output. Such properties that enhance model explainability and transparency have been recommended by the UK government [Joshi and Morley, 2019] as well as international organisations like the World Health Organisation [WHO, 2021], which aim to speed up adoption of these machine learning techniques into healthcare systems.

A common task that is performed by radiologists is landmark localisation, a process which

entails pinpointing the coordinates of specific anatomical structures of interest in images. These landmarks can be used for downstream tasks such as image registration [Han et al., 2014; Johnson and Christensen, 2002; Miao et al., 2012; Murphy et al., 2011], image segmentation [Beichel et al., 2005], and the derivation of surgical or diagnostic measures [Al et al., 2018; Bier et al., 2019; Kasel et al., 2013; Torosdagli et al., 2018; Vrtovec et al., 2009; Wang et al., 2016]. Given the laborious, repetitive nature of locating landmarks in images, this task is ripe for automation. To achieve this, we can train a model to extract features from the image, learning to identify which regions are discriminative to the landmark's location. However, even for a trained radiologist, the task of accurately defining anatomical landmarks poses significant challenges, further compounded by ambiguity when the structure of interest spans beyond a single pixel - for instance, the "corner" of the jaw. This ambiguity in landmark definition leads to aleatoric uncertainty in datasets of expertly labelled landmarks, an uncertainty caused from inherent randomness in the data or the task itself [Kendall and Gal, 2017]. Furthermore, there will exist uncertainty over the parameters of the model itself, a phenomena called epistemic uncertainty [Gal, 2016]. Reliably measuring these predictive uncertainties is crucial in the medical domain, where uncertain predictions can be flagged and manually corrected by a human-in-the-loop [Holzinger, 2016].

The field of uncertainty estimation for DNNs is still in its formative stages. Contemporary literature points to substantial *miscalibration* in DNNs i.e. a large discrepancy in deep models' perceived uncertainty and their actual error rates [Guo et al., 2017]. In the medical imaging domain, much research energy has been poured into overcoming this problem for the task of biomedical image segmentation, a sister task to landmark localisation that aims to identify large structures of an image, rather than a single point [Jungo et al., 2020]. However, the same is not true of landmark localisation, presenting an opportunity for a contribution to the field. Many State-of-the-Art methods for landmark localisation frame the task as an image-to-image regression, predicting Gaussian Heatmaps centred around the landmark of interest. Post-hoc techniques for uncertainty estimation for these methods would be an invaluable contribution to the field, presenting potential for high impact due to their plug-and-play nature. Since the task involves the regression of a 2D continuous image, traditional evaluation designed for uncertainty estimation in classification tasks like image segmentation cannot be directly

applied [Guo et al., 2017]. Therefore, evaluation metrics for uncertainty measures in landmark localisation beyond simply measuring their correlation with localisation error are needed to more effectively benchmark them [Drevickỳ and Kodym, 2020; Thaler et al., 2021].

Furthermore, a particular challenge for Deep Learning techniques in the medical imaging domain is a distinct lack of training data. The longstanding and well supported maxim of machine learning is that more data means better models [Sun et al., 2017]. Yet, datasets in the medical are limited to hundreds of samples, a stark contrast to the billions of data points utilized in recent Large Language Models [Brown et al., 2020]. Compounding this challenge is the practicality of on-site deployment, where practitioners are reluctant to send sensitive data to cloud-based systems capable of handling large models. Addressing these two critical issues simultaneously necessitates the development of lightweight models that can perform inference on low-compute devices and learn effectively from limited data. These models provide an effective solution, capable of making the most out of limited resources while being viable for on-site implementation in real-world medical scenarios.

An interesting and less obvious alternative to Deep Learning that excels in low-data regimes are Gaussian Processes (GPs) [Rasmussen and Williams, 2006]. GPs also bestow the property of truly Bayesian uncertainty estimation in their predictions, presenting a more mathematically rigorous and therefore trustable uncertainty than Deep Learning approaches. However, due to the computational demands when using high-dimensional data, GPs are rarely used in medical imaging applications, and to the best of our knowledge, never in a multi-task image regression task, regardless of domain. Progress in this area would represent an exciting milestone in the field for truly Bayesian techniques in medical image analysis.

Looking ahead, let us consider the evolving significance of “open-source” within the research landscape. Open-source software has become the backbone of machine learning innovation, fostering a collaborative environment that accelerates the rate of discovery. While the research community values reproducibility as a cornerstone of good practice, there remains an undervaluation of reusability. Reproducibility ensures that results can be consistently achieved across different settings, while reusability focuses on developing research that is designed to be accessible and easy for third parties to use [Lu et al., 2022b]. Further, software that is accessible to non-experts, including healthcare professionals and research clinicians, is

vital to see application outside of the research lab. Hence, for research to truly make an impact, the transition from code to application must be as seamless as possible. Embracing low-code or no-code approaches can be pivotal in this context, democratising access and enabling a broader spectrum of users to leverage the power of machine learning tools. Therefore, research must not only be robust and reproducible, but also reusable and readily accessible to those on the frontlines of patient care as well as to users with entry-level software skills.

1.2 Research Questions

The key research questions this thesis aims to address are as follows:

- **Q1:** How can we develop a lightweight and data-efficient Deep Neural Network for landmark localisation?
- **Q2:** Can heuristic uncertainties in landmark localisation be formalised using a data-driven, Frequentist framework?
- **Q3:** How can we better benchmark uncertainty measures in landmark localisation within a Frequentist context?
- **Q4:** How can we overcome the computational challenges with Gaussian Processes for rigorous, Bayesian uncertainty estimation for landmark localisation?

1.3 Thesis Outline

In this thesis we journey through the notoriously arduous endeavour of uncertainty estimation for DNNs, through the lens of landmark localisation. Figure 1.1 shows a typical clinical workflow involving machine learning in practice, highlighting the various aspects of the pipeline the contributions of the thesis impacts. We start our journey with Q1, tackling the challenge of improving landmark localisation with parameter and data-efficient models. We uncover an exciting thread of uncertainty estimation using a heuristic property of our model, motivating Q2. We follow this thread and extend the heuristic to general heatmap-based localisation methods and approximate Bayesian inference, building a Frequentist framework

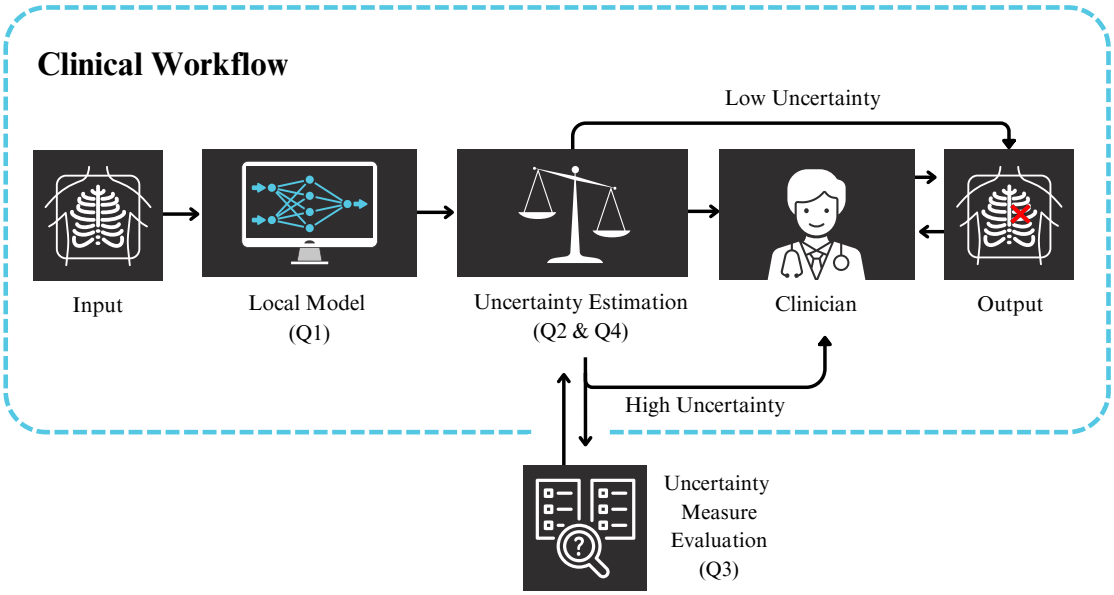


Figure 1.1: A representation of a typical clinical workflow involving machine learning with a human-in-the-loop [Wu et al., 2022], specifying which parts are relevant to our research questions. In Chapter 4 we address Q1, proposing a high-performing, lightweight model that can run on local machines. Throughout Chapter 4, 5 & 6 we propose various methods to estimate model uncertainty based on heuristic, Frequentist (Q2) and Bayesian (Q4) paradigms, allowing a human-in-the-loop to identify and correct poor predictions. In Chapter 5, we approach Q3 by proposing *Quantile Binning*, a framework to evaluate the quality of uncertainty metrics used to estimate predictive uncertainty.

to benchmark uncertainty estimation methods for any regression task, addressing Q2 and Q3. The final destination of this work answers Q4: an ambitious application of the fully Bayesian Gaussian Process framework to our task. We will show that uncertainty estimation in landmark localisation is an entirely achievable endeavour, with our contributions able to provide practical benefits to clinicians.To increase the impact of the work in this thesis, the code is freely available and open-source under the MIT license in the software repositories MediMarker [Schobs, 2022] and PyKale [Lu et al., 2022b].

Chapter 2 sets the stage by providing a brief overview of machine learning concepts pertinent to this thesis, with a focus on Deep Neural Networks and Gaussian Processes. We also introduce the concept of uncertainty estimation in machine learning, framing it in the Bayesian and Frequentist perspectives. We review and discuss several common methods for uncertainty estimation in DNNs.

Chapter 3 provides a review of the literature in landmark localisation approaches and uncertainty estimation in landmark localisation, where current gaps in the literature are pointed out. We outline the datasets used in this thesis: A standard collection of Cardiac Magnetic Resonance (CMR) images covering two views with accurate annotations (ASPIRE-S), a larger collection of CMR images from the same source with a single view (ASPIRE-L), and a public dataset of Cephalometric images (Cephalometric) we use to test the generalisability of our methods. Finally, we present a short study motivating our use of LannU-Net, a U-Net based model inspired by nnU-Net [Isensee et al., 2021], as a baseline for large capacity models representing State-of-the-Art localisation accuracy.

In Chapter 4, we present a lightweight, data-efficient model for landmark localisation that provides a heuristic-based uncertainty score alongside its prediction. Focusing on Q1, we are primarily motivated by improving localisation accuracy, which is reflected in our accuracy-based evaluation metrics. To achieve a lightweight model that localises accurately, we introduce the multi-task learning network, PHD-Net, which jointly performs **P**atch-based **H**eatmap and **D**isplacement regression for landmark localisation. We propose two methods to obtain coordinate prediction alongside a heuristic uncertainty score based on "patch votes", shown in Figure 1.2.(1). We show how to use these scores to separate predictions effectively into high and low error categories using Frequentist approaches. Further, we show that PHD-Net performs comparably to State-of-the-Art models of similar parameter counts in the literature, while retaining a small memory footprint and a unique uncertainty estimation heuristic. We will also show that our proposed training regime scales effectively with model capacity, attaining competitive performance with large State-of-the-Art models. We benchmark our vanilla PHD-Net on the smaller ASPIRE-S dataset, and its higher capacity extensions on the larger ASPIRE-L dataset.

In Chapter 5, we extend the heuristic-based approach for uncertainty estimation beyond our custom architecture proposed in Chapter 4, and ground it in an approximate Bayesian Inference framework using Deep Ensembles. Our primary contribution of this chapter, shown in Figure 1.2.(2), is the proposal of a Frequentist approach to uncertainty estimation, *Quantile Binning*. This is a data-driven method to bin any set of continuous uncertainty measure and continuous error pairs, estimating error bounds for each bin. With a focus on Q2 and Q3, our

evaluation is concerned with the quality of uncertainty estimation rather than localisation accuracy alone. Therefore, we use the smaller ASPIRE-S dataset and smaller capacity models to exemplify our methods, testing the generalisability using the Cephalometric dataset. We develop evaluation metrics for binning-based uncertainty methods and benchmark uncertainty measures with them. We use *Quantile Binning* to compare and evaluate three uncertainty measures across the three datasets, uncovering insights on their relationship to aleatoric uncertainty. Our thorough investigation of these uncertainty measures gives us practical recommendations on their use and suggests *Quantile Binning's* utility as a framework for evaluating future uncertainty estimation approaches. *Quantile Binning* is application agnostic, and can be used in any regression problem that provides continuous uncertainty measures for each sample.

In Chapter 6, we take a more rigorous approach to uncertainty estimation using the fully Bayesian machine learning framework of Gaussian Processes (GPs), as shown in Figure 1.2.(3). Specifically, we use Convolutional Gaussian Processes (CGPs): a variant of GPs which use a covariance function inspired by the convolutional structure of the kernels used in Convolutional Neural Networks (CNNs). To overcome the issues with computational complexity, we propose a two-stage approach. At Stage 1 we use a CNN to make a coarse prediction and refine it in Stage 2 with the CGP. The final prediction is Bayesian, giving the distribution of likely landmark locations, quantifying model uncertainty. We focus on a subset of the Cephalometric dataset, due to limited availability of ground-truth information on aleatoric uncertainty. Given the nature of Gaussian Processes, we introduce another evaluation method to reflect the Gaussian nature of the outputs. Despite a somewhat anticipated decrease in localisation accuracy, we demonstrate promising results in the area of uncertainty estimation.

Finally, in Chapter 7 we summarise the contributions of the thesis to the research and open-source community, and discuss potential for future work.

1.4 Contributions

All research presented in this thesis is publicly available under a unified, fully-documented, open-source framework, called *MediMarker* [Schobs, 2022]. The framework is built with a low-code/no-code usability in mind, meaning that training and inference can be customised

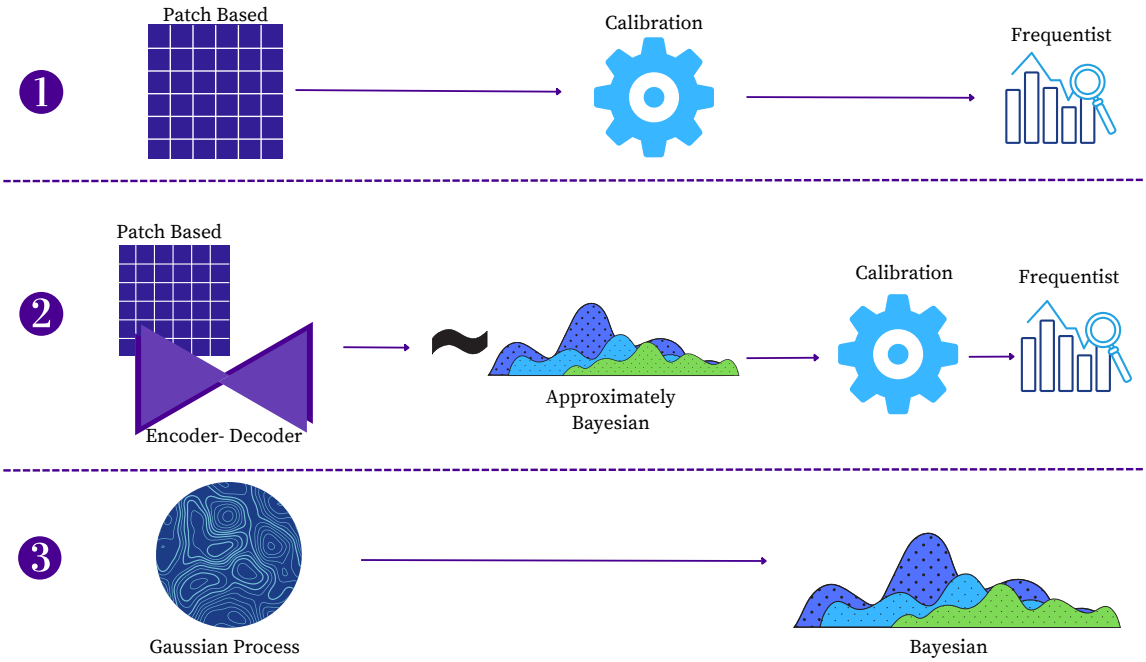


Figure 1.2: **Illustrative diagram outlining the approaches to uncertainty covered in this thesis.** 1) A patch-based approach to landmark localisation with a Frequentist approach to uncertainty, PHD-Net (Chapter 4). 2) A general framework applied to landmark localisation that is model-agnostic, utilising approximate Bayesian uncertainty with a Frequentist approach, Quantile Binning (Chapter 5). 3) A purely Bayesian approach to landmark localisation using Gaussian Processes (Chapter 6).

and performed without writing a line of code. We hope such an accessible approach to the research in this thesis can be a contribution in itself, facilitating straightforward use and improvement of the work in this thesis. Since the methods proposed in Chapter 5 are agnostic of application, the method is fully integrated into *PyKale*, an open-source framework that is officially a member of the PyTorch ecosystem [Lu et al., 2022b]. The two software repositories are summarised as follows:

1. **MediMarker** [Schobs, 2022]: A low-code/no-code standardised framework that provides DNN and GP models for landmark localisation. It contains the work from Chapters 3, 4 & 6: `https://github.com/Schobs/MediMarker`.

2. **PyKale** [Lu et al., 2022b]: A framework for accessible machine learning from multiple sources: containing a fully standardised, documented and tested implementation of *Quantile Binning*: `https://github.com/pykale/pykale`. A working, reproducible

example of the work from Chapter 5 can be found: `https://github.com/pykale/pykale/tree/main/examples/landmark_uncertainty`.

The contributions of this thesis summarised in Figure 1.1 and Figure 1.2, are as follows:

1. **Landmark Localisation Models**. We propose and validate improvements for patch-based landmark localisation models, offering a lightweight, computationally inexpensive solution: PHD-Net. We show our contributions scale favourably with model capacity for single landmark localisation. We also propose the first fully Bayesian approach to anatomical landmark localisation using Convolutional Gaussian Processes, introducing a two-stage approach and a novel inducing point initialisation to overcome computational issues. To the best of our knowledge, this is the first application of Gaussian Processes to multi-output regression on images, thus represents a significant contribution to the field. All models are available to train, deploy and augment in a low-code/no-code capacity in a single repository, MediMarker.

2. **Uncertainty Estimation Methods**. We study Frequentist and Bayesian approaches to uncertainty estimation in landmark localisation, summarised in Figure 1.2. Specifically, we propose uncertainty measures heuristically derived from heatmap-based landmark localisation methods. We use Frequentist approaches with calibration sets to improve such heuristics, namely *Candidate Smoothing* and *Adaptive Prediction* for patch-based methods, and the *Quantile Binning* framework for general heatmap-based methods. We demonstrate practical clinical utility of our methods by showing how to filter out poor predictions for manual correction based on learned thresholds. Under our *Quantile Binning* framework we provide learned error bound estimations for predictions based on their uncertainty value. The framework is application agnostic, relevant to any regression problem with per-sample uncertainty estimates. Further, using our Gaussian Process method for landmark localisation, we offer mathematically rigorous, fully Bayesian uncertainty estimation.

3. **Uncertainty Estimation Metrics**: We introduce novel evaluation metrics based on Quantile Binning, establishing a Frequentist framework that serves as a benchmark for uncertainty estimation techniques for general regression problems. Our contribution fills

a critical gap by extending the evaluation of models beyond accuracy metrics towards a more holistic perspective. In the application to landmark localisation, our proposed metrics address the existing limitations by enabling a comprehensive assessment of uncertainty estimation techniques beyond correlation with error.

1.5 Relationship to Published Work

The content of the thesis is primarily based on the following publications produced during my PhD journey.

1. **Schöbs, L.**, Zhou, S., Cogliano, M., Swift, A., & Lu, H. (2019). A Biased Sampling Network to Localise Landmarks for Automated Disease Diagnosis. In Medical Imaging Meets NeurIPS Workshop, NeurIPS 2019.

2. **Schöbs, L.**, Zhou, S., Cogliano, M., Swift, A., & Lu, H. (2021). Confidence-Quantifying Landmark Localisation for Cardiac MRI. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 985–988).

3. **Schöbs, L.**, Swift, A., & Lu, H. (2022). Uncertainty Estimation for Heatmap-Based Landmark Localisation. IEEE Transactions on Medical Imaging, 42(4), (pp. 1021–1034).

4. **Schöbs, L.**, McDonald, T., & Lu, H. (2023). Bayesian Uncertainty Estimation in Landmark Localization using Convolutional Gaussian Processes. In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging at MICCAI 2023, *Spotlight Talk*, MICCAI 2023.

Other Work and Publications

Throughout my PhD I was involved in other collaborative projects, which involved work related to this thesis:

1. Alabed, S., Uthoff, J., Zhou, S., Garg, P., Dwivedi, K., Alandejani, F., Gosling, R., **Schöbs, L**., Brook, M., Shahin, Y. and Capener, D. (2022). Machine learning cardiac-MRI features predict mortality in newly diagnosed pulmonary arterial hypertension. European Heart Journal-Digital Health, 3(2), (pp. 265-275).

2. Lu, H., Liu, X., Zhou, S., Turner, R., Bai, P., Koot, R., Chasmai, M., **Schobs, L.**, and Xu, H. (2022). PyKale: Knowledge-Aware Machine Learning from Multiple Sources in Python. In 2022 ACM 31st International Conference on Information & Knowledge Management (CIKM) (pp. 4274–4278).

3. Tripathi, P.C., Suvon, M.N., **Schobs, L.**, Zhou, S., Alabed, S., Swift, A.J. and Lu, H. (2023). Tensor-based Multimodal Learning for Prediction of Pulmonary Arterial Wedge Pressure from Cardiac MRI. In 2023 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Forthcoming.

I also co-supervised three Bachelor of Computer Science Students for their final year dissertations:

1. Jones, E (2023). Improving Epistemic and Aleatoric Uncertainty Estimation in Cephalometric Landmark Localisation Tasks. BSc. Thesis. University of Sheffield.

2. Smith, T (2023). Introducing Multi-Task Training to Vision Transformers for Landmark Localisation. BSc. Thesis. University of Sheffield. - *Some of this work is included in Section 4.5.*

3. Gavin, O (2023). Advancing Landmark Localisation with UNETR and Novel Heatmap Augmentation. BSc. Thesis. University of Sheffield.

Chapter 2

Basics of Deep Learning, Uncertainty, and Gaussian Processes

The chapter aims to equip the reader with the background knowledge relevant to my PhD research on landmark localisation using machine learning, with a focus on deep learning and uncertainty estimation. The chapter begins by introducing the fundamental building blocks of deep learning, particularly the techniques used for image processing. This paves the way for an exploration of common blueprints of network architectures used in medical image analysis. The chapter then changes focus to the topic of uncertainty estimation for machine learning models. We define uncertainty, differentiating between the data-derived aleatoric and model-based epistemic uncertainty. We present the two perspectives we will view uncertainty through: Bayesian and Frequentist; accompanying the discussion with common uncertainty estimation methods. Finally, we briefly outline the basics of Gaussian Processes, a Bayesian approach to machine learning.

2.1 Fundamentals of Deep Learning

Inspired by neuroscience, Deep Learning is a subset of machine learning that uses deep neural networks. In the following section, we will review the building blocks of these networks, and how to arrange them to create some of the neural network architectures that are used in State-of-the-Art medical imaging techniques. We will subsequently discuss the relevant

technical details on training these networks, which will be relied on later in the thesis.

2.1.1 Deep Neural Networks (DNNs)

A Neural Network is a computational model that is inspired by neurons in the brain [McCulloch and Pitts, 1943]. In a supervised setting, we are given an input matrix $\mathbf{X}$ and a label matrix $\mathbf{Y}$, where the matrices consist of of N input-output vector pairs $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$. We aim to approximate the function that maps the input vectors to the output vectors, i.e. the function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space.

To achieve this, we design a flexible system of learnable weights, with the aim to learn the ideal weights to transform each input vector $\mathbf{x}^{(i)}$ to its paired label $\mathbf{y}^{(i)}$. This system is called a neural network, and the artificial neuron is its basic building block. The neuron takes some input vector $\mathbf{x}^{(i)}$ and computes a weighted sum of this vector using its weights $\mathbf{w}$ and an added bias b:

$$z = \mathbf{w}^T \mathbf{x}^{(i)} + b, \tag{2.1}$$

where $\mathbf{w}$ and b are learned. This is followed by some non-linear activation function ϕ to obtain the output a:

$$a = \phi(z). \tag{2.2}$$

A neural network with a single layer of neurons (nodes) between its input and output, is only capable of learning a linear function. To learn non-linear functions, multiple layers of neurons are needed. The simplest instantiation of this concept is the FeedForward Neural Network (FFNN), shown in Figure 2.1. It consists of multiple layers of nodes (neurons) between the input and output layer. These layers are called hidden layers, with each node having unique, trainable weights followed by some non-linear function that transforms the neuron outputs. In classification tasks like image classification, the output layer is transformed into a vector of probabilities with a sigmoid function, and in regression tasks such as landmark localisation, the values in the output layer are simply returned. The identifier FeedForward refers to the fact that there are no cycles between the nodes i.e. information is only moving forward through the network.

The key to training a FFNN is through back-propagation, which involves calculating the

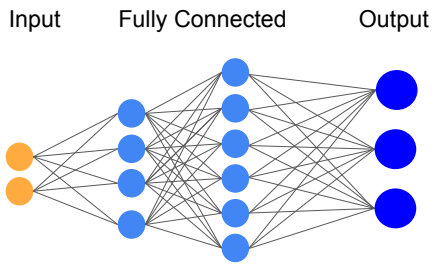


Figure 2.1: An example of a FeedForward Neural Network.

derivative of a given loss function (a measure of the FFNN's error) and using this to tune the weights back through the network. This is an iterative process, continually sending training data through the network and performing back propagation until the loss function converges to a small value.

Deep Learning simply refers to a Neural Network with multiple layers of interconnecting nodes (hidden layers) between the input and output layers [Goodfellow et al., 2016].

2.1.2 Convolutional Neural Networks (CNNs)

A Convolutional Neural Network (CNN) is a Deep FeedForward Neural Network founded on the architecture of the visual cortex system in mammals. It drew inspiration from specific subsets of neurons in the brain that fire when perceiving different shapes (e.g., horizontal lines vs. vertical lines [Hubel and Wiesel, 1962]). CNNs remain the most popular choice of method in medical image analysis, alongside the increasingly popular Vision Transformer [Dosovitskiy et al., 2020].

A basic CNN consists of a series of layers, with each layer sequentially applying four core functions: Convolution, Pooling, Activation Function and Normalisation. To determine the final prediction, the last layer is either a Fully Connected Layer or a Fully Convolutional Layer.

At a high-level overview of the process, each layer of a CNN employs a set of kernels (filters) to convolve over the input image or the output of the previous layer, transforming the information from a previous layer into a new feature representation. A kernel is an

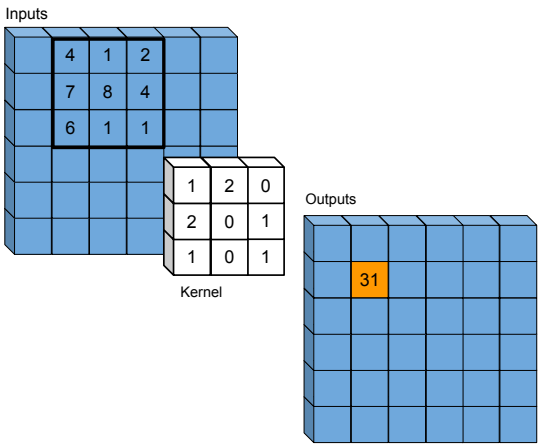


Figure 2.2: An example of a convolution operation using a kernel of size 3×3

n-dimsensional array (a tensor) made up of weights, that are tuned over time by the network. For example, a kernel on the first layer of the network may have its weights tuned to detect edges of the images, while the deeper kernels will learn relevant features that are progressively more abstract. As with Feedforward Neural Networks, the derivative of the error is back-propagated through the network with each pass, tuning these filters, until finally, the network outputs a classification or regression values(s).

Convolutional Layers

The convolutional layers of a CNN take an input image and a set of kernels with learnable weights and biases to produce an activation map. The kernel input is an n-dimensional tensor where n is the dimensionality of the input image (e.g 2D for an image), and the kernel is a tensor of weights and biases of the same dimensionality. The kernel slides across the input in fixed steps known as strides (typically set to 1) and computes the dot product between the patch of input it is over and the weights of the kernel. This produces an activation map which is the approximately the same size of the input, often padded with zeros to match the input size. A smaller stride is generally preferable, as it allows the network to retain more information from the input [Coates et al., 2011]. A larger stride can be used if there is a desire for the activation map to be significantly smaller than the input size, without the use of pooling.

The convolution process is depicted in Figure 2.2. The sliding window approach gives

CNN a form of global translational invariance, since the kernel will pick up the same feature regardless of its position in the image.

Pooling

Pooling layers are typically applied after convolutions, with the aim to reduce the spatial dimensions of the input for the next layer. This not only reduces the computational cost of the operations but also gives the system a local translational and deformation invariance. Pooling aims to retain the most informative features from the previous layer. The most common types of pooling operations are max pooling and average pooling. Max pooling returns the maximum value from the section of the image covered by the filter, while average pooling returns the average of all the values from that section. Figure 2.3 shows the max pooling operation.

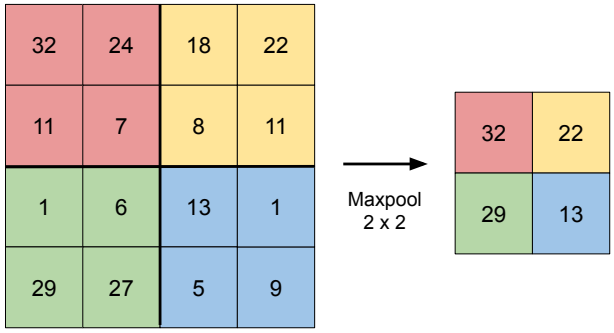


Figure 2.3: An example of max pooling with a pooling window of size 2×2.

Activation Functions

Activation Functions are typically implemented after pooling, and introduce non-linearity into the model, allowing us to approximate more complex functions.

The most common activation function in CNNs is the Rectified Linear Unit (ReLU), given by:

$$\text{ReLU}(x) = \max(0, x) \tag{2.3}$$

where x is the input to the ReLU function. This function retains any positive value and sets all negative values to zero, introducing non-linearity without affecting the receptive fields of the convolutional layer. Other activation functions such as the Sigmoid function are commonly used, but more relevant to classification problems.

Normalisation

Normalisation layers are typically nested between the convolution and activation function layers. The aim is to standardize the input distribution to the next layer, reducing internal covariate shift [Ioffe and Szegedy, 2015]. This standardisation reduces the effects of the vanishing and exploding gradient problem, improving the stability and speed of training. Batch Normalisation achieves this by scaling the batch of input data to have a mean of zero and standard deviation of 1 [Szegedy et al., 2015]. Instance Normalisation standardizes the input data a single training sample at a time, which is preferable when computational resources demand the use of smaller batch sizes, which would make batch normalisation less reliable [Ulyanov et al., 2016].

Fully Connected Layer/ Fully Convolutional Layer

Typically, the final layers in a CNN are fully connected, where each neuron is connected to all neurons in the previous layer. Although more computationally intensive than the convolution, these layers consolidate all information to calculate scores for the output, be it class predictions or regression values.

However, since the number of neurons in the fully connected layer is fixed, the size of the input image is forced to be a particular resolution to match the architecture. In methods relevant to these studies, the fully connected layer is often replaced by a Fully Convolutional Layer. This is a set of 1×1 convolutional kernels which slide across all the activations of the previous layer, akin to a single node in a fully-connected layer. This operation comes with the advantage of allowing an input of any size to be used.

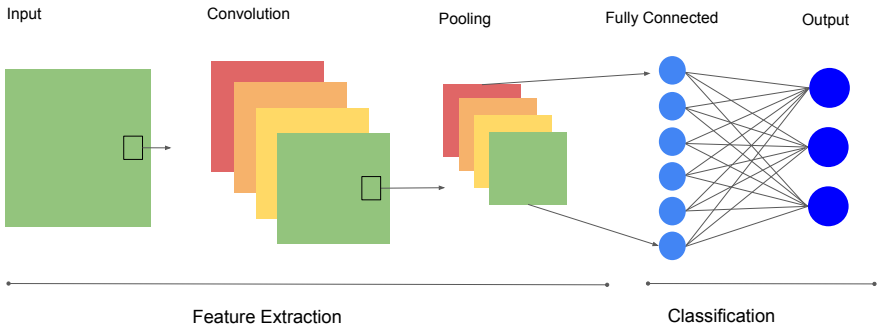


Figure 2.4: An example of a Convolutional Network Architecture. It involves a convolution operation with 4 filters, a maxpooling operation reducing the dimensionality, and a fully connected layer.

Designing a CNN

Using these simple building blocks as a foundation, we can create a CNN, as shown in Figure 2.4. Extensive research has been conducted to uncover architectures that can solve many tasks in computer vision. The model that thrust CNNs into the limelight by winning the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2012 was largely based on an architecture from the 1990s. This early model was LeNet, a CNN used to classify handwriting [LeCun et al., 1998]. The 2012 ILSVRC task was significantly more challenging, asking participants to classify hundreds of thousands of higher resolution photos, spanning 1000 categories. Surprisingly, the core building blocks of CNNs demonstrated decades prior in LeNet had to only be scaled up to win this task. AlexNet increased the network depth and size, and stacked convolutional layers on top of each other and won the competition by a significant margin [Krizhevsky et al., 2017].

The proceeding years saw fervent research to improve this style of model, demonstrating the importance of network depth [Simonyan and Zisserman, 2014; Szegedy et al., 2015] and the power data-driven models have on computer-vision tasks. However, with every increase of depth in these models, the silhouette of a new problem loomed closer: the vanishing gradient problem [Hochreiter, 1998]. As back-propagation is applied from the final layer to the first

one, the derivatives of each layer are multiplied down the network. If the derivatives are small, then the gradient will exponentially decrease as we move down the layers until it vanishes. Conversely, if the derivatives are large, the gradient will exponentially increase (an *exploding gradient*). Both problems significantly impact the ability of the network to learn, as the error cannot be propagated to the early layers effectively. We can see that this problem is compounded by more layers added to the network. To solve this problem, ResNet was introduced [He et al., 2016]. The authors introduced the deceptively simple residual block - a module that splits the path flow in two: (1) the regular convolutional path and (2) A skip connection that acts as an identity mapping, adding the block's input to the output of path (1). This is depicted in Figure 2.5. The skip connection ensures the derivative does not vanish due to the activation function in the convolutional path. As discussed, normalisation throughout the network also alleviates this problem.

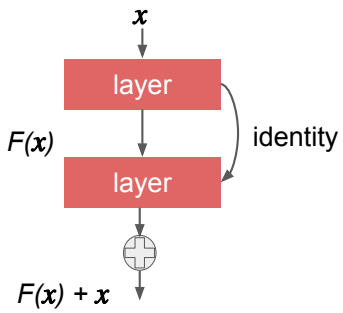


Figure 2.5: The Residual Block [He et al., 2016]. Note the identity mapping concatenating the input of the block to the output of the block.

2.1.3 Fully Convolutional Networks (FCNs)

The models discussed so far generally fit well to problems like classification or regression tasks, where the aim is to predict a class label or regression value from an image. However, CNNs can also be used for more holistic tasks, like image segmentation, or its sister problem that is the focus of this thesis - landmark localisation. Image segmentation is the process of partitioning an image into multiple segments e.g. identifying all pixels that relate to the left ventricle in a CMR image. Given an input, the model is asked to produce a mapping where only the pixels in the target structure are activated. Landmark localisation can also

be approached through a similar lens: producing a mapping where the highest activation is on the landmark of interest. Segmentation was tackled first using FCNs, when Long et al. [2015] proposed a Fully Convolutional Network (FCN), that removed fully connected layers in favour of many convolutions of size 1×1, allowing the input and corresponding output size of the network to be arbitrary. Therefore, using FCNs we can make pixel-wise predictions or predictions relating to patches of the image. This approach was key to image-to-image tasks like image segmentation and landmark localisation.

2.1.4 Encoder-Decoder Networks

Encoder-Decoder Networks are a special case of FCN that employ a mirrored downsample-upsample structure. The convolutional and pooling layers constitute the *encoder*, which compresses the input into a compact, lower-dimensional form. This encoded representation is then expanded back to its original size by the *decoder*, using transposed convolutions, unpooling layers, or upsampling layers. Central to the design of these networks is the concept of analyzing an image at multiple resolutions and subsequently reconstructing it to its full detail.

Ronneberger et al. [2015] introduced U-Net, concatenating the outputs of the decoder steps with the features from the encoder steps at the same level, exemplified in Figure 2.6. These "skip" connections allow information from the image features from the encoder side of the network to feed information that may be pertinent when building the image back up in the decoder side. This improved flow of information facilitated by the "skip" connections leads to higher performance in image-to-image regression task. This style of architecture has become the default for many computer vision tasks including image segmentation [Alom et al., 2019], pose estimation [Newell et al., 2016] and many anatomical landmark localisation tasks [Davison et al., 2018; Tiulpin et al., 2019], which will be discussed in more detail in Section 3.1.1.

2.1.5 Vision Transformers (ViTs)

Vision Transformers (ViTs) represent the next evolution in network architecture for vision problems, challenging the dominance of CNNs. First introduced by Dosovitskiy et al. [2020],

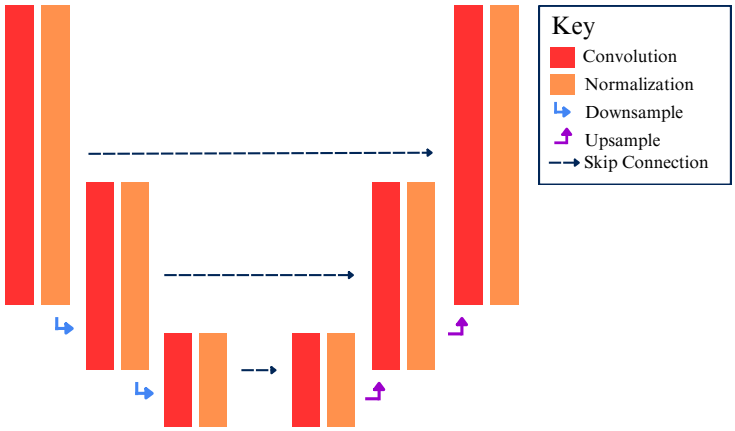


Figure 2.6: A representative diagram of the architecture of U-Net [Ronneberger et al., 2015].

the ViT applies the transformer architecture, originally developed for natural language processing [Vaswani et al., 2017] to image analysis tasks. The ViT abandons the translational invariant window-based inductive bias encoded into the CNN, in favour of an architecture representing more of a blank canvas.

At a high level, the ViT operates by dividing the input image into a sequence of fixed-size patches, each linearly embedded into a vector with a positional encoding. Then, the model applies self-attention mechanisms, modeling the dependencies between any pair of positions in the input sequence. The ViT forgoes the visual-system inspired inductive biases of the CNN, arguing that an architecture with fewer inbuilt biases is more expressive. Another key strength of the transformer model is its ability to model long-range interactions between the patches from opposite corners of the image, which traditional CNNs struggle with.

2.1.6 Training Neural Networks

Training Loop

We have covered the fundamental concepts of how to build a Deep Neural Network, highlighting a plethora of architectural choices. The next step is to train the network, in which we provide some signal to the network to tune the weights from random to successfully approximating some objective function. In this thesis, we are interested in *Supervised Learning*, in which we

generate the signal using known *{Data, Label}* pairs from a training dataset.

In a typical in landmark landmark localisation training loop where we wish to regress the coordinates of a single landmark contained in each image, our input data is a tensor of 2D images, $\boldsymbol{X}$, where $\mathbf{X}^{(i)}$ is the ith 2D image matrix. Our label data is a matrix of coordinates, $\mathbf{Y}$, where $\mathbf{y}^{(i)}$ is the coordinate vector of the landmark we want to localise, present in $\mathbf{X}^{(i)}$. We create a set of N training examples $\left\{\left(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}\right), \ldots, \left(\mathbf{X}^{(N)}, \mathbf{y}^{(N)}\right)\right\}$. In some cases, we may represent our landmark label as an image matrix encoding the coordinate i.e. the label data is a $3D$ tensor $\boldsymbol{Y}$, with the ith sample label matrix denoted as $\mathbf{Y}^{(i)}$.

Using our set of training examples $(\boldsymbol{X}, \mathbf{Y})$, the process of training a neural network using mini-batch Stochastic Gradient Descent (SGD) [Amari, 1993; LeCun et al., 1998] involves the following steps:

1. Initialisation: The network's weights $\theta = (\mathbf{W}, \mathbf{b})$, where $\mathbf{W}$ are the weights and $\mathbf{b}$ the biases, are initialised randomly.

2. Forward Propagation: A batch of c inputs $\left\{\mathbf{X}^{(i)}, \ldots, \mathbf{X}^{(i+c)}, \right\}$ is fed through the networks layers, producing outputs, $\left\{\hat{\mathbf{y}}^{(i)}, \ldots, \hat{\mathbf{y}}^{(i+c)}, \right\}$.

3. Cost Calculation: We use a predefined loss function, $\mathcal{L}(\mathbf{Y}, \widehat{\mathbf{Y}})$ to measure the difference between predicted outputs $\widehat{\mathbf{Y}}$ and the true targets $\mathbf{Y}$. The total loss is typically the average value over the entire batch.

4. Backward Propagation: We compute the gradients of the loss function with respect to $\mathbf{W}$ and $\mathbf{b}$ by applying the chain rule. This gives us $\frac{\partial \mathcal{L}}{\partial \mathbf{W}}$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{b}}$ at each layer, indicating the direction of steepest ascent in the loss function.

5. Weight Update: Using our computed gradient and a chosen gradient descent algorithm (e.g. SGD), we update the weights and biases of the network in the opposite direction of the gradient, minimising the function. Defining a learning rate, α, the weight and bias updates are given by $\mathbf{W} = \mathbf{W} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{W}}$ and $\mathbf{b} = \mathbf{b} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{b}}$.

We repeat steps 2-5 iteratively, using batches of training data to gradually tune the weights and biases of the network towards some local minima of the loss function. Each full pass over the entire dataset is known as an epoch.

Relevant Loss Functions

For landmark localisation, we are typically performing the task of *regression* i.e. regressing the coordinate location or some representation of the landmark. However, in some cases we may want to perform *classification* on whether some part of the image contains the landmark. Below, we cover loss functions relevant to this thesis.

Mean Squared Error (MSE): Mean Squared Error (MSE) is a commonly used loss function in regression tasks, where the goal is to predict a continuous value. MSE is the average of the squared differences between the actual and predicted values, putting more weight on large deviations due to the squaring operation:

$$\mathcal{L}_{MSE}(\mathbf{Y}, \widehat{\mathbf{Y}}) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\|_F^2, \tag{2.4}$$

where $\mathbf{Y}$ is the target value and $\widehat{\mathbf{Y}}$ is the model prediction, averaged over a matrix of N elements. We can use MSE to train models to regress the coordinates of landmarks directly, in which $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ are matrices of coordinate vectors.

Alternatively, we can use MSE as a *reconstruction loss* in image-image regression. In landmark localisation, it is common to frame the landmark localisation task as heatmap regression. Rather than regressing coordinates directly, the objective of the model is to learn a Gaussian heatmap image for each landmark, with the centre of the heatmap on the target landmark. The network learns to generate a high response near the landmark, smoothly attenuating the responses in a small radius around it. For each landmark L_i with 2D coordinate position $\tilde{\mathbf{c}}^{(i)}$, the 2D heatmap image is defined as the 2D Gaussian function:

$$g_i\left(\mathbf{x} \mid\mid \mu = \tilde{\mathbf{c}}^{(i)}; \sigma\right) = \frac{1}{(2\pi)\sigma^2} \exp\left(-\frac{\|\mathbf{x} - \mu\|_2^2}{2\sigma^2}\right), \tag{2.5}$$

where $\mathbf{x}$ is the 2D coordinate vector of each pixel and σ is a user-defined standard deviation. The network learns weights $\mathbf{w}$ and biases $\mathbf{b}$ to predict the heatmap $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$. During inference, we can interpret the activation of each pixel in the predicted heatmap as the pseudo-probability of that pixel being the landmark.

The network learns to regress N heatmaps simultaneously by minimising the MSE between predicted heatmaps $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$ and the corresponding target heatmaps $g_i\left(\mathbf{x} \mid\mid \mu = \tilde{\mathbf{c}}^{(i)}; \sigma\right)$

for all landmarks L_i. In this case, we can extend MSE in Equation (2.4) to the n-D case $\mathcal{L}_{MSE}(\boldsymbol{Y}, \widehat{\mathbf{Y}})$, where $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ are tensors representing N Gaussian heatmap labels and predictions, respectively.

We can use **Weighted Mean Squared Error (WMSE)** to weight certain pixels as more important to the loss function:

$$\mathcal{L}_{WMSE}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}, \boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{W}^{(i)} \odot \mathbf{Y}^{(i)} - \widehat{\mathbf{Y}}^{(i)}\|_F^2, \tag{2.6}$$

where $\mathbf{W}^{(i)}$ is a matrix of weights, where $\mathbf{W}^{(i)}_{j,k}$ is the scalar weight for the loss at the pixel on the jth row and kth the column. This is typically used to weight parts of the image closer to the landmark higher than distant parts of the image.

Binary Cross Entropy (BCE): Binary Cross Entropy (BCE) is typically used for classification tasks, in which the positive class is delineated by the label 1 and the negative class is 0:

$$\mathcal{L}_{BCE}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}), \tag{2.7}$$

where $\mathbf{y}$ is a vector of classification labels and $\hat{\mathbf{y}}$ is the vector of predictions. In landmark localisation, we can partition the image into N *patches*, and assign the patch containing the landmark the positive class label and the rest the negative class label.

Once again, we can weight certain patches as more or less important using **Weighted Binary Cross Entropy (WBCE)**:

$$\mathcal{L}_{WBCE}(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{w}) = -\frac{1}{N}\sum_{i=1}^{N} w^{(i)} \left(y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right), \tag{2.8}$$

where $w^{(i)}$ is the scalar weight for the loss at the ith patch.

Training Techniques

Below, we list a few techniques to enhance the training of DNNs:

- **Data Augmentation** [Chlap et al., 2021; Pérez-García et al., 2021]. This aims to improve model generalisation. We randomly transform images in the training set, artificially increasing the size and diversity. Therefore, the model is exposed to a wider

variety of data scenarios, reducing overfitting and improving generalisation to unseen data. This is particularly important in medical imaging, where datasets are limited.

- **Deep Supervision** [Lee et al., 2015]. This is utilised in Encoder-Decoder models, where which we ask the model to make predictions at multiple resolution scales, injecting the losses at each. This technique facilitates a coarse prediction at lower feature resolutions and helps alleviate the vanishing gradient problem.

- **Multi-Task Networks**. This is a general architectural design that simultaneously optimises two or more objective functions. The intuition behind the idea is that if the tasks are complementary, the signal provided by each task will benefit both [Zhang and Yang, 2018]. They can be used to learn two complementary tasks such as joint motion estimation and segmentation [Qin et al., 2018], or two representations of the same task, such as patch-wise classification and regression of a landmark [Noothout et al., 2018].

2.2 Uncertainty Estimation

We have covered how to create and train our models in a supervised fashion, but we are yet unable to quantify the uncertainty of our model's prediction. This subsection will first define what we mean by *uncertainty*, before moving on to discussing the current literature.

2.2.1 Defining Uncertainty

Uncertainty can be divided into two categories: (1) Aleatoric uncertainty, which is *irreducible*, caused by noise inherent in the data we are modeling; and (2) Epistemic uncertainty, which is caused by uncertainty in the model's parameters, *reducible* with more data [Der Kiureghian and Ditlevsen, 2009; Gal, 2016; Kendall and Gal, 2017].

In landmark localisation, aleatoric uncertainty can be caused by ambiguity in the ground truth labels [Thaler et al., 2021] or imaging artifacts. This can be subcategorised again into *homoscedastic uncertainty*, which is constant for all inputs and *heteroscedastic uncertainty*, which is different for each input. In landmark localisation, homoscedastic uncertainty could by caused by trying to label a pixel-precise landmark that is inherently ambiguous (i.e. the structure of interest spans multiple pixels with no clear centre). On the other hand

heteroscedastic uncertainty may be high for a particular poor quality image containing imaging artifacts from the scanner. Epistemic uncertainty is a direct expression of model uncertainty, arising when the model cannot approximate the true function between the imaging features and landmark of interest correctly. If given an infinite amount of training data, the epistemic uncertainty can be reduced to 0.

In practice, both forms of uncertainty are present in landmark localisation. Let us illustrate this with an example. Consider the case where an expert radiologist is asked to provide ground truth landmark annotations for "the bottom of the chin" for images in a Cephalometric dataset. Since the chin does not have a defined "bottom", the landmark label will drift along the horizontal axis of the chin across the entire dataset. This represents aleatoric uncertainty: random noise inherent in the task itself. Aleatoric uncertainty cannot be minimised unless we alter the dataset itself, such as improving the quality of the images, or reannotating landmarks with a more precise definition in this case.

Furthermore, the annotation of medical images by experts suffers from both inter-observer and intra-observer variability [Warfield et al., 2008]. Therefore, training and evaluating a model using only a single annotation per image is prone to the bias of a single annotator [Lampert et al., 2016]. In landmark localisation, we can observe inter-observer variability through the lens of aleatoric uncertainty, inferring that the higher the variation in annotator opinion, the greater the ambiguity of the landmark Thaler et al. [2021]. Training a model to reflect this uncertainty would be key in providing more useful, interpretable results.

Epistemic uncertainty is of particular interest in our domain since medical imaging datasets are often small and the images complex. Therefore, at test time it is likely the model will encounter an image outside of the distribution it was training on. For example, our training may consist of cardiac scans of patients with a particular disease. If the model was deployed in the general population, it may perform poorly since the features it associated with a particular landmark were only present when a patient has the disease. Even within the same distribution of patients, a dataset in the order of hundreds is not sufficient to train a model to recognise a landmark in a deformable and varied structure like the heart perfectly. The solution to this form of uncertainty over model parameters is to train our model on a sufficiently large, representative dataset.

The resulting combination of epistemic and aleatoric uncertainty can be used to induce *predictive uncertainty*, the model's confidence in its prediction. Furthermore, the distinction between aleatoric and epistemic uncertainty is not entirely objective, and the lines between can blur [Der Kiureghian and Ditlevsen, 2009]. In this thesis, we will narrow our definitions. We will consider *aleatoric uncertainty* as uncertainty caused by noise in our fixed-sized dataset. Specifically, imaging artifacts and annotator variability caused by landmark definition ambiguity. We will consider *epistemic uncertainty* the uncertainty about the model parameters given a fixed-sized dataset.

2.2.2 Bayesian and Frequentist Perspectives to Uncertainty

Similarly, for the purpose of this thesis we will broadly view uncertainty estimation techniques through two lenses: Frequentist and Bayesian.

The Bayesian paradigm in statistics differentiates from Frequentist paradigm in two fundamental philosophies. First, Bayesian statistics views probability as a measure of belief in the occurrence of events, in contrast to the Frequentist paradigm that treats probability as the limit of occurrence frequency as the number of samples approaches infinity. The Bayesian distinction acknowledges that probability is inherently subjective, reflecting our beliefs and uncertainties about the underlying phenomena. Secondly, Bayesian statistics recognizes the influence of prior beliefs on posterior beliefs. By incorporating prior knowledge or beliefs about the phenomenon of interest, Bayesian inference allows for the integration of existing information into the analysis. This ability to formally incorporate prior beliefs sets Bayesian statistics apart from Frequentist approaches, which rely solely on observed data.

We can explain both frameworks using Bayes' theorem:

$$\mathbb{P}(H|D) = \frac{\mathbb{P}(D|H) \cdot \mathbb{P}(H)}{\mathbb{P}(D)}, \tag{2.9}$$

where

$$\mathbb{P}(D) = \int_H \mathbb{P}(D|H)\mathbb{P}(H)dH. \tag{2.10}$$

Here, D is our training data, and H is our hypothesis. $\mathbb{P}(H|D)$ represents the posterior probability of hypothesis H given the observed data D. $\mathbb{P}(D|H)$ is the likelihood of the data

given the hypothesis, $\mathbb{P}(H)$ denotes the prior probability distribution over the hypotheses, and $\mathbb{P}(D)$ is the marginal likelihood or evidence. In the context of machine learning, the weights θ of our network represent our hypotheses, and $\mathbb{P}(D)$ is the probability of our data integrated over all possible model parameters. This term ensures that the posterior distribution $\mathbb{P}(H|D)$ is a valid probability distribution that integrates to 1.

From a Bayesian perspective, all unknown quantities are treated as random variables, and probabilities represent degrees of belief. We start with an initial belief about our parameters by assigning a prior distribution over the weights, $\mathbb{P}(H)$. We update our beliefs by observing data and comparing it to our initial belief, $\mathbb{P}(D|H)$, forming our posterior distribution, $\mathbb{P}(H|D)$. We use $\mathbb{P}(D)$ as a normalising constant, which is often difficult to compute for complex models and data. Therefore, at inference, we have a distribution of model parameters $\mathbb{P}(H|D)$.

In the Frequentist interpretation, unknown parameters are considered fixed but unknown, and probabilities represent long-run frequencies of events. Therefore, the concepts of "prior" and "posterior" are not explicitly considered, instead focusing finding a hypothesis that maximises the likelihood, $\mathbb{P}(D|H)$. Therefore, the issue of calculating/approximating $\mathbb{P}(D)$ is sidestepped, and the prior beliefs about the intial parameters $\mathbb{P}(H)$ is not considered. We do not have a distribution over our model parameters, but instead construct confidence intervals around the prediction. The interval is constructed such that if we were to repeat the experiment many times, the true parameter value would fall within this interval a certain percentage of the time (say, 95% of the time for a 95% confidence interval).

In essence, Bayesians consider hypotheses as random variables and data as fixed, while Frequentists consider hypotheses as fixed and data as random.

The final term to introduce is *calibration*, a Frequentist concept, that quantifies the disparity between model predictions and long-term empirical frequencies. The quality of calibration, distinct from accuracy, can be evaluated using proper scoring rules such as the Negative Log Likelihood [Friedman et al., 2001] and techniques like the Expected Calibration Error [Naeini et al., 2015]. Therefore, a model can be accurate yet miscalibrated, and vice versa. In classification, a model is said to be *well calibrated* if its uncertainty score is reliable i.e. if there are 10 positive classifications each with a confidence score of 0.4, we would only expect 4 of those 10 to be the positive class. For classification problems, softmax outputs,

which are typically miscalibrated [Guo et al., 2017] can be calibrated in a straight-forward fashion using Platt Scaling [Platt et al., 1999] or Temperature scaling [Guo et al., 2017]. Essentially, these methods use a hold-out calibration set to learn a function to transform the miscalibrated softmax scores into better calibrated probability scores. This is not as straight forward in regression tasks like landmark localisation, as we are predicting continuous values. For regression, calibration can only be directly measured if we output a predictive distribution (e.g. mean and variance) [Kuleshov et al., 2018]. A model can be considered well calibrated if the true target values fall within the predicted distribution in a way that is consistent with that distribution. For example, for a predicted Gaussian distribution with a predicted mean and standard deviation, the actual target should all within one standard deviation of the mean $\sim$68% of the time, within two standard deviations $\sim$95% of the time, etc.

2.2.3 Uncertainty Estimation for DNNs

Next, we will introduce common uncertainty estimation approaches for DNNs.

Bayesian Neural Networks

Bayesian neural networks (BNNs) are the most direct application of the Bayesian notion of uncertainty to DNNs. BNNs model network weights as a probabilistic distribution rather than deterministic scalar values. Initially proposed in the early '90s and extensively researched thereafter [Lampinen and Vehtari, 2001; Neal, 2012], BNNs alleviate the issues of overfitting, and importantly provide a measure of epistemic uncertainty.

BNNs can be described through the lens of Bayes' theorm, as defined in Equation (2.9). Ideally, we aim to compute the posterior distribution $\mathbb{P}(H|D)$, from our training data D. We begin by defining a prior distribution over the weights of some neural network architecture, $\mathbb{P}(H)$. Since our models are often learning a complex function, we set $\mathbb{P}(H)$ to be a very loose prior, such as a Gaussian distribution with a mean of zero and a large variance. The aim is to iteratively update this prior into a posterior that better explains the observed data, in essence, transitioning towards the true distribution of model parameters $\mathbb{P}(H|D)$. To achieve this, we observe the likelihood of observing our training data given our parameters $\mathbb{P}(D|H)$, combine it with our prior belief $\mathbb{P}(H)$ and normalize with our marginal likelihood $\mathbb{P}(D)$, and

update our hypothesis H.

The concept sounds promising, but calculating $\mathbb{P}(D)$ is intractable, since it is computationally infeasible to integrate over the entire set of possible model weights. Therefore, we turn to approximation.

Markov Chain Monte Carlo (MCMC) methods, particularly the Hamiltonian Monte Carlo (HMC) method, are considered the best solutions for sampling from exact posterior distributions. These methods generate a series of weight samples by initiating a stochastic process from our initialisation of H, saving the samples along the way. Over time, these samples converge towards the true posterior distribution, providing a good approximation of $\mathbb{P}(H|D)$. However, they are slow to train since they are taking a more meandering approach to model optimisation compared to the bee-line behaviour of gradient descent towards a local minima.

Therefore, Variational Inference (VI) methods are more feasible, approximating the true posterior with a simpler distribution. The aim is to then find the parameters of this simpler distribution that minimizes the divergence from the true posterior [Blei et al., 2017; Graves, 2011]. In VI, we directly update the distribution of each of the model weights, rather than sample a series of them like in MCMC. VI methods are more scalable than MCMC but can under-estimate the uncertainty due to the simplifying assumptions made about the posterior [Blei et al., 2017].

Regardless of method, BNNs are still significantly more expensive to train compared to traditional DNNs since we are effectively learning a distribution over each weight rather than a singular value, and the optimisation procedure is more complex.

At inference, we can sample from the distribution of model weights (sampled or learned), generating a mean prediction with variance. Specifically, given a BNN and a test input $\mathbf{x}$, we can produce T predictions by running T forward passes through T sets of randomly sampled weights. We denote each prediction as $y^{(t)}$, where t ranges from 1 to T. For a regression problem the predictive mean is given by:

$$\mu = \frac{1}{T} \sum_{t=1}^{T} y^{(t)}. \tag{2.11}$$

The predictive variance is given by:

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(y^{(t)} - \mu)^2. \tag{2.12}$$

The higher the variance, the higher the uncertainty in the model weights, so a higher epistemic uncertainty. Furthermore, BNNs can also be trained to predict a mean and variance during training. The variance here represents the model's uncertainty about the output, given the input. This provides a prediction for aleatoric uncertainty, too. If the model has correctly learned this variance, it should be higher for inputs where the observed outputs are more variable (due to noise, inherent randomness, etc.), and lower for inputs where the observed outputs are more consistent.

Monte Carlo Dropout (MCD)

To address the scalability issue presented by BNNs while keeping the desirable Bayesian properties of the system, Monte Carlo Dropout (MCD) for approximate Bayesian Inference was introduced by Gal and Ghahramani [2016]. The method remains one of the most popular forms of epistemic uncertainty estimation in DNNs, since it can be easily introduced into an existing model with minimal effort. "Dropout" is the process of randomly turning "off" a neuron during training and inference. Initially used as a form of regularisation, ensuring a neural network learned robust representations of its inputs, Gal and Ghahramani [2016] showed dropout can be seen as a variational approximation to Bayesian uncertainty from a Deep Gaussian Process [Damianou and Lawrence, 2013]. They argued that performing dropout during testing can be seen as performing Monte Carlo integration over this approximate posterior. Intuitively, each configuration of the network with random dropout represents a different network, and we can sample many networks from the space of possible networks.

Therefore, at test time, predictions are averaged over multiple forward passes and the variance can be used to estimate model uncertainty. Specifically, given a network with dropout, and a test input $\mathbf{x}$, we can produce T predictions by running T stochastic forward passes. We can use Equations (2.11) and (2.12) to generate our mean and variance predictions, respectively. The higher the variance, the higher the uncertainty in the model weights, so a

higher epistemic uncertainty.

Deep Ensembles

While not Bayesian in the traditional sense, deep ensembles can also be interpreted as a form of approximate Bayesian inference, again modelling epistemic uncertainty. A deep ensemble consists of T identical models, trained with different initialisations. By passing our input through our T trained models, we can use Equations (2.11) and (2.12) to generate our mean and variance predictions, respectively. A higher variance represents a greater epistemic uncertainty.

In practice, Deep Ensembles achieve greater accuracy and better uncertainty estimates than BNNs and MCD. On the surface, deep ensembles appear to be a Frequentist approach to uncertainty, but there is a growing body of work arguing deep ensembles are approximately Bayesian [D'Angelo and Fortuin, 2021; Fort et al., 2019; Hoffmann and Elster, 2021; Wilson and Izmailov, 2020]. Lakshminarayanan et al. [2017] found using ensembles of deep neural networks with different initialisations produced well-calibrated uncertainties that were comparable to Bayesian approaches. In a large scale empirical study, Ovadia et al. [2019] found that an ensemble of identical networks with random initialisations outperformed all other methods in terms of accuracy and uncertainty estimation. They also found that using as little as 5 models in the ensemble was sufficient. The study performed by Fort et al. [2019] suggests this behaviour is due to random initialisations exploring entirely different modes of the loss landscape, whereas BNNs and MCD tend to focus on a single mode. The paper presents a series of insightful experiments exploring the subspace of loss landscape, finding that ensembling using random initialisations facilitates a powerful decorrelation effect between the models. We also see extensive use of ensembles of identical models with random initialisations in the domain of medical image segmentation to improve accuracy and estimate uncertainty [Jungo et al., 2020; Karimi et al., 2019; Mehrtash et al., 2020; Mehta et al., 2022]. In terms of fusion strategy, averaging the models in the ensemble is the standard technique [Jungo et al., 2020; Lakshminarayanan et al., 2017].

Conformal Prediction

Conformal prediction is a Frequentist approach of uncertainty, learning from data to give a strong theoretical guarantee: the construction of a prediction set that contains the true output with a user-specified probability [Shafer and Vovk, 2008]. The method is distribution-free; making no assumptions about prior or posterior distributions, nor the underlying machine learning algorithm. Using any pretrained model, an uncertainty heuristic, a calibration set, and a user chosen error rate α, the process aims to create a prediction set $\mathbb{C}$ that is valid in the following sense:

$$1-\alpha \leq \mathbb{P}\left(\mathbb{Y}_{\text{test}} \in \mathbb{C}\left(\mathbb{X}_{\text{test}}\right)\right) \leq 1-\alpha+\frac{1}{n+1}, \tag{2.13}$$

with a calibration set of size n. Essentially, for a regression problem, we want to create confidence intervals around our prediction that guarantee the true label sits in within the intervals at a probability of $1-\alpha$. Therefore, conformal prediction promises perfect marginal calibration for the prediction intervals it produces.

This is achieved as follows [Angelopoulos and Bates, 2023]:

1. Identify a heuristic notion of uncertainty using any pre-trained model e.g. softmax, or predicted variance (derived from Deep Ensembles, MCD etc.).

2. Define the score function $s(x, y) \in \mathbb{R}$. (Larger scores encode worse agreement between x and y).

3. Compute $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the calibration scores $s_1 = s\left(X^{(1)}, Y^{(1)}\right), \ldots, s_n = s\left(X^{(n)}, Y^{(n)}\right)$.

4. Use this quantile to form the prediction sets for new examples:

$$\mathbb{C}\left(\mathbb{X}_{\text{test}}\right) = \left\{y : s\left(\mathbb{X}_{\text{test}}, y\right) \leq \hat{q}\right\}.$$

Intuitively, Conformal Prediction observes the distribution of some heuristic uncertainty over a calibration set, making the assumption that the calibration and test sets are from the same distribution i.e. all data is independent and identically distributed (i.i.d.). By examining

the relationship between model error and the uncertainty heuristic, conformal prediction infers the likely error associated with a given level of uncertainty for a sample at test time. Given a user-chosen error rate α, this learned information allows us to assign confidence bounds to predictions.

In recent years, conformal prediction has been integrated with DNNs [Angelopoulos and Bates, 2023; Balasubramanian et al., 2014; Stankeviciute et al., 2021; Zhang et al., 2021] However, despite its strong theoretical guarantees, conformal prediction has limitations. Firstly, the method requires a separate calibration set, reducing the size of the available training data for the model itself. Secondly, if the chosen heuristic measure of uncertainty is poor and contains limited or noisy information about error, the method will devolve towards unhelpfully large confidence intervals.

While conformal prediction is a well-established approach, it is only recently experiencing a resurgence within the medical imaging community, seeing applications in image segmentation [Csillag et al., 2023; Wieslander et al., 2020] and disease classification [Lu et al., 2022a].

2.3 Gaussian Processes (GPs)

Finally, we outline the basics of Gaussian Processes (GPs), a form of Bayesian machine learning that gives us true Bayesian uncertainty estimates. A deep dive into GPs is beyond the scope of this thesis, since we are primarily interested in the application of GPs to the task of landmark localisation. Nevertheless, the following section gives a brief overview for an intuitive understanding of GPs, and the reader is pointed towards Rasmussen and Williams [2006] and Murphy [2013] for extensive coverage.

Overview of Gaussian Processes

Gaussian Processes offer a Bayesian nonparametric approach to machine learning. A GP can be understood as a distribution over functions, where any finite set of function values is jointly Gaussian distributed [Rasmussen and Williams, 2006]. The properties of this distribution over functions is informed by the training data, where predictions are certain near the training data, and less certain far from the training data. Formally, a GP is defined by its mean

function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$, and it is denoted as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{2.14}$$

The mean function represents the expected value of the function at $\mathbf{x}$, and the covariance function (or kernel) encodes the similarity between function values at different inputs. The kernel chosen is dependent on the assumptions we have about the underlying problem, capturing various properties such as smoothness, periodicity, and linearity. For instance, the Matérn kernel provides a generalised class of functions that sits between the very smooth functions and rougher, less regular functions.

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{d}{\rho} \right), \tag{2.15}$$

where d is the distance between two points, Γ is the gamma function, K_ν is the modified Bessel function of the second kind, and ρ and ν are positive parameters of covariance. The parameter ν determines the smoothness. For instance, $\nu = \frac{3}{2}$ provides once differentiable functions, and $\nu = \frac{5}{2}$ gives twice differentiable functions. This kernel can be employed when the underlying function's smoothness is in question and allows for a more flexible model.

Sparse Variational Gaussian Processes (SVGPs)

One of the major challenges with GPs is their computational complexity. In a standard GP, the computational demand for inference scales $\mathrm{O}(n^3)$ with the number of data points, making it challenging to deploy GPs on large datasets. To address this, approximation methods such as sparse variational Gaussian processes (SVGP) have been introduced and worked on [Snelson and Ghahramani, 2005; Titsias, 2009]. Rather than model the entire training dataset, SVGPs utilize a smaller set of inducing points to represent the GP, reducing the computational overhead.

The primary idea is to approximate the true posterior with a variational distribution by introducing M inducing points $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$. The variational distribution is then given by:

$$q(f) = \int p(f|u)q(u)du, \tag{2.16}$$

where u represents the function values at the inducing points. The optimal $q(u)$ is found by minimising the Kullback-Leibler divergence between the true posterior and the variational approximation.

Chapter 3

Prior Work, Data, and a Baseline Case Study

3.1 Related Work

3.1.1 Landmark Localisation Methods

The explosion of progress in the computer vision domain over the recent years has caused a radical change in the methods used for medical imaging analysis, leading deep learning to become the dominant tool for landmark localisation. Before this, many early machine learning approaches used hand crafted graphical models to encode spatial relationships alongside image features [Cootes et al., 1995; Ibragimov et al., 2014; Lindner et al., 2014; Liu et al., 2010]. The inductive biases injected into these approaches, such as spatially constrained models describing how landmarks connected, facilitated impressive performance even when the amount of training data was limited. However, after achieving great success in various computer vision tasks [Krizhevsky et al., 2017; Simonyan and Zisserman, 2014], attention shifted to deep learning for landmark localisation. Utilizing the building blocks outlined in Section 2.1, it was quickly realised that Deep Neural Networks (DNNs) were not only capable of implicitly learning these inductive biases when given enough data, but their additional expressiveness even allowed them to outperform handcrafted models. Early deep learning approaches used Convolutional Neural Networks (CNNs) to directly regress landmark coordinates, achieving enough success to solidify the application of deep learning [Zhang et al.,

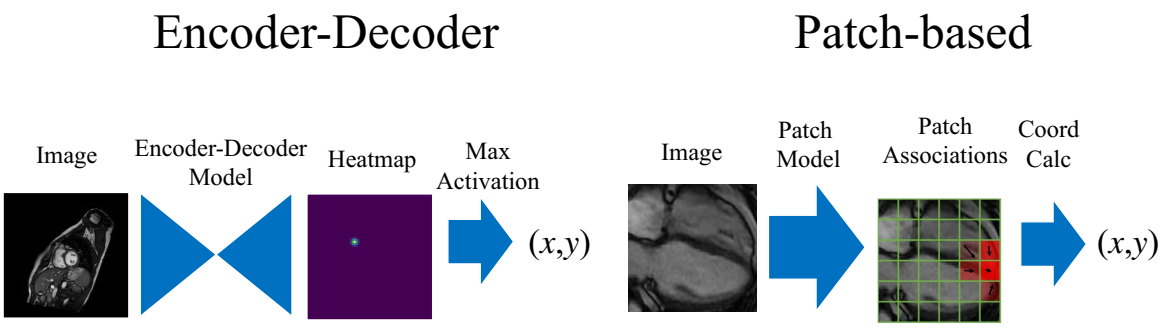


Figure 3.1: A visualisation of encoder-decoder and patch-based methods. Typically, encoder-decoder methods input the entire image and holistically analyse it, regressing a heatmap centred around the target landmark. On the other hand, patch-based methods learn associations between patches of the image and the target landmark. To obtain the final coordinates, the patch-wise predictions are fused.

2017]. However, regressing coordinates directly from an image is a highly non-linear task, and the method's performance limit was quickly realised. Since then, landmark localisation can be broadly categorised into two groups: heatmap regression using encoder-decoder models, and patch-based models, shown in Figure 3.1.

Encoder-Decoder Methods

Inspired by works in pose estimation [Newell et al., 2016; Tompson et al., 2015, 2014], approaches pivoted to formulate the coordinate regression problem as a heatmap estimation problem. Using encoder-decoder style CNNs described in Section 2.1.5 such as U-Nets [Ronneberger et al., 2015] or Hourglass networks [Yang et al., 2017b], a Gaussian heatmap (Equation (2.5)) is predicted for each landmark [Payer et al., 2016, 2019; Tiulpin et al., 2019; Yang et al., 2017a; Zhong et al., 2019]. Regressing heatmaps prove more effective than regressing coordinates, as they offer a smoother supervision, while also allowing some uncertainty in the prediction. There have been various methods using encoder-decoder style models, with some incorporating attention [Zhong et al., 2019], pyramid networks [Chen et al., 2019; Gilmour and Ray, 2020], specialised losses [Chen et al., 2019; Oh et al., 2020], heatmap property regression [Payer et al., 2020; Thaler et al., 2021], multiple stages [Gilmour and Ray, 2020] and most recently transformers [Jiang et al., 2022; Yueyuan and Hong, 2021].

An interesting but less common approach combines coordinate regression and heatmap

prediction into a single multi-task network. Davison et al. [2018] jointly predict the displacement to the landmark alongside the heatmap value for each pixel using an encoder-decoder style architecture. The final coordinate is obtained using a voting scheme, where each pixel's contribution is defined by its heatmap activation strength. This proved to be more robust than simply taking the peak value of the heatmap. Chen et al. [2019] also combine heatmap regression with displacement regression, finding the multi-task approach improves localisation accuracy significantly. Zhou et al. [2021] use a Reinforcement Learning (RL) framework in tandem with encoder-decoder networks to search for the optimal size of Gaussian Heatmap, side-stepping the need to pre-define the variance hyperparameter of Equation (2.5) at the cost of greater computation.

Patch-based Methods

In medical imaging, the number of available training samples is often small so the encoder-decoder network is often forced to be shallow, compromising its performance [Zhang et al., 2017]. One method to overcome this is via a *patch-based* approach; alleviating the problem by sampling many small 'patches' from an image, and predicting the displacement from the patch to the target landmark [Arik et al., 2017; Emad et al., 2015; Li et al., 2018]. This approach can generate orders of magnitude more training samples from a single image compared to the encoder-decoder style methods.

However, it is difficult to take into account high level, global features because each patch makes a prediction based only on local features. For example, the two-stage method by Zheng et al. [2015] first extracts candidate points, and then analyses image patches around these points to obtain the landmark, not considering global contextual information. Zhang et al. [2017] improve this method by first using patches to learn local information, before then using the entire image to learn global information. However, these approaches are costly to train, requiring two training phases. Furthermore, this method regresses coordinates directly, rather than the more effective and smoother heatmap.

Even less common is a patch-based approach that combines displacement regression and heatmaps. Noothout et al. [2018] come close to this method, applying multitask learning using a Fully Convolutional Neural Network (FCN) to jointly perform classification and regression

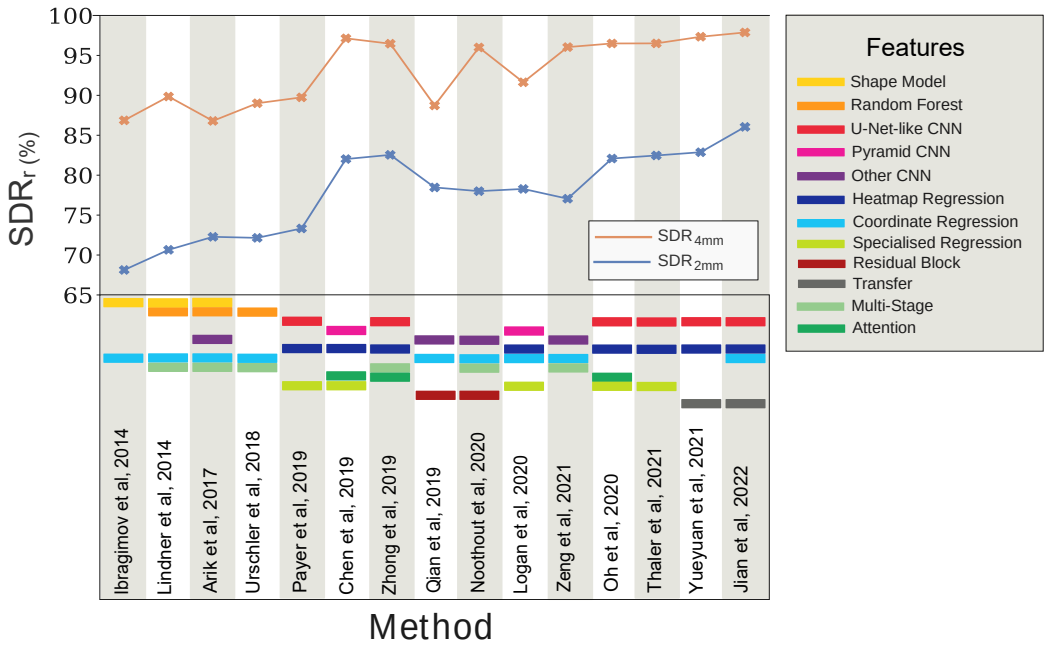


Figure 3.2: Landmark localisation performance on a Cephalometric dataset using the *biased* ISBI 2015 evaluation protocol [Wang et al., 2016]. The success detection rate (SDR_r) shows the percentage of predictions within a $r$mm radius of the target landmark. Features utilised by the methods are indicated by the presence of the coloured bands.

on each patch. The classification task determines whether a patch contains the landmark, and the regression task estimates the 3D displacement from the patch. To dampen the effect of parts of the image distant to the landmark the log displacement was used, meaning the further the patch, the smaller its effect on the loss function. Similar to Donner et al. [2013], *only* the patch classified as containing the landmark is used to determine the final coordinates, filtering out the rest. This multi-task, joint learning leads to a light-weight network and enhanced localisation performance, with the two tasks sharing a feature representation that improves the performance of both [Zhang and Yang, 2018]. However, the resulting network has a strong local focus and is also susceptible to failure if the predicted containing patch is incorrect. In a followup work, Noothout et al. [2020] extend their method into a two stage method: they first train a CNN to provide global estimates for the landmarks, then employ specialised CNNs for each landmark for the final prediction. This method improves upon the first in terms of localisation error, but has the drawback of requiring multiple training stages.

Trends

Figure 3.2 shows performance over time on a public Cephalometric landmark localisation dataset [Wang et al., 2016] (reviewed in Section 3.2.3), breaking down the features of each method. With the exception of a combing an encoder-decoder style architecture (U-Net/pyramid) and heatmap regression, there is no clear correlation between any architectural modification and performance, with most methods tuning parameters and reporting the best results. Even in the case of the best-performing method using a transformer [Jiang et al., 2022], the authors provide results from a hyperparameter search which shows the method has large fluctuations in performance based on implementation details. However, it is clear from this case study that traditional patch-based methods have fallen out of favour, firmly replaced by the more globally focused, large encoder-decoder style models. It is interesting to note that Vision Transformers are a form of patch-based model, albeit more expressive than their CNN cousins due to sequence encoding and attention.

In summary, Deep Neural Networks using supervised learning have concretely set themselves as the State-of-the-Art paradigm to landmark localisation. Specific architectures have closely followed the general trends of Computer Vision research, evolving from simple FCNs, to the Vision Transformer. However as we have seen, little attention has been paid to parameter-efficient models that are cheap to train and small enough to be deployed on low-resource machines.

3.1.2 Uncertainty Estimation in Landmark Localisation

It is necessary to look beyond accuracy if we aspire for large-scale adoption of our models. Uncertainty estimation is vital. Despite the exciting progress in this area in the community, there remains many challenges remaining for uncertainty estimation in landmark localisation. A concentrated effort in uncertainty estimation has been applied to image segmentation by the community, a task similar to landmark localisation that instead aims to predict a mask for an entire structure rather than a single point.

Segmentation aims to produce a binary map, with pixels activated on the structure of interest and inactive elsewhere. In traditional heatmap-based landmark localisation, only the pixel pertaining to the landmark coordinate has a magnitude of 1, smoothly attenuating

to 0 in a set radius. The loss function used by landmark localisation is the Mean Squared Error (MSE) between the target and predicted heatmap, whereas segmentation uses pixel-wise classification-based losses [Jungo et al., 2020]. Nevertheless, the tasks are similar in that the magnitude of the activation of any given pixel in each image can be leveraged for information on the epistemic "confidence" (inverse of uncertainty) of the model. Jungo et al. [2020] use pixel activation to measure the uncertainty of each pixel segmentation class, including using the average activation of an ensemble. They found that this naive approach was surprisingly well calibrated and that ensembling 5+ identical but randomly initialised models significantly improved calibration. These results have been corroborated by Mehrtash et al. [2020] who also used pixel activation for confidence as well as an ensemble of 5 identical, randomly initialised networks to convincing effect.

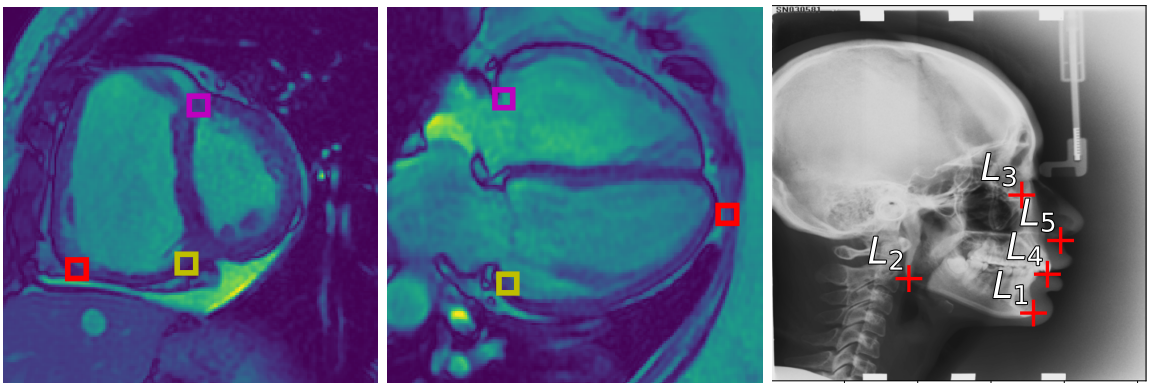
Other successful approaches for epistemic uncertainty estimation in segmentation use Bayesian Neural Networks [Kwon et al., 2020] or Bayesian approximation methods like Monte-Carlo dropout [Nair et al., 2020]. However, the prevailing approach is an ensemble of identical, randomly initialised networks. This method affords better performance [Mehrtash et al., 2020] and a more accurate mechanism for Bayesian marginalisation [Wilson and Izmailov, 2020] compared to a single model using Monte-Carlo dropout.

In landmark localisation we are ultimately predicting a single coordinate point rather than a mask, but similar uncertainty estimation approaches can be utilised. However, there are limited works exploring uncertainty in landmark localisation. Payer et al. [2019] directly modeled aleatoric uncertainty during training by learning the isotropic Gaussian covariances of target heatmaps, and predicting the distribution of likely locations of the landmark at test time. Thaler et al. [2021] took this approach further, learning anisotropic (directionally skewed) Gaussian heatmaps for each landmark, demonstrating that the learned heatmap shapes correspond to inter-observer variability from multiple annotators. However, this method only models the homoscedastic aleatoric uncertainty of the dataset, whereby a single covariance matrix is learned over the entire dataset for each landmark during training. At inference, a Gaussian function is fitted to each individual prediction to model heteroscedastic aleatoric uncertainty, however this measure heavily depends on the learned homoscedastic uncertainty. In terms of epistemic uncertainty, Lee et al. [2020] borrowed from image segmentation

approaches by proposing to use Monte-Carlo dropout to predict the location and subject-level uncertainty of Cephalometric landmarks.

Another method to measure the subject-level, epistemic uncertainty of a heatmap-based landmark prediction is to measure the maximum heatmap activation (MHA) of the predicted heatmap. Since the activation of a Gaussian heatmap at a particular pixel represents the pseudo-probability of the pixel being the landmark, we can use this pseudo-probability as an uncertainty measure: the higher the activation, the more certain the prediction. Drevickỳ and Kodym [2020] compared MHA with ensemble and Monte-Carlo dropout methods, finding MHA surprisingly effective given its simplicity. However, similarly to image segmentation, they found using an ensemble of models was best at predicting uncertainty. They calculated the coordinate prediction variance between an ensemble of models, and found this method performed best at estimating prediction uncertainty. McCouat and Voiculescu [2022] shift the task from regression to classification by learning binary heatmaps rather than Gaussian heatmaps. This shift allowed the authors to calibrate the heatmap activations using Temperature Scaling [Guo et al., 2017], at the cost of reduced accuracy.

Overall, there has been limited exploration into exploiting information from regressed Gaussian Heatmaps, which are prevalent in most State-of-the-Art (SOTA) methods. Given that the Gaussian heatmap regression produces a 2D continuous output, traditional calibration techniques and evaluation metrics tailored for classification aren't immediately suitable. Existing evaluation metrics for uncertainty in landmark localisation are somewhat limited. Often, the primary method is to measure the correlation between an uncertainty measure and localisation error [Drevickỳ and Kodym, 2020; Thaler et al., 2021]. This landscape underscores the need to delve deeper into uncertainty estimation within the Gaussian Heatmap regression framework, alongside improved evaluation metrics for such methods. A final intriguing avenue of exploration is the application of purely Bayesian techniques to landmark localisation, which are notably absent in the literature, likely due to the computational complexity. However, the pursuit of these techniques promises mathematically rigorous and therefore more trustable uncertainty estimation.



(a) SA CMR. (b) 4CH CMR. (c) Cephalometric.

Figure 3.3: **(a)** Landmarks for Short Axis (SA) CMR: Magenta = superior right ventricle insertion point valve; Yellow = inferior right ventricle insertion point; Red = inferior lateral reflection of right ventricle free wall. **(b)** Landmarks for 4 chamber (4CH) CMR: Magenta = tricuspid valve; Yellow = mitral valve; Red = apex of left ventricle. **(c)** Subset of Landmarks included in the Cephalometric dataset [Wang et al., 2016]. Displayed landmarks are a subset of the total 19 landmarks, for better visibility.

3.2 Datasets

We use two modalities of data in this thesis: Cardiac Magnetic Resonance (CMR) images and Cephalometric Radiographs. These modalities are distinct in structure as well as methodology of capture, so are used at different points in this thesis to evaluate our method's generalisability. Table 3.1 details the attributes of the datasets used in this thesis.

3.2.1 ASPIRE Cardiac MRI (Standard): ASPIRE-S

ASPIRE-S is a subset of data from the ASPIRE Registry [Hurdman et al., 2012], with CMR sequences containing a mix of subjects suffering from pulmonary arterial hypertension (PAH) patients and no pulmonary hypertension (PH). Each subject has a four chamber (4CH) view and/or a short axis view (SA). Each CMR sequence has a spatial resolution of 512×512 pixels, where each pixel represents 0.9375mm of the organ, and 20 frames (we use only the first frame for landmark localisation in this thesis). There are 303 SA images, each with three annotated landmarks: the inferior right ventricle insertion point (infSA), the superior right ventricle insertion point (supSA), and the inferior lateral reflection of the right ventricle free wall (RVSA). There are 422 4CH images, each with three annotated landmarks: the apex of the left ventricle at end diastole (LVDEV Apex), the mitral valve (mitral), and

Dataset	Modality	View	Pixel Resolution	Land-marks	Annota-tors	Samples
ASPIRE-S [Hurdman et al., 2012]	CMR	4CH	512 × 512	3	1	422
ASPIRE-S [Hurdman et al., 2012]	CMR	SA	512 × 512	3	1	303
ASPIRE-L [Hurdman et al., 2012]	CMR	4CH	512 × 512	4	1	789
Cephalograms [Wang et al., 2016]	Cephalometric Radiographs	N/A	1935 × 2400	19	1	400
Cephalograms [Wang et al., 2016]	Cephalometric Radiographs	N/A	1935 × 2400	5	19	100

Table 3.1: A summary of datasets used in this thesis. **CMR** is Cardiac Magnetic Resonance, **4CH** is Four Chamber, and **SA** is Short Axis.

tricuspid valve (tricuspid). Annotated examples of the SA and 4CH images are showing in Figure 3.3a and Figure 3.3b, respectively. The 4CH dataset represents a more challenging landmark localisation task as the images have much higher variability than the SA dataset. The landmarks were decided and manually labelled by a radiologist.

3.2.2 ASPIRE Cardiac MRI (Large): ASPIRE-L

ASPIRE-L is a larger subset of the ASPIRE dataset registry [Hurdman et al., 2012], which has 789 CMR images. Again, each image has a resolution of 512×512 pixels, where each pixel constitutes 0.9375mm of the organ. Each subject in this dataset has a four-chamber (4CH) view. Each image has four landmarks: the Left Ventricular Apex (LV), Lateral Mitral Annulus (LMA), Lateral Tricuspid Annulus (LTA) and Spinal Cord (SA).

Although the dataset is larger than ASPIRE-S, this dataset represents a more challenging task since the annotations in the data are noisier and less reliable. Unfortunately, only 1 annotation is available for each landmark, so the aleatoric noise in the labels cannot be measured.

3.2.3 Cephalometric Radiographs

The third and final dataset used in this thesis consists of Cephalometric Radiographs (Cephalograms), which contain 400 images with repetitive structures [Wang et al., 2016]. The dataset

has a total of 19 annotated landmarks, where we use the junior annotator as the ground truth (following the convention of [Lindner et al., 2016; Thaler et al., 2021; Zhong et al., 2019]). For our study of aleatoric uncertainty in Section 5.5.7, we use subset of 5 landmarks which have a total of 11 annotations provided by Thaler et al. [2021] for 100 images. The images have a spatial resolution of 1935×2400 pixels, where each pixel represents 0.1mm of the structure. Fig. 3.3c shows an example image annotated with the aleatoric uncertainty landmark subset.

3.3 Case Study of Baseline Method: LannU-Net

In this section, we provide a brief experiment validating one of the large capacity baseline models we use as a comparison in the main body of the thesis. We show LannU-Net is a representative method of State-of-the-Art landmark localisation models.

As described in Section 3.1.1, recent developments demonstrate consistent but increasingly incremental improvements in landmark localisation on medical images. However, much of this progress is benchmarked on a Cephalometric dataset (Section 3.2.3) using a known biased evaluation protocol [Lindner et al., 2016; Payer et al., 2020; Thaler et al., 2021]. In this short study, we perform a litmus test on the validity of the reported progress by comparing State-of-the-Art methods with a vanilla U-Net in an unbiased evaluation regime. We find our model performs comparably or better than tailored architectures with sophisticated modifications. Specifically, we achieve fewer gross mispredictions than the previous State-of-the-Art method for Cephalometric landmark localisation, and show that a vanilla U-Net using data augmentation is robust in environments with limited training data.

3.3.1 Methods

Inspired by nnU-Net [Isensee et al., 2021] for image segmentation, we hypothesise that a vanilla U-Net with a careful training regime contains enough expressive power to perform on par with tailored architectures.

Architecture We design a vanilla U-Net inspired by the template of the original U-Net [Ronneberger et al., 2015] and nnU-Net [Isensee et al., 2021], we call LannU-Net. The LannU-Net follows the standard configuration of two blocks per resolution layer, with each block

consisting of a 3×3 convolution, Instance Normalisation [Ulyanov et al., 2016], and Leaky ReLU (negative slope, 0.01). Downsampling is achieved through strided convolutions and upsampling through transposed convolutions. The initial number of feature maps is set to 32, doubling with each downsample to a maximum of 512 and halving at each upsample step. We automatically configure the number of resolution layers by adding encoder steps until any dimension of the feature map resolution hits a minimum of 4.

Training scheme We use the standard landmark localisation objective function: the Mean Squared Error (MSE) between the Gaussian target heatmap and the predicted heatmap (σ is a hyperparameter representing the standard deviation of the 2D Gaussian function, Equation (2.5)). The MSE loss function is detailed in Section 2.1.6, Equation (2.6). We implement deep supervision, injecting losses at every resolution of the network except the lowest two. This technique facilitates a coarse localisation at lower feature resolutions, achieving similar effects to the attention mechanism in Zhong et al. [2019] and *Spatial Configuration* component of Payer et al. [2019]; Thaler et al. [2021]. Unlike the patch-based sampling of nnU-Net, we force images to be a size no larger than 512×512 and perform training and inference on the entire down-sized image.

Evaluation regime We evaluate localisation performance using point-to-point error, defined by the Euclidean distance/Frobenius norm from a predicated coordinate $\hat{\mathbf{c}}$, to a target coordinate $\tilde{\mathbf{c}}$:

$$\mathcal{D}_{\mathrm{PE}}(\hat{\mathbf{c}} - \tilde{\mathbf{c}}) = ||\hat{\mathbf{c}} - \tilde{\mathbf{c}}||_F. \tag{3.1}$$

We also measure the Success Detection Rate (SDR), the percentage of predicted landmarks within a defined point-to-point error radius:

$$SDR_r = \frac{100}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (\mathcal{D}_{\mathrm{PE}}(\hat{\mathbf{c}}^{(i,j)} - \tilde{\mathbf{c}}^{(i,j)}) \leq r), \tag{3.2}$$

where M is the number of images, each with N landmarks and $\widetilde{\boldsymbol{C}}$ is the tensor encoding the coordinate labels over all images and landmarks, and $\widehat{\boldsymbol{C}}$ is the corresponding predicted tensor. SDR is the standard evaluation regime for this dataset, facilitating easy comparison to

Method	PE (mm) mean ± std	SDR_r 2mm	2.5mm	3mm	4mm
Lindner et al. [2016]	1.2 ± NA	84.7 %	89.38 %	92.62 %	96.3 %
Zhong et al. [2019]	1.22 ± 2.45	86.06 %	90.84 %	94.04 %	97.28 %
Gilmour and Ray [2020]	1.07 ± 0.95	86.72 %	92.03 %	94.93 %	97.82 %
Thaler et al. [2021]	**0.99 ± 1.07**	**89.7 %**	**93.74 %**	**95.83 %**	97.82 %
Ours ($\sigma = 8$)	1.15 ± 1.39	86.47 %	92.38 %	95.46 %	**98.11 %**
Ours ($\sigma = 2$)	1.28 ± 1.23	82.76 %	89.21 %	93.12 %	96.62 %
Ours ($\sigma = 3$)	1.21 ± 1.46	85.43 %	91.04 %	94.41 %	97.50 %
Ours ($\sigma = 5$)	1.17 ± 2.00	86.74 %	92.32 %	95.37 %	98.03 %
Ours ($\sigma = 8$)	1.15 ± 1.39	86.47 %	92.38 %	95.46 %	98.11 %
Ours ($\sigma = 10$)	1.17 ± 1.74	86.29 %	92.34 %	95.43 %	98.03 %
Ours ($\sigma = 12$)	1.19 ± 1.47	85.88 %	91.92 %	95.29 %	97.93 %
Ours ($\sigma = 15$)	1.19 ± 1.18	85.34 %	91.99 %	95.26 %	98.14 %
Ours ($T = 25$ %, $\sigma = 8$)	1.38 ± 2.93	82.46 %	89.29 %	93.55 %	97.17 %
Ours ($T = 50$ %, $\sigma = 8$)	1.23 ± 2.23	85.75 %	91.58 %	94.79 %	97.88 %
Ours ($T = 75$ %, $\sigma = 8$)	1.17 ± 1.12	86.24 %	91.99 %	95.36%	97.92 %

Table 3.2: Localisation results from the Cephalometric dataset [Wang et al., 2016] over a 4-fold CV. The point-to-point error (PE) is reported in mm, alongside the success detection rate (SDR_r). We show the results of our method using different percentages of the available training data (T), and varying the size of the Gaussian target heatmap function (σ). **Bold** indicates best results, underlining indicates second-best.

other methods. We consider SDR_{2mm} a measure of the precision of a model, and SDR_{4mm} a measure of robustness against outliers i.e. how many samples are not gross mispredictions.

We train models using 100%, 75%, 50% and 25% of the available training data to simulate the effect of limited datasets.

3.3.2 Dataset

We show results on the publicly available Cephalometric dataset [Wang et al., 2016], outlined in Section 3.2.3. We report results of a 4-fold cross validation (CV) over all 400 images, using the junior annotations. We intentionally do not report results on the standard ISBI 2015 Challenge regime [Wang et al., 2016] since there is a systematic shift in annotation between the training set and test set, which has been widely reported [Lindner et al., 2016; Payer et al., 2020; Thaler et al., 2021]. For each fold, we select a random 10% of that fold's training

data as a validation set.

3.3.3 Experiments and Results

Experimental Setup and Training Details

We train for 500 epochs using Stochastic gradient descent with an initial learning rate of 0.01, decaying it using the "poly" scheme, $(1 - epoch/epoch_{max})^{0.9}$ [Chen et al., 2017]. One epoch consists of 150 mini-batches, where each mini-batch is 12 samples. We employ early stopping using a hold-out validation set (10% of training set), stopping training if the validation set's localisation error does not drop for 150 epochs. We employ data augmentations with a probability of 0.5, uniformly sampling from a continuous range [α, ω]: Random scaling [0.8, 1.2], translation [-0.07%, 0.07%], rotation [-45°, 45°], shearing [-16, 16] and vertical flipping.

Localisation Results

Table 3.2 shows LannU-Net produces fewer gross mispredictions than the previous approaches, achieving a SDR_{4mm} result of 98.11% - a slight improvement over 97.82% [Thaler et al., 2021]. However, our method is not the best in terms of precision, with an SDR_{2mm} of 86.47% compared to 89.7% from the approach by Thaler et al. [2021]. Overall, we achieve better or comparable results to tailored architectures [Gilmour and Ray, 2020; Thaler et al., 2021] and an approach using an attention mechanism [Zhong et al., 2019].

Our model is trained on images of size 512×512, and we upscale the resulting heatmap to the original image size of 1935×2400. Consequently, a localisation error of a single pixel in the model output propagates to a $\sim$4.2 pixel error (0.42mm) in the final upscaled prediction. Therefore, due to the convincing performance in the coarser SDR_{4mm} evaluation metric, we hypothesise that the slightly worse performance in SDR_{2mm} is in part caused by the quantisation error from transforming the low resolution output image to the final coordinate prediction.

Table 3.2 also shows results varying the value of σ in our target heatmap. When the performance of recent models is so close, this hyperparameter alone has a significant impact on the standings of our method. It would be reasonable to expect that an extensive hyperparameter search would achieve SOTA on this dataset. Finally, Table 3.2 shows the performance

of our model under the constraint of smaller training set sizes. We show that even with 25% of the training data, our vanilla U-Net performs surprisingly robustly.

3.3.4 Conclusion

In summary, our results suggest that a vanilla U-Net holds enough expressive power for the task of landmark localisation, held back in terms of precision by the bottleneck of the input image size. For the purpose of this thesis, we will use LannU-Net to represent a baseline, high capacity model.

Chapter 4

PHD-Net: Lightweight Landmark Localisation with Uncertainty

4.1 Introduction

In this chapter, our primary focus revolves around tackling **Q1**, as outlined in Section 1.2, which centers on enhancing lightweight models with limited data availability. To this end we propose **PHD-Net**: a lightweight, multi-task **P**atch-based network combining **H**eatmap and **D**isplacement regression. We propose two strategies to fuse the network's outputs to generate a final coordinate prediction. We evaluate PHD-Net on hundreds of Short Axis and Four Chamber Cardiac Magnetic Resonance (CMR) images, showing promising results. Using a calibration set with our branch fusion strategies, we present preliminary heuristics to categorise predictions as high or low uncertainty. We also show that our patch-based training regime scales with model capacity.

In more detail, we address three core challenges in this chapter. First, we confront the challenge of limited data through the training regime itself by using patch-based sampling. As discussed in Chapter 2, an almost universal observation has been made in the literature: more data means better results [Sun et al., 2017]. Unfortunately, this shortcut to success is hard to follow in the domain of medical imaging since datasets are often limited in size due to ethical grounds and the costly process of expert annotations.

Second, a key challenge for landmark localisation in medical imaging is the prevalence of

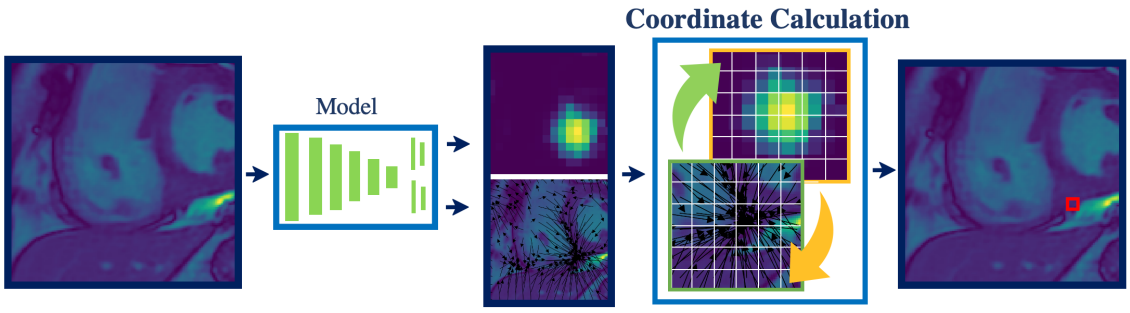


Figure 4.1: General Framework of the proposed PHD-Net. The image (cropped for clarity) is passed to a multi-branch network, predicting a heatmap and displacement value (the black arrows) for each patch. These are then combined with *Adaptive Prediction* or *Candidate Smoothing* to produce the final coordinates and associated uncertainty value.

locally similar structures in the image, leading to misidentifications of the landmark [Thaler et al., 2021]. This challenge is compounded when training with patch-based sampling compared to full image sampling, since identifiers that would distinguish similar structures are not always present in a given patch of the image.

The final challenge we touch on in this chapter is the need for clinicians to distinguish between *high uncertainty* and *low uncertainty* predictions [Tonekaboni et al., 2019]. As discussed in Section 3.1, this capability is largely absent in existing solutions. The preliminary work on uncertainty in this chapter lays the groundwork for the subsequent investigations in Chapter 5.

4.2 Contributions

To address the aforementioned challenges, we require a compact, uncertainty-estimating model that can learn rich feature representations while efficiently making use of the training data available. To this end, we propose a **P**atch based method that combines **H**eatmap and **D**isplacement regression: PHD-Net. Our contributions, outlined in Figure 4.1 are four-fold:

1. We improve the capabilities of patch-based approaches to landmark localisation. We build on Noothout et al. [2018], proposing a multi-task patch-based framework in which

one branch of the network focuses on generating *locally accurate* candidate predictions, regularised by another branch focusing on the *globally likely* landmark location using heatmap regression.

2. We propose and evaluate two branch-fusion strategies, *Adaptive Prediction* and *Candidate Smoothing*, better leveraging the outputs of both branches to generate the final landmark coordinates compared to the baseline approach of Noothout et al. [2018].

3. We reveal and exploit a useful property of the mappings produced by *Candidate Smoothing* which allows us to estimate the *predictive uncertainty* of a prediction. We use a Frequentist approach to categorise predictions into *high uncertainty* and *low uncertainty* groups. We also use a Frequentist approach to categorise by uncertainty when using *Adaptive Prediction*.

4. We demonstrate our patch-based multi-task training regime scales with model capacity, experimentally showing that localisation accuracy improves with increases to model parameters. Our study indicate our proposed approach is a high-performing alternative to traditional Gaussian heatmap regression, with the added benefit of a unique, patch-based notion of uncertainty.

We compare our proposed method to several State-of-the-Art (SOTA) methods. During analysis, we not only compare localisation accuracy of the approaches, but we also use our comparison methods to explore the subtleties of the uncertainty categorisations given by *Candidate Smoothing* and *Adaptive Prediction*, discussing whether they are identifying aleatoric or epistemic uncertainty. Furthermore, we scrutinise our patch-based multi-task framework using a diversity of model architectures, including a Vision Transformer [Dosovitskiy et al., 2020]. We demonstrate our method scales well, with performance improving as we increase model capacity.

We provide an open-source implementation of the models presented in this work at `https://github.com/Schobs/MediMarker`.

4.3 Methods

Section 4.3.1 describes the patch-based multi-task FCN design and architecture, PHD-Net. Section 4.3.2 outlines our *Adaptive Prediction* and *Candidate Smoothing* strategies to fuse the model outputs for the landmark coordinates. Section 4.3.3 describes how we leverage the mappings produced by *Adaptive Prediction* and *Candidate Smoothing* to estimate the uncertainty of the prediction. Finally, Section 4.3.4 outlines the evaluation metrics to assess the landmark localisation accuracy used in this chapter.

4.3.1 PHD-Net: The Patch-based Multi-Task Network

In the multi-task FCN of Noothout et al. [2018], the regression and classification tasks share parameters in the convolutional layers. The network processes images patch-wise, with the regression task predicting the log-transformed 2D displacement from the centre of each patch to the landmark location. The classification task predicts whether the landmark is contained in the patch or not using a binary map. During training, subimages are randomly sampled from the image and used as training samples. In testing, the whole image is taken as input, and the patch with the highest classification score is used to calculate the landmark's predicted location [Noothout et al., 2018].

In PHD-Net, we formulate the model in a similar fashion, with the following key differences:

- **Heatmap regression**: Instead of considering the classification task as binary, we instead regress a Gaussian heatmap centered around the landmark-containing patch, providing smoother supervision [Payer et al., 2019].
- **Weighted Loss**: We further suppress the influence of distant patches by weighting the displacement regression branch loss by the Gaussian heatmap label. This places more importance on patches close to the landmark, which are more informative compared to distant patches.

Our key idea is to consider the pixel-precise predictions from the displacement branch as *locally accurate* candidate predictions. However, since each candidate prediction is only informed by the small area in its patch, there may be misidentifications caused by locally

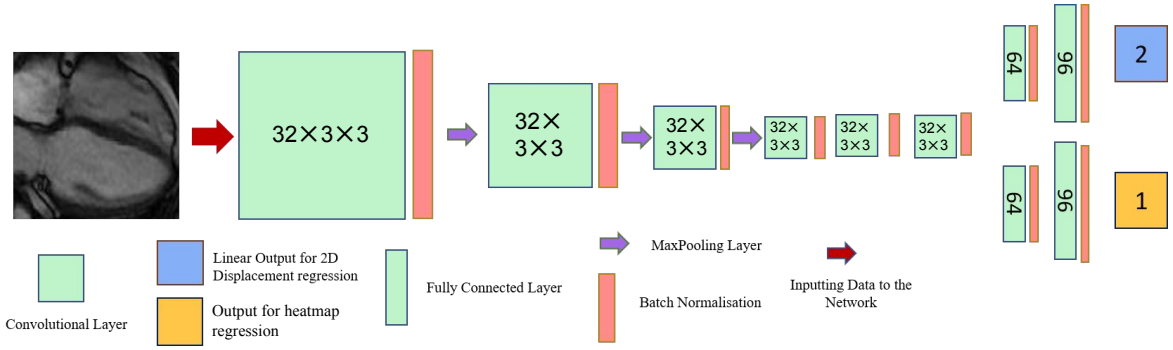


Figure 4.2: The network architecture of the proposed PHD-Net. The image is analysed patch-wise to produce two predictions for each patch: the displacement to the landmark, and a heatmap value.

similar structures in the image. To regularise this, we leverage the prediction from the heatmap branch, which indicates the *globally likely* location for the landmark.

In testing, PHD-Net takes as input the entire image and analyses it in a patch-wise manner. We present the details below.

Architecture

Figure 4.2 shows the architecture of PHD-Net adopted from [Noothout et al., 2018], composing three convolutional layers of 32 filters with 3×3 kernels, each followed by a batch normalisation and a 2×2 maxpooling layer. Now, each 1×1 feature in the current representation corresponds to an 8×8 patch in original input space. Thus the size of each patch, denoted as a two-dimensional vector $\mathbf{s} = (s^{(x)}, s^{(y)})$, is directly determined by the number of maxpooling layers in the network architecture. As the feature dimensionality decreases due to the pooling, the pixel space each feature now represents increases. Given a 2×2 maxpooling operation, each component of the patch size, $\mathbf{s}$ is:

$$s^{(x)} = s^{(y)} = 2^z \text{ pixels}, \tag{4.1}$$

where z is the number of maxpooling layers. Since the network is fully convolutional any image size can be used, as long as the input image dimensions are a factor of the patch size.

After these layers a further 3 convolutional layers with the same properties as the previous ones are added, followed finally by two branching sets of two fully connected layers with 64 and

96 filters. These are modelled as convolutional layers with 1×1 kernels. Each convolutional layer employs batch normalisation and the Rectified Linear Unit (ReLU) activation function. The output of the first branch is for the regression, employing a linear activation function to output the log-transformed 2D displacement, and the output from the second is passed through a sigmoid function to produce the heatmap value. The total number of parameters of PHD-Net is 0.06M.

Joint Displacement and Heatmap Regression

To learn a richer, more robust feature representation we use a multi-task approach, predicting two attributes that we can leverage to obtain coordinates for the landmark. Each input image to PHD-Net outputs two predictions for each patch: the heatmap value and the log-transformed displacement from the centre of the patch to the landmark.

To best learn an expressive feature representation of the image, weights are shared at the beginning of the model, before branching to separate additional layers to generate the two predictions. This provides the model with two opportunities to discover the landmark: the displacement branch focuses on generating pixel-precise candidate coordinates, and the classification branch focuses on the more coarse object-detection task. Framing the task in this fashion facilitates predictions that are pixel-precise despite the output map's low resolution compared to the full image (due to patch-wise predictions, not pixel-wise). The total loss $\mathcal{L}_A$, consists of the displacement loss $\mathcal{L}_d$ and the heatmap loss $\mathcal{L}_h$:

$$\mathcal{L}_A = \mathcal{L}_d + \mathcal{L}_h. \tag{4.2}$$

Heatmap Regression ($\mathcal{L}_h$):

The heatmap branch of PHD-Net focuses on the sub-task of a coarse prediction, generating a probability map over the patches. When producing the final landmark location, this coarse prediction can be used to regularise the more precise but potentially ambiguous candidate locations from the displacement branch. Following the popular encoder-decoder style models [Newell et al., 2016; Ronneberger et al., 2015], we generate a patch-wise Gaussian heatmap: the mean, $\boldsymbol{\mu}$, is the matrix indices of the patch containing the landmark, with a predefined standard deviation σ. For a landmark L_i with coordinates $\tilde{\mathbf{c}}^{(i}$ contained in the patch with

matrix indices $\tilde{\mathbf{p}}^{(i)} \in \mathbb{R}^2$, the 2D Gaussian heatmap image is defined as the 2D Gaussian function: $g_i\left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \tilde{\mathbf{p}}^{(i)}; \sigma\right) : \mathbb{R}^2 \to \mathbb{R}$:

$$g_i\left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \tilde{\mathbf{p}}^{(i)}; \sigma\right) = \frac{1}{(2\pi)\sigma^2} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_2^2}{2\sigma^2}\right), \tag{4.3}$$

where $\mathbf{x}$ is the vector of each patch's position in the matrix and σ is a hyperparameter. The Gaussian heatmap naturally assigns the peak value to the patch containing the landmark, with values smoothly attenuating over the patches with distance. Compared to a binary map, this method provides a smoother supervision, with each patch's value now representing a psuedo-probability of the landmark being contained in it.

Let $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$ denote the predicted patch heatmap produced by our model for landmark L_i. The classification loss for landmark L_i is then the mean squared error (MSE) between $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$ and the target heatmap $g_i\left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \tilde{\mathbf{p}}^{(i)}; \sigma\right)$:

$$\mathcal{L}_h = \left\| h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}) - g_i\left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \tilde{\mathbf{p}}^{(i)}; \sigma\right) \right\|_F^2. \tag{4.4}$$

Displacement Regression ($\mathcal{L}_d$):

This sub-task of the model concentrates on predicting precise candidate landmark locations. Despite the trend of forgoing the use of regressing coordinates in landmark detection tasks in favour of pure heatmap regression, pushing the model to predict both forces it to learn a richer feature representation of the image for the task. Given an entire high-resolution image, the task of regressing a single coordinate value is extremely non-linear and complex. However, the task becomes much simpler when given a small patch of an image that contains the landmark. The further the patch is from the landmark, the lower its predictive power. Therefore, we ask the model to predict the displacement from the centre of each patch to the landmark, but we dampen the affect of distant patches in two ways: (1) We apply the log function to the displacement labels and (2) We weight closer patches as more important than distant patches by multiplying the error of the patch-wise predictions by a Gaussian heatmap centered around the landmark.

Put simply, the regression loss is the sum of the weighted mean squared error (WMSE) between the predicted 2D log displacement and the ground truth 2D log displacements of

each patch, weighted by a Gaussian heatmap with variance σ centered around the annotated landmark:

$$\mathcal{L}_r = \frac{1}{2P} \sum_{j=1}^{P} \left(g_i \left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \tilde{\mathbf{p}}^{(i)}; \sigma \right)^{\mathbf{p}^{(j)}} \left\| \tilde{\mathbf{d}}^{(j)} - \hat{\mathbf{d}}^{(j)} \right\|_F^2 \right), \tag{4.5}$$

where $\mathbf{x}$ is the patch matrix indices vector, P is the number of patches (flattened), and $\mathbf{p}^{(j)}$ is matrix indices vector of the jth patch. $\widetilde{\mathbf{D}}$ and $\widehat{\mathbf{D}}$ are the matrices encoding the actual and predicted log displacement vectors, respectively. Here, $\tilde{\mathbf{d}}^{(j)}$ and $\hat{\mathbf{d}}^{(j)}$ are the actual and predicted displacement vectors from the centre of patch j to landmark coordinates $\tilde{\mathbf{c}}^{(i)}$, which is contained in the patch with matrix indices $\tilde{\mathbf{p}}^{(i)}$.

4.3.2 Landmark Coordinate Retrieval

After designing our multi-task network to predict patch-wise heatmap and displacement values, we need to process the outputs to obtain the landmark's coordinates. First, we define strategies to obtain the landmark from a single branch. Then, we outline a baseline approach to combine both branch outputs [Noothout et al., 2018]. We propose a data-driven approach, *Adaptive Prediction*, which learns which patches to use for the final prediction using a validation set. We also propose *Candidate Smoothing*, which uses all patches and does not require a validation set. Finally, we show a data-driven Frequentist heuristic for both *Adaptive Prediction* and *Candidate Smoothing* strategy to estimate the uncertainty of the prediction.

Single Branch Strategies

Displacement Only: We define a method to obtain predicted coordinates, $\hat{\mathbf{c}}^{(i)}$, from the regression branch *only*. Since the model predicts log-transformed displacements, we first transform these predictions back to the original scale by taking the exponential of the predicted log displacement. We then obtain the final coordinates by calculating a weighted average of all the patch predictions. Each patch's contribution is weighted inversely to the magnitude of its predicted displacement:

$$\hat{\mathbf{c}}^{(i)} = \sum_{j=1}^{P} \frac{\mathbf{a}^{(j)} + e^{\hat{\mathbf{d}}^{(j)}}}{\|e^{\hat{\mathbf{d}}^{(j)}}\|_F \left(\sum_{k=1}^{P} \frac{1}{\|e^{\hat{\mathbf{d}}^{(k)}}\|_F} \right)}, \tag{4.6}$$

for P patches, where $\mathbf{a}^{(j)}$ is the vector of the centre of patch j, $\hat{\mathbf{d}}^{(j)}$ is the predicted log-transformed displacement vector from the centre of patch j to the landmark L_i, and $e^{\hat{\mathbf{d}}^{(j)}}$ is the displacement in the original scale. The terms $\|e^{\hat{\mathbf{d}}^{(j)}}\|_F$ and $\|e^{\hat{\mathbf{d}}^{(k)}}\|_F$ represent the Euclidean/Frobenius norms (or magnitude) of the predicted displacement vectors in their original scale. Each patch contributes to the final prediction, but the influence of patches with larger displacement magnitudes is dampened in the final prediction.

Heatmap Only: We also define a method to obtain the coordinates, $\hat{\mathbf{c}}^{(i)}$, from the heatmap branch *only*. We simply return the centre of the patch with the highest activation from the classification output, $\hat{\mathbf{a}}^{(i)}$. The precision of this method is constrained to the resolution of the patch-size, $\mathbf{s}$, unable to produce pixel-precise predictions. First, we retrieve the predicted patch indices:

$$\hat{\mathbf{p}}^{(i)} = \arg\max_{\mathbf{x}} h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}). \tag{4.7}$$

Next, we obtain the centre of the predicted patch in coordinate space:

$$\hat{\mathbf{c}}^{(i)} = \hat{\mathbf{a}}^{(i)} = \left(\hat{\mathbf{p}}^{(i)} \times \mathbf{s} \right) + \frac{\mathbf{s}}{2}, \tag{4.8}$$

where $\mathbf{s}$ is the patch-size, described in Equation (4.1)

Multi-branch Strategies

Baseline: Adopted from [Noothout et al., 2018], our baseline multi-branch strategy to calculate the final coordinates $\hat{\mathbf{c}}^{(i)}$ is as follows:

1. Identify the patch with the highest heatmap score.
2. Add the displacement prediction from the identified patch to the patch's centre coordinates to obtain the final landmark location:

$$\hat{\mathbf{c}}^{(i)} = \hat{\mathbf{a}}^{(i)} + e^{\hat{\mathbf{d}}^{\left(\hat{\mathbf{p}}^{(i)}\right)}}, \tag{4.9}$$

where $\hat{\mathbf{a}}^{(i)}$ is the centre of the peak patch (Equation. (4.8)) and $e^{\hat{\mathbf{d}}^{\left(\hat{\mathbf{p}}^{(i)}\right)}}$ is the inverse log-transformed predicted displacement from this patch to the landmark (indexed by $\hat{\mathbf{p}}^{(i)}$, Equation (4.7)).

Adaptive Prediction Strategy: The central idea behind *Adaptive Prediction* is to learn which patches should be used for the final prediction. Since some landmarks are more difficult to localise than others, each landmark should adapt the patch selection process. If the landmark is difficult to localise, the prediction should be regularised by using more patches, whereas if it is easy to localise we should eliminate noise by using only the predicted containing patch.

Therefore, for each landmark, during the inner loops of cross validation while training, PHD-Net learns an uncertainty threshold (T), between 0 and 1, and a parameter representing the number of selected patches (P) through a grid search, optimising for the minimum localisation error on the validation (calibration) set. This algorithm is shown in Algorithm. 4.1. If a patch's heatmap activation exceeds T, the patch should either contain the landmark or be part of the same visually similar anatomical structure, so is used to calculate the landmark's position. If no patch exceeds T, then the model failed to confidently find the patch the landmark is contained in, and the top classifying P patches are used. When multiple patches are selected, a weighted average based on the inverse of their predicted displacements is calculated (see Equation (4.6)).

Algorithm 4.1 Grid Search for Optimal T and P

procedure GRIDSEARCH($PHDNet$, $\mathbb{S}_{\text{valid}}$, $\mathbb{T}, \mathbb{P}$)
  $\delta^p_{\min} \leftarrow \delta^t_{\min} \leftarrow \infty$
  $T_{\text{best}} \leftarrow P_{\text{best}} \leftarrow None$
  for $T \in \mathbb{T}$ **do**
    for $P \in \mathbb{P}$ **do**
      $\delta^p_{curr} = \delta^t_{curr} = 0$
      for $\mathbf{X}, \mathbf{y} \in \mathbb{S}_{\text{valid}}$ **do**
        $\boldsymbol{O}_{disp}, \boldsymbol{O}_{class} = PHDNet(\mathbf{X})$
        $\hat{\mathbf{y}}_p = \text{PredictionByP}(\boldsymbol{O}_{disp}, \boldsymbol{O}_{class}, P)$ ▷ Using Algorithm 4.2
        $\hat{\mathbf{y}}_t = \text{PredictionByT}(\boldsymbol{O}_{disp}, \boldsymbol{O}_{class}, T)$ ▷ Using Algorithm 4.3
        $\delta^p_{curr} = \delta^p_{curr} + \mathcal{D}_{\text{PE}}(\mathbf{y}, \hat{\mathbf{y}}_p)$ ▷ Using Equation (4.12)
        $\delta^t_{curr} = \delta^t_{curr} + \mathcal{D}_{\text{PE}}(\mathbf{y}, \hat{\mathbf{y}}_t)$ ▷ Using Equation (4.12)
      end for
      $\delta^p_{avg} = \frac{\delta^p_{curr}}{\|\mathbb{S}_{\text{valid}}\|}$,
      $\delta^t_{avg} = \frac{\delta^t_{curr}}{\|\mathbb{S}_{\text{valid}}\|}$,
      if $\delta^p_{avg} < \delta^p_{min}$ **then**
        $\delta^p_{min} = \delta^p_{avg}$
        $P_{\text{best}} = P$
      end if
      if $\delta^t_{avg} < \delta^t_{min}$ **then**
        $\delta^t_{min} = \delta^t_{avg}$
        $T_{\text{best}} = T$
      end if
    end for
  end for
  return $T_{\text{best}}, P_{\text{best}}$
end procedure

Algorithm 4.2 Generate Coordinates from PHD-Net outputs by P

procedure PREDICTIONBYP($\boldsymbol{O}_{disp}, \boldsymbol{O}_{class}$, P)
  coords_list ← empty_list()
  $\boldsymbol{O}^{sort}_{class}$ ← sort_descending($\boldsymbol{O}_{class}$) ▷ Sort $\boldsymbol{O}_{class}$ in descending order of activations.
  for i in range(P) **do**
    $h \leftarrow \boldsymbol{O}^{sort}_{class}[i]$
    $index \leftarrow$ get_index_of_element($\boldsymbol{O}_{class}, h$)
    coords_list.append(get_coords($h, \boldsymbol{O}_{disp}[index]$)) ▷ Using Equation (4.9).
  end for
  coordinates ← weighted_average(**coords_list**) ▷ Using Equation (4.6).
  return coordinates
end procedure

Algorithm 4.3 Generate Coordinates from PHD-Net outputs by T

```
procedure PredictionByT(O_disp, O_class, T)
    coords_list ← empty_list()
    for i in range(|O_class|) do
        t ← O_class[i]
        if t ≥ T then
            coords_list.append(get_coords(t, O_disp[i]))        ▷ Using Equation (4.9).
        end if
    end for
    coordinates ← weighted_average(coords_list)                 ▷ Using Equation (4.6).
    return coordinates
end procedure
```

Candidate Smoothing Strategy: Despite *Adaptive Prediction*'s ability to tailor a strategy for each landmark, it still only takes into account a limited number of patches. Therefore, we also propose the *Candidate Smoothing* strategy. The central idea behind this strategy is to use a large number of patches to produce *locally precise* but ambiguous candidate predictions, which are then regularised to filter out the *globally unlikely* locations. The process is shown in Figure 4.3.

First, we find the $V \times B$ area of the image which correlates to the part of the Gaussian heatmap with the largest centre of mass (i.e. the area with the largest summed activations). The choice of V and B is for the user, with the ideal values being the resolution of the full image, but smaller values being less computationally expensive. In this study, we choose $V = B = 128$, for greater computational speed. Second, for every patch contained in this area, we plot the prediction from the displacement branch as a small Gaussian blob with a standard deviation of 1. The mapping is additive, meaning if multiple patch's predictions overlap, the heatmap values add on to eachother.

This produces a $V \times B$ Candidate Mapping for landmark L_i containing pixel-precise candidate locations for the landmark, $\mathbf{C}^{(i)}_{map}$:

$$\mathbf{C}^{(i)}_{map} = \sum_{j=1}^{P} g_j \left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \mathbf{a}^{(j)} + e^{\hat{\mathbf{d}}^{(j)}}; \sigma = 1 \right), \tag{4.10}$$

where $\mathbf{x}$ is the coordinate vector of each pixel in the $V \times B$ subimage, $\mathbf{a}^{(j)}$ is the centre of the patch j, $e^{\hat{\mathbf{d}}^{(j)}}$ is the inverse log predicted displacement from patch j to the landmark L_i and

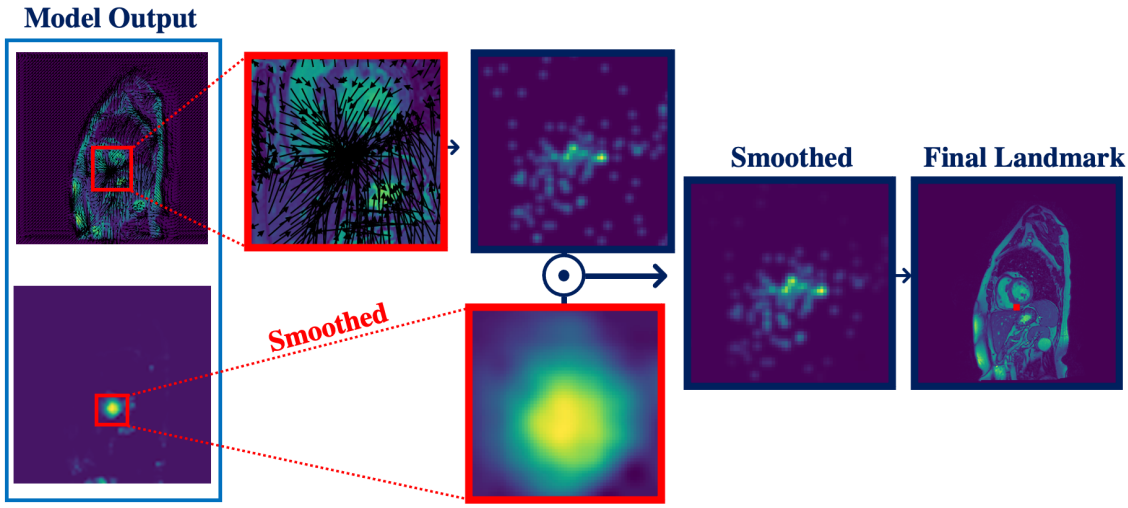


Figure 4.3: **Candidate Smoothing** method to produce a final prediction from model outputs. First, we isolate the part of the image with the highest heatmap activations. We additively map each displacement prediction (black arrows) in this area as a small Gaussian blob. This mapping is multiplied by the upsampled and smoothed predicted Gaussian heatmap. The final coordinate is obtained by taking the peak activation in the new mapping. Note the suppressed activations in the final mapping.

P is the number of patches. The candidate points are precise to a local degree, but since each patch predicts a location blind to its surroundings, it can fail due to locally similar structures.

To solve this, we next smooth these predictions by multiplying the mapping with the up-sampled corresponding Gaussian heatmap predicted by the heatmap branch, $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$ to create a Candidate Smoothed Map:

$$\mathbf{C}^{(i)}_{smooth} = \upsilon(h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})) \odot \mathbf{C}^{(i)}_{map}, \tag{4.11}$$

where υ is the upsampling function (bilinear interpolation). Multiplying the mapping by the predicted Gaussian heatmap suppresses the globally unfeasible predictions determined by the classification branch, while retaining pixel-precise predictions from the regression branch.

To obtain the final predicted coordinate value, $\hat{\mathbf{c}}^{(i)}$, we take the coordinates of the pixel with the highest heatmap activation.

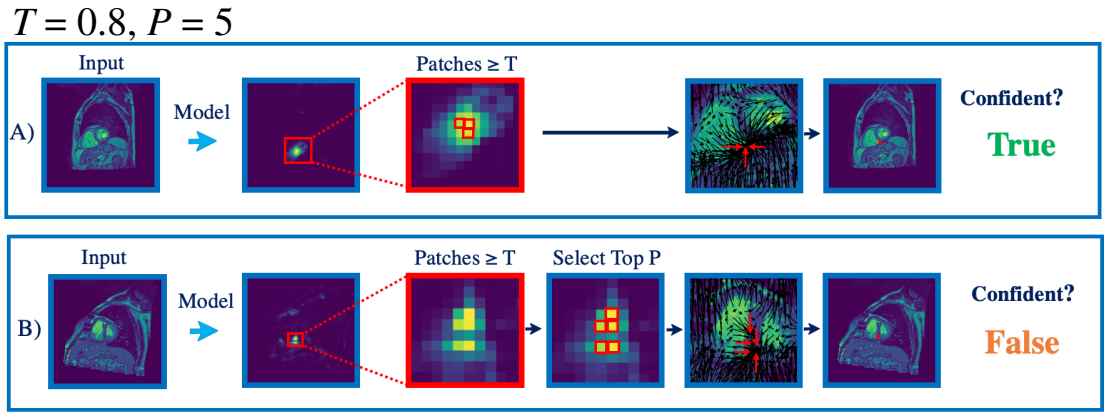


Figure 4.4: Adaptive prediction strategy to produce a final prediction from model outputs. The parameters T and P are learned in the inner loops of cross validation. If a patch's heatmap value exceeds T, the patch's displacement output is used for the final prediction. If zero patches exceed T, the P patches with the highest heatmap values are used. The black arrows show the predicted displacement from each patch to the landmark. The red arrows originate from the selected patches. **Case A)** Depicts a *low uncertainty* prediction, where the model detects patches as likely to contain the landmark. **Case B)** Shows a *high uncertainty* prediction, where no patches exceed the learned threshold, T.

4.3.3 Estimating Prediction Uncertainty

After we obtain a coordinate value for the landmark, we need to estimate that prediction's uncertainty.

Adaptive Prediction Uncertainty

For *Adaptive Prediction*, we consider images where no patch's activation from the heatmap branch exceeds T as *High Uncertainty*, and *Low Uncertainty* otherwise. This is depicted in Figure 4.4.

Candidate Smoothing Uncertainty

We hypothesise that the *Candidate Smoothed* maps obtained from PHD-Net contain information about the uncertainty of a prediction, which can be leveraged for subject-level uncertainty estimation. Since the activation of the pixels in the candidate map is defined by the amount of *"patch votes"* it achieved, we hypothesise a peak pixel with a low activation is more likely to be a worse prediction than one with a high activation. We call this peak pixel the Maximum

Candidate Smoothed Map Activation (CSMA).

We propose a data-driven (Frequentist) method using a hold-out calibration set to analyse the mappings and give a binary prediction of *Low Uncertainty* or *High Uncertainty*. Recall that our *Candidate Smoothing* maps are generated from the multi-task FCN, and are used to give us the final coordinate values. Therefore, we can use a hold-out calibration set to extract the localisation error from each *Candidate Smoothing* map alongside its CSMA value. We calculate a CSMA threshold through a weighted average of the CSMA of the 10% least accurate predictions' peak values, weighted according to the magnitude of the error.

In testing, if the final heatmap's peak value is below this threshold, we can infer that there was no clear consensus among the patches of the landmark's location, and consider it *High Uncertainty* prediction. Otherwise, the prediction is considered *Low Uncertainty*.

4.3.4 Evaluation Metrics

We evaluate our method using two common metrics in the literature, to capture both the accuracy and the robustness of the predictions.

We define the point-to-point error as the euclidean distance/Frobenius norm between a predicted landmark coordinates, $\hat{\mathbf{c}}$ and annotated landmark coordinates $\tilde{\mathbf{c}}$:

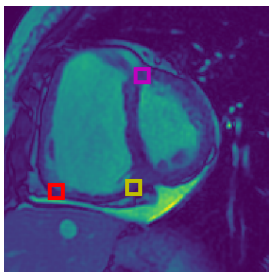
$$\mathcal{D}_{\text{PE}}(\hat{\mathbf{c}} - \tilde{\mathbf{c}}) = ||\hat{\mathbf{c}} - \tilde{\mathbf{c}}||_F. \tag{4.12}$$

We can calculate the mean standard deviation over all landmarks over all images from this metric.

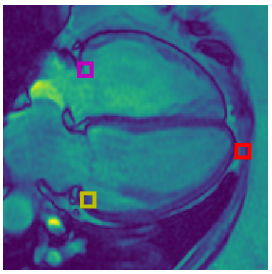
To quantify robustness we define the image-specific point-to-point error ($\mathcal{D}_{\text{IPE}}$) for an image j:

$$\mathcal{D}_{\text{IPE}}(\widetilde{\mathbf{C}}^{(j)}, \widehat{\mathbf{C}}^{(j)}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{D}_{\text{PE}}(\hat{\mathbf{c}}^{(j,i)}, \tilde{\mathbf{c}}^{(j,i)}), \tag{4.13}$$

over N landmarks, where $\widetilde{\mathbf{C}}^{(j)} = \{\tilde{\mathbf{c}}^{(j,1)}, \ldots, \tilde{\mathbf{c}}^{(j,N)}\}$ and $\widehat{\mathbf{C}}^{(j)} = \{\hat{\mathbf{c}}^{(j,1)}, \ldots, \hat{\mathbf{c}}^{(j,N)}\}$. $\mathcal{D}_{\text{IPE}}$ can illustrate the number of errors past a certain radius, visualising outliers. We present plots of the cumulative $\mathcal{D}_{\text{IPE}}$ distributions, which give the proportion of test images that achieve a localisation accuracy up to a certain error.



(a) SA Landmarks



(b) 4Ch landmarks.

Figure 4.5: **(a)** Landmarks for 4 chamber (4Ch) CMR: Magenta = tricuspid valve; Yellow = mitral valve; Red = apex of left ventricle. **(b)** Landmarks for Short Axis (SA) CMR: Magenta = superior right ventricle insertion point valve; Yellow = inferior right ventricle insertion point; Red = inferior lateral reflection of right ventricle free wall.

4.4 Empirical Validation of PHD-Net and Comparative Study

In this section, we perform experiments validating our proposed contributions to the patch-based training regime and compare our method to similar low-capacity approaches in the literature.

4.4.1 Datasets: ASPIRE-S

We evaluate PHD-Net on two datasets from the ASPIRE Registry [Hurdman et al., 2012], ASPIRE-S, outlined in Section 3.2.1 . To remind the reader, ASPIRE-S consists of Cardiac Magnetic Resonance (CMR) sequences, with 303 Short Axis (SA) view images, each with three annotated landmarks and 422 Four-Chamber (4CH) view images, each with three annotated landmarks. The 4CH dataset represents a more challenging landmark localisation task as the images have much higher variability than the SA dataset. The landmarks were decided and manually labelled by a radiologist, as shown in Figure 4.5.

Each dataset was split into 8 equally sized folds. Unless stated otherwise, all experiments were performed using 8-fold cross validation. At each iteration, 6 folds were used for training, 1 for validation, and 1 for testing.

4.4.2 Ablation Study

In this subsection we focus on our multi-task network, presenting a series of experiments demonstrating the value of our contributions, while exploring and evaluating the effects different choices have on the model. Building on the baseline network by [Noothout et al., 2018], we first explore and evaluate our choice of objective function in the classification branch of the network. Then, we qualitatively and quantitatively demonstrate key failures with the baseline coordinate resolution method before demonstrating the effectiveness of our proposed methods.

Experimental Setup and Training Details

PHD-Net was trained for a maximum of 500 epochs using a batch size of 32 and a learning rate of 0.001, using the Adam Optimiser. Early stopping was employed if the validation set's loss had not improved for 75 epochs. The sizes of the sub-images used in training were 128×128 pixels. During training, the sub-images were sampled randomly from the full image.

Effect of Multi-Task Learning

First, we demonstrate the effectiveness of multi-task learning, using the original binary classification mask alongside Binary Cross-Entropy [Noothout et al., 2018] rather than our proposed Gaussian heatmap with MSE. We train variants of PHD-Net culling either the classification branch or regression branch and removing the corresponding term from the loss function. For the regression-only variant, we use the weighted average method (Equation (4.6)) to obtain the final coordinates. For the classification-only variant we use the baseline binary mapping as the classification label, and use the peak patch method (Equation (4.8)) to obtain the final coordinates at test time.

Table 4.1 and Figure 4.6 shows the results from these experiments. Multi-task learning provides a convincing reduction in error compared to the single branch model, doubling performance compared to the single branch strategy. However, localisation error is still high. Upon inspection, the majority of failures stem from failures in the classification branch; the network cannot effectively identify the patch containing the landmark.

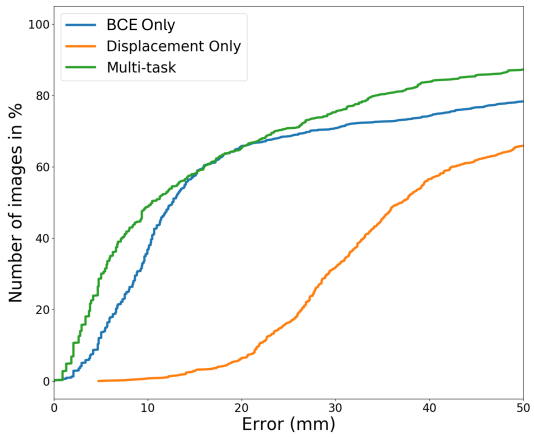


Figure 4.6: Cumulative $\mathcal{D}_{\text{IPE}}$ (mm) over all landmarks in SA images over a fixed 8-fold cross validation, comparing single branch to multi-task learning. **BCE** refers to Binary Cross Entropy, from Noothout et al. [2018].

Branch	Error (mm)
BCE	42.48 ± 58.07
Displacement	47.66 ± 11.64
Multi-task	**24.79 ± 31.82**

Table 4.1: Comparison of different branch strategies. Localisation error (mm) over all landmarks in SA images, over a fixed 8-fold cross validation. **BCE** refers to Binary Cross Entropy, from Noothout et al. [2018].

Effect of Objective in Heatmap Branch

After demonstrating that our choice of multi-task learning is beneficial, we analyse the effect changing the objective in the heatmap branch had on the task. Motivated by the success of discussed heatmap approaches in landmark localisation, we hypothesise the smoother supervision will overcome the challenges the binary mapping faced. We use the baseline multi-task strategy to resolve the coordinates (Equation (4.9)). Table 4.2 shows the results comparing using a binary map to a Gaussian heatmap of varying size. We find using a Gaussian heatmap noticeably outperforms a simple binary map, due to its smoother supervision and ability to encode some uncertainty in the prediction. However, the larger the Gaussian's standard deviation gets, the worse the performance. This is because the large Gaussian blob does not clearly indicate the landmark, morphing the task into a much coarser object detection task. We highlight a sweet spot of a standard deviation of two, which we settle on for all future experiments. Despite a drastic improvement over the baseline binary mapping, reducing the error from 24.28mm to 6.30mm, the localisation error remains subpar. Specifically, the model is still not robust to gross misidentifications, as indicated by the high standard deviation of

Branch Type	σ	Error (mm)
Binary	NA	24.79 ± 31.82
Gaussian	0.25	150.66 ± 122.80
Gaussian	0.5	52.19 ± 83.91
Gaussian	0.75	19.65 ± 58.38
Gaussian	1	13.00± 42.80
Gaussian	1.5	8.27± 26.75
Gaussian	2	**6.30± 19.32**
Gaussian	3	8.98± 20.18
Gaussian	4	8.02 ± 18.25
Gaussian	5	8.45± 14.52
Gaussian	6	17.97± 33.13
Gaussian	7	17.07± 28.07
Gaussian	8	21.76 ± 32.46

Table 4.2: PHD-Net results between a binary map & varying Gaussian maps for the heatmap branch. σ refers to the standard deviation parameter in Equation (4.3). Mean error and standard deviation in mm across landmarks, over a fixed 8-fold cross validation for the SA images is reported.

19.32mm.

Effect of Coordinate Calculation Strategy

To improve the robustness of the predictions and reduce gross misidentifications, we next demonstrate our coordinate resolution strategies:*Adaptive Prediction* (AP) (Algorithm 4.1), and *Candidate Smoothing* (CS) (Equation (4.11)), compared with the baseline coordinate resolution (Equation (4.9)). Following our results in the previous subsection, we learn patch-wise Gaussian heatmaps using $\sigma = 2$ in Equation (4.3).

For each landmark and fold, a single model was trained and all coordinate resolution strategies were performed on the same model to fairly compare them. Table 4.3 (*All Error* column) shows the localisation error for both the SA and 4CH images using the baseline, *Candidate Smoothing* and *Adaptive Prediction* approaches. It is clear that *Candidate Smoothing* and *Adaptive Prediction* outperform the baseline approach in terms of overall localisation error, achieving a lower mean error with less variance, indicating an increased robustness compared to the baseline classification branch. Further, *Candidate Smoothing* outperforms *Adaptive Prediction*, with a slight 1.5% improvement for the SA images and a significant

Strategy	All Error	% LowU	LowU Error
Baseline	6.30 ± 19.32	N/A	N/A
AP	4.80 ± 12.37	68%	4.43 ± 16.74
CS	4.73 ± 15.39	56%	**2.97 ± 2.20**

(a) Short Axis Images

Strategy	All Error	% LowU	LowU Error
Baseline	14.31 ± 34.12	N/A	N/A
AP	12.02 ± 31.73	55%	5.46 ± 8.82
CS	9.51 ± 25.89	42%	**4.40 ± 4.58**

(b) Four Chamber Images

Table 4.3: Comparing localisation error (mm) between PHD-Net's coordinate calculation strategies. *LowU* refers to images either *Candidate Smoothing* (CS) or *Adaptive Prediction* (AP) considered *Low Uncertainty*. The accompanying *%* is the percentage of images considered *Low Uncertainty* by each strategy. We report results over a fixed 8-fold cross validation.

23.3% improvement for the harder 4CH images. This suggests that in the more homogeneous SA dataset, using a constant number of patches with high activation was sufficient to locate the landmark. However, with the more heterogeneous 4CH dataset, regularising using a large number of patches using *Candidate Smoothing* significantly improves performance.

A common error with the baseline coordinate resolution is landmark misidentification, causing outliers. Since the strategy only considers the patch with the highest classification prediction value, the information from the surrounding patches is ignored. A small error in the classification branch can lead to a complete misidentification. This vulnerability is qualitatively shown in Figure 4.7, where it is clear that if the surrounding context was taken into account the error would have been avoided. Using *Adaptive Prediction* also fails in this case, falling victim to the same classification path failure. Using the *Candidate Smoothing* strategy, the model first plots each patch's independent prediction. We can see there are many high activations in the image representing locally accurate but globally ambiguous candidate points. The globally focused heatmap suppresses the unfeasible predictions, leading to an accurate final prediction.

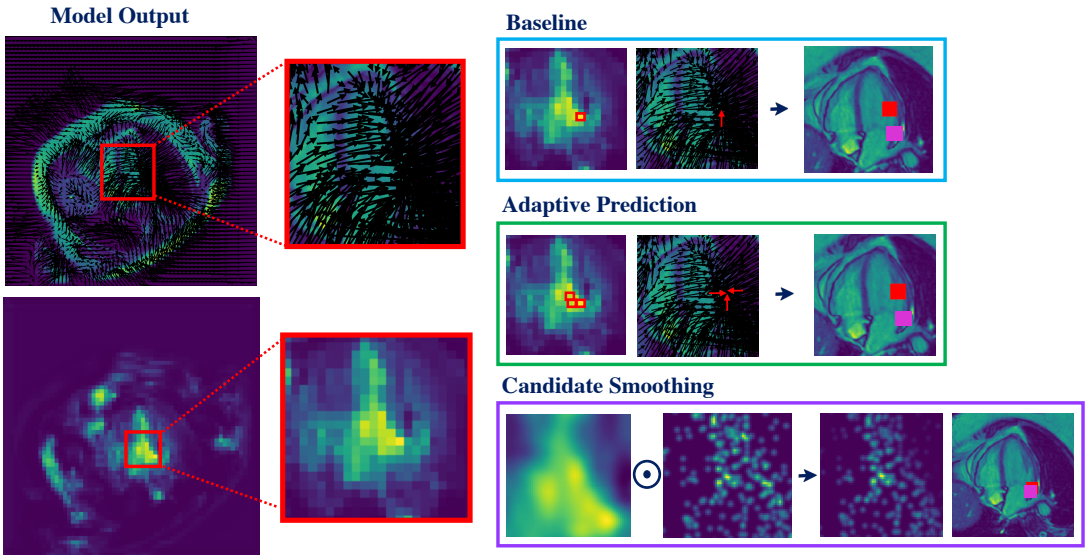


Figure 4.7: Visualisation of an example where using *Candidate Smoothing* is preferable. On the left are the patch-wise displacement (top) and heatmap (bottom) predictions. The **Baseline** [Noothout et al., 2018] coordinate calculation strategy and our **Adaptive Prediction** strategy fail, where **Candidate Smoothing** succeeds, due to the more global focus of the image. The red square on the last images represent the model's prediction, and the purple square is the ground truth landmark.

Localisation Accuracy Across Uncertainty Strategies

The uncertainty estimation of the two strategies gives further texture to the evaluation. The *LowU Error* column of Table 4.3 shows the localisation error for the predictions the two strategies flagged as *Low Uncertainty*, compared to the results over the entire set of landmarks (*All Error* column). Across both the SA and 4CH images, *Adaptive Prediction* flags more predictions as *Low Uncertainty* compared to *Candidate Smoothing*, with 12.5% more *Low Uncertainty* predictions on average. In the easier task of the SA images, the error reduction in the *Low Uncertainty* is slight for *Adaptive Prediction* compared to *Candidate Smoothing* (0.37mm vs. 1.76mm). In the 4CH task the improvement is more significant, at 5.11mm lower for *Candidate Smoothing* and 6.56mm lower for *Adaptive Prediction*. Notably,

the standard deviation of errors in the *LowU Error* columns for *Candidate Smoothing* is consistently significantly lower than the *All Error* column. This suggests *Candidate Smoothing* is successfully filtering out the gross misidentifications of the model, retaining only the accurate predictions. *Adaptive Prediction* achieves a similar but less significant reduction for the 4CH images, but the standard deviation for the SA dataset remains high, suggesting it is not filtering outliers as effectively.

These results highlight a precision/recall trade-off between the two strategies: *Candidate Smoothing* gives lower proportions of more accurate *Low Uncertainty* predictions, whereas *Adaptive Prediction* gives a high proportion of less accurate *Low Uncertainty* predictions. However, this result is dependent on the heuristic used to set the threshold of *Candidate Smoothing*. We explore this relationship in a generalised framework in Chapter 5.

Furthermore, it is difficult to pinpoint whether the samples filtered from the *Low Uncertainty* subset were flagged due to high epistemic uncertainty within the model or aleatoric uncertainty from the difficult data samples themselves. Upon inspection, many of the samples flagged as *High Uncertainty* were poor quality with imagining artifacts, hinting the methods are categorising by aleatoric uncertainty. We explore this further through the assistance of comparison methods in the next subsection.

4.4.3 Comparison to State-of-the-Art (SOTA)

Next, we compare PHD-Net's performance to several SOTA approaches, focusing our comparisons on relatively lightweight models. For all approaches, we tune hyper-parameters on the first fold of the 4CH dataset. We employ early stopping if validation error does not improve.

Experimental Setup and Training Details

Proposed Network

PHD-Net: For each landmark, PHD-Net was trained using the same settings described in Section 4.4.2. Again, all landmark localisation experiments were conducted using a fixed 8-fold cross validation. For all experiments, the standard deviation (σ) for the Gaussian heatmap label as defined in Equation (4.3) is two.

Comparison Networks

Hourglass Model: We compare PHD-Net to a popular encoder-decoder style work, the Hourglass Model [Newell et al., 2016]. We follow the implementation of the model as described by the author, downscaling the input images to 256×256 pixels, and training the model to output a 64×64 heatmap for each landmark. Due to the small amount of training data, we only use a single stacked Hourglass, leading to 6M trainable parameters. Despite the lower resolution input compared to PHD-Net, the capacity of the model in terms of parameters is much larger, even when taking into account its ability for simultaneous landmark localisation. We train for a maximum of 1000 epochs using the Adam optimiser. We select a learning rate of 0.001 and a batch size of 3.

We use the standard heatmap regression objective function, regressing a Gaussian heatmap, with the centre of the heatmap on the target landmark. For each landmark L_i with 2D coordinate position $\tilde{\mathbf{c}}^{(i)}$, the 2D heatmap image is defined as the 2D Gaussian function:

$$g_i\left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \tilde{\mathbf{c}}^{(i)}; \sigma\right) = \frac{1}{(2\pi)\sigma^2} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_2^2}{2\sigma^2}\right), \tag{4.14}$$

where $\mathbf{x}$ is the 2D coordinate vector of each pixel and σ is a user-defined standard deviation. The network learns weights $\mathbf{w}$ and biases $\mathbf{b}$ to predict the heatmap $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$.

U-Net Model: We also compare PHD-Net to a U-Net model [Ronneberger et al., 2015] provided by the Monai package [1]. We design the model with 5 encoding-decoding levels, creating 1.63M learnable parameters. The model uses two residual blocks per layer, with each block consisting of a 3×3 convolution, Batch Normalisation and ReLU. Downsampling is achieved through strided convolutions and upsampling through transposed convolutions. The initial number of feature maps is 16, doubling with each downsample and halving with each upsample. We modify the objective function from image segmentation to simultaneous landmark localisation, jointly predicting a heatmap for each landmark (Equation (4.14)). We downsample the input image to 256×256 pixels in order to create as much parity to PHD-Net in terms of model capacity, but unlike the Hourglass Network, the output heatmaps are a full size of 256×256. We train for a maximum of 1000 epochs with a selected batch size of 2, and a learning rate of 0.001 using the Adam Optimiser.

Patch-based Method: PIN: We also compare PHD-Net to a patch-based method, PIN

[1] Project MONAI, www.github.com/Project-MONAI

		Short Axis	4 Chamber
Model	# Parameters	All	All
Noothout et al. [2018]	0.06M	24.79 ± 31.82	52.90 ± 35.58
Newell et al. [2016]	6M	5.76 ± 8.48	13.33 ± 21.63
Ronneberger et al. [2015]	1.63M	5.93 ± 12.75	**7.78 ± 9.82**
Li et al. [2018]	4.7M	17.22 ± 12.00	25.17 ± 15.83
PHD-Net CS	0.06M	**4.73 ± 15.39**	9.51 ± 25.89
PHD-Net AP	0.06M	4.80 ± 12.37	12.02 ± 31.73

Table 4.4: Comparing localisation error (mm) between PHD-Net and comparison models. CS is *Candidate Smoothing* and AP is *Adaptive Prediction*. We report results over a fixed 8-fold cross validation. The best results for each dataset are highlighted in **bold**.

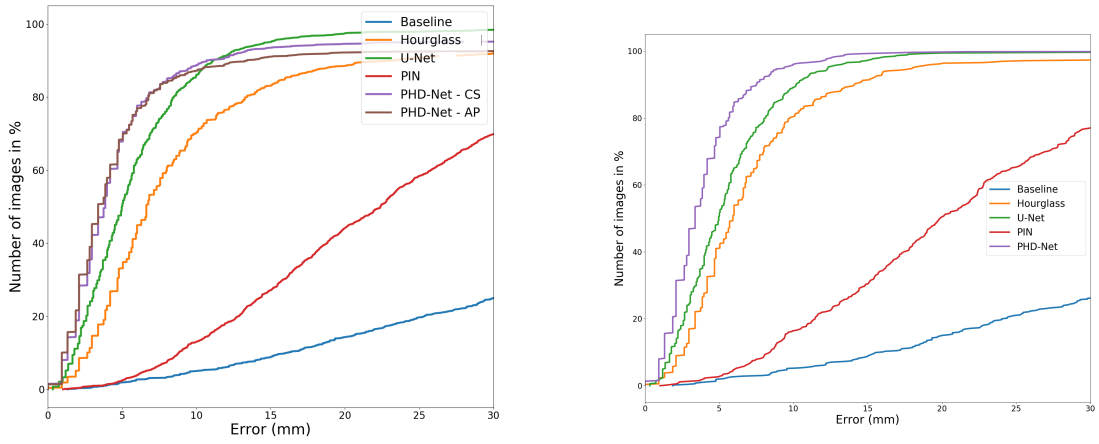
[Li et al., 2018]. PIN uses a **P**atch-based **I**terative **N**etwork that also combines classification and regression in a multi-task framework. Patches are repeatedly passed to the CNN until the estimated landmark position converges to the true landmark location. To the best of our ability, we follow the implementation made available by the authors, modifying it for our pipeline. We settle on using the Adam Optimiser with a learning rate of 0.0005, drop out rate of 0.1 and batch size of 64. We train with the full 512×512 resolution images. PIN has 4.72M traininable parameters.

Localisation Error Comparison

In Table 4.4, we present comprehensive results over all SA and 4CH images, using our two strategies: *Candidate Smoothing* and *Adaptive Prediction*. These are contrasted against our comparison methods detailed above. The cumulative $\mathcal{D}_{\text{IPE}}$ for all 4CH and SA images are shown in Figure 4.8a and Figure 4.9a, respectively. From this data, we glean three observations.

First, we observe that our proposed approach substantially surpasses the performance of the baseline method by Noothout et al. [2018]. This substantiates our claim that the smoother supervision facilitated by the heatmap label, as opposed to a binary map, alongside the improved coordinate retrieval strategies improves performance.

Second, it is discernible that PHD-Net consistently outperforms the comparison networks on the SA images. Figure 4.9a shows the encoder-decoder methods of Ronneberger et al. [2015] and Newell et al. [2016] converge to a similar performance as PHD-Net at a cumulative

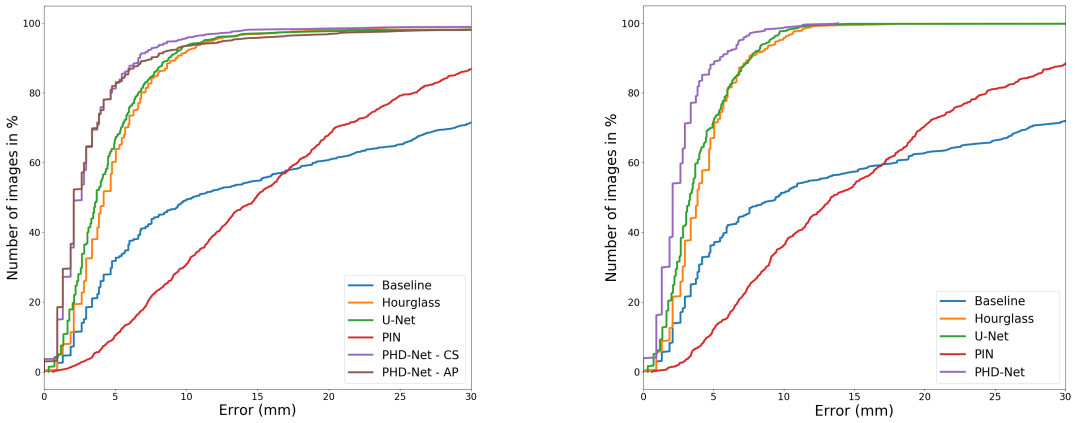


(a) (b)

Figure 4.8: Cumulative $\mathcal{D}_{\text{IPE}}$ (mm) over a fixed 8-fold cross validation for 4CH images. **(a)** All 4CH images. **(b)** Subset of 4CH images PHD-Net considered *Low Uncertainty* in *Candidate Smoothing*. PHD-Net uses the *Candidate Smoothing* strategy in reported results. Baseline is Noothout et al. [2018], Hourglass Model is Newell et al. [2016], U-Net Model is Ronneberger et al. [2015] and PIN Model is Li et al. [2018].

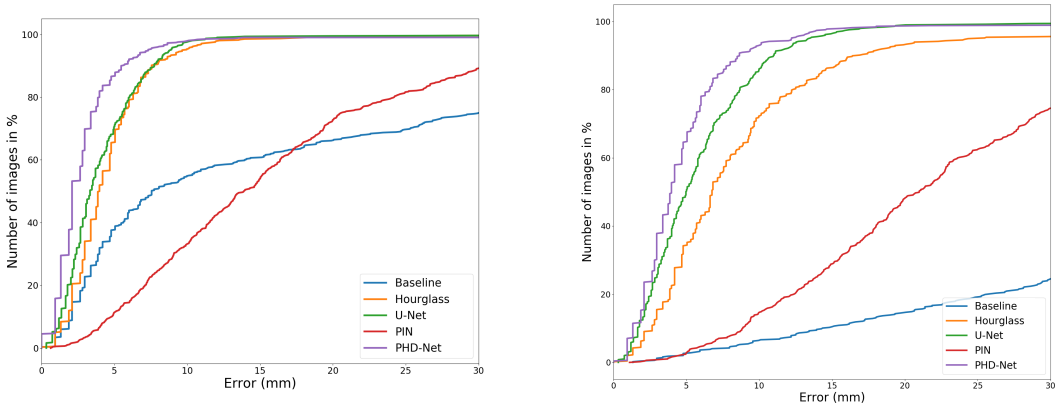
error threshold of 10mm. However, the results from 4CH images show the U-Net model [Ronneberger et al., 2015] manifests superior performance, suggesting the higher capacity model is more robust to datasets with large variations. The other encoder-decoder method by Newell et al. [2016] does not match performance with U-Net in this dataset, highlighting the importance of high-resolution outputs in its final decoder layer. Despite U-Net's superior performance, it is noteworthy how impressively PHD-Net performs considering its significantly smaller model capacity.

Lastly, our *Candidate Smoothing* strategy has been demonstrated to outperform the *Adaptive Prediction* strategy on both the SA and 4CH datasets. This shows the advantage a more globally focused model provides. The efficacy of the *Candidate Smoothing* strategy points towards the benefits of implementing models that have a wider, more holistic perspective in processing data.



(a) (b)

Figure 4.9: Cumulative $\mathcal{D}_{\text{IPE}}$ (mm) over a fixed 8-fold cross validation for SA images. **(a)** All SA images. **(b)** Subset of SA images PHD-Net considered *Low Uncertainty* in *Candidate Smoothing*. PHD-Net uses the *Candidate Smoothing* strategy in reported results. Baseline is Noothout et al. [2018], Hourglass Model is Newell et al. [2016], U-Net Model is Ronneberger et al. [2015] and PIN Model is Li et al. [2018].



(a) (b)

Figure 4.10: Cumulative $\mathcal{D}_{\text{IPE}}$ (mm) over a fixed 8-fold cross validation for SA and 4CH images using *Adaptive Prediction*. **(a)** Subset of SA images PHD-Net considered *Low Uncertainty* in *Adaptive Prediction*. **(b)** Subset of 4CH images PHD-Net considered *Low Uncertainty* in *Adaptive Prediction*. Baseline is Noothout et al. [2018], Hourglass Model is Newell et al. [2016], U-Net Model is Ronneberger et al. [2015] and PIN Model is Li et al. [2018].

	Short Axis Images			
Model	LowU CS (56 %)	Error Red	LowU AP (68 %)	Error Red
Noothout et al. [2018]	24.98 ± 33.98	+ 0.8%	22.39 ± 31.12	- 10.2%
Newell et al. [2016]	4.54 ± 4.61	- 23.7%	4.73 ± 4.40	- 19.6%
Ronneberger et al. [2015]	4.22 ± 6.52	- 33.7%	4.51 ± 9.07	- 33.7%
Li et al. [2018]	16.21 ± 11.17	- 6.0%	16.00 ±10.40	- 7.3%
Average Comparison Models	12.48 ± 14.07	- 15.7%	11.91 ± 13.75	- 17.7%
PHD-Net	**2.97 ± 2.20**	- 45.7%	4.43 ± 16.74	- 8.0%
	Four Chamber Images			
Model	LowU CS (42%)	Error Red	LowU AP (55 % %)	Error Red
Noothout et al. [2018]	49.84 ± 32.30	- 6.0%	51.52 ± 34.15	- 2.6%
Newell et al. [2016]	8.40 ± 12.71	- 45.4%	10.96 ± 17.81	- 19.5%
Ronneberger et al. [2015]	5.72 ± 4.50	- 30.5%	6.23 ± 7.67	- 22.1%
Li et al. [2018]	22.17 ± 12.19	- 12.6%	23.21 ± 13.54	- 8.1%
Average Comparison Models	21.5 ± 15.4	- 23.6%	22.98 ± 18.29	- 13.1%
PHD-Net	**4.40 ± 4.58**	- 73.5%	5.46 ± 8.82	- 75.1%

Table 4.5: Comparing localisation error (mm) between PHD-Net and comparison models, **only** on images PHD-Net considered *Low Uncertainty*. *LowU CS* refers to images *Candidate Smoothing* considered *Low Uncertainty* and *LowU AP* refers to images *Adaptive Prediction* considered *Low Uncertainty*. The accompanying *%* is the percentage of images considered *Low Uncertainty* by each strategy. The column *Error Red* refers to the reduction in error from *All* images to the subset of *LowU* images. We report results over a fixed 8-fold cross validation. **Bold** indicates best results.

Dissecting Sources of Uncertainty: Epistemic or Aleatoric?

We also evaluate performance of the comparison models on the images PHD-Net's two coordinate calculation strategies flagged as *Low Uncertainty*. The purpose of this study is to identify whether the uncertainty categorisations are primarily due to epistemic uncertainty in the model or aleatoric uncertainty in the data itself. If the uncertainty is epistemic in nature, we would expect to see no significant reduction in error for the other models. However, if the uncertainty scores are driven by aleatoric uncertainty, the subset of images flagged as *High Uncertainty* should be difficult cases for all models, leading to a greater reduction in localisation error for all models in the *Low Uncertainty* subset of images. To this end, we isolate and report results only on the subsets of samples each coordinate calculation strategy labelled as *Low Uncertainty*.

Table 4.5 shows the full results, with the reduction in error between *All* images and the images flagged as *Low Uncertainty*. We can see that almost all models performed better on both *Low Uncertainty* subsets, indicating both strategies are discriminating between difficult and easy images. Further, the standard deviation in the error for the *LowU* results in Table 4.5 is significantly lower than the standard deviation in Table 4.4 for all models, particularly the higher performing ones (PHD-Net, Newell et al. [2016]; Ronneberger et al. [2015]). Upon inspection, this was caused by anomalous images (e.g. scanning artifacts) leading to gross errors, which were successfully identified and filtered out of the *Low Uncertainty* subset by *Candidate Smoothing* and *Adaptive Prediction*. Averaged over both datasets and comparison models, *Candidate Smoothing* reduced localisation error by 19.7% and *Adaptive Prediction* reduced error by 15.4%. However, this is a significantly lower reduction in error compared to PHD-Net's average of 59.6% for *Candidate Smoothing* and 41.6% *Adaptive Prediction*. Therefore, we can conclude that epistemic uncertainty is a more significant factor in the uncertainty estimation strategies than aleatoric uncertainty. Further, it highlights the difficulty in decoupling these concepts of uncertainty in learning based models, since a dataset with high levels of inherent aleatoric uncertainty will undoubtedly influence the epistemic uncertainty of the learned model.

To gain more insights on the nuance of the uncertainty scores, Figure 4.8 shows a comparison of the cumulative $\mathcal{D}_{\text{IPE}}$ for all 4CH images and those flagged as *Low Uncertainty* by the *Candidate Smoothing* strategy, and Figure 4.9 shows the same for the SA images. Figure 4.10 shows the cumulative $\mathcal{D}_{\text{IPE}}$ for the images flagged as *Low Uncertainty* in the *Adaptive Prediction* strategy for both the SA and 4CH images. For the *Candidate Smoothing Low Uncertainty* SA images (Figure 4.9b), localisation errors above 15mm are entirely filtered out for PHD-Net and the encoder-decoder methods ([Newell et al., 2016; Ronneberger et al., 2015]). For the *Low Uncertainty* 4CH images, (Figure 4.8b), the same elimination of gross-identifications is not replicated, but we see significantly steeper curves for PHD-Net and the encoder-decoder methods compared to the set of *All* images (Figure 4.8a). We observe similar trends for *Adaptive Prediction* in Figure 4.10. However, *Adaptive Prediction* catches fewer gross misidentifications compared to *Candidate Smoothing* for PHD-Net and the encoder-decoder methods, shown by the Cumulative $\mathcal{D}_{\text{IPE}}$ curves converging to 100% later than those

in the *Candidate Smoothing* graphs (Figures 4.9b, 4.8b). Furthermore, PHD-Net significantly outperforms all comparison models when only considering *Low Uncertainty* predictions. This provides a compelling value for real world use as a user can decide to only use *Low Uncertainty* predictions in sensitive task, with the rest manually corrected.

Overall, PHD-Net outperforms the comparison methods for the SA images, but falls behind on the more challenging 4CH images. However, PHD-Net has the distinct advantage of being able to flag predictions as uncertain using "patch votes". When only considering certain predictions, PHD-Net demonstrates greater accuracy and robustness, evidenced by fewer outliers. Through deeper analysis of the uncertainty categorisations using our comparison models, we observe that the uncertainty estimation of our methods can be used as a marker for higher levels of aleatoric uncertainty in data samples. However, aleatoric uncertainty does not tell the whole story, with most of the uncertainty being attributed to epistemic uncertainty within the model itself.

4.5 Scaling Model Capacity with Patch-based Training Regime

For a final study, we explore pushing localisation performance under the patch-based, multi-task regime further using model architectures with larger capacity and introducing data augmentation. From the results of the previous subsection, we hypothesise that a higher capacity model has the capability to be more expressive than our original architecture, improving results. Therefore, we refine our training regime and design two larger models: a Residual Network and a Vision Transformer. Furthermore, we extend our patch-based training regime to multiple landmark prediction. We compare our method with the large capacity SOTA baseline, LannU-Net, detailed in Section 3.3.

4.5.1 Methods

Residual PHD-Net (PHD-Resnet)

First, we simply increase the model capacity of PHD-Net in terms of parameters. Inspired by the use of a Residual Network by Noothout et al. [2020], we opt to use a ResNet-34-esque network [He et al., 2016] as the backbone of the network.

First we use a convolutional block: involving 32 channels of 7×7 convolutional kernels with stride 2; followed by batch normalisation, ReLU and 3×3 maxpooling operation. 16 residual blocks then follow, each block consisting of batch normalisation, a 3×3 convolution, ReLU, a 3×3 convolution and a final batch normalisation. The convolutional operations in the first 3 blocks use 32 channels, the next 4 use 64 channels, proceeded by 6 blocks using 128 channels before a final 3 blocks with 256 channels. All convolutions use a stride of 1, bar the second group of 4 with 64 channels, which use a stride of two in order to fit the desired output shape of the network ($\frac{1}{8}$ the size of the input size, where each 1×1 feature corresponds to an 8×8 patch of the original input, as with the original PHD-Net). A skip connection is added from the beginning to the end of each block, concatenating the input of each block to the output of each block.

After the 16 residual blocks, the network again splits into two separate output branches, one for displacement regression and the other for heatmap regression. These branches are unchanged from PHD-Net, other than the increased input size of the first layer to match the now larger number of channels, which is now 256.

Transformer PHD-Net (PHD-Former)

Next, we evaluate the proposed training protocol using a Vision Transformer [Dosovitskiy et al., 2020]. Given the intrinsic mechanism of Vision Transformers, which segments images into patches, these models offer a seamless integration into the proposed task. Our implemented model was based on the "ViT-Base" [Dosovitskiy et al., 2020]. We opt to use the more compact "ViT-B/16" variant due to computational constraints.

For a thorough explanation of Vision Transformers, we point the reader to the original Vision Transformer paper [Dosovitskiy et al., 2020]. The specific details of our implementation are as follows.

Tokenisation: First, the full 512×512 image is first split into 1024 patches of size 16×16, flattened, and map to a $768D$ vector with a trainable linear projection to acquire the token sequence. This is followed by positional encoding, imparting the contextual spatial information of each patch in the sequence. This facilitates the learning of spatial relationships across the image. Here, the classification token is also integrated into the sequence. We use a dropout

rate of 0.1.

Encoder: The sequence is then input into the transformer encoder. The encoder constitutes eight consecutive blocks of an attention mechanism each followed by a feed-forward neural work. Layer normalisation is applied before each block, and residual connections are applied after each block. We use 12 heads for the multi-head self-attention mechanism. We use a dropout of 0.1, randomly omitting a portion of the neurons in the network. Each feed-forward neural network is implemented as a linear layer with 768 input neurons, GELU, 3072 neurons in the hidden layer with dropout of 0.1, and 768 output neuron with dropout of 0.1. These parameters were adopted from the original Vision Transformer paper [Dosovitskiy et al., 2020], with the only change being reducing the number of blocks from 12 to 8, due to computational the burden.

Multi-task Head: After the transformer encoder, the classification token is discarded. The remaining sequence of tokens, representing the patches of the image, are then reshaped to $768 \times 32 \times 32$. We then apply the same multi-task displacement and regression branches as used by the original PHD-Net and the Residual Neural Network. However, here the patch-wise predictions relate to patches of size 16×16 pixels rather than 8×8 pixels, again due to computational limitations in memory.

Multi-Landmark Transformer PHD-Net (PHD-Former)

We extend the multi-task patch-based training protocol to multiple landmarks, using the Vision Transformer network as the backbone architecture. In order to achieve this we simply increase the number of outputs for each patch of the network. For N landmarks, an input image size of 512×512, and a patch size of 16×16, the heatmap branch now outputs predictions of size $32 \times 32 \times N \times 1$ rather than $32 \times 32 \times 1 \times 1$. Similarly, the displacement branch now outputs a prediction of size $32 \times 32 \times N \times 2$ rather than $32 \times 32 \times 1 \times 2$.

The loss function is adjusted accordingly, using the mean of the landmark's individual losses.

Comparison: Multi-Landmark & Single Landmark U-Net (LannU-Net)

For a comparison encoder-decoder network that performs traditional heatmap regression, we use larger capacity LannU-Net, predicting full 512×512 resolution heatmaps. This model is a suitable benchmark for landmark localisation (see Section 3.3 for details, we use $\sigma = 8$ for Equation (4.14)).

For our **Multi-Landmark** variant, we predict outputs of size $512 \times 512 \times N$, where N is the number of landmarks. For our single landmark variant, $N = 1$.

4.5.2 Datasets: ASPIRE-L

To test our larger capacity models we opt to use the larger but more challenging ASPIRE-L dataset, detailed in Section 3.2.2. To remind the reader, this dataset consists of 789 4CH CMR images, each with four annotated landmarks. Although the dataset is larger than ASPIRE-S, it represents a more challenging task due to noiser annotations.

We split the dataset into 6 equally sized folds. For our experiments, we perform 6-fold cross validation. At each iteration, 4 folds are used for training, 1 for validation, and 1 for testing.

4.5.3 Experiments and Results

Experimental Setup and Training Details

We follow the same patch-based, multi-task training regime as the original PHD-Net with the objective function defined in Equation (4.2). We use a standard deviation of 2 for the Gaussian heatmap labels (Equation (4.3)). For the coordinate calculation, we use *Candidate Smoothing*, as described in Section 4.3.2. For PHD-Net and PHD-Resnet, we perform patch-based sampling using sub-images of size 128×128. Since PHD-Former does not have the property of being fully convolutional, full image sampling was used. Full image sampling was also used for LannU-Net.

We train for 500 epochs using Stochastic gradient descent with an initial learning rate of 0.01, decaying it using the 'poly' scheme, $(1 - epoch/epoch_{max})^{0.9}$ [Chen et al., 2017]. One epoch consists of 150 mini-batches, where each mini-batch is 12 samples. We employ early stopping if the validation set's localisation error does not drop for 150 epochs. We employ

Method	# Params	Proposed Patch Training	Multi-Landmark	Mean Error (mm)
PHD-Net	0.06M	✓	✗	9.84 ± 25.46
PHD-Resnet	6.57M	✓	✗	7.95 ± 21.41
PHD-Former	57.77M	✓	✗	5.48 ± **7.22**
LannU-Net	46.36M	✗	✗	**4.88** ± 18.39
PHD-Former	57.77M	✓	✓	9.01 ± **7.36**
LannU-Net	46.36M	✗	✓	**4.11** ± 15.10

Table 4.6: The results of all methods on four-chamber CMRI images, for all of the landmarks localised individually: Left Ventricular Apex (LV), Lateral Mitral Annulus (LMA), Lateral Tricuspid Annulus (LTA) and Spinal Cord (SA). The average of the results over all landmarks is taken for the mean error and standard deviation.

data augmentations with a probability of 0.5, uniformly sampling from a continuous range [α, ω]: Random scaling [0.8, 1.2], translation [-0.07%, 0.07%], rotation [-45°, 45°], shearing [-16, 16] and vertical flipping.

Localisation Results

Table 4.6 shows the localisation results over ASPIRE-L. For single-landmark prediction, we can see a significant correlation between the model capacity and localisation performance. This supports the results from the Section 4.4.3 which showed the larger capacity U-Net outperformed the smaller PHD-Net on the more challenging 4CH datatset. Furthermore, the introduction of the transformer in PHD-Former dramatically reduced gross mispredictions, as noted by the significantly lower standard deviation. The larger capacity U-Net, LannU-Net, outperformed the CNN variants of PHD-Net and slightly better than PHD-Former. However, LannU-Net was less robust than PHD-Former, making more gross mispredictions leading to a higher standard deviation in results. In the multiple landmark case, LannU-Net significantly outperformed PHD-Former. Unlike LannU-Net, predicting multiple landmarks at once did not help PHD-Former make better predictions by implicitly learning the spatial relationships between the landmarks. Upon inspection of the training loss, the loss for PHD-Former was significantly less stable compared to LannU-Net, particularly the displacement loss. This suggests learning the multi-task loss of each patch for all landmarks is a more complex, noisier function to optimise compared to the pixel-wise single-task Gaussian heatmap regression.

Overall, for the single landmark case, the multi-task patch-wise training regime scales

with model capacity, and offers a high-performing alternative to encoder-decoder heatmap regression models, with the added benefit of uncertainty estimation from the patch votes. In the multiple landmark case, the method lags behind LannU-Net due to a noisier and more complex objective function.

4.6 Discussion and Conclusion

4.6.1 Summary of Findings

In this chapter, we proposed a lightweight, uncertainty-estimating model for landmark localisation, PHD-Net. The method takes a patch-based, multi-task approach with joint heatmap and displacement regression. We presented two strategies to fuse the multi-task model outputs for a final prediction, concurrently estimating the prediction uncertainty. We performed evaluation on a CMR dataset covering two scanning protocols, using ablation studies to demonstrate the benefits our proposed contributions to the patch-based multi-task training regime.

In terms of uncertainty estimation, both our heuristic-based *Adaptive Prediction* and *Candidate Smoothing* approaches successfully discriminate between high and low error predictions when using Frequentist approaches to learn thresholds. Specifically, *Adaptive Prediction* strategy captures a higher proportion of images as *Low Uncertainty*, with lower localisation accuracy. *Candidate Smoothing* on the other hand is more conservative, capturing a smaller but more accurate cohort of *Low Uncertainty* predictions. These simple strategies combine the multi-branch outputs to provide a globally informed prediction, presenting a choice of trade-off between recall/precision.

Further, PHD-Net achieved localisation error better or similar to more expensive comparison models. In analysing uncertainty classifications with our comparison models, we find that uncertainty estimation can indicate heightened aleatoric uncertainty in data. Nevertheless, most of the categorisation can be explained by epistemic uncertainty, inherent within the model itself.

We also demonstrate our patch-based multi-task regime is scalable, applying the method to a large Residual Network and Vision Transformer, also extending it to predict multiple

landmarks simultaneously. In the arena of higher capacity models, PHD-Net variants perform comparably to the large encoder-decoder LannU-Net, with the added benefit of uncertainty estimation through "patch votes".

Despite this promising performance, the PHD-Net variants lagged behind U-Net in multiple landmark prediction, due to the complex objective function. Furthermore, both uncertainty estimation methods are based on a heuristic, lacking statistical rigour to guarantee or approximate either the percentage of images that will be flagged as *Low Uncertainty*, nor an upper or lower expected error bound.

4.6.2 Recommendations

We offer the following recommendations:

1. When constrained by memory or computation, our patch-based method PHD-Net can perform landmark localisation more effectively than similar lightweight models.

2. For the patch-based training regime, *Candidate Smoothing* should be used as the coordinate extraction and uncertainty estimation method, since it offers the best localisation performance.

3. Using "patch votes" can be used as an effective heuristic for uncertainty, but lacks statistical rigour. Enhance this with Frequentist methods using calibration sets like *Candidate Smoothing*.

4. Patch-based training regimes for single landmark prediction scale well with model capacity and should be used. However, due to an inherent local focus and complex objective functions, fall behind State-of-the-art encoder-decoder methods in multiple-landmark prediction.

4.6.3 Conclusion

In this chapter, we significantly improved the localisation performance of patch-based models, achieving a lightweight solution with comparable results to larger models. Furthermore, the scalability of our patch-based training regime with varying model capacities was experimentally

demonstrated. We proposed Frequentist approaches to our patch-derived heuristic uncertainties, demonstrating practical utility by discriminating between high and low error predictions. In the next chapter, we will build on our findings, generalising the uncertainty estimation to any heatmap-based method. We extend it beyond heuristics to approximate Bayesian inference, and offer a data-driven Frequentist framework to incorporate more fine-grained categorisations with estimated error bounds.

Chapter 5

Quantifying Uncertainty Estimation Methods with Quantile Binning

5.1 Introduction

In the previous chapter, we introduced PHD-Net - a light-weight multi-task network that provides a heuristic-based uncertainty estimate of its prediction. However, as discussed in Section 3.1.2, there is a need for a general approach to uncertainty estimation in the Gaussian Heatmap regression framework for landmark localisation. Furthermore, as outlined in research question **Q3** in Section 1.2, more holistic evaluation metrics for uncertainty estimation in landmark localisation are needed to provide a standard benchmark on which to compare them,

In this chapter we address these issues by first extending our heatmap-based uncertainty estimation concept introduced in the previous chapter to any heatmap-based landmark localisation model. We address our research question **Q2** in Section 1.2, which calls for more statistical rigour, by proposing Quantile Binning. This is a data-driven framework to estimate a prediction's quality by explicitly approximating the relationship between any continuous uncertainty measure and localisation error using Isotonic Regression. Using the framework, we place predictions into bins of increasing subject-level, predictive uncertainty and assign each bin a pair of estimated localisation error bounds. These bins can be used to identify the subsets of predictions with expected high or low localisation errors, allowing the user to make a choice of which subset of predictions to review and reannotate based on their expected error

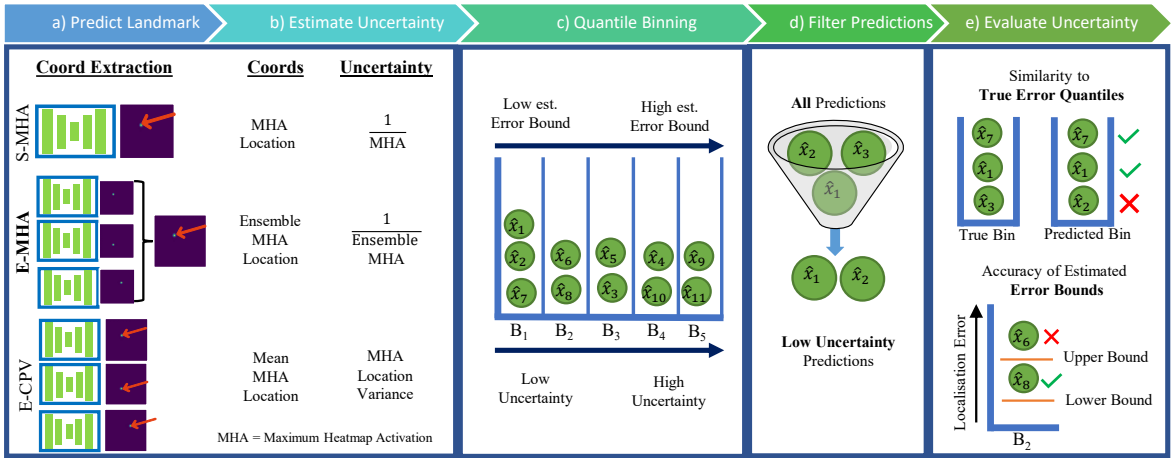


Figure 5.1: Overview of our general Quantile Binning framework. **a)** We make a prediction using a heatmap-based landmark localisation model, and **b)** extract a continuous uncertainty measure. **c)** We learn thresholds to categorise predictions into bins of increasing uncertainty, estimating error bounds for each bin. **d)** We filter out predictions from high uncertainty bins to improve the proportion of acceptable predictions. **e)** Finally, we evaluate each uncertainty measure's ability to capture the true error quantiles and the accuracy of the estimated error bounds.

bounds. We develop evaluation metrics for binning-based methods, facilitating a textured and comprehensive comparison between uncertainty measures.

Our approach is rooted in similar Frequentist principles to Conformal Prediction, an uncertainty estimation method discussed in Section 2.2.3. Conformal Prediction is also general method which also uses hold-out calibration set to improve uncertainty estimation for trained models. However, Conformal Prediction does not perform well with small calibration sets, particularly if the uncertainty metric is noisy [Angelopoulos and Bates, 2023]. This limitation becomes particularly pronounced in the medical domain, where training sets are often limited in size. As a result, the derived confidence intervals can be overly broad, diminishing their practical utility.

Our Quantile Binning method is generalisable to any continuous uncertainty measure, and the examples we investigate in this study cannot only be applied as a post-processing step to any heatmap-based landmark localisation method, but any regression problem that gives a sample-wise uncertainty measure. We aspire that this method can be used as a framework to build, evaluate and compare uncertainty metrics in landmark localisation beyond those demonstrated in this chapter.

5.2 Contributions

Our contributions, depicted in Figure 5.1, are threefold:

- We propose Quantile Binning, a Frequentist method to categorise predictions by any continuous uncertainty measure, and estimate error bounds for each bin (Figure 5.1c, Section 5.3.3).
- We construct two evaluation metrics for uncertainty estimation methods from Quantile Binning: 1) Similarity between predicted bins and true error quantiles; 2) Accuracy of estimated error bounds (Figure 5.1e, Section 5.3.4).
- We evaluate three heatmap-derived uncertainty measures and recommend our proposed method Ensemble Maximum Heatmap Activation (E-MHA) to extract landmark coordinates from an ensemble of heatmaps and estimate uncertainty (Figure 5.1a, 5.1b, Section 5.3.2).

We demonstrate the impact of our contributions by using our proposed Quantile Binning to compare E-MHA to two existing coordinate extraction and uncertainty estimation methods: a weak baseline of Single Maximum Heatmap Activation (S-MHA), and a stronger baseline of Ensemble Coordinate Prediction Variance (E-CPV). In Section 5.5.2, we compare the baseline coordinate extraction performance of the three approaches, followed by the uncertainty estimation performance in Section 5.5.3. We explore the reach of heatmap-based uncertainty measures by demonstrating they are applicable to both U-Net regressed Gaussian heatmaps and patch-based voting heatmaps. We show each uncertainty measure can identify a subset of predictions with significantly lower mean error than the full set by filtering out predictions from high uncertainty bins (Figure 5.1c). In Section 5.5.5 we demonstrate the generalisability of our method by applying Quantile Binning to a publicly available Cephalometric dataset [Wang et al., 2016], with significantly more annotated landmarks and images containing some repetitive structures. We show the flexibility of our method by reporting results over a range of binning resolutions in Section 5.5.6. Furthermore, in Section 5.5.7 we select a subset of landmarks from the Cephalometric dataset with multiple annotations (provided by Thaler et al. [2021]) to explore the effect of aleatoric uncertainty caused by landmark ambiguity on

Quantile Binning using our three uncertainty measures. Finally, in Section 5.7.2 we make recommendations for which uncertainty measure to use, and how to use it.

We provide an open-source implementation of this work and the tabular data obtained from the landmark localisation models to reproduce our results, alongside extensive experimental results at `https://github.com/pykale/pykale/tree/main/examples/landmark_uncertainty`.

5.3 Methods

5.3.1 Landmark Localisation Models

First, we briefly review the two models we will use for landmark localisation, allowing us to compare the generalisability of our uncertainty measures across different heatmap generation approaches. We implement a variation of the popular encoder-decoder networks that regresses Gaussian heatmaps, U-Net [Ronneberger et al., 2015]. We also implement a patch-based method introduced in the previous chapter, PHD-Net, which produces a heatmap from patch votes.

Encoder-Decoder Model (U-Net)

The vast majority of state-of-the-art landmark localisation approaches are based on the foundation of a U-Net style encoder-decoder architecture, described in Section 2.1.5.

Rather than regressing coordinates directly, the objective of the model is to learn a Gaussian heatmap image for each landmark, with the centre of the heatmap on the target landmark. For each landmark L_i with 2D coordinate position $\tilde{\mathbf{c}}^{(i)}$, the 2D heatmap image is defined as the 2D Gaussian function:

$$g_i\left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \tilde{\mathbf{c}}^{(i)}; \sigma\right) = \frac{1}{(2\pi)\sigma^2} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_2^2}{2\sigma^2}\right), \tag{5.1}$$

where $\mathbf{x}$ is the 2D coordinate vector of each pixel and σ is a user-defined standard deviation. The network learns weights $\mathbf{w}$ and biases $\mathbf{b}$ to predict the heatmap $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$. During inference, we can interpret the activation of each pixel in the predicted heatmap as the pseudo-probability of that pixel being the landmark. We will exploit this in our uncertainty

estimation methods.

Patch-based Model (PHD-Net)

Patch-based models use a Fully Convolutional Network (FCN), with the architecture resembling the first half of an encoder-decoder architecture. Therefore, they are more light-weight than encoder-decoder networks, with significantly less parameters leading to faster training.

In the previous chapter, we proposed PHD-Net: a multi-task patch-based network. We incorporated a variant of the heatmap objective function from encoder-decoder networks into the objective function, predicting the 2D displacement from each patch to the landmark alongside the coarse Gaussian pseudo-probability of each patch.

PHD-Net aggregates the patch-wise predictions to obtain a heatmap by plotting candidate predictions from the displacement branch as small Gaussian blobs, then regularising the map by the upsampled Gaussian from the heatmap branch.

Again, we can consider the activation of each pixel in heatmap as an indicator for uncertainty, where instead of the pseudo-probability, the activation represents the amount of "patch votes".

Ensemble Models

Using an ensemble of identical but randomly initialised models is more robust than using a single model, as it reduces the effect of a single model becoming stuck in a local minima during training. Furthermore, random initialisations explore different modes of the loss landscape, facilitating a powerful decorrelation effect between models [Fort et al., 2019]. As discussed in Section 2.2.3, there is a growing body of work arguing deep ensembles are approximately Bayesian [D'Angelo and Fortuin, 2021; Fort et al., 2019; Hoffmann and Elster, 2021; Wilson and Izmailov, 2020]. Therefore, we use the variance in the predictions of each model to estimate the uncertainty of the prediction, using an use an ensemble of T models.

5.3.2 Estimating Uncertainty and Coordinate Extraction

Although generated differently, we hypothesise both U-Net and PHD-Net produce heatmaps containing useful information to quantify a prediction's uncertainty - but are they equally

effective? To this end, we compare the performance of both models under three uncertainty estimation methods: two baseline approaches, and a proposed approach extending one of the baselines to an ensemble of networks. Each method extracts coordinate values from the predicted heatmap, and estimates the prediction's uncertainty.

Single Maximum Heatmap Activation (S-MHA)

We introduce the baseline coordinate extraction and uncertainty measure. For the ith landmark, we use the standard method to obtain the predicted coordinates $\hat{\mathbf{c}}^{(i)}$ from the predicted heatmap $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$, by finding the pixel with the highest activation:

$$\hat{\mathbf{c}}^{(i)} = \arg\max_{\mathbf{x}} h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}). \tag{5.2}$$

We hypothesise that the pixel activation at the coordinates $\hat{\mathbf{c}}^{(i)}$ can describe the model's uncertainty: the higher the activation, the lower the uncertainty, and the lower the prediction error. However, due to this inverse relationship, this measures "confidence", not uncertainty.

We transform our confidence metric to an "uncertainty" metric $\hat{u}^{(i)}$, by applying the following transformation to the pixel activation at the predicted landmark location:

$$\hat{u}^{(i)} = \frac{1}{\max_{\mathbf{x}} h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}) + \epsilon}, \tag{5.3}$$

where ϵ is a small constant scalar that prevents $\frac{1}{0}$. Now, as the pixel activation at $\hat{\mathbf{c}}^{(i)}$ increases, $\hat{u}^{(i)}$ decreases.

We call the transformed activation of this peak pixel Single Maximum Heatmap Activation (S-MHA). This is a continuous value bounded between $[\frac{1}{\epsilon}, \frac{1}{1+\epsilon}]$ for U-Net, and bounded between $[\frac{1}{\epsilon}, \frac{1}{N+\epsilon}]$ for PHD-Net, where N is the number of patches. The lower the S-MHA, the lower the uncertainty. It is important to note that the uncertainty encoded in S-MHA does not distinguish between aleatoric and epistemic uncertainty.

Ensemble Maximum Heatmap Activation (E-MHA)

In this chapter we extend the S-MHA uncertainty measure to ensemble models. We hypothesise that E-MHA should hold a stronger correlation with error than S-MHA due to the additional

robustness an ensemble of models affords. We use our deep ensemble as approximate Bayesian inference [Wilson and Izmailov, 2020], and generate the mean heatmap of the T models in the ensemble, obtaining the predicted landmark coordinates as the pixel with the highest activation:

$$\hat{\mathbf{c}}^{(i)} = \arg\max_{\mathbf{x}} \frac{1}{T} \sum_{t=1}^{T} h_i^t(\mathbf{x}; \mathbf{w}, \mathbf{b}). \tag{5.4}$$

Using the average prediction of an ensemble is the simplest, low-cost, standard form of ensemble fusion [Jungo et al., 2020; Karimi et al., 2019; Mehrtash et al., 2020; Mehta et al., 2022]. Again, we hypothesise the activation of the pixel $\hat{\mathbf{c}}^{(i)}$ correlates with model confidence. Similar to S-MHA, we inverse the pixel activation and add a small ϵ to the activation of $\hat{\mathbf{c}}^{(i)}$ to give us our uncertainty measure, $\hat{u}^{(i)}$:

$$\hat{u}^{(i)} = \frac{1}{\left(\max_{\mathbf{x}} \frac{1}{T} \sum_{t=1}^{T} h_i^t(\mathbf{x}; \mathbf{w}, \mathbf{b})\right) + \epsilon}. \tag{5.5}$$

E-MHA is a continuous value constrained to the same bounds as S-MHA. This is a form of late feature fusion, combining features from all models before a decision is made. E-MHA directly captures the uncertainty in the model parameters since we are using an ensemble, so it is a truer measure of epistemic uncertainty than S-MHA.

Ensemble Coordinate Prediction Variance (E-CPV)

We also implement an additional strong baseline for approximately Bayesian uncertainty estimation: Ensemble Coordinate Prediction Variance (E-CPV) [Drevickỳ and Kodym, 2020]. The more the models disagree on where the landmark is, the higher the uncertainty.

To extract a landmark's coordinates we first use the traditional S-MHA coordinate extraction method on each of the T models' predicted heatmaps. Then, we use decision-level fusion to calculate the mean coordinate of the individual predictions to compute the final coordinate predictions $\hat{\mathbf{c}}^{(i)}$:

$$\hat{\mathbf{c}}^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \arg\max_{\mathbf{x}} h_i^t(\mathbf{x}; \mathbf{w}, \mathbf{b}). \tag{5.6}$$

We generate the E-CPV by calculating the mean absolute difference between the T model predictions $\widehat{\mathbf{C}}_{ens}$ and $\hat{\mathbf{c}}^{(i)}$:

$$\hat{u}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} |\hat{\mathbf{c}}_{ens}^{(t,i)} - \hat{\mathbf{c}}^{(i)}|. \tag{5.7}$$

This is a continuous value bounded between $[0, \sqrt{H^2 + W^2}]$, where H and W are the height and width of the original image, respectively. The more the models disagree on the landmark location, the higher the coordinate prediction variance, and the higher the uncertainty.

Unlike S-MHA and E-MHA, this metric completely ignores the value of the heatmap activations. This makes E-CPV the truest measure of epistemic uncertainty. However, it potentially loses useful uncertainty information encoded in the MHA, but avoids possible bias caused by model miscalibration [Guo et al., 2017] or the Gaussian assumptions of the target heatmap.

5.3.3 Quantile Binning: Categorising Predictions by Uncertainty and Estimating Error Bounds

We leverage the described uncertainty measures to inform the predictive, subject-level uncertainty of any given prediction, i.e. *is the model's prediction likely to be accurate, or inaccurate based on this uncertainty value?* We propose a data-driven Frequentist approach, Quantile Binning, using a hold-out validation set to establish thresholds delineating varying levels of uncertainty specific to each trained model. We use these learned thresholds to categorise our predictions into bins and estimate error bounds for each bin. We opt for a data-driven approach over using static, pre-defined thresholds to increase robustness. For example, two identical models with randomly initialised weights trained on the same training set will converge to different modes [Fort et al., 2019], with a different distribution of MHA on the same test set. Furthermore, the difficulty of the landmark will also influence the characteristics of the resulting localisation model as well as the distribution of the uncertainty measures. Therefore, establishing a set of thresholds for each model is more invariant to training differences compared to using the same thresholds for all models.

Quantile Binning is application agnostic; applicable to any data as long as it consists of continuous tuples of `<Uncertainty Measure, Evaluation Metric>`. In this context, a

continuous tuple is a pair of continuous variables output by the prediction model, relating to a single sample.

In this instance, we generate these pairings after the landmark localisation model is trained. We use a hold-out validation set and make coordinate predictions and uncertainty estimates using each of our three uncertainty measures described in Section 5.3.2. Since we have the ground truth annotations of the validation set we can produce continuous `<Uncertainty Measure, Localisation Error>` tuples for each uncertainty measure.

Establishing Quantile Thresholds

We aim to categorise predictions using our continuous uncertainty metrics into Q bins. We make the following assumption: *The true function between a good uncertainty measure and localisation error is monotonically increasing (i.e. the higher the uncertainty, the higher the error).*

Quantile binning is a non-parametric method that fits well with these assumptions - a variant of histogram binning which is commonly used for calibration of predictive models [Guo et al., 2017; Naeini et al., 2015]. By considering the data in quantiles rather than intervals, we can better capture a skewed distribution as the outliers in the tail of the distribution can be grouped into the same group. In other words, *quantiles divide the probability distribution into areas of approximately equal probability.*

This property allows us to interrogate model-specific uncertainties. Rather than compute uncertainty thresholds based on predefined error thresholds for each bin, we use Quantile Binning to create thresholds that group our samples in relative terms. This enables the user to flag the worst $X\%$ of predictions. We describe the steps below.

First, for any given uncertainty measure we sort our validation set `<Uncertainty Measure, Localisation Error>` tuples in ascending order of their uncertainty value and sequentially group them into Q equal-sized bins $\mathbb{B}_1, ..., \mathbb{B}_Q$. We assign each bin $\mathbb{B}_q$ a pair of boundaries defined by the uncertainty values of the tuples at the edges of the bin to create an interval: $[\alpha_{q-1}, \alpha_q)$. To capture all predictions at the tail ends of the distribution, we set $\alpha_0 = 0$, and $\alpha_Q = \infty$.

During inference, we use these boundaries to bin our predictions into Q bins ($\mathbb{B}_1 ... \mathbb{B}_Q$), with

uncertainty increasing with each bin. For each predicted landmark $\hat{\mathbf{c}}^{(i)}$ with uncertainty $\hat{u}^{(i)}$ where $\alpha_{q-1} \leq \hat{u}^{(i)} < \alpha_q$, $\hat{\mathbf{c}}^{(i)}$ is binned into $\mathbb{B}_q$. As long as the validation set is representative of the true distribution, the distribution of samples should be uniform across the bins due to the quantile method we used to obtain thresholds.

The higher the value of Q, the more fine-grained we can categorise our uncertainty estimates. However, as Q increases the method becomes more sensitive to any noise present in the uncertainty measure, leading to less accurate prediction binnings. We demonstrate this trade-off in Section 5.5.6.

Since the uncertainty boundaries are defined by the density of the validation set distribution, the method is agnostic to the absolute range of the uncertainty measure. Therefore it is applicable to any continuous uncertainty measure.

Estimating Error Bounds using Isotonic Regression

Establishing thresholds has allowed us to filter predictions by uncertainty in relative terms, but we lack a method to estimate absolute localisation error for each bin. For example, for an easy landmark, the samples in $\mathbb{B}_1$ may have a very low localisation error in absolute terms, but for a more difficult landmark even the lowest relative uncertainty samples in $\mathbb{B}_1$ may have a high error. Therefore, in order to offer users information about the expected error for each group, we present a data-driven approach to predict error bounds.

A simple approach would be to observe the localisation error of the tuple at the quantile boundaries [α_{q-1} and α_q). However, observing a single sample from the validation set is subject to noise and may produce a poor estimate for an error bound. Therefore, on our hold-out validation set, we first use Isotonic Regression to approximate the function between uncertainty and error, constraining it to be monotonically increasing. Isotonic regression is a method to fit a free-form, non-decreasing line to a set of observations, also commonly used for predictive model calibration [Guo et al., 2017; Zadrozny and Elkan, 2002]. It is non-parametric, so can learn the true distribution if given enough i.i.d. data. Given a list of n observations $\{(\eta_1, \beta_1), \ldots, (\eta_n, \beta_n)\}$, the regression seeks a weighted least squares fit $\widehat{\beta}_i \approx \beta_i$ subject to the constraint that $\hat{\beta}_i \leq \hat{\beta}_j$ whenever $\eta_i \leq \eta_j$:

$$\min \sum_{i=1}^{n} \left(\hat{\beta}_i - \beta_i\right)^2 \text{ s.t. } \hat{\beta}_i \leq \hat{\beta}_j \text{ for all } (i,j) \in E, \tag{5.8}$$

where $E = \{(i,j) : \eta_i \leq \eta_j\}$. In our case, the observations are the ($UncertaintyMeasure$, $LocalisationError$) tuples.

Next, we use our isotonically regressed line to estimate error bounds for each of our quantile bins. We input each bin's threshold intervals $[\alpha_{q-1}, \alpha_q)$ into our fitted Isotonic Regression function, obtaining error predictions for each threshold, $[\gamma_{q-1}, \gamma_q)$. We use these values as the estimated lower and upper error bounds, respectively, of the predictions in bin $\mathbb{B}_q$. Note, that for $\mathbb{B}_1$ we only estimate an upper bound, and for $\mathbb{B}_Q$ we only estimate a lower bound.

In summary, we use a data-driven approach to learn thresholds to progressively filter predictions at inference into Q bins of increasing uncertainty, and assign each bin estimated error bounds.

5.3.4 Evaluation Metrics for Uncertainty Measures

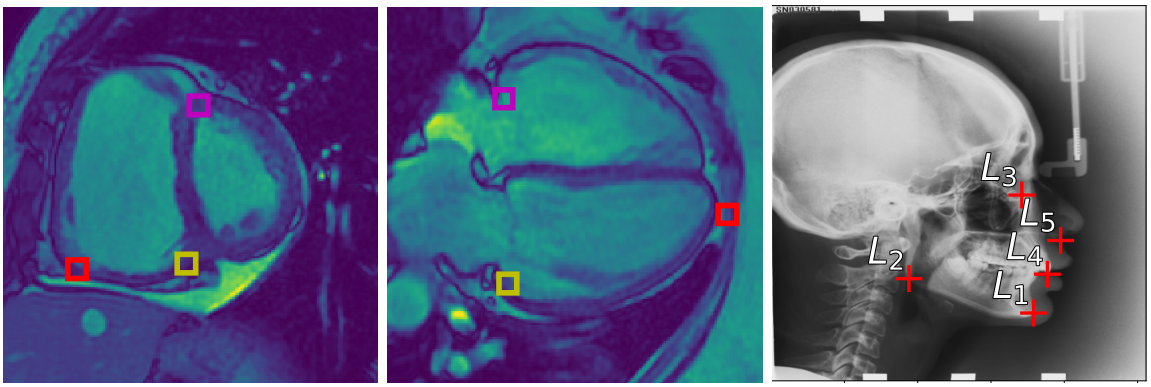
Next, we construct methods to evaluate how well an uncertainty measure's predicted bins represent the true error quantiles, and how accurate each bin's estimated error bounds are.

Evaluating the Predicted Bins

A good uncertainty measure will have a strong correlation with localisation error. Therefore, it should provide quantile thresholds that correspond to the true error quantiles. For example, since Bin $\mathbb{B}_1$ contains the predictions with the uncertainties at the lowest $\frac{1}{Q}$ quantile, the localisation errors of the predictions in $\mathbb{B}_1$ should be the lowest $\frac{1}{Q}$ quantile of the test set. This can be generalised to each group, until $\mathbb{B}_Q$, which should contain the errors in the $\frac{Q-1}{Q}$ quantile.

To evaluate this desired property, we propose to measure the similarity between each predicted bin and its respective theoretically perfect bin.

We create the ground truth (GT) bins by ordering the test set samples in ascending order of error. Then, we sequentially bin them into Q equally sized bins: $\widetilde{\mathbb{B}}_1 ... \widetilde{\mathbb{B}}_Q$.



(a) SA CMR. (b) 4CH CMR. (c) Cephalometric.

Figure 5.2: **(a)** Landmarks for Short Axis (SA) CMR: Magenta = superior right ventricle insertion point valve; Yellow = inferior right ventricle insertion point; Red = inferior lateral reflection of right ventricle free wall. **(b)** Landmarks for 4 chamber (4CH) CMR: Magenta = tricuspid valve; Yellow = mitral valve; Red = apex of left ventricle. **(c)** Subset of Landmarks included in the Cephalometric dataset [Wang et al., 2016]. Displayed landmarks are used in the aleatoric uncertainty analysis (Section 5.5.7).

For each predicted and GT bin pair $\widehat{\mathbb{B}}_q$ & $\widetilde{\mathbb{B}}_q$, we calculate the Jaccard Index (JI) between them and report the mean measure of each bin across all folds:

$$J_q(\widehat{\mathbb{B}}_q, \widetilde{\mathbb{B}}_q) = \frac{|\widehat{\mathbb{B}}_q \cap \widetilde{\mathbb{B}}_q|}{|\widehat{\mathbb{B}}_q \cup \widetilde{\mathbb{B}}_q|}. \tag{5.9}$$

The higher the JI, the better the uncertainty measure has binned predictions by localisation error. Therefore, it follows that the higher the JI, the better the uncertainty measure predicts localisation error.

Accuracy of Estimated Error bounds

A good uncertainty measure will have a monotonically increasing relationship with localisation error. Therefore, estimating the true function using isotonic regression should provide accurate error bound estimations.

To measure this, for each predicted bin $\widehat{\mathbb{B}}_q$, we calculate the percentage of predictions whose error falls between the estimated error bound interval, $[\gamma_{q-1}, \gamma_q)$. The higher the percentage, the higher the accuracy of our estimated error bounds.

5.4 Datasets

We perform our experiments using three datasets. To remind the reader, example images are shown in Figure 5.2.

5.4.1 ASPIRE-S

The first two datasets are from the ASPIRE Registry [Hurdman et al., 2012], the ASPIRE-S datasets, introduced in Section 3.2.1. Again, the 4CH dataset represents a more challenging landmark localisation task as the images have much higher variability than the SA dataset. The landmarks were decided and manually labelled by a radiologist, as shown in Figures 5.2a & 5.2b. For this study, we consider the SA images the **EASY** dataset, and the 4CH images the **HARD** dataset.

Once again, we split both CMR datasets into 8 folds, and perform 8-fold cross validation for both U-Net and PHD-Net. For each of the eight iterations, we select one fold as our testing set, one our hold-out validation set and the remaining 6 as our training set. These splits are differently initialised to those used in Chapter 4.

5.4.2 Cephalometric Radiographs

To test generalisablity across imaging modalities, we use a third dataset consisting of Cephalometric Radiographs [Wang et al., 2016]. The details are described in Section 3.2.3. To remind the reader, the dataset has a total of 19 annotated landmarks, and the images contain repetitive structures. For our study of aleatoric uncertainty in Section 5.5.7, we use subset of 5 landmarks which have a total of 11 annotations provided by Thaler et al. [2021]. The images have a spatial resolution of 1935×2400 pixels, where each pixel represents 0.1mm of the structure. Figure 5.2c shows an example image annotated with the aleatoric uncertainty landmark subset.

For the Cephalometric dataset we perform 4-fold cross validation using junior annotations, setting aside a random 20% of each fold's training set as our hold-out validation-set.

	4 Chamber Images		Short Axis Images	
Method	U-Net	PHD-Net	U-Net	PHD-Net
S-MHA All	**10.00 ± 18.99**	11.07 ± 21.33	5.86 ± 14.19	**3.58 ± 3.52**
S-MHA $\mathbb{B}_1$	6.79 ± 6.09	**5.80 ± 9.03**	3.62 ± 2.45	**2.78 ± 1.99**
E-MHA All	**6.36 ± 8.01**	9.14 ± 18.11	4.37 ± 8.86	**3.36 ± 3.50**
E-MHA $\mathbb{B}_1$	4.93 ± 2.85	**4.70 ± 3.21**	2.98 ± 2.09	**2.39 ± 1.90**
E-CPV All	**8.13 ± 10.16**	9.42 ± 13.07	4.97 ± 7.51	**3.22 ± 2.93**
E-CPV $\mathbb{B}_1$	5.34 ± 3.00	**5.10 ± 6.76**	3.75 ± 2.13	**2.47 ± 2.08**

Table 5.1: Localisation errors (mm) for the uncertainty methods outlined. *All* indicates entire set of predictions; $\mathbb{B}_1$ indicates subset with the *lowest uncertainties*. Mean error and standard deviation are reported across all folds & all landmarks. **Bold** indicates best results in row for the given dataset.

5.5 Experiments and Results

First, in Section 5.5.2 we present the baseline landmark localisation performance of PHD-Net and U-Net over both SA and 4CH datasets using the S-MHA, E-CPV, and E-MHA methods to extract coordinates. This gives us a comparison of the coordinate extraction performance from each of our methods, and a baseline to measure the effectiveness of each method's uncertainty estimation. Second, in Section 5.5.3 we interrogate how using Quantile Binning with our uncertainty measures delineates predictions in terms of their localisation error, and compare the predicted bins to the ground truth error quantiles. We show a practical example of how filtering out highly uncertain predictions can dramatically increase the proportion of acceptable localisation predictions. In Section 5.5.4 we assess how well the uncertainty measures can predict error bounds for each bin. Next, we demonstrate the generalisability of Quantile Binning in Section 5.5.5 on the more diverse Cephalometric dataset. In Section 5.5.6, we highlight the flexibility of the method by quantifying the effects of varying the number of quantile bins (Q) used. Finally, in Section 5.5.7 we explore aleatoric uncertainty, demonstrating Quantile Binning's effectiveness on landmarks with high ambiguity, as well as sharing insights on our studied uncertainty measure's relationship with aleatoric uncertainty. When comparing between $\mathbb{B}_1, \mathbb{B}_{2-4}, \mathbb{B}_5$ we use an unpaired t-test ($p \leq 0.05$) to test for significance. When comparing uncertainty metrics among the same Bin category and model, we use a paired t-test ($p \leq 0.05$) to test for significance.

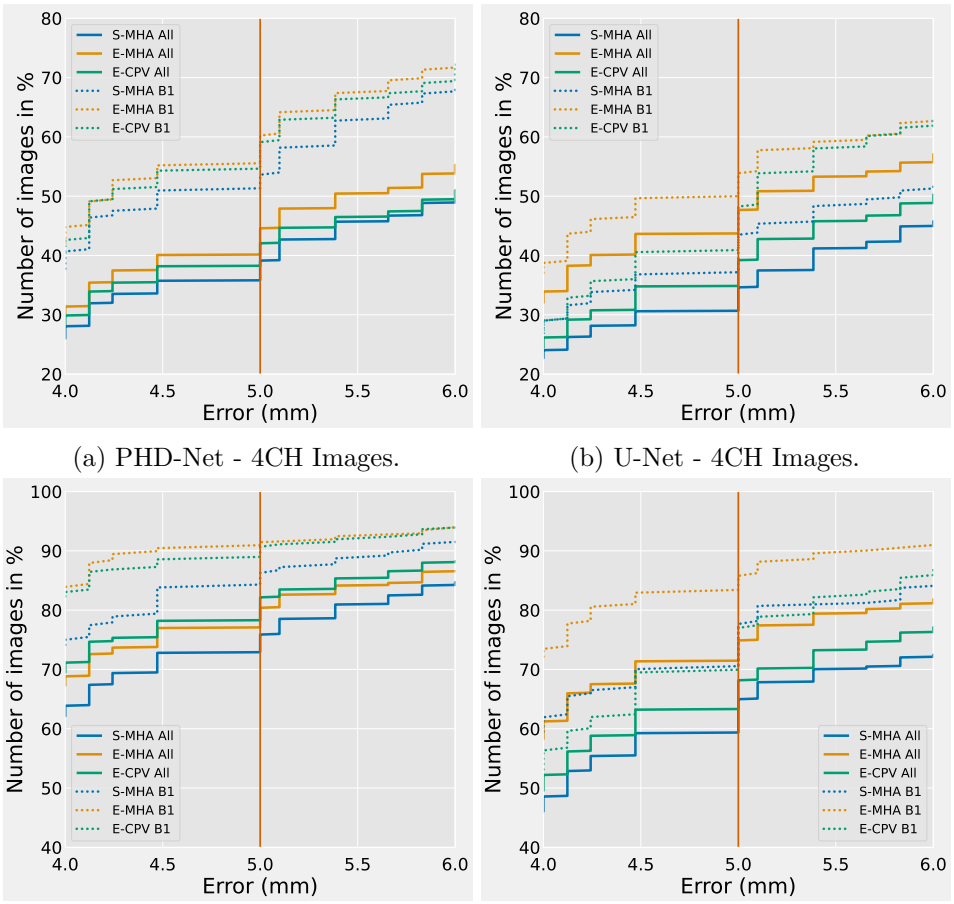
5.5.1 Experimental Setup and Training Details

For all data, we resize the images to 512×512 pixels and upsample the predicted heatmaps to the original image size before coordinate extraction. We select $T = 5$ for the ensemble methods, training 5 identical, randomly initialised models at each iteration. We chose $T = 5$ to compromise with computational constraints, asserting that 5 models are representative to compare the uncertainty methods for our purposes. We randomly select a model from our trained ensemble for our S-MHA uncertainty measure. For our Quantile Binning method, we select $Q = 5$ for 5 bins, striking a balance between the resolution of separation of the data and the limited size of our hold-out validation set ($\sim$30 samples for the CMR datasets, $\sim$60 samples for the Cephalometric dataset). We explore the effect of changing Q in Section 5.5.6.

since the focus of this study is uncertainty estimation rather than localisation accuracy, we implement a vanilla U-Net model [Ronneberger et al., 2015], opting for a computationally less expensive model. We design the architecture with 5 encoding-decoding levels, creating 1.63M learnable parameters. Each level contains 2 residual units, where each residual unit applies a 3×3 convolution, instance normalisation, and ReLU to the input, before concatenating the resulting output with the unit input. As we descend down the five levels of the encoder we use (16, 32, 64, 128, 256) input channels respective to each layer, mirroring this in the decoder path. On the encoder path we use maxpooling after each level to reduce spatial dimensions, and on the decoder path we use transposed convolutions to upsample the spatial resolution. We modify the objective function from image segmentation to simultaneous landmark localisation, minimising the mean squared error between the target and predicted heatmaps. We use the full 512×512 pixel image as input, and learn heatmaps of the same size. We train for 1000 epochs with a batch size of 2, and a learning rate of 0.001 using the Adam Optimiser (settings from [Schobs et al., 2021]). We generate target heatmaps using Equation (5.1) with a standard deviation of 8 for our CMR datasets and 2 for our Cephalometric dataset (chosen experimentally using the first fold of each dataset). We do not use data augmentation.

We implement our PHD-Net model as described in Chapter 4, creating a model with 0.06M learnable parameters. For all experiments we trained PHD-Net for 1000 epochs using a batch size of 32 and a learning rate of 0.001, using the Adam Optimiser. We train one landmark at a time. Note, the only difference in setup from the previous this chapter is

different fold splits and training for an additional 500 epochs (same as U-Net) with no early stopping, since we now use our validation set for the calibration set in Quantile Binning. We do not use data augmentation.



(a) PHD-Net - 4CH Images.

(b) U-Net - 4CH Images.

(c) PHD-Net - SA Images.

(d) U-Net - SA Images.

Figure 5.3: Cumulative distribution of localisation errors showing the % of predictions under a given error threshold, comparing all predictions (*All*) to the lowest uncertainty subset ($\mathbb{B}_1$) for the uncertainty methods across all folds & landmarks. The vertical line is the acceptable error threshold, chosen by a radiologist. Higher percentage is better.

5.5.2 Baseline Landmark Localisation Performance

Table 5.1 shows the baseline performance for U-Net and PHD-Net at localising landmarks in our 4CH and SA datasets. We make the following observations:

- When considering localisation error for the entire set of landmarks (*All*), performance

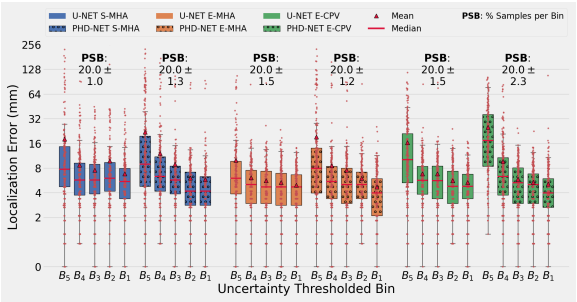
is better on the SA dataset for both models, with PHD-Net outperforming U-Net. On the 4CH dataset, U-Net outperforms PHD-Net in terms of fewer gross mispredictions, suggesting the higher capacity model of U-Net is more robust to datasets with large variations.

- Simply using a single model with our S-MHA strategy is predictably less robust than ensemble approaches.
- E-MHA outperforms the previous strong baseline of E-CPV for coordinate extraction. However, does it outperform E-CPV in terms of uncertainty estimation? We explore this in Section 5.5.3.
- The standard deviation in the error for the baseline *All* results in Table 5.1 is high for all models. We aspire to catch these bad predictions using Quantile Binning in Section 5.5.3.

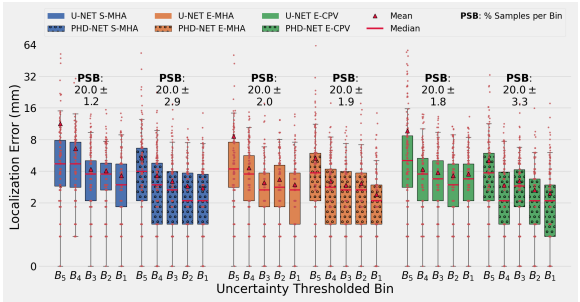
5.5.3 Analysis of the Predicted Quantile Bins

We apply quantile binning to each uncertainty measure: S-MHA, E-MHA and E-CPV. We compare results over U-Net and PHD-Net for both the SA and 4CH datasets.

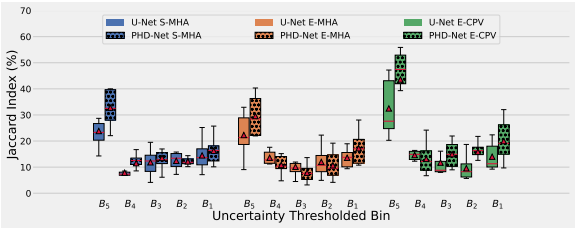
First, we test our assumption that our uncertainty measures correlate with localisation error. We report the Spearman's Rank Correlation Coefficient (ρ) since we are measuring a monotonic correlation which is not necessarily linear. All correlations are reported from the aggregated test sets across all eight folds of our CMR datasets, using a significance level of $p < 0.001$. For our 4CH dataset, S-MHA achieves correlations of 0.33 (weak-moderate) & 0.47 (moderate), E-MHA shows weak-moderate correlations of 0.39 & 0.39, and E-CPV shows moderate correlations of 0.42 & 0.53; for U-Net and PHD-Net respectively. For our SA dataset, S-MHA achieves correlations of 0.27 (weak) & 0.33 (weak-moderate), E-MHA weak-moderate correlations of 0.38 & 0.38, and E-CPV correlations of 0.27 (weak), 0.36 (weak-moderate); for U-Net and PHD-Net, respectively. The correlation strength of S-MHA has high variance, whereas E-MHA shows a stable correlation across datasets and localisation models. E-CPV achieves the strongest correlation with error across both models for our harder 4CH dataset, but a weaker correlation than E-MHA for our easier SA dataset. Overall, these results show



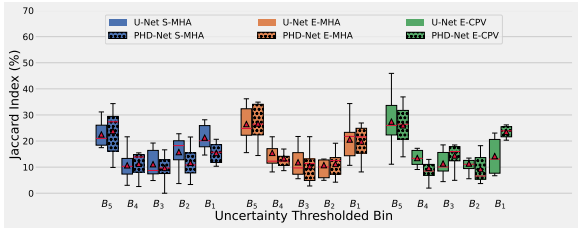
(a) Localisation error for each Bin - 4CH dataset (Lower is better).



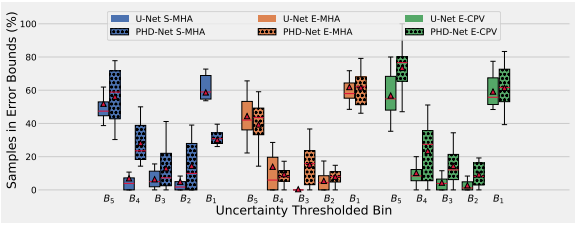
(b) Localisation error for each Bin - SA dataset (Lower is better).



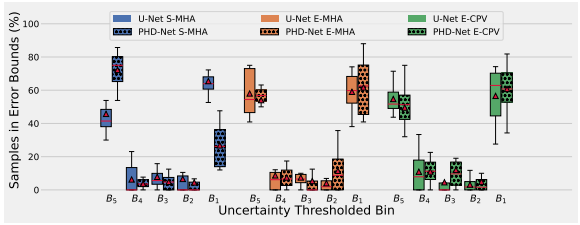
(c) Jaccard Index for each Bin - 4CH dataset (Higher is better).



(d) Jaccard Index for each Bin - SA dataset (Higher is better).



(e) Estimated Error Bound Accuracies- 4CH dataset (Higher is better).



(f) Estimated Error Bound Accuracies- SA dataset (Higher is better).

Figure 5.4: Results from Quantile Binning for U-Net and PHD-Net across all landmarks & folds, using our three coordinate extraction & uncertainty estimation methods. Bins are in descending order of uncertainty ($\mathbb{B}_5$ highest uncertainty, $\mathbb{B}_1$ lowest uncertainty). (a) and (b) show the mean localisation error of each bin, with error decreasing as we move towards the bins with lower uncertainty. (c) and (d) present the Jaccard Index, showing how similar the predicted bins are to the ground truth error quantiles. (e) and (f) visualise the estimated error bound accuracy, showing the percentage of predictions within the estimated error bounds for each bin. Best viewed on screen.

that MHA and E-CPV contain information that can be exploited to estimate the uncertainty of our predictions.

Next, we compare how our uncertainty measures can predict the true error quantiles. We found the most useful information is at the tail ends of the uncertainty distributions. Figures 5.4c & 5.4d plot the Jaccard Index between ground truth error quantiles and predicted error quantiles. We notice a parabolic trend, where the outer bins are closer to the true error quantiles than the middle bins. The highest uncertainty quantile bin ($\mathbb{B}_5$) is significantly better at capturing the correct subset of predictions than the intermediate bins ($\mathbb{B}_2 - \mathbb{B}_4$). Similarly, in some cases the bin representing the lowest uncertainties ($\mathbb{B}_1$) had a significantly higher Jaccard Index than the intermediate bins, but still lower than $\mathbb{B}_5$. Figures 5.4a & 5.4b show the mean error (▲) of the samples of each quantile bin over both datasets. The most significant reduction in localisation error is from $\mathbb{B}_5$ to $\mathbb{B}_4$ for all uncertainty measures. The sample distribution over the bins, indicated by the red dots, confirms that $\mathbb{B}_5$ captures more gross mispredictions than the remaining bins, particularly for the 4CH dataset. A tabular representation of this data is available in Appendix A, Table A.1. These findings suggest that most of the utility in the uncertainty measures investigated can be found at the tail ends of the scale. This is an intuitive finding, as the predictions in $\mathbb{B}_5$ are *certainly uncertain*, and the predictions in $\mathbb{B}_1$ are *certainly certain*. Figures 5.4a & 5.4b show that each bin contains ∼20% of the predictions, confirming our data-driven approach to setting uncertainty thresholds successfully approximates the true uncertainty distribution.

The worse trained the landmark localisation model, the more useful the uncertainty measure. Table 5.1 shows the localisation error of all methods, models and datasets for the entire set (*All*) and lowest uncertainty subset ($\mathbb{B}_1$) of predictions. PHD-Net's baseline localisation performance on the 4CH dataset was worse than U-Net. However, when we consider the lowest uncertainty subset of predictions ($\mathbb{B}_1$), PHD-Net sees a 47% average reduction in error from all predictions (*All*), compared to U-Net's average reduction of 30%. Similarly, U-Net performed worse than PHD-Net for the SA dataset, but saw an average error reduction of 31% compared to PHD-Net's 25%. This suggests that all investigated uncertainty measures are more effective at identifying gross mispredictions when models are poorly trained. When we separate results per-landmark, we find similar trends; landmarks

with worse localisation performance overall see the largest proportional reduction in error from $\mathbb{B}_5$ to $\mathbb{B}_1$ (see Appendix A, Figure A.2).

Using heatmap-based uncertainty measures is generalisable across heatmap generation approaches. The bin similarities in Figures 5.4c & 5.4d show that using S-MHA and E-MHA yields similar performance with PHD-Net and U-Net, despite their different heatmap derivations. Surprisingly using E-MHA does not give a significant increase in bin similarity compared to S-MHA, suggesting the thresholds remain relatively stable across models.

No investigated method is conclusively best for estimating uncertainty in all scenarios. For the more challenging 4CH data, Figure 5.4c shows E-CPV is significantly better than S-MHA and E-MHA for both models at capturing the true error quantiles, corroborating the findings of Drevickỳ and Kodym [2020]. E-CPV is particularly good at identifying the worst predictions ($\mathbb{B}_5$). For the easier SA data, no method has a significantly higher Jaccard Index. Therefore, when we generalise across both models and datasets, all uncertainty measures fared broadly similar on average in terms of error reduction between the entire set and the $\mathbb{B}_1$ subset of predictions. S-MHA had an average error reduction of 35.07%, E-MHA 32.94% and E-CPV 32%.

Despite similar performances in uncertainty estimation, we found E-MHA yields the greatest localisation performance overall. Table 5.1 shows E-MHA offers the best localisation performance for $\mathbb{B}_1$ across both datasets and models. This is due to the combination of offering the most robust coordinate extraction on average (Table. 5.1), and similar uncertainty estimation performance (Figure 5.4c, Figure 5.4d). We more concretely demonstrate Quantile Binning's ability to identify low uncertainty predictions in Figure 5.3. We clearly observe a significant increase in the percentage of images below the acceptable error threshold of 5mm when considering only predictions in $\mathbb{B}_1$ - with E-MHA giving the greatest proportion of acceptable predictions.

5.5.4 Analysis of Error Bound Estimation

We analyse how accurate the isotonically regressed estimated error bounds are for our quantile bins. Figures 5.4e & 5.4f show the percentage of samples in each bin that fall between the estimated error bounds. Figure 5.5 shows the results from an example fold.

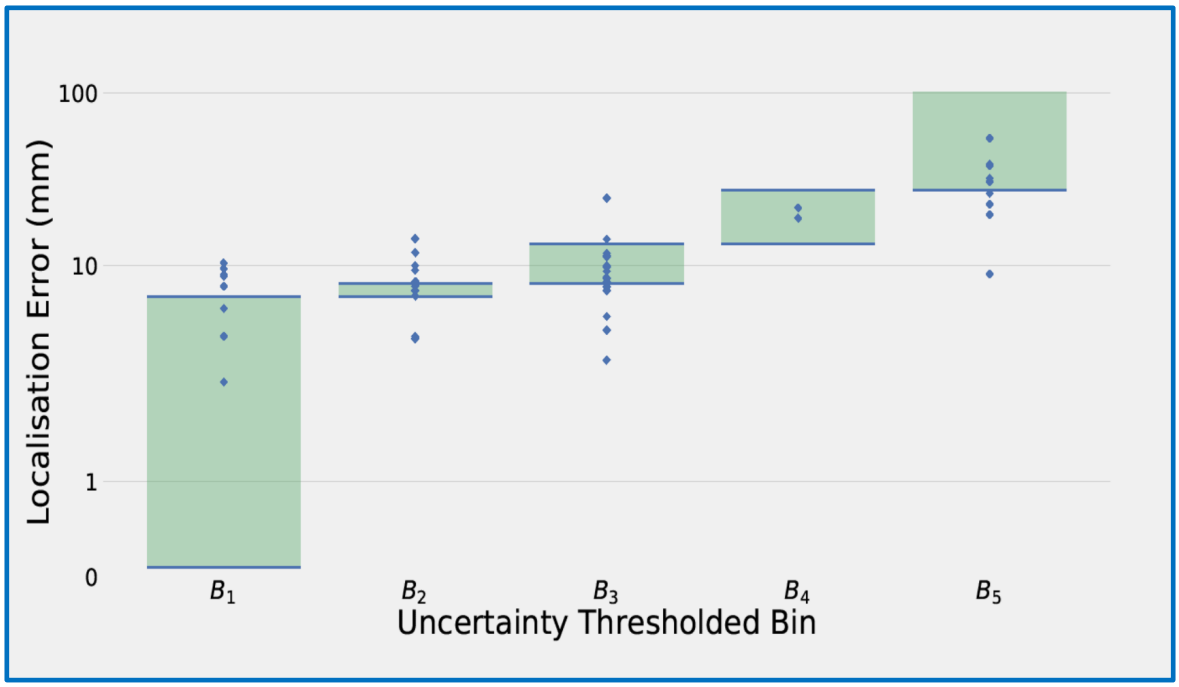


Figure 5.5: Results from an example fold of E-MHA for the 4CH dataset. The blue bars reprsent the estimated error bounds for each bin, and the blue diamonds represent the observed error of each sample in the fold.

We found we can predict the error bounds for the two extreme bins better than the intermediate bins. Figures 5.4e & 5.4f show a similar parabolic pattern to the Jaccard Index Figures 5.4c & 5.4d, with the two extreme bins $\mathbb{B}_5$ and $\mathbb{B}_1$ predicting error bounds significantly more accurately than the inner bins. Again, this indicates the most useful uncertainty information is present at the extremes of the uncertainty distribution, with the predicted uncertainty-error function unable to capture a consistent relationship for the inner quantiles. Further, the increased accuracy of the outer bins can be explained by the fact that it is easier to predict a single lower/upper bound than a pair of tighter bounds for the middling bins.

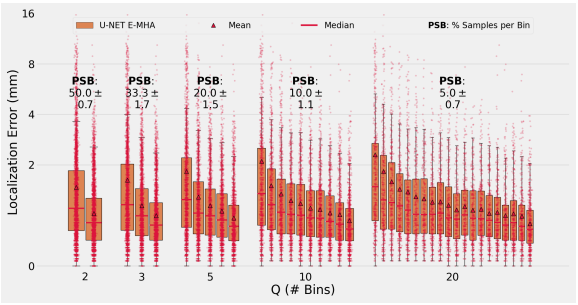
We also found that a well defined upper bound for heatmap activations is important for error bound estimates. For both the 4CH and SA datasets, S-MHA for PHD-Net is significantly more accurate at predicting error bounds for the highest uncertainty quantile $\mathbb{B}_5$ compared to the lowest uncertainty quantile $\mathbb{B}_1$ (56% & 72% compared to 30% & 27% for 4CH & SA, respectively), correlating with S-MHA capturing a greater proportion of those

bins (Jaccard Indexes of 32% & 24% compared to 16% & 15%). On the other hand, U-Net using S-MHA predicts error bounds for low uncertainty bins better than high uncertainty bins. This suggests that although PHD-Net's heatmap activation is a robust indicator of gross mispredictions, the upper error bound of $\mathbb{B}_1$ (γ_1) cannot be accurately predicted due to the loose upper bound of the heatmap activations causing high variance. This is alleviated by using an ensemble of networks in E-MHA, where the $\mathbb{B}_1$ bound accuracy is improved to 62%.

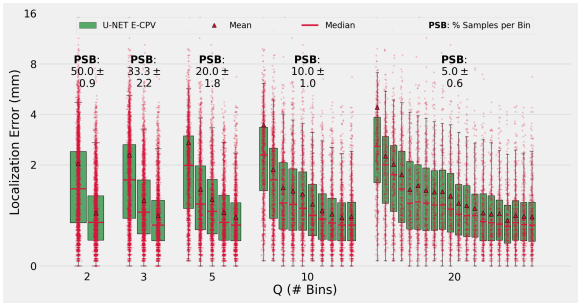
E-MHA and E-CPV are more consistent than S-MHA. Overall, there is no significant difference between the error bound estimation accuracy of E-MHA and S-MHA, but Figures 5.4e & 5.4f show E-MHA has less variation in performance between U-Net and PHD-Net compared to S-MHA, suggesting an ensemble of models is more robust. For the 4CH dataset, PHD-Net using E-CPV is on average significantly more accurate at predicting error bounds than S-MHA and E-MHA. However, there are no significant differences for PHD-Net on the easier SA dataset, nor U-Net on either dataset. There are also no significant differences between U-Net and PHD-Net in error bound estimation accuracy, with each method broadly equally effective for both models.

5.5.5 Generalisability

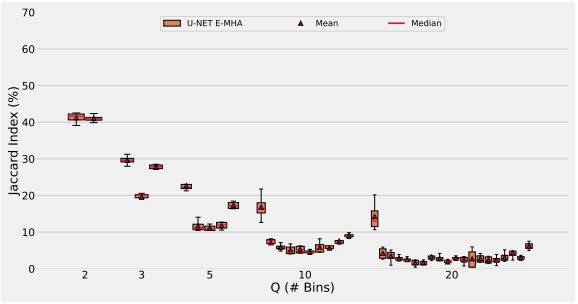
We train U-Net on the Cephalometric dataset, applying Quantile Binning to E-MHA and E-CPV to test their generalisability across imaging modalities. First, we measure the correlation between out uncertainty measures and localisation errors. Similar levels of correlation are seen as with the 4CH and SA datasets for S-MHA and E-MHA, and a slightly stronger correlation for E-CPV (more detail can be found in Appex Section A.1). For $Q = 5$, Figures 5.6c & 5.6d show a predictive power of true error quantiles comparable with the CMR datasets. The mean Jaccard Index (JI) for $\mathbb{B}_5$ is 22% for E-MHA and 34% for E-CPV on the Cephalometric dataset, compared to 22% & 32% for U-Net on the 4CH dataset. $\mathbb{B}_1$ shows a better result than the CMR datasets, achieving a JI of 18% for E-MHA and 19% for E-CPV, compared to 15% & 14% for U-Net on the 4CH dataset. E-CPV more effectively identifies the extreme mis-predictions compared to E-MHA, as evidenced by a higher JI for $\mathbb{B}_5$ (left-most bin) in Figures 5.6c & 5.6d, supporting the results from the challenging 4CH dataset. For $Q = 5$, Figures 5.6a and 5.6b show a gradual reduction in error from $\mathbb{B}_5$ (left-most) to $\mathbb{B}_1$. Overall,



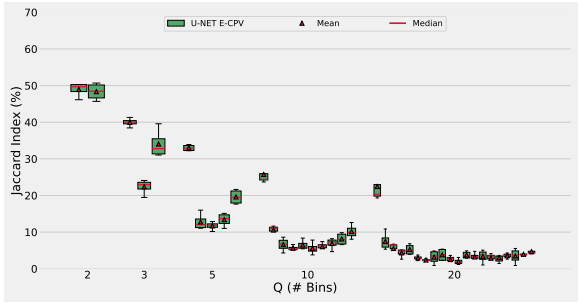
(a) Localisation error - E-MHA (Lower is better).



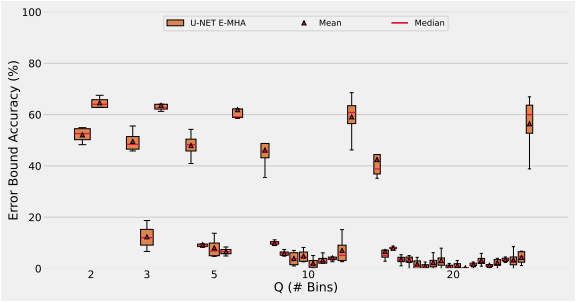
(b) Localisation error - E-CPV (Lower is better).



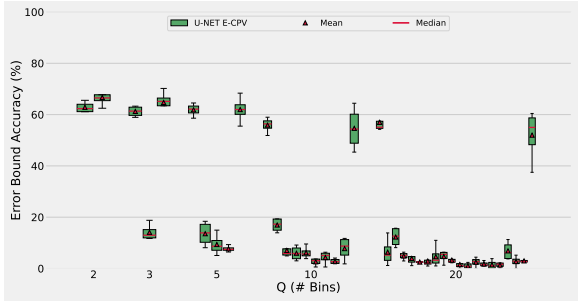
(c) Jaccard Index - E-MHA (Higher is better).



(d) Jaccard Index - E-CPV (Higher is better).



(e) Estimated Error Bound Accuracies - E-MHA (Higher is better).



(f) Estimated Error Bound Accuracies - E-CPV (Higher is better).

Figure 5.6: Quantile Binning varying Q (Number of Quantile Bins) on the Cephalometric dataset. We show results for the uncertainty measures E-MHA and E-CPV, over all landmarks from a 4-fold CV, trained on the U-Net model. Red dots represent the errors of individual samples, best viewed on screen.

the larger Cephalometric dataset (19 landmarks) shows a more consistent downward trend in error across bins compared to our smaller CMR datasets (3 landmarks).

Next, to test the robustness of using MHA as an uncertainty measure across target heatmaps of varying sizes, we repeated these experiments changing the standard deviation of the target heatmap from Equation (5.1) to 2, 4, 8 and 12, shown in Figure 5.7. We found the trends of our Quantile Binning results hold, with only the localisation error deteriorating as we increased the size of the standard deviation of the Gaussian heatmap. We conclude that as long as the standard deviation leads to a learnable heatmap, similar uncertainty estimation properties are exhibited by MHA. Further supporting experiments showing can be found in Section A.4.

5.5.6 Varying Quantile Binning Resolution

We vary the number of Quantile Bins ($Q = \{2, 3, 5, 10, 20\}$) for the larger Cephalometric dataset to gain deeper insights on the flexibility of Quantile Binning. Figures 5.6a & 5.6b show the localisation error quantiles across Q for the Cephalometric dataset, with a gradual reduction in mean localisation error (▲) from $\mathbb{B}_Q$ to $\mathbb{B}_1$ for all values of Q. We find that the edge bins are most useful for all values of Q, with the Jaccard Indexes in Figures 5.6c & 5.6d and error bound accuracies in Figures 5.6e & 5.6f showing parabolic trends, confirming our results from the CMR datasets.

Further, Quantile Binning provides utility for a range of Q values. First, consider the extreme case of $Q = 2$, where the threshold is the median uncertainty of the validation set. Here, $\mathbb{B}_2$ (the high uncertainty bin, left) captures the majority of the gross mispredictions and $\mathbb{B}_1$ (the low uncertainty bin, right) captures the majority of the best predictions. Now, consider the effect of increasing Q, shown in Figures 5.6a & 5.6b. As we increase the number of Quantile Bins, the mean error (▲) of $\mathbb{B}_Q$ (far left bin of each set) increases. This is because as Q increases, $\mathbb{B}_Q$ is pushed farther towards the edge of the uncertainty measure distribution, capturing progressively more extreme outliers. Therefore, as Q increases, we observe an increasingly logarithmic trend of the mean error across the bins as poor predictions are filtered out more gradually.

In practice, the higher the value of Q, the greater the resolution of separation of the data.

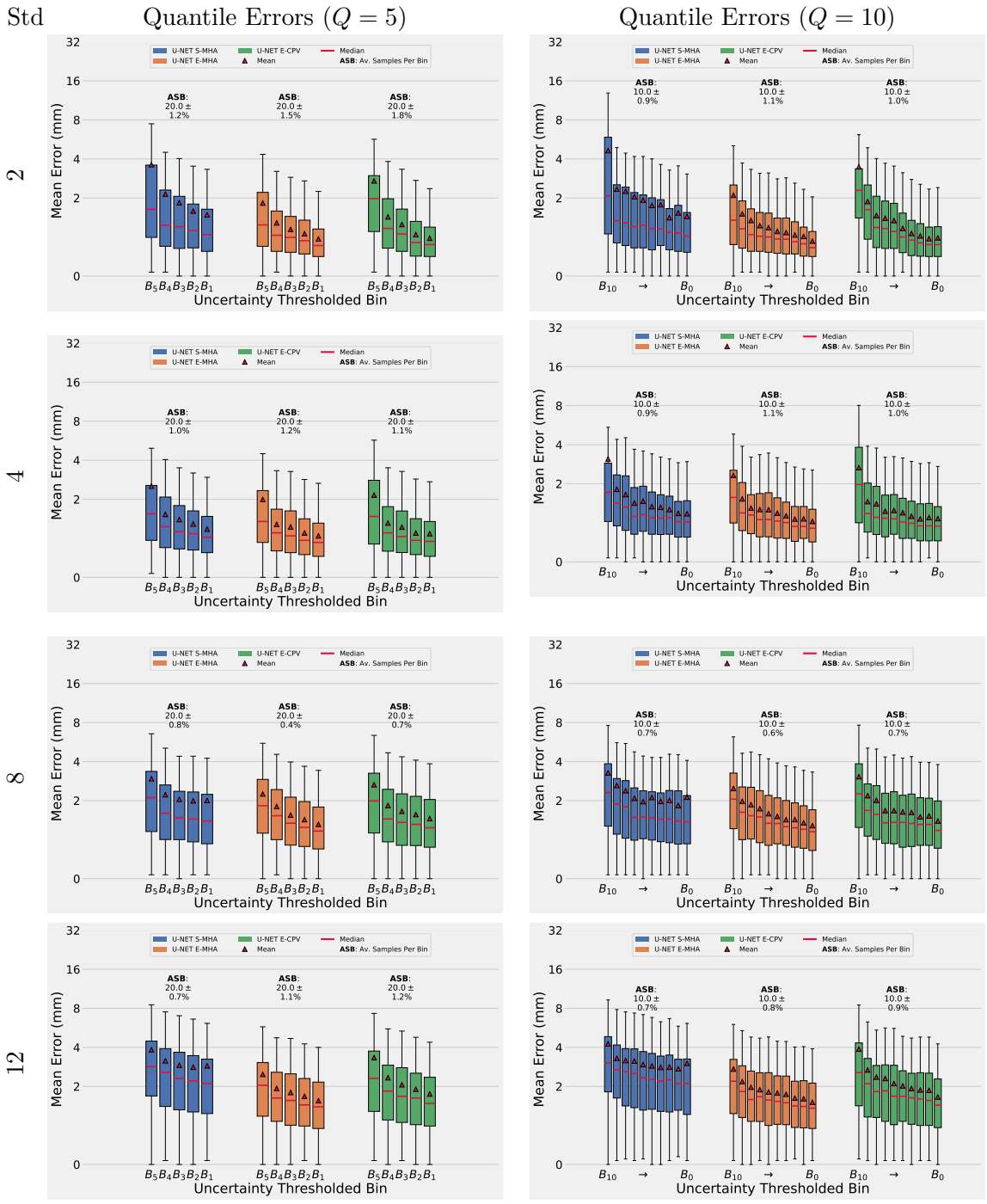


Figure 5.7: Comparing results for models using different standard deviation values for the ground truth heatmap labels. We show the Quantile Localization Errors using 5 & 10 Quantile bins. We present results on all landmarks from a 4-fold CV on the Cephalometric dataset Wang et al. [2016].

For example, consider the task of flagging up uncertain landmark predictions for manual review. Using $Q = 2$ and flagging predictions from the highest uncertainty bin will lead to 50% of predictions requiring review and re-annotation. On the other hand, filtering out the highest uncertainty bin using $Q = 10$ leaves only 10% of predictions to be reviewed. In each case, the user will have an upper error bound estimate for the remaining predictions with reasonable accuracy ($\sim$50% for E-MHA and $\sim$60% for E-CPV, the left-most bins, $\mathbb{B}_Q$, in Figures 5.6e & 5.6f). However, the contents of $\mathbb{B}_Q$ are more accurate when Q is small, with a Jaccard Index of 50% for $Q = 2$ compared to 25% for $Q = 10$ for E-CPV (Figure 5.6d). Therefore, this trade-off between true error quantile accuracy and binning resolution means Q is a subjective choice that depends on the specificity of the downstream task and the resources available for reannotation.

Similar trends are present for our 4CH dataset and SA dataset, but we note that results are poor for $Q >= 10$ compared to the Cephalometric dataset. This is because when fitting the data for Quantile Binning, our CMR datasets had access to a much smaller validation set compared to the Cephalometric dataset ($\sim$30 samples compared to $\sim$60 samples) and could not accurately estimate the quantile uncertainty distribution for large values of Q. Therefore, the larger the available validation set, the larger Q can be set. Full experimental results varying the number of bins for PHD-Net and U-Net over the SA and 4CH datasets can be found in Section A.5.

To address the observation that the intermediate bins ($\mathbb{B}_2 - \mathbb{B}_{Q-1}$) capture less reliable information about localisation error than the two outer bins, we perform an experiment combining all but the edge bins into one large super bin. The results are shown in Figure 5.8. We compared using (a) 20 Quantile Bins; (b) 3 Bins where the edge bins are the same as (a) and the middle bin is a super bin from merging $\mathbb{B}_{19} - \mathbb{B}_2$; and (c) 3 Quantile Bins. From Figure 5.8 we can see that merging middle bins in (b) achieves a higher error bound accuracy due to the greatly relaxed uncertainty bounds. Furthermore, compared to the 3 Quantile bins in (c), (b) retains the benefits of a lower mean error for $\mathbb{B}_1$ and more discriminative outlier detection in $\mathbb{B}_3$.

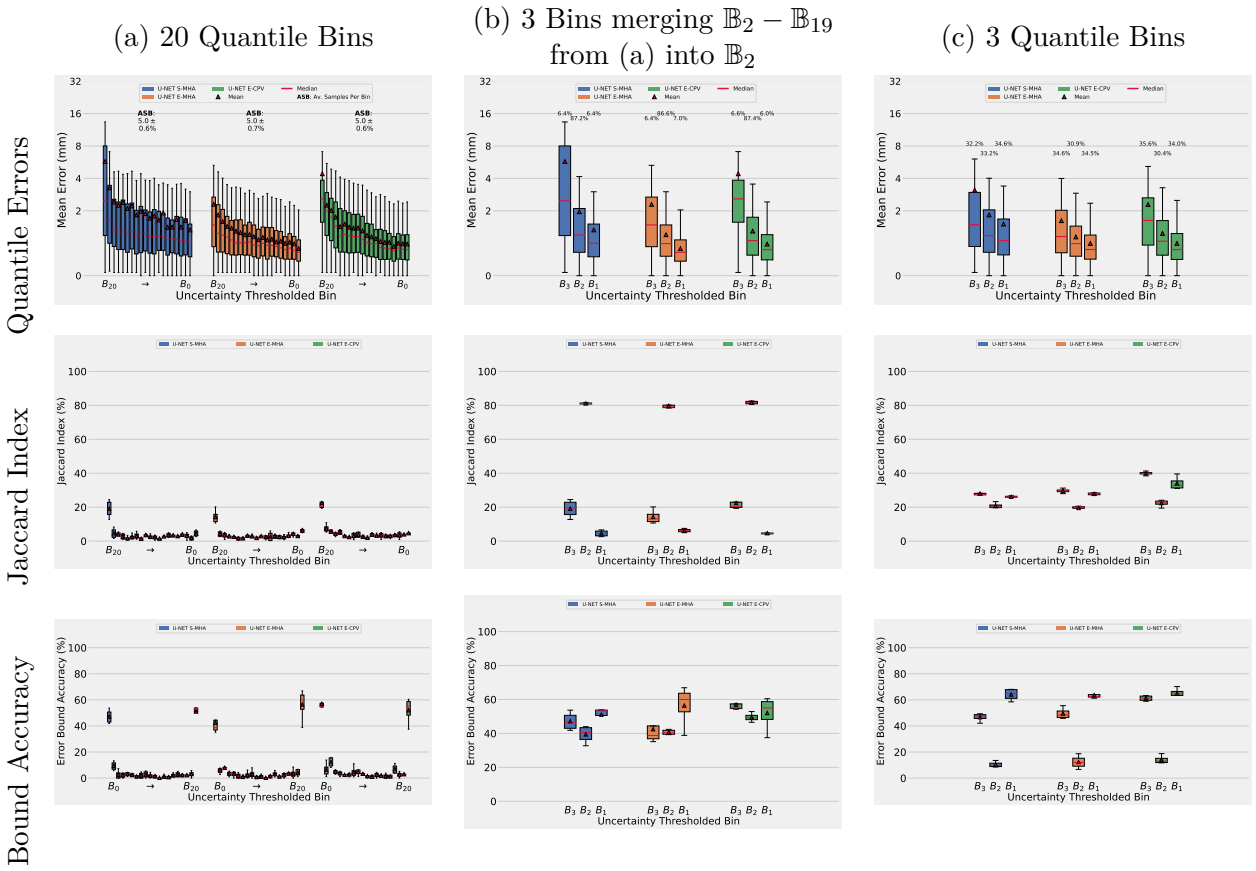


Figure 5.8: Comparing using **(a)** 20 quantile bins, **(b)** 3 Bins where the edge bins are the same as (a) and the middle bin is a super bin from merging $\mathbb{B}_{19} - \mathbb{B}_2$, and **(c)** 3 Quantile Bins. We show the distribution of localization errors in each bin, the Jaccard index of each bin compared to the ground truth error quantiles the estimated error bound accuracies.

5.5.7 Relationship with Aleatoric Uncertainty

Lastly, we study aleatoric uncertainty, which refers to uncertainty caused by internal randomness in the data. Using Quantile Binning, we explore how our predictive uncertainty measures deal with landmarks of varying levels of aleatoric uncertainty. In landmark localisation, one way to measure aleatoric uncertainty is from the inherent ambiguity of the landmark, quantified by the inter-observer variability of multiple annotators. We can infer that the higher the variation in annotator opinion, the greater the ambiguity of the landmark. We can observe the directional ambiguity of the landmark by fitting an anistotropic (directionally skewed) Gaussian function to the distribution of the annotations, seen in the *Annotator Dist.* column of Figure 5.9. Thaler et al. [2021] provide this ground truth measure of the aleatoric uncertainty, using a total of 11 annotators to label a subset of five landmarks (Figure 5.2c) from 100 images of the Cephalometric dataset. We assume the landmark-specific ambiguities hold for the full Cephalometric dataset.

Figure 5.9 shows that all studied coordinate extraction methods are best for landmarks with low aleatoric uncertainty. The mean errors (▲) over the boxplots in the *Quantile Errors* column in Figure 5.9 confirm that landmarks with higher aleatoric uncertainty (L_3, L_2) had worse localisation performance than landmarks with low aleatoric uncertainty (L_4, L_1). However, the distribution of individual samples (represented by red dots, best seen on screen) show that E-MHA and E-CPV reliably capture the majority of gross mispredictions ($\mathbb{B}_5$) regardless of landmark ambiguity. S-MHA performs poorly on some landmarks (L_5, L_2) due to the reliance on a single model capacity. In terms of filtering out poor predictions, we see the best results for all uncertainty methods for the landmark with the tightest annotation distribution (L_4), with $\mathbb{B}_5$ Jaccard Index's showing a mean of 40% and 45% similarity with the true quantile bin for E-MHA and E-CPV, respectively.

However, MHA methods falter for landmarks with directional ambiguity, whereas E-CPV estimates uncertainty well for all types of ambiguity. The *Annotator Dist.* column of Figure 5.9 shows that the annotation distribution of landmarks L_1 and L_2 have a distinct directional skew. The Jaccard Indexes of E-MHA for these landmarks ($L_1 = 16\%$, $L_2 = 20\%$ for $\mathbb{B}_5$) are lower than the other landmarks with more isotropic annotation distributions ($L_4 = 40\%$, $L_5 = 24\%$, $L_3 = 23\%$ for $\mathbb{B}_5$). Furthermore, the mean and median localisation errors do not

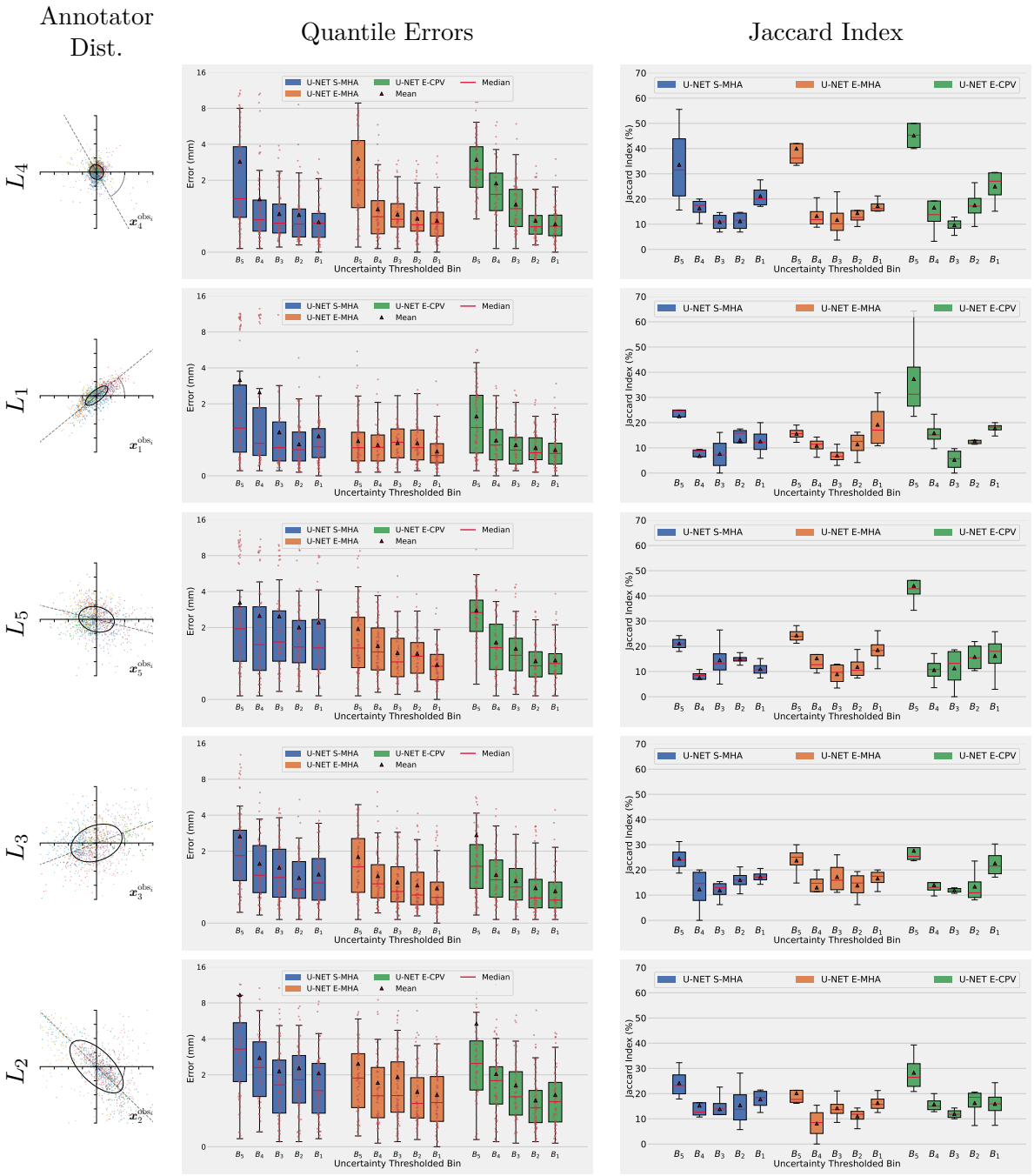


Figure 5.9: Column *Annotator Dist.* shows the individual offsets from each of the 11 annotators to the mean annotation of each landmark [Thaler et al., 2021]. The larger the fitted Gaussian, the more variance between annotators and the higher the aleatoric uncertainty. *Quantile Errors* column shows the boxplots of localisation errors for each quantile bin, showing the landmarks across all folds. The *Jaccard Index* column shows the similarity between the predicted Quantiles and the true error quantiles.

consistently trend down across bins for E-MHA for the anisotropic landmarks (L_1, L_2).

On the other hand, Quantile Binning shows E-CPV is consistently effective regardless of directional ambiguity, with mean Jaccard Indexes for $\mathbb{B}_5$ no less than 28% across all landmarks. This is likely because the objective function (Equation (5.1)) encourages the model to predict isotropic Gaussian Heatmaps, which better match isotropic annotator distributions. When we calculate the mean heatmap to extract the peak pixel using E-MHA, the resulting map will still be constrained to the isotropic properties defined by the objective function. This explains why E-MHA even performs well on the ambiguous yet isotropic landmarks L_3 and L_5, but poorly on the directionally ambiguous, anisotropic landmarks L_1 and L_2. In contrast, E-CPV calculates the variance between peak pixel activations of a group of individual models, where sampling enough independent predictions of an ensemble can effectively approximate the anisotropic distribution. In practice, if the Quantile Error bins for E-MHA show uniformity as they do in L_1 and L_2, this is an indication to the user that the landmark may contain some directional ambiguity.

5.6 Application to Pulmonary Arterial Wedge Pressure Prediction

Quantile Binning was applied to a cardiac classification problem for training sample selection, improving performance. The pipeline consisted of three stages: (1) landmark localisation, which was used to register images to a common orientation, before (2) tensor feature learning using Multilinear Principal Component Analysis (MPCA) [Lu et al., 2008] was applied to four resolution scales of the image followed by an Support Vector Machine (SVM) using the top k-ranked MPCA features for classification of high or low wedge pressure. We hypothesised an incorrect landmark prediction would cause poor image registration, in turn impeding feature extraction and classification performance. This concern was due to unlike CNNs, tensor-based learning is not translationally invariant.

For the landmark localisation and uncertainty estimates, we used E-MHA using an ensemble of 5 LannU-Net models (architecture and training regime detailed in Section 3.3). We chose the larger capacity LannU-Net rather than U-Net since we had access to more training data

Method	Resolution	AUC	Accuracy	MCC
Proposed method	64×64	0.8146 ± 0.04	0.7774 ± 0.03	0.4460 ± 0.02
with uncertainty binning	128×128	$\mathbf{0.8327 \pm 0.06}$	$\mathbf{0.8038 \pm 0.05}$	$\mathbf{0.5099 \pm 0.04}$
Proposed method	64×64	0.7892 ± 0.04	0.7513 ± 0.05	0.4278 ± 0.02
without uncertainty binning	128×128	0.8036 ± 0.03	0.7820 ± 0.04	0.4779 ± 0.01

Table 5.2: Performance comparison using three metrics (with **best** in bold). The standard deviations of methods were obtained by dividing the test set into 5 parts based on the diagnosis time. AUC is Area Under the Curve, a method to measure classification performance. is Matthew's Correlation Coefficient [Chicco and Jurman, 2020], also used for classifier evaluation.

for this task and localisation accuracy was a priority. For the target heatmaps, we used $\sigma = 2$. We trained on a larger cohort from the ASPIRE registry [Hurdman et al., 2012] made available especially for this study, with 1446 SA scans and 1329 4CH scans. The landmarks were the same three as detailed in Section 3.2.1. The dataset for classification consisted of 1346 patients from the ASPIRE registry [Hurdman et al., 2012], and was non-overlapping with the dataset used for the landmark localisation training. It was split into 1081 cases for training and 264 cases for testing. Each patient had a 4CH and SA scan, and landmark localisation was performed using the E-MHA strategy.

To tackle the quality control problem, we first partitioned the training scans based on the uncertainty values of the landmarks. The predicted landmarks of the classification training set were divided into 50 quantiles, i.e., $Q = \{q_1, q_2, ..., q_{50}\}$, based on the epistemic uncertainty values. We then iteratively filtered out training samples starting from the highest uncertainty quantile. A sample is discarded if the uncertainty of any of its landmarks lies in quantile q_k where $k = \{1, 2, ..., 50\}$. The samples are discarded iteratively until there is no improvement in the validation performance, as measured by the area under the curve (AUC), for two subsequent iterations.

Figure 5.10 depicts the results of binning using 10-fold cross-validation on the training set, where the performance improves consistently over the four scales when removed bins ≤ 5. Based on the results, we removed 5 bins (129 out of 1081 samples) from the training set, and used the remaining 952 training samples for the rest of the study. Table 5.2 shows results with and without the uncertainty-based quality control, with a significant increase in classification performance when using the quality control. The reader is encouraged to read the paper for more details [Tripathi et al., 2023].

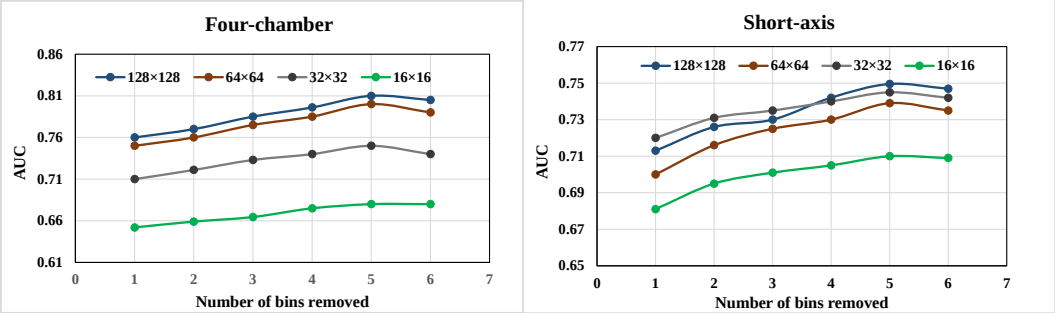


Figure 5.10: Performance comparison of removing a different number of bins of training data on 10-fold cross-validation. Each line corresponds to a different resolution scale of image used. AUC is Area Under the Curve, a method to measure classification performance.

5.7 Discussion and Conclusion

5.7.1 Summary of Findings

This chapter presented a general Frequentist framework to assess any continuous uncertainty measure in landmark localisation, demonstrating its use on three uncertainty metrics and two paradigms of landmark localisation model. We introduced a new coordinate extraction and uncertainty estimation method, E-MHA, offering the best baseline localisation performance and competitive uncertainty estimation.

Our experiments indicate that both heatmap-based uncertainty metrics (S-MHA, E-MHA), as well as the strong baseline of coordinate variance uncertainty metric (E-CPV) are applicable to both U-Net and PHD-Net. Despite the two models' distinctly different approaches to generating heatmaps, using the maximum heatmap activation as an indicator for uncertainty is effective for both models. We showed that all investigated uncertainty metrics were effective at filtering out the gross mispredictions ($\mathbb{B}_Q$) and identifying the most certain predictions ($\mathbb{B}_1$), but struggled to capture useful information for the intermediate uncertainty bins ($\mathbb{B}_2$-$\mathbb{B}_{Q-1}$).

Our experiments also showed that E-MHA and S-MHA had a surprisingly similar ability to capture the true error quantiles of the best and worst 20% of predictions (Figures 5.4c & 5.4d), but E-MHA was more consistent with its performance predicting the error bounds of those bins across models (Figures 5.4e & 5.4f). This suggests that the correlation with localisation error at the head and tail ends of the heatmap distributions are stable across our ensemble of models, but susceptible to variance when fitting our isotonically regressed line to

predict error bounds. On the more challenging 4CH dataset, E-CPV broadly remained the strongest method for filtering out the worst predictions, but this trend did not continue in the easier SA dataset (Fig 5.3).

In terms of error bound estimation, we found bins $\mathbb{B}_Q$ and $\mathbb{B}_1$ could offer good error bound estimates, but the intermediate bins could not (Figures 5.4e & 5.4f). We found all uncertainty methods performed broadly the same: effective at predicting error bounds for $\mathbb{B}_1$ and $\mathbb{B}_Q$, but poor at predicting error bounds for $\mathbb{B}_2$-$\mathbb{B}_{Q-1}$. The one exception was PHD-Net using S-MHA, which could not accurately predict error bounds for $\mathbb{B}_1$ due to the high variance in pixel activations of highly certain predictions.

We demonstrated our Quantile Binning and the three uncertainty metrics are generalisable across imaging modalities by reporting effective results on the Cephalometric dataset in Figure 5.6. Here, we also showed the flexibility of Quantile Binning by varying the number of bins (Q), illustrating the trade-off between true error quantile accuracy and binning resolution as Q increases.

Next, in Section 5.5.7 we explored the effect of aleatoric uncertainty on our predictive uncertainty measures, using Quantile Binning to uncover weaknesses of E-MHA when dealing with landmarks with high directional ambiguity under conventional isotropic heatmap regression.

Finally in Section 5.6 we showed an example usecase of our method, in which we use Quantile Binning and E-MHA as a quality control for landmark predictions, improving downsteam performance.

5.7.2 Recommendations

We offer the following recommendations:

- When resources are available, E-MHA should be used as the coordinate extraction and uncertainty estimation method since it offers the best baseline localisation performance with a sufficient ability to filter out the gross mispredictions.
- If the definition of the landmark is known to be directionally ambiguous, use E-CPV over E-MHA for uncertainty estimation. If this is unknown, uniformity in the E-MHA Quantile Bins can be an indication of directional ambiguity in the landmark.

- When resources are constrained, S-MHA is surprisingly effective at capturing the true error quantiles for bins $\mathbb{B}_1$ and $\mathbb{B}_Q$, but note that when using a patch-based voting heatmap that is not strictly bounded, the error bound estimation for $\mathbb{B}_1$ is not robust.
- The number of Quantile Bins used (Q) is a trade-off, with a larger Q offering a finer binning resolution at the cost of less accurate bins. Q is constrained by the size of the hold-out validation set and can perform poorly when $Q > 10$ and the validation set is smaller than 60 samples.

5.7.3 Conclusion

Beyond the above recommendations, we hope our Frequentist framework described in this chapter can be used to assess refined or novel uncertainty metrics for landmark localisation, and act as a baseline for future work. Not only this, but *Quantile Binning* is application agnostic, relevant to any regression problem that provides sample-wise uncertainty values. Furthermore, we have shown that both the voting derived heatmap of PHD-Net, and the regressed Gaussian heatmap of U-Net can be exploited for uncertainty estimation. In this chapter, we only explored the activation of the peak pixel, but it is likely that more informative measures can be extracted from the broader structure of the heatmap, promising greater potential for uncertainty estimation in landmark localisation waiting to be uncovered.

Chapter 6

Bayesian Uncertainty Estimation with Convolutional Gaussian Processes

6.1 Introduction

So far in this thesis, we have tackled landmark localisation using deep learning approaches. As discussed in Chapter 2, and confirmed by our studies in the previous chapters, it is difficult to obtain consistently reliable uncertainty estimates from deep neural networks. In Chapter 4 we developed a heuristic-based approach, which we extended to a Frequentist approach using deep ensembles and *Quantile Binning* in Chapter 5. However, for the penultimate chapter of this thesis we tackle the research question **Q4**, presenting the first Bayesian framework using Gaussian Processes to anatomical landmark localisation. As far as we are aware, this is the first work to apply Gaussian Processes to an image regression task of this nature.

As evident from previous chapters, current State-of-the-Art approaches for landmark localisation are deep-learning based, the most popular paradigm being heatmap regression. It is noteworthy that conventional training methodologies in heatmap regression models fix the variance of the Gaussian function, preventing the model from modifying its output to reflect elevated or diminished uncertainty in its predictions.

Thaler et al. [2021] incorporated aleatoric uncertainty directly during training by learning

anisotropic Gaussian heatmaps for each landmark. The study demonstrated that the learned heatmap shapes correspond to inter-observer variability from multiple annotators [Thaler et al., 2021]. However, this method only models the homoscedastic aleatoric uncertainty of the dataset, whereby a single covariance matrix is learned over the entire dataset for each landmark during training. At inference, a Gaussian function is fitted to each individual prediction to model heteroscedastic aleatoric uncertainty. Nonetheless, this measure heavily depends on the learned homoscedastic uncertainty, and its post-hoc nature makes it challenging to rely on for a true reflection of model uncertainty. As discussed in Chapter 2 and 4 , other existing approaches to estimate landmark uncertainty include approximate Bayesian inference like deep ensembles [Drevickỳ and Kodym, 2020], and Monte-Carlo Dropout [Lee et al., 2020]. However, these methods rely on deep neural networks and are not truly Bayesian. As far as we know, there are no fully Bayesian methods applied to this dataset.

In this study, we depart from the conventional practice of utilising deep learning and instead employ a Bayesian methodology for landmark localisation, relying on Gaussian processes (GPs). GPs are nonparametric statistical models which are robust to both the presence of noisy data and overfitting, even in low-data regimes which can prove challenging for neural network-based techniques [Rasmussen and Williams, 2006]. Specifically, we use Convolutional Gaussian Processes (CGPs), which offer an attractive alternative to deep neural networks for the task of landmark localisation. CGPs are constructed using a covariance function which is heavily inspired by the efficient convolutional structure of the kernels used in Convolutional Neural Networks (CNNs) [Van der Wilk et al., 2017]. CGPs offer us a mathematically rigorous Bayesian framework for predicting the distribution of likely landmark locations and quantifying model uncertainty.

Due to limitations with the scalability of Convolutional Gaussian Processes (CGPs), we use a two stage coarse-to-fine approach, outlined in Figure 6.1. The first stage uses a CNN to obtain a coarse prediction of the landmark location. Then, the CGP predicts the final landmark distribution using the corresponding cropped patch of the image. The intrinsically Gaussian nature of the uncertainty estimates generated by CGPs render them an intriguing alternative to the conventional deep learning approaches that aim to predict a Gaussian Heatmap (either fixed or learned).

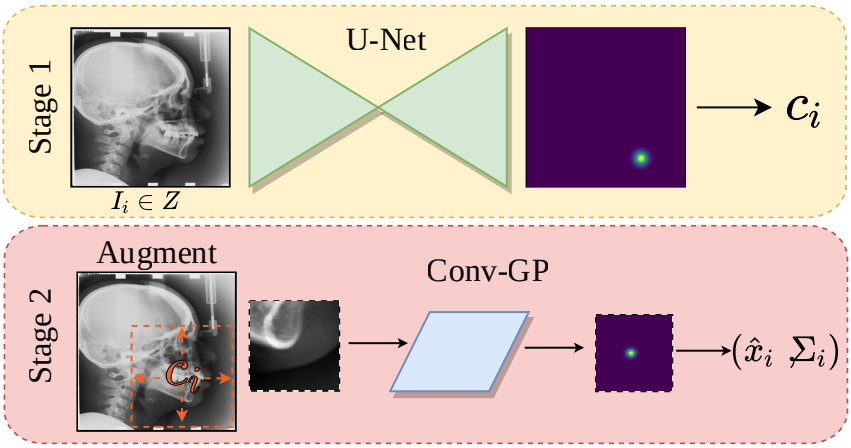


Figure 6.1: Overview of our two stage coarse-to-fine framework. We utilise Deep Learning in Stage 1 to obtain coarse predictions, and refine them in Stage 2 with a Convolutional Gaussian Process and obtain an uncertainty estimate (covariance).

6.2 Contributions

Our contributions are threefold:

- We present a two stage coarse-to-fine approach using a multi-task CGP for fine predictions and uncertainty estimates, which is the first of its kind for regression. The two stage approach alleviates issues with high-resolution data when using CGPs.
- We address optimisation issues by devising a novel approach to initialise inducing patches for the CGP using first-stage prediction information.
- We provide an evaluation of CGP uncertainty estimates against a deep learning baseline CNN method.

We provide an open-source implementation of the model presented in this work in our repository, *MediMarker*: `https://github.com/Schobs/MediMarker`.

6.3 Methods

Due to limitations with the scalability of Convolutional Gaussian Processes (CGPs), we use a two stage coarse-to-fine approach. The first stage uses a CNN to obtain a coarse prediction of

the landmark location. Then, the CGP predicts the final landmark distribution using the corresponding cropped patch of the image.

6.3.1 Stage 1: Coarse Prediction using U-Net

To obtain our coarse predictions, we use LannU-Net, detailed in Section 3.3. To remind the reader, LannU-Net follows the standard configuration of two blocks per resolution layer, with each block consisting of a 3×3 convolution, Instance Normalisation [Ulyanov et al., 2016], and Leaky ReLU (negative slope, 0.01). Downsampling is achieved through strided convolutions and upsampling through transposed convolutions. The initial number of feature maps is set to 32, doubling with each downsample to a maximum of 512 and halving at each upsample step. We automatically configure the number of resolution layers by adding encoder steps until any dimension of the feature map resolution hits a minimum of 4. The objective for the model is to learn a Gaussian heatmap image for each landmark, with the centre of the heatmap on the target landmark. For a landmark L_i with 2D coordinate position $\tilde{\mathbf{c}}^{(i)}$, the 2D heatmap image is defined as the 2D Gaussian function:

$$g_i\left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \tilde{\mathbf{c}}^{(i)}; \sigma\right) = \frac{1}{(2\pi)\sigma^2} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_2^2}{2\sigma^2}\right), \tag{6.1}$$

where $\mathbf{x}$ is the 2D coordinate vector of each pixel and σ is a user-defined standard deviation. The network learns weights $\mathbf{w}$ and biases $\mathbf{b}$ to predict the heatmap $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$. The objective function is the Mean Squared Error (MSE) between the Gaussian target heatmap and the predicted heatmap. We train on images of size $H \times W$ at this stage, obtaining the coarse predictions of our landmarks, $\widehat{\mathbf{C}}_c$. For each landmark L_i, we obtain $\hat{\mathbf{c}}_c^{(i)}$ by selecting the pixel with the highest activation for heatmap g_i.

6.3.2 Stage 2: Fine Prediction using a Convolutional Gaussian Process

Following the initial prediction, we extract cropped patches of size $H' \times W', (H' < H,\ W' < W)$ around each image using our *Stage 1* predictions, $\widehat{\mathbf{C}}_c$. We then use a multi-task CGP for the final, sub-pixel prediction and uncertainty estimate.

Multi-task Convolutional Gaussian Process

Gaussian processes (GPs) are powerful nonparametric Bayesian models, gaining popularity due to their ability to provide a rigorous quantification of predictive uncertainty. A GP defines a distribution over functions and is completely specified by a covariance function $k(\cdot, \cdot)$ and a mean function which is commonly assumed to be zero. For an input $\mathbf{I} \in \mathbb{R}^{D_{\text{in}}}$ (e.g. an image), a GP is denoted by $u(\mathbf{I}) \sim \mathcal{GP}\left(0, k(\mathbf{I}, \mathbf{I}')\right)$ [Rasmussen and Williams, 2006]. The choice of covariance function (also called the *kernel*) affects the variation of the function over the input domain, and many options for the kernel exist. The CNN-inspired *image convolutional kernel* [Van der Wilk et al., 2017] has become the standard tool for applying GPs to computer vision tasks. A GP with this kernel (i.e. a CGP) can be written as,

$$u \sim \mathcal{GP}\left(0, \sum_{p=1}^{P} \sum_{p'=1}^{P} k_g\left(\mathbf{I}^{[p]}, \mathbf{I}'^{[p']}\right)\right), \tag{6.2}$$

where k_g is a base kernel (e.g. RBF or Matérn) which generates a real-valued response value for a square patch of pixels within the image $\mathbf{I}$. P denotes the total number of patches we can extract from our input image, and is therefore determined by the patch size, which we define to be (5×5). Specifically, we use the weighted kernel proposed by the authors, whereby each patch is additionally assigned a learnable weighting parameter. As in the original work, we take a stochastic variational approach to performing inference, using a set of *inducing patches*. The intuition behind the overall approach is that the inducing patches can be considered as analogous to the filters in a conventional CNN. For further details on the approach we refer the reader to the original work [Van der Wilk et al., 2017] and material on variational inducing points-based inference in GPs [Leibfried et al., 2020].

Typically, CGPs are used with a single output. However, for our problem setting, we require a multi-output GP as we wish to predict a 2D coordinate associated with each landmark. To achieve this, we firstly model each output using independent GPs, each with their own separate convolutional kernel, sharing the inducing patches across both outputs. We instantaneously mix the outputs of these two GPs $\mathbf{u} \in \mathbb{R}^2$ using the *linear model of coregionalisation* (LMC) [Alvarez and Lawrence, 2011; Journel and Huijbregts, 1976], such that $\mathbf{f} = \mathbf{W}\mathbf{u}$, where $\mathbf{f} \in \mathbb{R}^2$ are our correlated outputs, and $\mathbf{W} \in \mathbb{R}^{2\times 2}$ is a learnable mixing

matrix.

As this is a regression problem, we use a Gaussian likelihood with independent likelihood noise for each spatial dimension, denoted by l_x and l_y. Therefore, for each prediction we obtain the final sub-pixel coordinate prediction $\hat{\mathbf{c}}^{(i)}$ from the CGP mean prediction $\mathbf{m}^{(i)}$ and a covariance matrix $\mathbf{\Sigma}^{(i)}$, which is the summation of the LMC covariance and the likelihood noise:

$$\hat{\mathbf{c}}^{(i)} = \mathbf{m}^{(i)}, \quad \mathbf{\Sigma}^{(i)} = \mathbf{W}\mathbf{u} + \begin{pmatrix} l_x & 0 \\ 0 & l_y \end{pmatrix}. \tag{6.3}$$

Inducing Patch Initialisation

To initialise the inducing patches, we introduce a bias towards inducing patches proximal to the anticipated landmark location, given by *Stage 1*. We do so because regions in the image near the landmark tend to contain more salient information about its location as compared to regions further away. Therefore, this initialisation sets up a simpler optimisation problem as opposed to choosing the initial patches purely at random. For each inducing patch P, we randomly select an image from the training set and sample a 5×5 patch, selecting a patch centre point $\mathbf{p}_c$ by sampling from a Gaussian distribution over the image:

$$\mathbf{p}_c \sim g\left(\mathbf{x} \mid\mid \boldsymbol{\mu} = \hat{\mathbf{c}}_c^{(i)}; \sigma\right), \tag{6.4}$$

where g is a 2D Gaussian Distribution (see Equation (6.1)), $\hat{\mathbf{c}}_c^{(i)}$ is the coarse prediction for landmark L_i from *Stage 1*, and σ is defined by the user.

Neural Network Baseline

To serve as a Deep Learning baseline for *Stage 2* predictions, we replace the CGP with a compact U-Net architecture. Specifically, we adopt the identical design to LannU-Net used in *Stage 1*, with the only difference being a smaller number of layers (due to the use of smaller 64×64 input images). To obtain a comparable likelihood distribution to the CGP, we follow the approach by Thaler et al. [2021] to fit a 2D Gaussian function to the heatmap prediction, robust least squares fitting method [Branch et al., 1999], obtaining the sub-pixel mean prediction $\hat{\mathbf{c}}^{(i)}$ and covariance matrix $\mathbf{\Sigma}^{(i)}$.

Training Procedure

We train using mini-batching, maxinimising the evidence lower bound (ELBO) at each iteration step (or the MSE for the CNN baseline). At each step, we extract a batch of $H' \times W'$ patches from the image batch, centred around the *Stage 1* predictions, $\widehat{\mathbf{C}}_c$. To prevent overfitting and improve robustness, we implement data augmentation by randomly selecting a new centre point for each patch, $\mathbf{o^{(i)}}$, each time the data is seen using the following equation:

$$\mathbf{o^{(i)}} = \hat{\mathbf{c}}_c^{(i)} + \mathbf{r}, \quad \mathbf{r} \sim R(-D, D)^2, \tag{6.5}$$

where R is Uniform distribution and $D < \frac{H'}{2}, \frac{W'}{2}$, ensuring the *Stage 1* prediction $\hat{\mathbf{c}}_c^{(i)}$ is always present in the image patch.

6.3.3 Evaluation Metrics

To evaluate the quality of uncertainty estimates in our predictive model, we report the Negative Log Predictive Density (NLPD). This metric is a proper-scoring method, able to assess the calibration of our Gaussian outputs. NLPD is defined as:

$$NLPD = -\log p\left(\widetilde{\mathbf{C}}^{(i)} | \widehat{\mathbf{C}}^{(i)}, \mathbf{\Sigma}^{(i)}\right) = -\sum_{j=1}^{N} \log \mathbb{P}\left(\tilde{\mathbf{c}}^{(i,j)} | \hat{\mathbf{c}}^{(i,j)}, \mathbf{\Sigma}^{(i,j)}\right), \tag{6.6}$$

where $\hat{\mathbf{c}}^{(i,j)}$ is the predicated coordinate for landmark L_i in image j, $\tilde{\mathbf{c}}^{(i,j)}$ represents the target coordinates for landmark L_i, and $\mathbf{\Sigma}^{(i)}$ the predicted covariances. The NLPD quantifies the likelihood that the predicted distribution contains the true landmark location. Lower NLPD values indicate a better fit of the model fit to the data, and thus better uncertainty estimates. For completeness, we evaluate localisation performance using point-to-point error, defined by the Euclidean distance/Frobenius norm from a predicated coordinate $\hat{\mathbf{c}}$, to a target coordinate $\tilde{\mathbf{c}}$:

$$\mathcal{D}_{\mathrm{PE}}(\hat{\mathbf{c}} - \tilde{\mathbf{c}}) = ||\hat{\mathbf{c}} - \tilde{\mathbf{c}}||_F. \tag{6.7}$$

6.4 Datasets

6.4.1 Cephalometric Radiographs Subset

We show results on the publicly available Cephalometric dataset [Wang et al., 2016], introduced in Chapter 3, Section 3.2.3. Using the annotator variability made available by Thaler et al. [2021] as a indicator for annotation difficulty caused by aleatoric uncertainty, we select 3 landmarks representing various difficulty levels: the tip of the chin, the corner of the jaw, and the tip of the incisor (exemplified in Figure 6.2). The chin has the smallest annotator disagreement (lowest inter-observer variance) of the subset, and the jaw has the largest annotator disagreement (highest inter-observer variance) of the subset. The images are resized to a resolution of 512×512 pixels, and the final sub-pixel coordinate predictions are scaled to the original resolution. We report results of a 4-fold cross validation (CV) over all 400 images, using the junior annotations only, following convention [Lindner et al., 2016; Payer et al., 2020; Thaler et al., 2021]. We set aside 20% of each fold's training data for our validation set.

6.5 Experiments and Results

6.5.1 Experimental Setup and Training Details

All hyperparameter tuning was performed on the first fold of landmark L_1, the tip of the chin.

Stage 1: For the target Gaussian Heatmap in Equation (6.1), we use $\sigma = 8$. We train for 500 epochs using stochastic gradient descent with an initial learning rate of 0.01, decaying it using the 'poly' scheme, $(1 - epoch/epoch_{max})^{0.9}$ [Chen et al., 2017]. One epoch consists of 150 mini-batches, where each mini-batch is 12 samples. We employ early stopping using a hold-out validation set (20% of training set), stopping training if the validation set's localisation error does not drop for 150 epochs. We employ data augmentations with a probability of 0.5, uniformly sampling from a continuous range $[\alpha, \omega]$: Random scaling [0.8, 1.2], translation [-0.07%, 0.07%], rotation [-45°, 45°], shearing [-16, 16] and vertical flipping.

Stage 2: We set $H' = W' = 64$ to select patches of size 64×64 around the *Stage 1* predictions $\widehat{\mathbf{C}}_c$, reaching a compromise between capturing enough context in the image patch

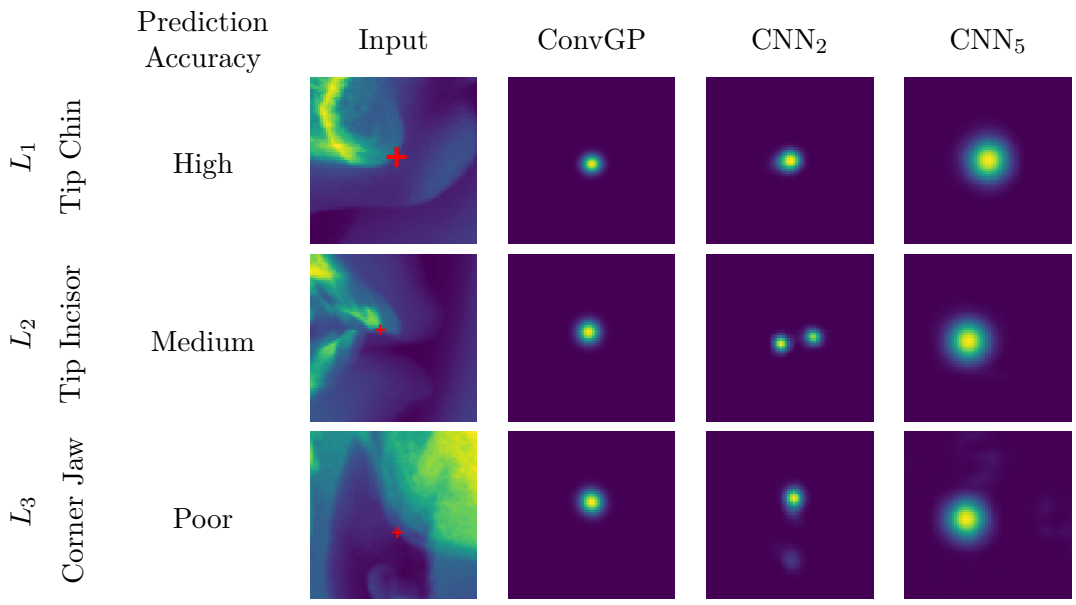


Figure 6.2: Figure showing 3 landmarks of varying difficulty and predictions from our method with learned covariances (ConvGP), and 2 deep learning baselines with fixed covariances (CNN_5 and CNN_2, which use σ values of 5 and 2 in Equation 6.1, respectively.).

and computational limitations. For the data augmentation, we select $D = 32$ for Equation (6.5), enabling the model to train on a diverse set of images. We do not perform further data augmentation. At test time, we choose the model checkpoint with the lowest validation loss during training.

For the Convolutional Gaussian Process, we use a stride of 5 when extracting the 5×5 patches, since a stride of 1 is to demanding in terms of memory. We train for 6000 epochs with minibatches of size 6, using the Adam optimiser [Kingma and Ba, 2014] for stochastic gradient descent with an initial learning rate of 0.01, reducing it after 3000 epochs to 0.001. To prevent the likelihood variance from growing large during optimisation and dominating the posterior covariance, we fix it for 3000 epochs. For the inducing patch sampling, we set $\sigma = 1$ for the sampler in Equation (6.4), heavily biasing the patch initialisations to parts of the image near the *Stage 1* prediction. For the base kernel within the CGP, we use an automatic relevance determination (ARD) Matérn 1/2 kernel, initialised with a lengthscale of 1 for each input dimension, and a variance of 20.

For the CNN baseline, we use the same minibatch size, training length and learning rate schedule as the CGP. We experiment using $\sigma = 2$ and $\sigma = 5$ for Equation (6.1) to compare

	Metric	CGP **(Learned Covariance)**	CNN_2 (Fixed Covariance)	CNN_5 (Fixed Covariance)
All	**NLPD**(↓)	5.54 ± 2.24	6.83 ± 17.64	5.51 ± 1.87
	PE (↓)	1.81 ± 1.11	1.36 ± 1.84	1.15 ± 1.28
L_1	**NLPD** (↓)	4.67 ± 1.33	4.30 ± 3.22	5.25 ± 0.23
	PE (↓)	1.26 ± 0.76	0.91 ± 0.75	0.87 ± 0.58
L_2	**NLPD** (↓	5.35 ± 1.73	5.50 ± 14.14	5.27 ± 0.85
	PE (↓)	1.72 ± 1.00	0.84 ± 1.58	0.68 ± 0.93
L_3	**NLPD**(↓)	6.57 ± 3.55	10.67 ± 35.56	6.02 ± 4.52
	PE (↓)	2.46 ± 1.55	2.33 ± 3.19	1.90 ± 2.34

Table 6.1: Localisation results from 3 landmarks of the Cepalmetric dataset [Wang et al., 2016] over a 4-fold CV. The Negative Log Predictive Density is reported (NLPD, lower is better), the mean point-to-point error (PE), in millimeters. Our non-Deep Learning method, Convolutional Gaussian Process (CGP), is compared to two Deep Learning baseline methods: CNN_2, CNN_5, which use a heatmap label $\sigma = 2$, $\sigma = 5$ in Equation 6.1, respectively.

the sensitivity of uncertainty estimates depending on the target heatmap size. The values respectively represent a high and low precision choice for the target heatmap, typically selected through a hyper-parameter search.

6.5.2 Results and Analysis

Table 6.1 shows that our CGP is capable of reliably quantifying predictive uncertainty, achieving an NLPD of 5.54. The CNN with a target heatmap of $\sigma = 5$ (CNN_5) achieves a marginally lower NLPD score of 5.51, despite achieving much better localisation error of 1.15mm compared to the CGP's error of 1.84mm. Notably, an improved mean estimate (PE) will in turn improve the Negative Log Predictive Density (NLPD). Therefore, the fact that the CGP achieves an NLPD within approximately 0.03 of CNN_5 is encouraging.

Moreover, the performance of the CNN model, in terms of both the PE and NLPD, is highly dependent on the hyperparameter σ, resulting in significant variations between CNN_5 and CNN_2. This reflects the limitations of deep learning methods in measuring uncertainty, as the NLPD score is almost entirely dependent on the mean estimate. This is exemplified by the results of CNN_2 on L_3, where a poor PE of 2.33mm leads to extremely poor uncertainty estimations, resulting in an NLPD score of 10.67. In contrast, the CGP, with a higher PE error of 2.46mm, still manages to produce more reliable uncertainty estimates, achieving an

NLPD score of 6.57. Unlike the CNNs, the covariance of the CGP is learned during training and mathematically grounded, making it a more trustable and stable measure of the model's confidence.

Figure 6.2 highlights how the distribution of the CNN's output heatmap is highly dependent on the value of σ from the target function in Equation (6.1), and is uniform across all landmarks. Note, that CNN_2 even has multiple hallucinated covariances. In contrast, the covariances learned by the CGP are distinct for each landmark, but unfortunately show a relatively uniform pattern across predictions of the same landmark. This uniformity is due to the fact that the covariance function was dominated by the likelihood noise during training, resulting in more homogeneous uncertainty estimates than optimal.

6.6 Discussion and Conclusion

6.6.1 Summary of Findings

We showed that Convolutional Gaussian Proccesses (CGPs) can be applied to the complex vision task of landmark localisation, allowing us to quantify the uncertainty associated with predictions using a nonparametric approach. Empirically, the localisation error obtained with CGPs are not yet competitive with those of a CNN. This is attributed to optimisation challenges related to the GP likelihood noise, which dominates the output covariance matrices, resulting in relatively uniform uncertainty estimates.

Despite this, the CGP obtains a similar NLPD to the CNN baseline, suggesting that if this optimisation issues can be addressed and the predictive error decreased, the uncertainty quantification provided by the CGP would be far superior to that of the CNN. This highlights the primary challenge of using GPs for this task: the hyperparameter tuning is prohibitively expensive. Our Stage 2 CNN baseline was initialised and trained in less than an hour with no hyperparameter tuning, but achieving usable results with the CGP required weeks of careful hyperparameter tuning for the kernel and training regime. An exciting direction for future work which could simplify the task would be to apply Deep Kernel Learning [Wilson et al., 2016], in which we use a CNN backbone for feature extraction and a GP head for mean and covariance prediction.

On the other hand, a concrete benefit of the method is that the uncertainty estimates provided by GPs are more rigorous mathematically and more interpretable in practice than those of deep learning. A promising avenue for future work could address the uniformity of the uncertainty estimates by incorporating a heteroscedastic likelihood that produces different likelihood variance outputs for each image.

6.6.2 Conclusion

In this chapter, we presented a Bayesian approach to landmark localisation with uncertainty estimation. As far as we are aware, this is the first multi-output Convolutional Gaussian Process (CGP) approach to an image regression problem. Through careful experimental design, inducing patch initialisation and hyperparameter tuning, we achieved a CGP model that can localise landmarks Cephalometric landmarks to within a 1.3-2.5mm accuracy. Despite lagging behind in accuracy to a deep learning baseline, the uncertainty estimates our method provides are mathematically founded and therefore more reliable than those derived from deep learning methods. If the optimisation issues can be overcome, CGPs represent an exciting future for a Bayesian approach to trustable uncertainty estimation for landmark localisation.

Chapter 7

Discussion and Conclusions

In this thesis, we improved landmark localisation of lightweight models using a a multi-task patch-based framework with uncertainty estimation (*PHD-Net*), proposed a Frequentist uncertainty estimation framework for continuous regression and applied it to heatmap-based landmark localisation (*Quantile Binning*), and proposed the first Gaussian Process approach to landmark localisation providing truly Bayesian uncertainty estimation. All contributions are open-source and accessible to experts and non-experts as low-code/no-code solutions.

7.1 Contributions to Research

In this thesis, we set out to answer four principal research questions.

Q1: How can we develop a lightweight and data-efficient Deep Neural Network for landmark localisation? In real-world applications of AI in healthcare systems practitioners are reluctant to send sensitive data to third-party systems based on the cloud. Unable to harness the power of computationally powerful models based off-site, practitioners may be constrained to their less powerful local machines. Recognising this constraint, we highlighted the pivotal need for lightweight models with small memory footprints and fast inference. In Chapter 4 we tackled this challenge of improving parameter-efficient models, proposing a patch-based multi-task network: PHD-Net. We showed that our novel loss function that that trains the model to identify globally likely coarse locations as well as locally-focused, pixel precise locations improved the performance of our patch-based model. Furthermore,

our novel branch-fusion strategies *Adaptive Prediction* and *Candidate Smoothing* improved localisation accuracy and performed comparably or better to similarly sized State-of-the-Art models. Alongside this contribution to lightweight models, we showed the patch-based training regime scaled well with model capacity for single landmark localisation. Notably, it introduces the unique benefit of estimating uncertainty through "patch votes", allowing us to discern between high and low error predictions using a calibration set - a foundational principle that underpinned our subsequent uncertainty explorations.

Q2: Can heuristic uncertainties in landmark localisation be formalised using a data-driven, Frequentist framework? Building upon our insights in Chapter 4, we translated our concept of heuristic uncertainty derived from heatmaps to approximate Bayesian inference using Deep Ensembles. We proposed a novel method to fuse model outputs to calculate the mean prediction alongside prediction variance, *Ensemble Maximum Heatmap Activation*. Through extensive evaluations on heatmap-based uncertainty measures spanning patch-based and encoder-decoder style models, we showcased the broad applicability of our uncertainty measures. These can seamlessly be integrated into any State-of-the-Art landmark localisation model that regresses heatmaps as its objective function. Furthermore, we proposed a more structured paradigm to uncertainty using Frequentist principles: *Quantile Binning*. We proposed using Isotonic Regression with a hold-out calibration to approximate the true monotonic function between error and uncertainty. Using the learned function with and a user-defined number of bins, we calculated uncertainty thresholds for bins of increasing uncertainty, each bin accompanied by estimated error bounds. Our *Quantile Binning* approach is application agnostic, and can be used by any continuous regression problem that provides sample-wise uncertainty.

Q3: How can we better benchmark uncertainty measures in landmark localisation within a Frequentist context? Historically, uncertainty estimation from post-hoc methods in landmark localisation were primarily evaluated by assessing the measure's correlation with localisation error [Drevický and Kodym, 2020; Thaler et al., 2021]. Although a useful preliminary indicator, this one-dimensional measure offers only a narrow lens into the intricacies of the uncertainty measure. In Chapter 5, we were faced with the challenge of how to more holistically evaluate our uncertainty measures within our *Quantile Binning*

framework. To this end, we proposed two evaluation metrics for binning based uncertainty methods: 1) Jaccard Index Similarity, in which we evaluate the similarity to the theoretically perfect ground truth bins using the Jaccard Index similarity measure; and 2) Error Bound Accuracy, where we measure how well our uncertainty measures predict localisation error beyond simple correlation by measuring how accurate the estimated bin error bounds are.

Q4: How can we overcome the computational challenges with Gaussian Processes for rigorous, Bayesian uncertainty estimation for landmark localisation? Deep Learning based uncertainty is often miscalibrated [Guo et al., 2017], a challenge which can be partially remedied using data-driven, Frequentist methods as shown in Chapter 4 and Chapter 5. However, training a model within the Bayesian paradigm intrinsically provides a distribution of predictions without requiring post-hoc analysis using calibration sets. This Bayesian approach offers an entirely different and more mathematically rigorous uncertainty estimation compared to Frequentist approaches, a paradigm that is largely unexplored in landmark localisation. To this end, in Chapter 6 we proposed the first application of Gaussian Processes (GPs) to landmark localisation using a Convolutional Gaussian Process (CGP), achieving the first truly Bayesian uncertainty estimation for the task. Confronted with the notorious computational issues with GPs, we proposed a two stage approach. First, we used a Deep Learning model for coarse predictions, followed by a CGP centred around the coarse prediction for the final coordinate prediction with uncertainty estimation. To improve optimisation, we proposed a novel inducing patch initialisation based on predictions from the first stage. The final results gave trustable uncertainty measures, but compromised on localisation accuracy. Beyond landmark localisation, to the best of our knowledge, our work represents the first application of a multi-task CGP for image regression, which is a significant contribution to the GP and medical imaging community.

7.2 Contributions to Open-Source

7.2.1 MediMarker

The models presented in this thesis are freely available and open-source in a single unified framework for landmark localisation, *MediMarker* [Schobs, 2022]: `https://github.com/`

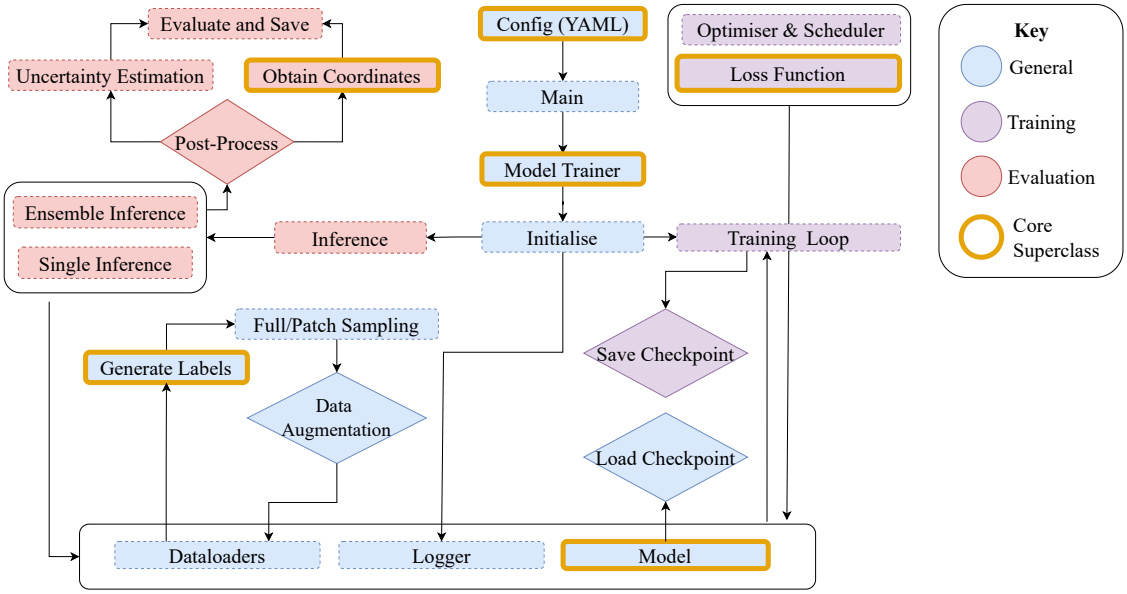


Figure 7.1: Program flow of *MediMarker*. The *Core Superclass* modules are the key configurable/extendable classes needed to build new methods into *MediMarker*.

`Schobs/MediMarker`. *MediMarker* is an out-of-the-box automated pipeline for landmark localisation, supporting uncertainty estimation for model predictions. For a user, the only requirement is to provide data in the correct format. Pre-trained models are available for inference. The dedication to collaborative, open research means research can have a life beyond acceptance of publication.

To ensure accessibility to non-experts, we use a low-code/no-code approach for system configuration, called YAML [Ben-Kiki et al., 2009]. Considering the complexity of configuring modern machine learning systems, using command lines or Python modules can be challenging. YAML offers a more readable and manageable alternative, making configuration changes much easier via a default configuration specified in a single file. Non-expert users can customise configurations by changing text in the file, where each option is comprehensively documented. We use version control using Git, following standard software practices.

The pipeline of *MediMarker* is simple, and the default model is based on the U-Net architecture, LannU-Net, described in Section 3.3. By default, we use heatmap regression for landmark localisation. As a non-expert user, after customising the YAML file to point towards the data, the system automatically configures the architecture based on the size of

the images. From there, training, validation and inference is available by simply changing a setting in the YAML file.

As a researcher/developer, it is trivial to add additional models, loss functions, training schemes etc. by extending a few classes, as shown in Figure 7.1. The framework reduces the need for thousands of lines of boiler plate code common to deep learning, medical imaging and specifically landmark localisation. Both PyTorch and TensorFlow are integrated seamlessly. At it's core, the framework runs on the concept of a *Model Trainer*. This is a superclass which automates the entire machine learning training and evaluation pipeline based on the YAML configuration file. To add another solution to landmark localisation into *MediMarker*, one should make a child class of *Model Trainer* and define/redefine four core functions/classes, indicated by the gold outlines in Figure 7.1:

1. `Model.` A PyTorch or TensorFlow model object with trainable weights e.g. PHD-Net, U-Net, a Convolutional Gaussian Process.

2. `Generate Labels.` A python function that given coordinates for a sample, generates a label for the model e.g. Gaussian Heatmap for U-Net or patch-wise heatmap & displacements for PHD-Net.

3. `Loss Function.` A PyTorch or TensorFlow loss function which calculates some error based on model output and target e.g. Mean Square Error.

4. `Obtain Coordinates.` A python function that returns a coordinate prediction from a model output e.g. Coordinates of the pixel with maximum activation for Heatmap regression or *Candidate Smoothing* for PHD-Net.

MediMarker was used by three BSc. students to facilitate their dissertation research, as detailed in Section 1.5. In each case, a Github *branch* was created from the main repository and each student extended the requisite classes for their experiments. For example, PHD-Former (detailed in Section 4.5.1) was implemented by creating a single `Model` class. Due to the modular nature of *Medimarker*, training and evaluation could be performed completely automatically. Furthermore, *Test Time Augmentation* and *Monte-Carlo Dropout* were integrated into the framework with ease due to extensive documentation and collaboration enabled by Github.

The open-source, standardised, collaborative and well documented framework of *MediMarker* promotes re-usability of research as well as reproducability, maximising the impact of the research in this thesis. Furthermore, the low-code/no-code options for non-experts improve the accessibility of the machine learning solutions presented in this thesis, again improving potential impact of the work.

7.2.2 PyKale

Beyond landmark localisation and *MediMarker*, we have integrated *Quantile Binning* (proposed in Chapter 5) into *PyKale*, an open-source framework that is officially a member of the PyTorch ecosystem [Lu et al., 2022b]. *PyKale* focuses on accessible machine learning from multiple sources and follows industry standard software engineering practices including standardisation, documentation and testing. Since *Quantile Binning* is applicable to any uncertainty estimation regression problem with sample-specific uncertainty scores, the integration of the method into *PyKale* promises greater impact outside of the domain of landmark localisation.

7.3 Future Developments

7.3.1 Transformer-Powered Patch-based Models for Landmark Localisation

In Chapter 4 we improved localisation accuracy for lightweight patch-based models for single landmark localisation. Yet, when scaling this patch-based training to larger capacity models for multi-landmark prediction, traditional heatmap regression methods surpassed them in performance. We hypothesise that the more complex multi-task objective function is noisier to optimise compared to the traditional Gaussian heatmap regression. However, using a Vision Transformer [Dosovitskiy et al., 2020] as the backbone network, there is undoubtedly room for further optimisation and tuning. Future work could continue down this patch-based paradigm with larger models, offering patch-wise landmark predictions with the intuitive heuristic uncertainty based on “patch votes”.

7.3.2 Uncertainty Estimation with Quantile Binning: Beyond Landmark Localisation

In Chapter 5 we introduced *Quantile Binning*, a general Frequentist framework for uncertainty estimation for continuous regression problems. While its efficacy in landmark localisation was established, its incorporation into *PyKale* [Lu et al., 2022b] hints at future opportunities for research beyond the domain explored here.

7.3.3 Convolutional Gaussian Processes for Landmark Localisation

In Chapter 6 we proposed the first Gaussian Process (GP) for landmark localisation using a multi-stage pipeline. We achieved comparable localisation accuracy to a Deep Learning baseline, although the latter demonstrated superior accuracy. However, the Bayesian uncertainty estimation showcased significant potential. Future work could improve the optimisation of the GP training, with a particular focus on preventing the likelihood noise dominating the covariance function. Furthermore, a clear direction for future work could address the uniformity of the uncertainty estimates by incorporating a heteroscedastic likelihood that produces different likelihood variance outputs for each image. One final avenue to point out is to combine the benefits of Deep Learning with Gaussian Processes using Deep Kernel Learning [Wilson et al., 2016]. Such an approach would involve performing feature extraction using a deep neural network, with a GP head making the final mean and variance prediction, trained end-to-end. This would have the advantage of being a single stage method, and could harness a pre-trained landmark localisation model to initialise the feature extractor. A limitation, however, is that uncertainty pertains to the latent image representation rather than the actual image itself.

7.3.4 Improving Practical Application and Interpretability

The practical benefits of the uncertainty estimation techniques introduced in this thesis include the ability to flag potentially erroneous model predictions, which can be corrected manually by a human-in-the-loop. However, no such study was undertaken to measure the efficacy or practicality of this application. While technical advancements are crucial, the integration of human-in-the-loop perspectives is equally important to ensure these technologies are effectively

translated into real-world medical scenarios.

Therefore, future work could focus on establishing collaborative frameworks with clinicians to evaluate the proposed models in practical medical scenarios. This could involve usability studies, where clinicians interact with the models in simulated or real clinical environments, providing feedback on the model's performance, interpretability of uncertainty estimates, and overall integration into the clinical workflow.

Furthermore, involving a human-in-the-loop in the model development process itself presents a promising avenue for exploration with active learning [Ren et al., 2021]. This technique involves a human during training acting as an oracle, annotating data when necessary. In our case, highly uncertain predictions could be passed to the oracle for correction before being fed back into model training, improving model accuracy.

Finally, the issue of interpretability is critical in the medical domain, where the demand for explainable and transparent AI models is particularly pronounced [WHO, 2021]. The ability of medical professionals to understand and trust the outputs of AI models is paramount, as these tools increasingly inform critical decisions regarding diagnosis, treatment planning, and patient management. While this thesis has made progress in quantifying model uncertainty, this is only one aspect of interpretability. At this stage in time, a fundamental problem of relying on Deep Neural Networks is that they are "black boxes"; their decision-making processes incomprehensible to a human. Therefore, future research should also prioritize the development of methodologies that enhance their interpretability, where a human-in-the-loop could allow practitioners to provide direct input and oversight.

7.3.5 Advocacy for Open-Source Software

Finally, I would like to end by emphasising the importance of open-source and reiterating the commitment to it throughout this thesis. By integrating the work in this thesis with the actively maintained package *PyKale* [Lu et al., 2022b], as well as *MediMarker*, it ensures the progress made in this journey has a life beyond it. Future researchers can easily reproduce, reuse and build on the ideas presented this thesis; an optimistic and fulfilling promise after the long and arduous journey of a PhD.

Bibliography

Al, W. A., Jung, H. Y., Yun, I. D., Jang, Y., Park, H.-B., and Chang, H.-J. (2018). Automatic aortic valve landmark localization in coronary CT angiography using colonial walk. *PloS one*, 13(7):e0200317.

Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., and Asari, V. K. (2019). Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging*, 6(1):014006.

Alvarez, M. A. and Lawrence, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. *The Journal of Machine Learning Research*, 12:1459–1500.

Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196.

Angelopoulos, A. N. and Bates, S. (2023). Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591.

Arik, S. Ö., Ibragimov, B., and Xing, L. (2017). Fully automated quantitative cephalometry using convolutional neural networks. *Journal of Medical Imaging*, 4(1):014501.

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Beichel, R., Bischof, H., Leberl, F., and Sonka, M. (2005). Robust active appearance models and their application to medical image analysis. *IEEE Transactions on Medical Imaging*, 24(9):1151–1169.

Ben-Kiki, O., Evans, C., and Ingerson, B. (2009). Yaml ain't markup language (yaml™) version 1.1. *Working Draft 2008-05*, 11.

Bier, B., Goldmann, F., Zaech, J.-N., Fotouhi, J., Hegeman, R., Grupp, R., Armand, M., Osgood, G., Navab, N., Maier, A., et al. (2019). Learning to detect anatomical landmarks of the pelvis in X-rays from arbitrary views. *International Journal of Computer Assisted Radiology and Surgery*, 14:1463–1473.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Branch, M. A., Coleman, T. F., and Li, Y. (1999). A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, 21(1):1–23.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848.

Chen, R., Ma, Y., Chen, N., Lee, D., and Wang, W. (2019). Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 873–881. Springer.

Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563.

Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59.

Csillag, D., Paes, L. M., Ramos, T., Romano, J. V., Schuller, R., Seixas, R. B., Oliveira, R. I., and Orenstein, P. (2023). AmnioML: Amniotic Fluid Segmentation and Volume Prediction with Uncertainty Quantification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15494–15502.

Damianou, A. and Lawrence, N. D. (2013). Deep Gaussian Processes. In *Artificial Intelligence and Statistics*, pages 207–215. PMLR.

D'Angelo, F. and Fortuin, V. (2021). Repulsive deep ensembles are bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465.

Davison, A. K., Lindner, C., Perry, D. C., Luo, W., Cootes, T. F., et al. (2018). Landmark localisation in radiographs using weighted heatmap displacement voting. In *International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, pages 73–85. Springer.

Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural safety*, 31(2):105–112.

Donner et al. (2013). Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization. *Medical Image Analysis*, 17(8):1304–1314.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Drevickỳ, D. and Kodym, O. (2020). Evaluating Deep Learning Uncertainty Measures in Cephalometric Landmark Localization. In *BIOIMAGING*, pages 213–220.

Emad, O., Yassine, I. A., and Fahmy, A. S. (2015). Automatic localization of the left ventricle in cardiac MRI images using deep learning. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 683–686. IEEE.

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Gal, Y. (2016). *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Gilmour, L. and Ray, N. (2020). Locating cephalometric X-Ray landmarks with foveated pyramid attention. In *Medical Imaging with Deep Learning*, pages 262–276. PMLR.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Graves, A. (2011). Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*, 24.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Han, D., Gao, Y., Wu, G., Yap, P.-T., and Shen, D. (2014). Robust anatomical landmark detection for MR brain image registration. In *17th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–193. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Bision and Pattern recognition*, pages 770–778.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

Hoffmann, L. and Elster, C. (2021). Deep ensembles from a Bayesian perspective. *arXiv preprint arXiv:2105.13283*.

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cats visual cortex. *The Journal of Physiology*, 160(1):106–154.

Hurdman, J., Condliffe, R., Elliot, C., Davies, C., Hill, C., Wild, J., Capener, D., Sephton, P., Hamilton, N., Armstrong, I., et al. (2012). ASPIRE registry: assessing the Spectrum of Pulmonary hypertension Identified at a REferral centre. *European Respiratory Journal*, 39(4):945–955.

Ibragimov, B., Likar, B., Pernuš, F., and Vrtovec, T. (2014). Shape representation for efficient landmark-based segmentation in 3-D. *IEEE Transactions on Medical Imaging*, 33(4):861–874.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

Jiang, Y., Li, Y., Wang, X., Tao, Y., Lin, J., and Lin, H. (2022). CephalFormer: Incorporating global structure constraint into visual features for general cephalometric landmark detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 227–237. Springer.

Johnson, H. J. and Christensen, G. E. (2002). Consistent landmark and intensity-based image registration. *IEEE Transactions on Medical Imaging*, 21(5):450–461.

Joshi, I. and Morley, J. (2019). *Artificial Intelligence: How to get it right*. NHSX, London, United Kingdom. Accessed: 7th August 2023.

Journel, A. G. and Huijbregts, C. J. (1976). Mining geostatistics.

Jungo, A., Balsiger, F., and Reyes, M. (2020). Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282.

Karimi, D., Zeng, Q., Mathur, P., Avinash, A., Mahdavi, S., Spadinger, I., Abolmaesumi, P., and Salcudean, S. E. (2019). Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Medical Image Analysis*, 57:186–196.

Kasel, A. M., Cassese, S., Bleiziffer, S., Amaki, M., Hahn, R. T., Kastrati, A., and Sengupta, P. P. (2013). Standardized imaging for aortic annular sizing: implications for transcatheter valve selection. *JACC: Cardiovascular Imaging*, 6(2):249–262.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2796–2804. PMLR.

Kwon, Y., Won, J.-H., Kim, B. J., and Paik, M. C. (2020). Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

Lampert, T. A., Stumpf, A., and Gançarski, P. (2016). An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Transactions on Image Processing*, 25(6):2557–2572.

Lampinen, J. and Vehtari, A. (2001). Bayesian approach for neural networks—review and case studies. *Neural networks*, 14(3):257–274.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. (2015). Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570. PMLR.

Lee, J.-H., Yu, H.-J., Kim, M.-j., Kim, J.-W., and Choi, J. (2020). Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. *BMC Oral Health*, 20(1):1–10.

Leibfried, F., Dutordoir, V., John, S., and Durrande, N. (2020). A tutorial on sparse Gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*.

Li, Y., Alansary, A., Cerrolaza, J. J., Khanal, B., Sinclair, M., Matthew, J., Gupta, C., Knight, C., Kainz, B., and Rueckert, D. (2018). Fast multiple landmark localisation using a patch-based iterative network. In *21st International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 563–571. Springer.

Lindner, C., Bromiley, P. A., Ionita, M. C., and Cootes, T. F. (2014). Robust and accurate shape model matching using random forest regression-voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1862–1874.

Lindner, C., Wang, C.-W., Huang, C.-T., Li, C.-H., Chang, S.-W., and Cootes, T. F. (2016). Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Scientific Reports*, 6(1):1–10.

Liu, D., Zhou, K. S., Bernhardt, D., and Comaniciu, D. (2010). Search strategies for multiple landmark detection by submodular maximization. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2831–2838. IEEE.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

Lu, C., Lemay, A., Chang, K., Höbel, K., and Kalpathy-Cramer, J. (2022a). Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12008–12016.

Lu, H., Liu, X., Zhou, S., Turner, R., Bai, P., Koot, R., Chasmai, M., Schobs, L., and Xu, H. (2022b). PyKale: Knowledge-aware machine learning from multiple sources in Python. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*.

Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks*, 19(1):18–39.

McCouat, J. and Voiculescu, I. (2022). Contour-hugging heatmaps for landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20597–20605.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*.

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., and Kapur, T. (2020). Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878.

Mehta, R., Filos, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Murugesan, G. K., et al. (2022). QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. *The Journal of Machine Learning for Biomedical Imaging*, 2022.

Miao, S., Lucas, J., and Liao, R. (2012). Automatic pose initialization for accurate 2D/3D registration applied to abdominal aortic aneurysm endovascular repair.

Murphy, K., van Ginneken, B., Klein, S., Staring, M., de Hoop, B. J., Viergever, M. A., and Pluim, J. P. (2011). Semi-automatic construction of reference standards for evaluation of image registration. *Medical Image Analysis*, 15(1):71–84.

Murphy, K. P. (2013). *Machine Learning: A Probabilistic Perspective, Chapter 15*. MIT Press.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59. Art. no. 101557.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer.

Noothout, J., de Vos, B., Wolterink, J., Leiner, T., and Isgum, I. (2018). CNN-based landmark detection in cardiac CTA scans. In *Medical Imaging with Deep Learning. MIDL Amsterdam*, pages 1–11.

Noothout, J. M., De Vos, B. D., Wolterink, J. M., Postma, E. M., Smeets, P. A., Takx, R. A., Leiner, T., Viergever, M. A., and Išgum, I. (2020). Deep learning-based regression and classification for automatic landmark localization in medical images. *IEEE Transactions on Medical Imaging*, 39(12):4011–4022.

Oh, K., Oh, I.-S., Lee, D.-W., et al. (2020). Deep anatomical context feature learning for cephalometric landmark detection. *IEEE Journal of Biomedical and Health Informatics*, 25(3):806–817.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32.

Payer, C., Štern, D., Bischof, H., and Urschler, M. (2016). Regressing heatmaps for multiple landmark localization using CNNs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–238. Springer.

Payer, C., Štern, D., Bischof, H., and Urschler, M. (2019). Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Medical Image Analysis*, 54:207–219.

Payer, C., Urschler, M., Bischof, H., and Štern, D. (2020). Uncertainty estimation in landmark localization based on Gaussian heatmaps. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pages 42–51. Springer.

Pérez-García, F., Sparks, R., and Ourselin, S. (2021). TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, page 106236.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Qian, J., Cheng, M., Tao, Y., Lin, J., and Lin, H. (2019). CephaNet: An improved faster R-CNN for cephalometric landmark detection. In *2019 IEEE 16th International Symposium on Biomedical Imaging*, pages 868–871. IEEE.

Qin, C., Bai, W., Schlemper, J., Petersen, S. E., Piechnik, S. K., Neubauer, S., and Rueckert, D. (2018). [joint learning of motion estimation and segmentation for cardiac mr image sequences]. In *21st International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 472–480. Springer.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning.* Adaptive computation and machine learning. MIT Press.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

Schobs, L. (2022). MediMarker. `https://github.com/Schobs/MediMarker`. GitHub repository.

Schobs, L., Zhou, S., Cogliano, M., Swift, A. J., and Lu, H. (2021). Confidence-Quantifying Landmark Localisation For Cardiac MRI. In *2021 IEEE 18th International Symposium on Biomedical Imaging*, pages 985–988. IEEE.

Shafer, G. and Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3).

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Snelson, E. and Ghahramani, Z. (2005). Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 18.

Stankeviciute, K., M Alaa, A., and van der Schaar, M. (2021). Conformal time-series forecasting. *Advances in Neural Information Processing Systems*, 34:6216–6228.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 843–852.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Thaler, F., Payer, C., Urschler, M., Štern, D., et al. (2021). Modeling Annotation Uncertainty with Gaussian Heatmaps in Landmark Localization. *Journal of Machine Learning for Biomedical Imaging*, 1:1–10. Art. no. 014.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574. PMLR.

Tiulpin, A., Melekhov, I., and Saarakkala, S. (2019). KNEEL: Knee Anatomical Landmark Localization Using Hourglass Networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656.

Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR.

Torosdagli, N., Liberton, D. K., Verma, P., Sincan, M., Lee, J. S., and Bagci, U. (2018). Deep geodesic learning for segmentation and anatomical landmarking. *IEEE Transactions on Medical Imaging*, 38(4):919–931.

Tripathi, P., Suvon, M., Schobs, L., Zhou, S., Alabed, S., Swift, A., and Lu, H. (2023). Tensor-based Multimodal Learning for Prediction of Pulmonary Arterial Wedge Pressure from Cardiac MRI. In *26th International Conference on Medical Image Computing and Computer Assisted Intervention*. Forthcoming.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Urschler, M., Ebner, T., and Štern, D. (2018). Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. *Medical Image Analysis*, 43:23–36.

Van der Wilk, M., Rasmussen, C. E., and Hensman, J. (2017). Convolutional Gaussian processes. *Advances in Neural Information Processing Systems*, 30.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Vrtovec, T., Pernuš, F., and Likar, B. (2009). A review of methods for quantitative evaluation of spinal curvature. *European Spine Journal*, 18:593–607.

Wang, C.-W., Huang, C.-T., Lee, J.-H., Li, C.-H., Chang, S.-W., Siao, M.-J., Lai, T.-M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al. (2016). A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Analysis*, 31:63–76.

Warfield, S. K., Zou, K. H., and Wells, W. M. (2008). Validation of image segmentation by estimating rater bias and variance. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1874):2361–2375.

WHO (2021). Who issues first global report on ai in health and six guiding principles for its design and use. World Health Organization Press Release. Accessed: 7th August 2023.

Wieslander, H., Harrison, P. J., Skogberg, G., Jackson, S., Fridén, M., Karlsson, J., Spjuth, O., and Wählby, C. (2020). Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. *IEEE Journal of Biomedical and Health Informatics*, 25(2):371–380.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378. PMLR.

Wilson, A. G. and Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 33:4697–4708.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Yang, D., Xiong, T., Xu, D., Huang, Q., Liu, D., Zhou, S. K., Xu, Z., Park, J., Chen, M., Tran, T. D., et al. (2017a). Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In *International Conference on Information Processing in Medical Imaging*, pages 633–644. Springer.

Yang, J., Liu, Q., and Zhang, K. (2017b). Stacked hourglass network for robust facial landmark localisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 79–87.

Yueyuan, A. and Hong, W. (2021). SWIN transformer combined with convolutional encoder for cephalometric landmarks detection. In *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing*, pages 184–187. IEEE.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM International Conference on Knowledge Discovery and Data Mining*, pages 694–699.

Zeng, M., Yan, Z., Liu, S., Zhou, Y., and Qiu, L. (2021). Cascaded convolutional networks for automatic cephalometric landmark detection. *Medical Image Analysis*, 68:101904.

Zhang, J., Liu, M., and Shen, D. (2017). Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Transactions on Image Processing*, 26(10):4753–4764.

Zhang, J., Norinder, U., and Svensson, F. (2021). Deep learning-based conformal prediction of toxicity. *Journal of Chemical Information and Modeling*, 61(6):2648–2657.

Zhang, Y. and Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, 5(1):30–43.

Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., and Comaniciu, D. (2015). 3D deep learning for efficient and robust landmark detection in volumetric data. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 565–572. Springer.

Zhong, Z., Li, J., Zhang, Z., Jiao, Z., and Gao, X. (2019). An attention-guided deep regression model for landmark detection in cephalograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer.

Zhou, G.-Q., Miao, J., Yang, X., Li, R., Huo, E.-Z., Shi, W., Huang, Y., Qian, J., Chen, C., and Ni, D. (2021). Learn fine-grained adaptive loss for multiple anatomical landmark detection in medical images. *IEEE Journal of Biomedical and Health Informatics*, 25(10):3854–3864.

Appendix A

Additional Experimental Results for Quantile Binning

We present additional experimental results for Quantile Binning, introduced in Chapter 5. Section A.1 motivates using our metrics as uncertainty measures, showing positive correlation between the metrics and localisation errors with piece-wise linear regression plots. Section A.2 shows the tabular data of Figures 5.4a & 5.4b. Section A.3 shows results per-landmark, allowing us to see our uncertainty measures are more effective at predicting error for landmarks with lower localisation accuracy overall. Section A.4 shows experimental results of using Quantile Binning on models trained with varying the standard deviation of the Gaussian heatmap objective function (Equation (2.5)). The results show our method performs similarly regardless of the hyperparameter chosen. Finally, Section A.5 shows our approach is general, with extensive results varying the number of Quantile Bins over our two cardiac datasets (Section 3.2.1) using PHD-Net and U-Net.

A.1 Uncertainty-Error Correlation

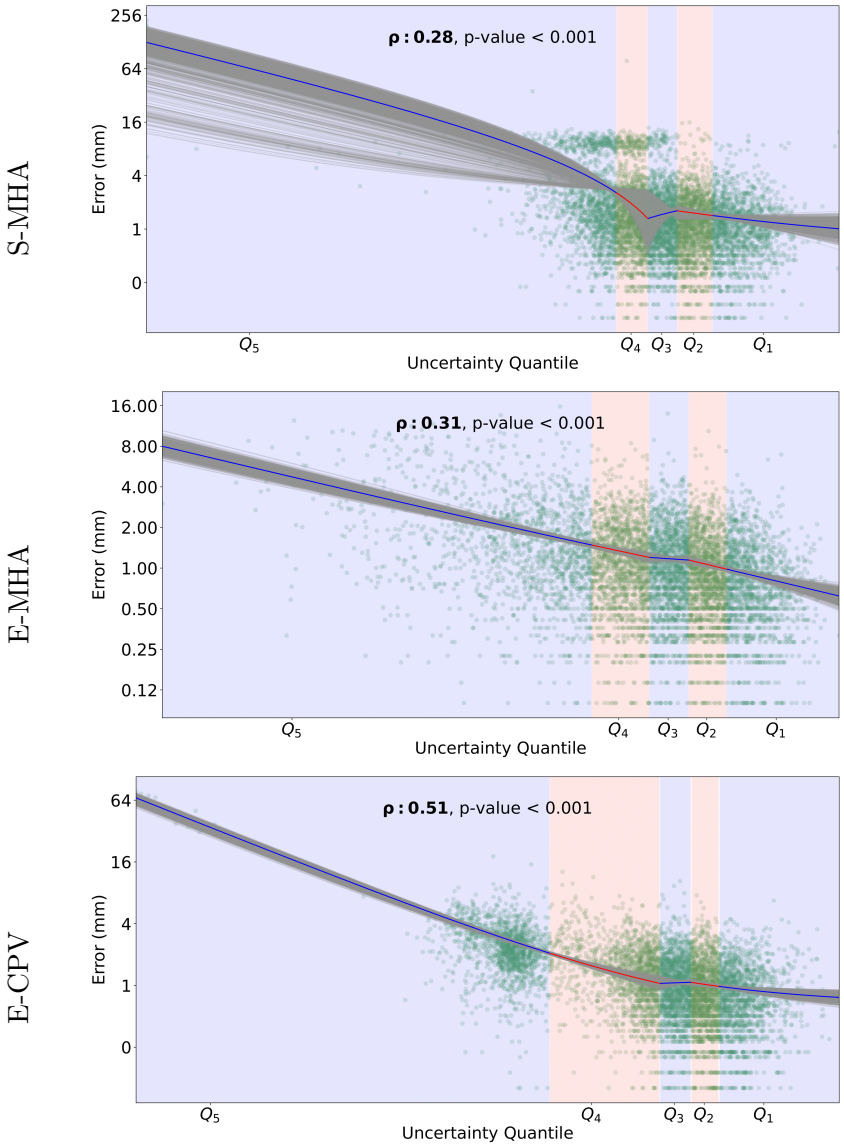


Figure A.1: Piece-wise linear regression of uncertainty with localisation error, with breakpoints at the uncertainty quantiles. Grey represents bootstrap confidence intervals. Data is reported on all data from 4-fold cross validation on the Cephalometric dataset [Wang et al., 2016] using the U-Net model. ρ is the Spearman's Rank Correlation Coefficient between the uncertainty measure and error. Both the x-axis and y-axis are log-transformed.

A.2 Localisation Results Over all Bins

	4 Chamber Images		Short Axis Images	
Method	U-Net	PHD-Net	U-Net	PHD-Net
S-MHA All	**10.00 ± 18.99**	11.07 ± 21.33	5.86 ± 14.19	**3.58 ± 3.52**
S-MHA B_5	18.09 ± 35.09	21.85 ± 33.77	11.28 ± 28.3	5.34 ± 5.87
S-MHA B_4	8.66 ± 10.63	12.05 ± 21.25	6.55 ± 13.21	3.62 ± 2.89
S-MHA B_3	7.52 ± 6.73	8.93 ± 17.05	4.17 ± 2.56	3.21 ± 2.58
S-MHA B_2	9.74 ± 19.48	6.05 ± 8.78	4.03 ± 2.36	2.88 ± 1.96
S-MHA B_1	6.79 ± 6.09	**5.80 ± 9.03**	3.62 ± 2.45	**2.78 ± 1.99**
E-MHA All	**6.36 ± 8.01**	9.14 ± 18.11	4.37 ± 8.86	**3.36 ± 3.50**
E-MHA B_5	9.99 ± 14.51	19.12 ± 32.75	8.56 ± 19.53	5.31 ± 6.18
E-MHA B_4	6.09 ± 6.18	8.56 ± 16.04	4.29 ± 2.71	3.21 ± 2.36
E-MHA B_3	5.61 ± 6.12	7.55 ± 11.69	3.11 ± 2.43	2.95 ± 1.93
E-MHA B_2	5.36 ± 3.72	5.99 ± 6.06	3.32 ± 2.28	3.05 ± 2.52
E-MHA B_1	4.93 ± 2.85	**4.70 ± 3.21**	2.98 ± 2.09	**2.39 ± 1.90**
E-CPV All	**8.13 ± 10.16**	9.42 ± 13.07	4.97 ± 7.51	**3.22 ± 2.93**
E-CPV B_5	16.3 ± 17.77	25.04 ±22.1	9.65 ± 15.5	5.08 ± 4.7
E-CPV B_4	6.82 ± 6.34	9.66 ±11.88	4.18 ± 2.69	2.95 ± 2.03
E-CPV B_3	6.84 ± 7.97	5.87 ± 3.61	3.88 ± 2.49	3.24 ± 2.38
E-CPV B_2	5.62 ± 3.45	5.38 ± 3.07	3.63 ± 2.22	2.65 ± 2.08
E-CPV B_1	5.34 ± 3.0	**5.10 ± 6.76**	3.75 ± 2.13	**2.47 ± 2.08**

Table A.1: Localization errors (mm) for the uncertainty methods outlined. *All* indicates entire set of predictions; B_1 indicates subset with the *lowest uncertainties*. Mean error and standard deviation are reported across all folds & all landmarks. **Bold** indicates best results in row for the given dataset for *All* and B_1.

A.3 Quantile Binning Separating Landmarks

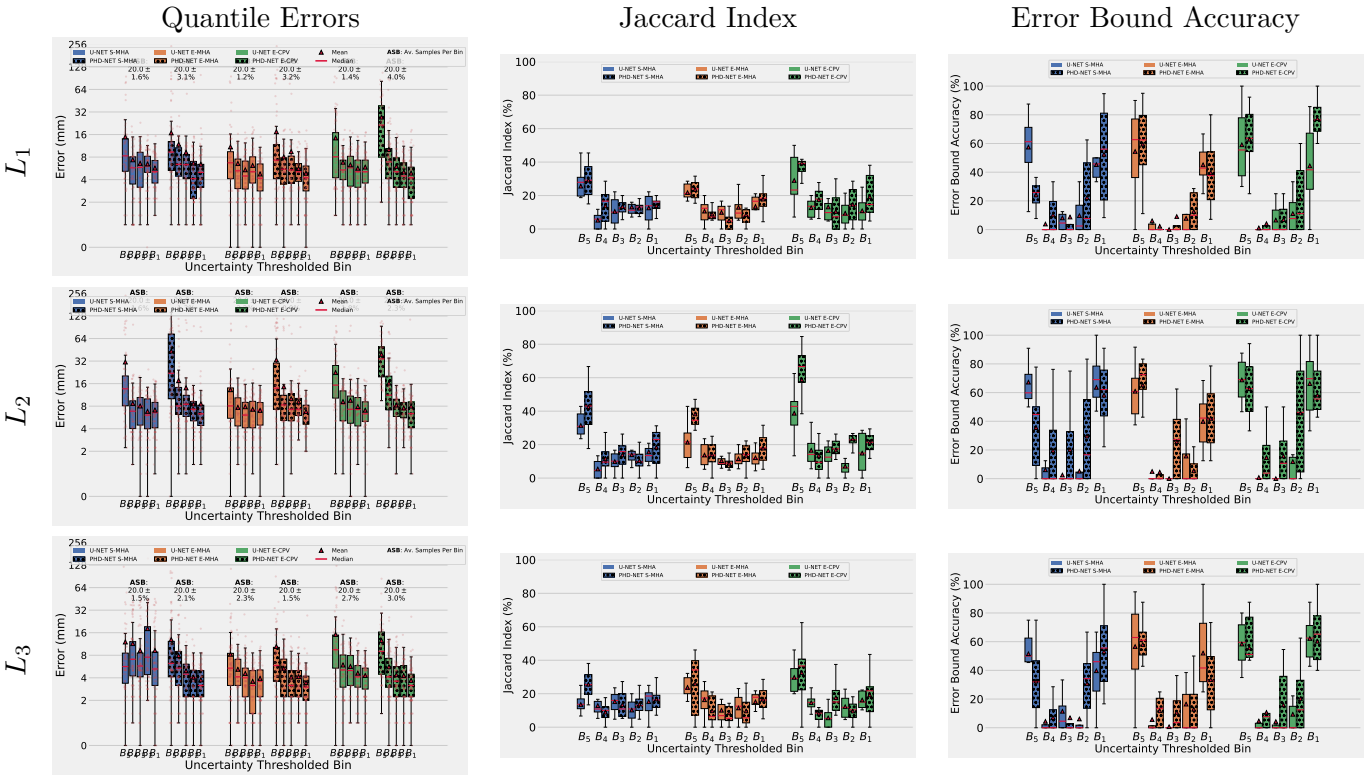


Figure A.2: The results from Quantile Binning for S-MHA, E-MHA and E-CPV uncertainty measures on individual landmarks from the 4CH dataset. The *Quantile Errors* column shows the boxplots of localization errors for each quantile bin, showing the landmarks across all folds. The *Jaccard Index* column shows the similarity between the predicted Quantiles and the true error quantiles, and the *Error Bound Accuracy* column shows the accuracy of the predicted error bounds for each quantile bin.

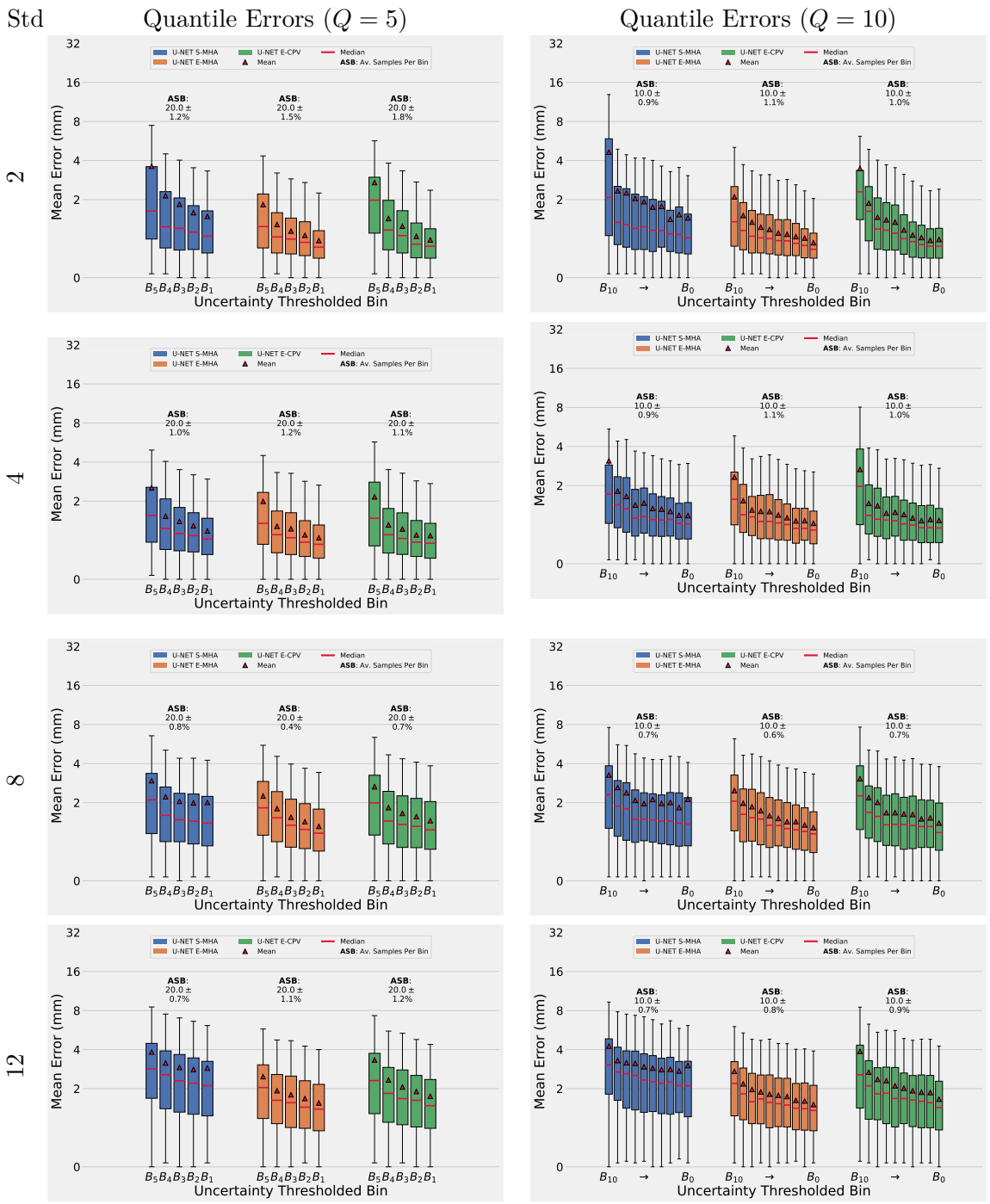


Figure A.3: Comparing results for models using different standard deviation values for the ground truth heatmap labels. We show the Quantile Localization Errors using 5 & 10 Quantile bins. We present results on all landmarks from a 4-fold CV on the Cephalometric dataset [Wang et al., 2016].

A.4 Variance of Target Heatmap Comparison

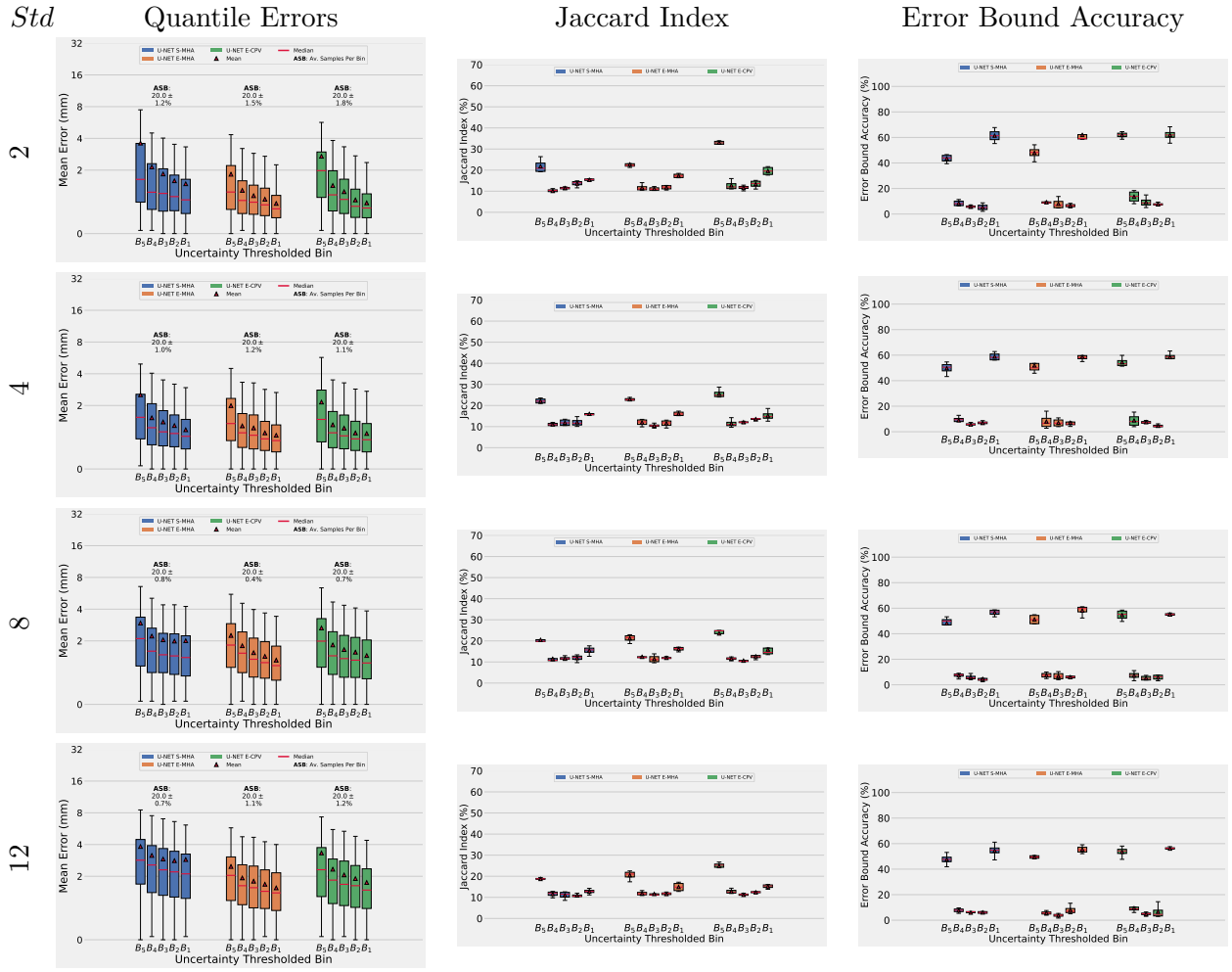


Figure A.4: Quantile Binning from a U-Net localization model trained using target heatmaps, varying the standard deviation of the Gaussian blob. The Quantile Errors, Jaccard index and Error Bound Accuracy are presented over a 4-fold Cross Validation on the Cephalometric dataset [Wang et al., 2016].

A.5 Comparing Q Values

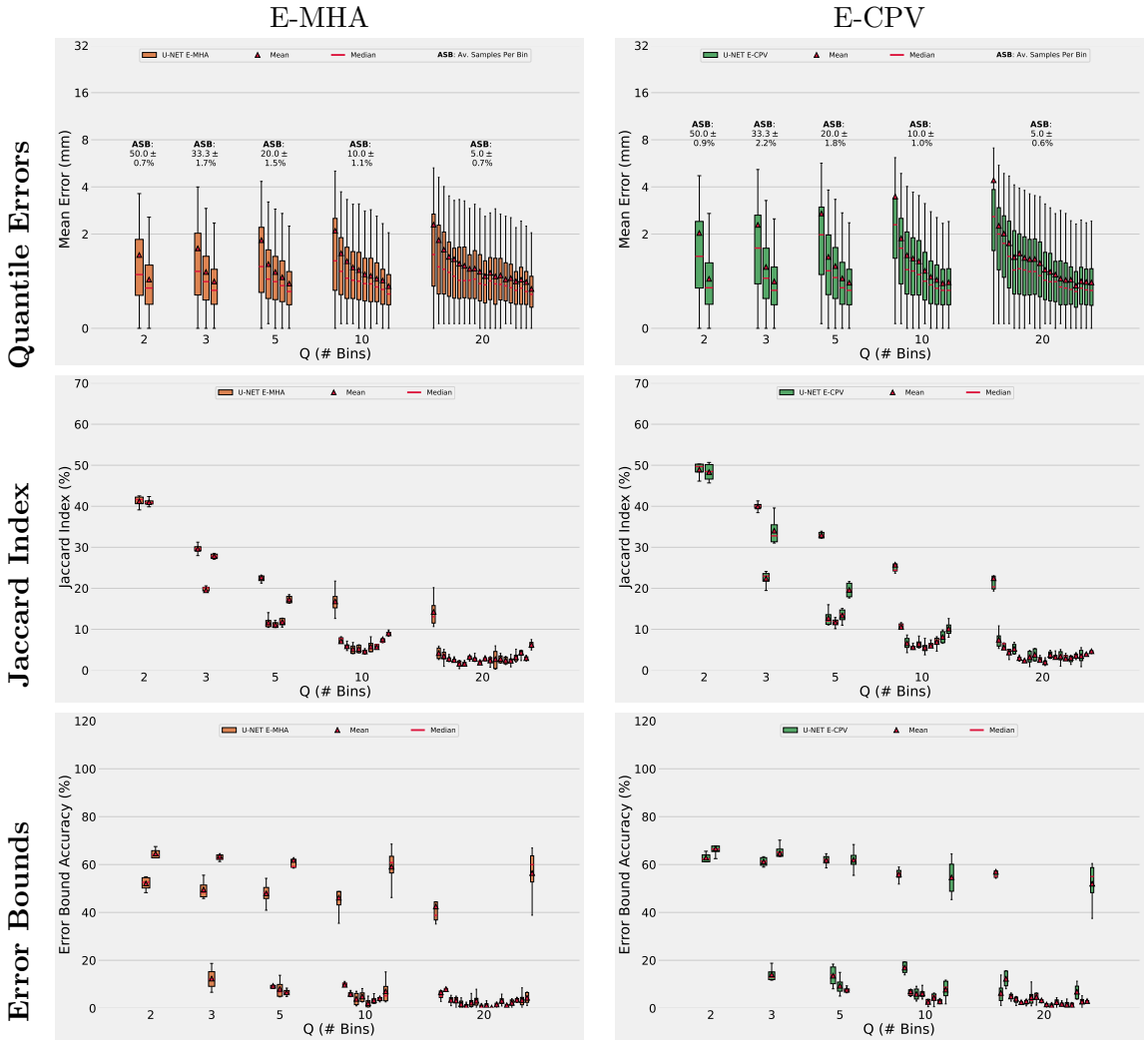


Figure A.5: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures E-MHA and E-CPV, over all landmarks from a 4-fold CV on the **Cephalometric dataset [Wang et al., 2016], trained on the U-Net model.**

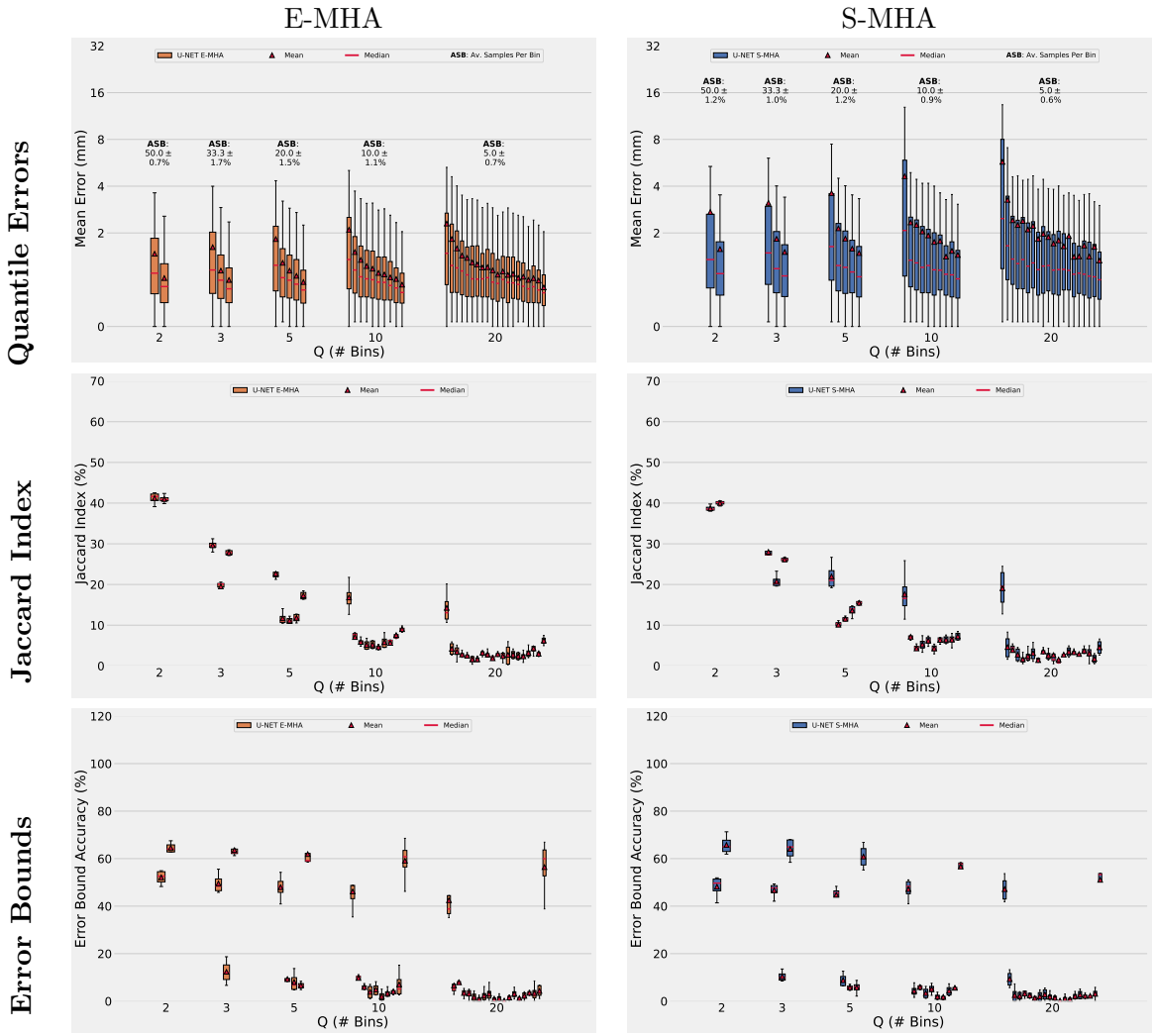


Figure A.6: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures E-MHA and S-MHA, over all landmarks from a 4-fold CV on the **Cephalometric dataset [Wang et al., 2016], trained on the U-Net model.**

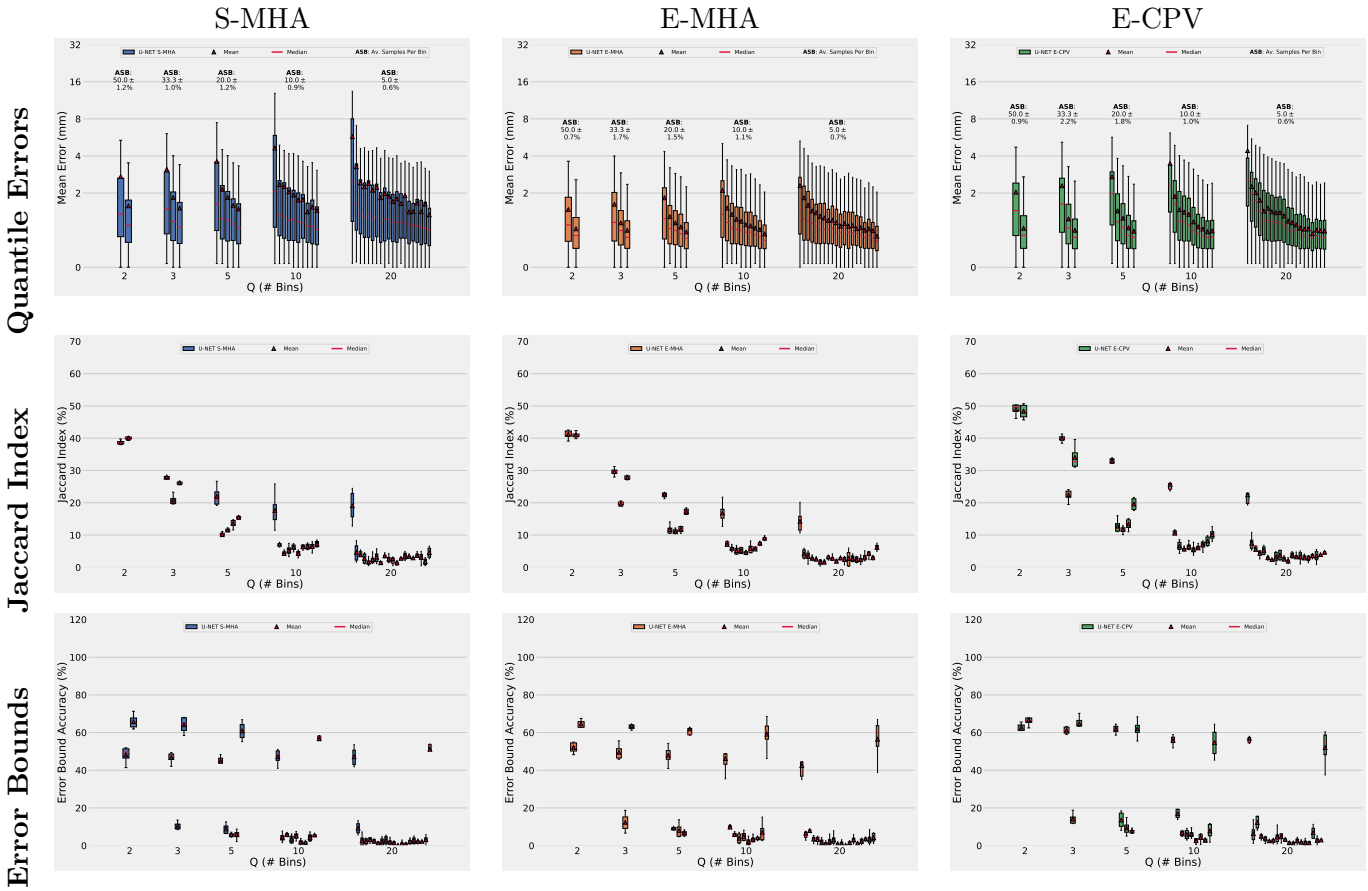


Figure A.7: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 4-fold CV on the **Cephalometric dataset [Wang et al., 2016], trained on the U-Net model.**

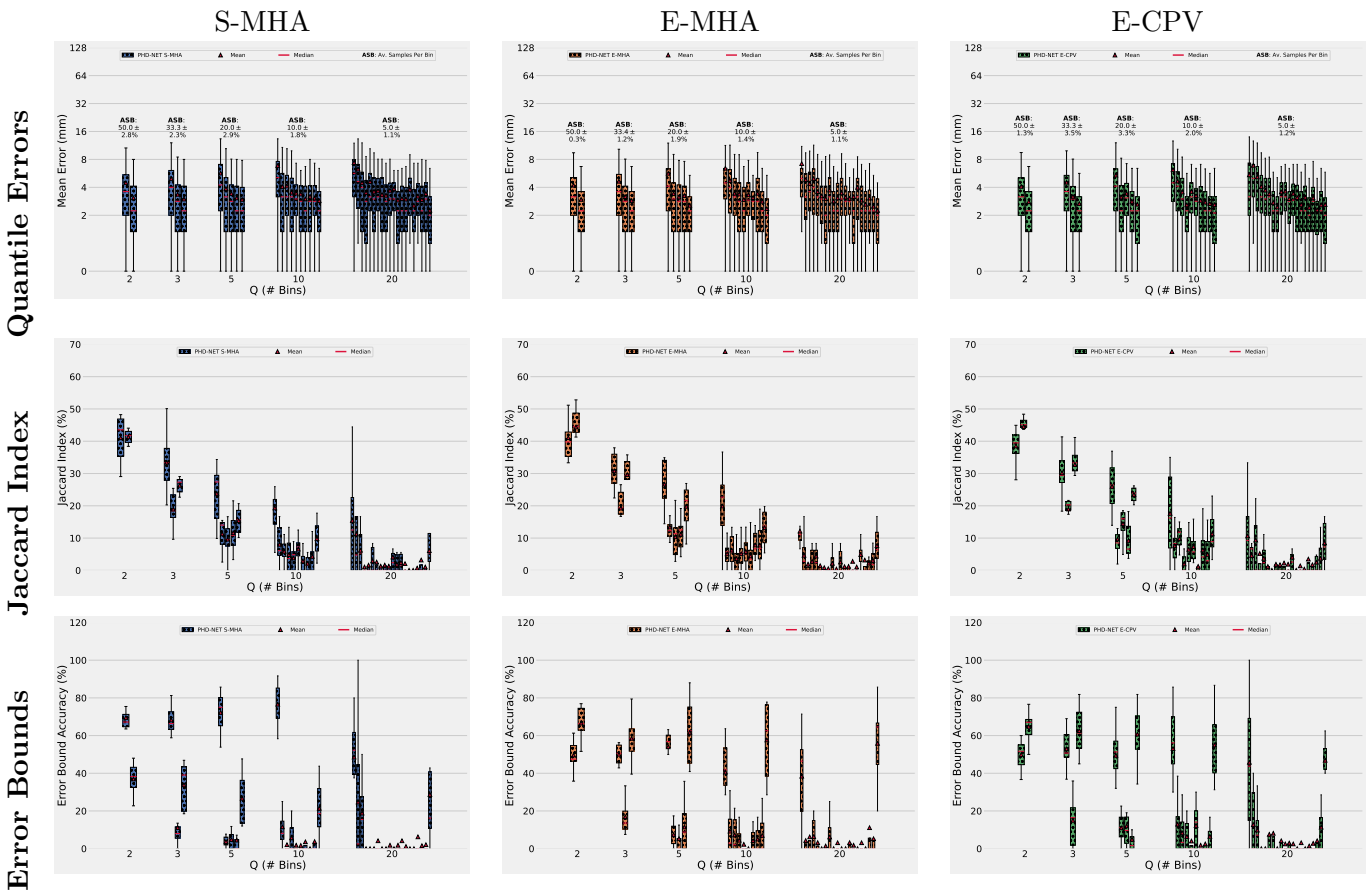


Figure A.8: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 8-fold CV on the **SA dataset, trained on the PHD-Net model.**

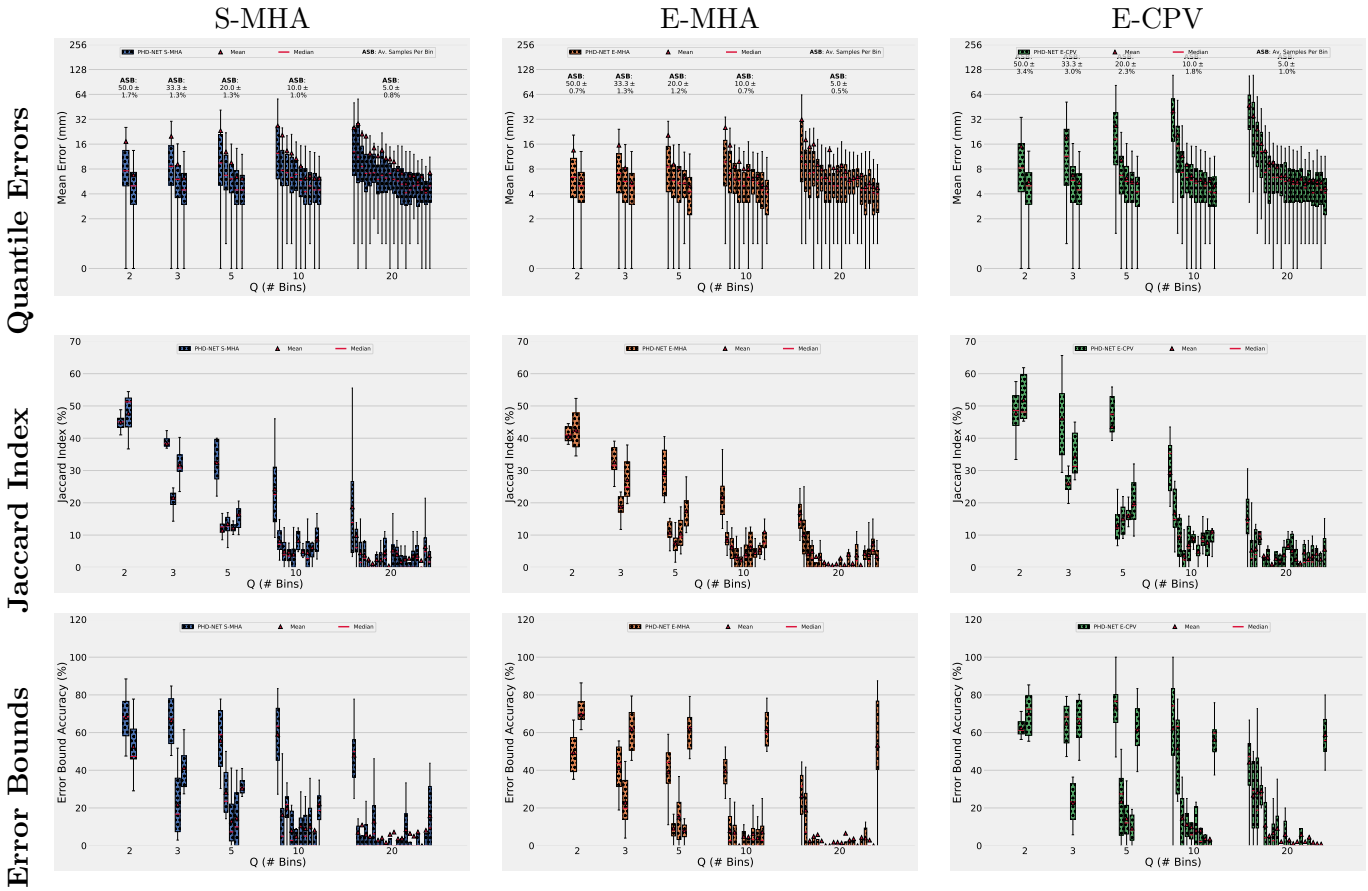


Figure A.9: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 8-fold CV on the **4CH dataset, trained on the PHD-Net model.**

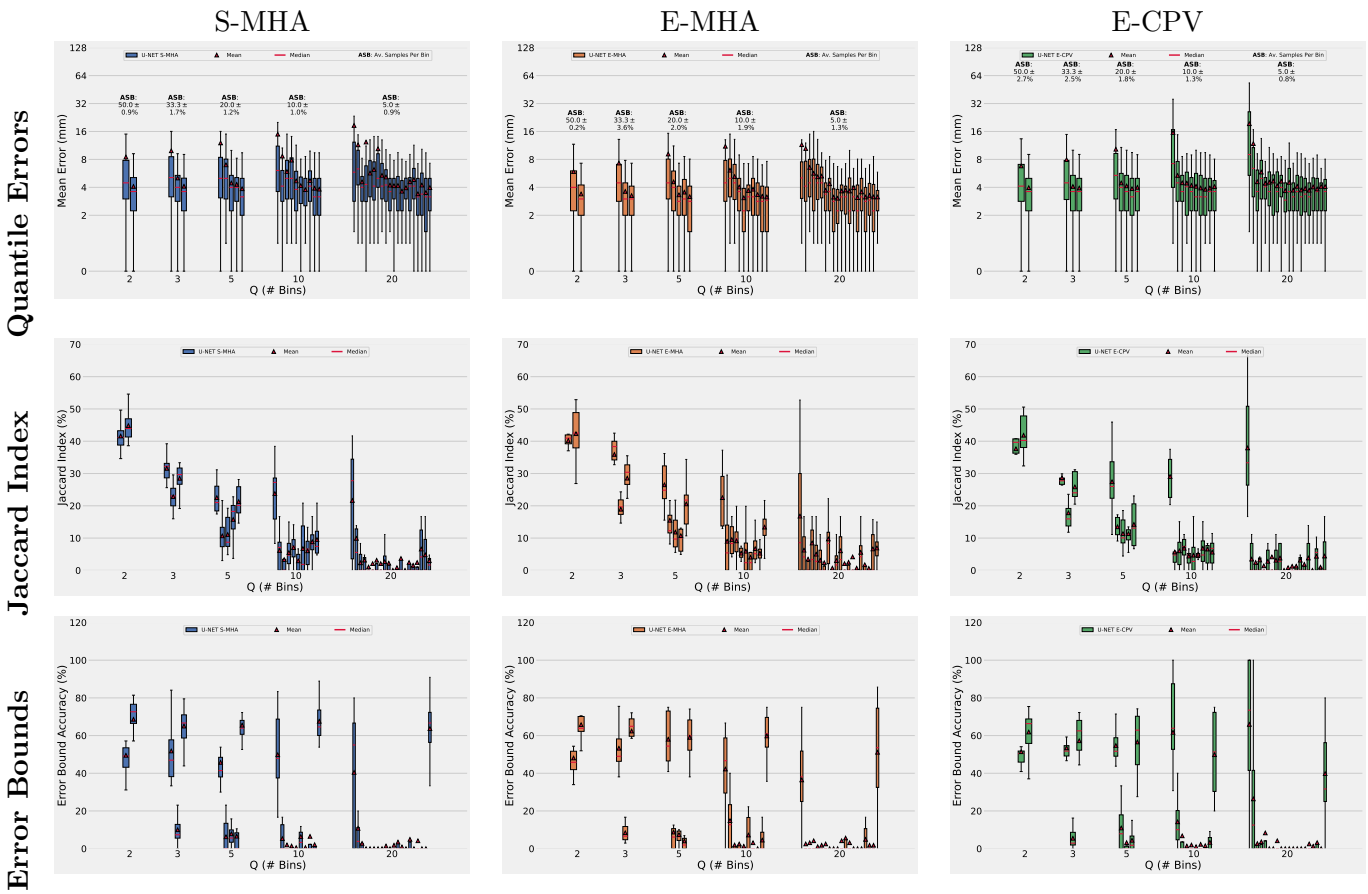


Figure A.10: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 8-fold CV on the **SA dataset, trained on the U-Net model.**

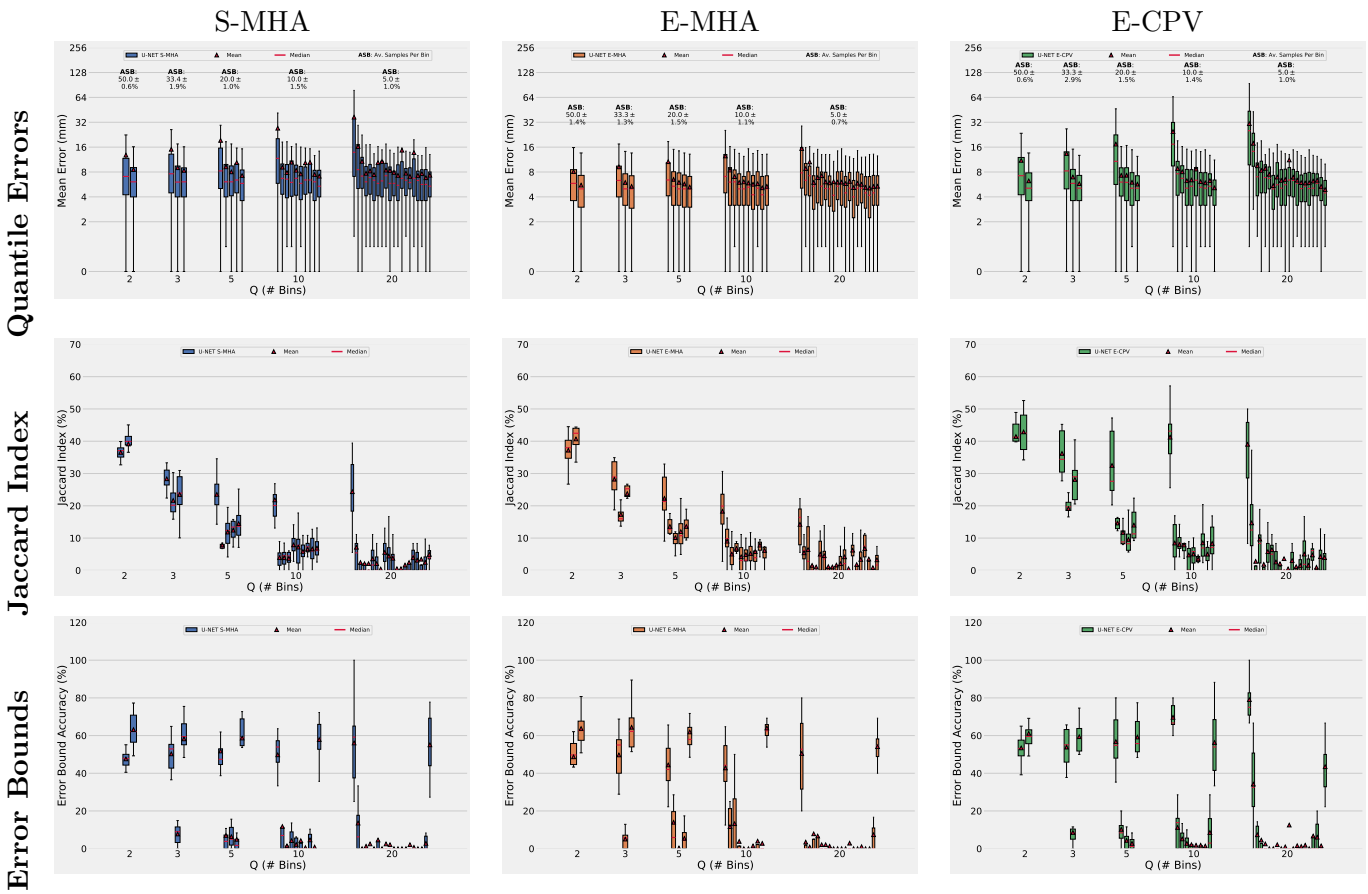


Figure A.11: Quantile Errors varying Q (Number of Quantile Bins). We show results for the uncertainty measures S-MHA, E-MHA and E-CPV, over all landmarks from a 8-fold CV on the **4CH dataset, trained on the U-Net model.**