

# What Can Computational Models Tell Us About Scene Memorability?

**Cameron Kyle-Davidson**

Doctor of Philosophy

University of York

Computer Science

February 2022

# Acknowledgements

I would like to thank my supervisors, Dr. Karla K. Evans and Dr. Adrian G. Bors for their continued support and advice during the course of my PhD, and my TAP panel members, Dr. William Smith and Dr. Aiden Horner for their valuable input on the research presented within this thesis. I would also like to extend thanks to the members of the Complex Cognitive Processing lab and the Vision, Graphics, and Learning lab at the University of York.

# Abstract

Computational memorability prediction has allowed significant advances in the understanding of human visual memory; and in turn, advances in understanding what makes an image memorable. Recently, this research has expanded to the second dimension, with Visual Memory Schemas (VMS) maps revealing the specific regions in a scene that lead to that scene being remembered. In this thesis, we explore the concept of VMS maps in detail, develop new VMS datasets, novel models for VMS prediction, explore whether human memory can be modulated with VMS maps, and finally investigate the relationship between scene memorability and scene complexity. We propose three new approaches for predicting visual memory schemas, starting with a variational autoencoder-based model, before exploring the role of self-attention, multi-scale information, and depth in the prediction of scene memorability. Based upon this work, we develop a novel "dual-feedback" model that uses both VMS datasets and pre-existing single-score memorability datasets to predict memorability maps for scene images, setting a new state-of-the-art for VMS prediction. This work is supported by our efforts in expanding VMS datasets; from the original 800 images, up to a dataset of over 4000+ scenes and VMS maps. We make use of our VMS predictors by integrating them with generative models with the goal of synthesising scene images of controllable memorability. We test our generated scenes against real-world human observers and find that images we synthesise to be more memorable have a greater hit-rate than images we synthesise to be less memorable. Finally, we investigate the relationship between scene complexity and scene memorability, developing novel techniques and architectures capable of predicting how complex a human finds a scene, and ultimately finding that the complexity of the scene plays a small, but significant role, in the memorability of that scene.

# Declaration

I declare that this dissertation represents my own original work; and that parts of this work developed in collaboration, the collaborators, and their contributions are identified herein. I am the sole author of this dissertation. However, chapter 5 is based on a paper developed in part as a collaboration with the University of Toronto. Details of this collaboration are given below. This work has not been presented for a previous award at the University of York or any other institution or University. All sources are acknowledged as References.

## Collaboration

Chapter 5 is based upon a submitted journal paper ‘Characterizing and Dissecting Human Perception of Scene Complexity’, written in majority by myself, with the following author list: Cameron Kyle-Davidson, Elizabeth Yue Zhou, Dirk B. Walther Adrian G. Bors, Karla K. Evans. In this chapter, the vast majority of work was conducted by myself, including the data gathering of the VISC-C/CI datasets, the data analysis of said datasets, development of the neural network model and analysis of the dissected network). Our collaborators, Elizabeth Y. Zhou and Dirk B. Walther provided the materials (dataset), procedure, and data cleaning presented in Section 5.5 ("Generalising to a Different Dataset"). I developed the results included in that section, in which I apply my developed methods and techniques to their dataset for generalisation purposes. Adrian G. Bors and Karla K. Evans are my supervisors.



# Abbreviations

**AE** Autoencoder.

**ANOVA** Analysis of Variance.

**BCE** Binary Cross Entropy.

**CE** Cross Entropy.

**CNN** Convolutional Neural Network.

**DF** Dual Feedback.

**DFVMS** Dual Feedback - Visual Memory Schema.

**ELBO** Evidence Lower Bound.

**EMD** Earth Mover Distance.

**FAR** False Alarm Rate.

**FID** Fréchet Inception Distance.

**GAN** Generative Adversarial Network.

**GBVS** Graph Based Visual Saliency.

**GPU** Graphics Processing Unit.

## Abbreviations

**HOG** Histogram of Gradients.

**HR** Hit Rate.

**HRA** Hierarchical Regression Analysis.

**HSV** Hue, Saturation, Value.

**KL** Kullback-Leibler.

**KLD** Kullback-Leibler Divergence.

**LOP** Level of Processing.

**LSTM** Long Short-Term Memory.

**MLP** Multi-layer Perceptron.

**MSB** Multi-scale block.

**MSE** Mean Squared Error.

**ONR** Old/New Recognition.

**RELU** Rectified Linear Unit.

**RGB** Red, Green, Blue.

**RNN** Recurrent Neural Network.

**ROC** Receiver Operating Characteristic.

**SGD** Stochastic Gradient Descent.

**SIFT** Scale Invariant Feature Transform.

**SIM** Histogram Similarity.

**SoTA** State of the Art.

## Abbreviations

**SSIM** Structural Similarity Index.

**SVM** Support Vector Machine.

**SVR** Support Vector Regression.

**VAE** Variational Autoencoder.

**VGG** Visual Geometry Group.

**VLTM** Visual Long-Term Memory.

**VMS** Visual Memory Schema.

**WGAN** Wasserstein GAN.

## Key Terms

**Complexity** An intrinsic property of images, of which humans are capable of judging consistently. Dependent upon both low-level textural features and high-level semantic content. There is not yet an agreed-upon definition for complexity in the context of human perception, though it's computational counterparts are well understood..

**Memorability** An intrinsic property of images (often represented as a scalar value between 0 and 1.0) that corresponds to how well that image is remembered, on average, by humans. Does not relate strongly to many other image properties such as interestingness or aesthetics, nor to human predictions of image memorability. Mostly driven by the semantic content of the image..

**Saliency** The likelihood of an image area to draw the attention of the observer. A *saliency map* reveals the areas of an image that humans fixate on first upon viewing the image..

**Scene** Scene, in this work, refers to *natural scenes*. Specifically, a still image of a common environment in which a person may reasonably be expected to have been immersed. Kitchen, living room, golf course, and playground images are examples of natural scene images. A collection of items (e.g, kitchen implements) arranged on a table and photographed do not represent a *scene* as defined by this work, nor do images of single objects (a toaster) or an image where a few prototypical elements of the scene consume the vast majority of the frame (an image of a fridge near a counter edge). The scene image typically displays the majority of

## Key Terms

prototypical elements common to that scene, where present, as well as captures the overall structure of the scene ('the whole kitchen') in the still image..

**Schema** A mental construct characterising concepts and the relationship between them..

**Semantics** The high-level features present in the image that compose the scene, distinct from low-level image features such as spatial frequencies, colour, or scene statistics. The objects present (e.g, a chair) are considered semantic scene elements, as are composite arrangements of objects (e.g, a dining table surrounded by chairs). The scene category itself may also be considered a semantic feature of the scene..

**Visual Memory Schema** A cognitive representation of a scene, containing semantic elements and the relationship between them, which facilitates encoding of said scene. Scene images that more strongly match held visual memory schemas are more strongly encoded. See Chapter 3.

# Contents

<b>Acknowledgements</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>Declaration</b>	<b>iv</b>
<b>Abbreviations</b>	<b>v</b>
<b>Key Terms</b>	<b>viii</b>
<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.1.1 A Computational Approach . . . . .	2
1.1.2 Beyond Single-Score Metrics . . . . .	3
1.1.3 Scene Complexity . . . . .	4
1.1.4 Research Direction . . . . .	6
1.2 Thesis Structure . . . . .	7
1.3 Publications & Presentations . . . . .	8
<b>2 Background</b>	<b>9</b>
2.1 A Brief Overview of Memory . . . . .	9
2.2 Remembering Images . . . . .	11

## CONTENTS

2.3	The Fallibility of Visual Memory . . . . .	15
2.4	Interestingness and Aesthetics . . . . .	17
2.4.1	Interestingness . . . . .	18
2.4.2	Aesthetics . . . . .	19
2.4.3	Relationship to Memorability . . . . .	20
2.5	Machine Learning . . . . .	21
2.5.1	Support Vector Machines . . . . .	22
2.5.2	Feature Extraction . . . . .	23
2.5.3	Neural Networks . . . . .	25
2.5.4	Convolutional Neural Networks . . . . .	28
2.5.5	Recurrent Neural Networks . . . . .	30
2.5.6	Additional Network Types . . . . .	31
2.5.7	Transfer Learning & Common Architectures . . . . .	31
2.6	A General Overview of Complexity . . . . .	33
2.7	Summary . . . . .	36
<b>3</b>	<b>Visual Memory Schemas</b> . . . . .	<b>37</b>
3.1	Background . . . . .	37
3.1.1	Computational Memorability . . . . .	38
3.1.2	Visual Memory Schemas . . . . .	45
3.2	Methodology . . . . .	47
3.2.1	Predicting Visual Memory Schemas with Variational Autoencoders	48
3.2.2	Exploring Visual Memory Schema Prediction with Multi-Scale In- formation, Depth, and Self-Attention . . . . .	51
3.2.3	A Dual-Feedback Approach to Visual Memory Schema Prediction	55
3.3	Experimental Results . . . . .	57
3.3.1	Visual Memory Schema Datasets . . . . .	58
3.3.2	Variational Autoencoder Approach . . . . .	67
3.3.3	Multi-Scale Information, Depth, and Self-Attention . . . . .	72
3.3.4	A Dual-Feedback Approach . . . . .	76
3.4	Conclusion . . . . .	80
<b>4</b>	<b>Modulating Human Memory</b> . . . . .	<b>83</b>
4.1	Introduction . . . . .	83

4.2	Results . . . . .	85
4.2.1	VMS Consistency and Memorability . . . . .	85
4.2.2	Generating Memorable Images Based on VMS Maps . . . . .	88
4.2.3	Human memory performance for generated images . . . . .	94
4.3	Discussion . . . . .	98
4.4	Methods . . . . .	102
4.4.1	Memorability Estimation Feedback Network . . . . .	102
4.4.2	W-MEMGAN Architecture & Training . . . . .	103
4.4.3	MEMGAN Architecture & Training . . . . .	104
4.4.4	Loss functions . . . . .	106
4.4.5	Generating Images for Human Observer Experiments . . . . .	106
4.4.6	Human Memory Experiment . . . . .	107
4.4.7	Evaluating Scene Recognition Differences . . . . .	108
4.5	Summary . . . . .	109
<b>5</b>	<b>Perceptual Scene Complexity</b>	<b>110</b>
5.1	Introduction . . . . .	110
5.2	Factorising Complexity . . . . .	113
5.2.1	Clutter . . . . .	113
5.2.2	Patch-based Symmetry . . . . .	114
5.2.3	Entropy . . . . .	115
5.2.4	Openness . . . . .	115
5.3	Experiment 1 - Two Dimensional Complexity . . . . .	115
5.3.1	Participants . . . . .	116
5.3.2	Materials . . . . .	116
5.3.3	Procedure . . . . .	117
5.3.4	Data Analysis . . . . .	117
5.3.5	Results . . . . .	118
5.4	Experiment 2 - The Effect of Semantics . . . . .	120
5.4.1	Participants . . . . .	121
5.4.2	Materials & Procedure . . . . .	121
5.4.3	Results . . . . .	121
5.5	Experiment 3 - Generalizing to a Different Dataset . . . . .	123
5.5.1	Participants . . . . .	123



## CONTENTS

5.5.2	Materials . . . . .	123
5.5.3	Procedure . . . . .	124
5.5.4	Data Cleaning . . . . .	124
5.5.5	Results . . . . .	125
5.6	Modelling Complexity . . . . .	126
5.6.1	Predicting Complexity Scores & Maps . . . . .	126
5.6.2	What Neural Networks Learn about Complexity . . . . .	128
5.7	Discussion . . . . .	129
<b>6</b>	<b>Complexity &amp; Memorability</b>	<b>133</b>
6.1	Introduction . . . . .	133
6.2	Complexity Ratings & Memorability . . . . .	134
6.3	Two-dimensional Statistics . . . . .	136
6.3.1	Complexity Annotations . . . . .	136
6.3.2	Simplicity Annotations . . . . .	138
6.3.3	Region Memorability . . . . .	139
6.4	Computational Methods . . . . .	140
6.4.1	Multi-factor Analysis . . . . .	140
6.4.2	Computational Metrics . . . . .	141
6.5	Discussion . . . . .	142
6.6	Summary . . . . .	145
<b>7</b>	<b>Conclusion</b>	<b>146</b>
7.1	Thesis Summary . . . . .	146
7.1.1	Visual Memory Schemas . . . . .	147
7.1.2	Modulating Human Memory . . . . .	147
7.1.3	Factorising Scene Complexity . . . . .	148
7.1.4	Complexity & Memorability . . . . .	148
7.2	Conclusions . . . . .	149
7.3	Limitations . . . . .	151
7.4	Future Work . . . . .	152
7.5	Summary . . . . .	153
	<b>Appendix</b>	<b>154</b>

CONTENTS

<b>A</b>	<b>Co-occurrences of Objects in Memorable Regions</b>	<b>154</b>
A.1	3 Objects . . . . .	155
A.2	5 Objects . . . . .	156
A.3	7 Objects . . . . .	157
<b>B</b>	<b>Images with Modulated Memorability</b>	<b>159</b>
B.1	Additional MEMGAN Architecture Details . . . . .	159
B.2	High-Memorability Images . . . . .	160
B.3	Low-Memorability Images . . . . .	161
<b>C</b>	<b>Complex and Simple Images</b>	<b>162</b>
C.1	High Complexity Images . . . . .	162
C.1.1	Upright Scenes . . . . .	163
C.1.2	Inverted Scenes . . . . .	164
C.2	Low Complexity Images . . . . .	164
C.2.1	Upright Scenes . . . . .	165
C.2.2	Inverted Scenes . . . . .	166
C.3	Additional Prediction Results . . . . .	167
<b>D</b>	<b>MLR Table for Complexity/Memorability</b>	<b>168</b>
	<b>Bibliography</b>	<b>169</b>

## List of Figures

2.1	LeCunn et al's LeNet - one of the first convolutional neural network architectures. The feature maps mentioned are analogous to filters. Dimensionality reduction is accomplished through subsampling. [Fig. 2 in [87]] . . . . .	29
3.1	The AlexNet Architecture [3]. The second GPU processing stream is truncated, but follows the same architecture as the first. . . . .	40
3.2	An example from the VISHEMA dataset produced in [2] . . . . .	46
3.3	Predicting VAEs in images using an autoencoder. . . . .	50
3.4	End-to-end deconvolutional network showing single and dual headed outputs. The height and width of the convolution filters is given above, while the channels are given below the diagram. The dimensions of the output is given below each output. . . . .	51
3.5	Multi-scale VMS predictor with multi-scale blocks (MSB) from [64]. . .	52
3.6	Multiscale architecture modified to embed depth-map information. . . .	53
3.7	Architecture of proposed Visual Memory Schema predictor with Dual Memorability Feedback. Colors refer to layer types and are given in the legend. . . . .	55
3.8	Examples from the VMS4k Dataset. Green areas indicate that region caused the image to be remembered, red areas indicate regions that caused an image to be falsely remembered; indicated as seen despite never being shown to a participant. . . . .	58
3.9	Repeat-recognition experiment structure . . . . .	59
3.10	Participant D-Primes reveal good memory performance for the shown images. No participants were excluded from the analysis. . . . .	60

LIST OF FIGURES

3.11 True and false memorability for the VISCHEMA image set. . . . . 61

3.12 There is no difference in memorability performance between categories as measured by hit-rate or VMS intensity (an analogue for participant consistency). . . . . 62

3.13 Outdoor scenes (right) show bias towards larger annotation areas compared to indoor scene images (left). . . . . 63

3.14 Outdoor scenes (right) show a greater bias towards fewer per-image annotations than indoor scenes (left). . . . . 63

3.15 The average annotations per-image is significantly greater for indoor, than outdoor scenes (left), and there is a significant difference between the sizes of annotations between indoor and outdoor regions (right). . . . . 64

3.16 MaskFormer architecture; a neural network that can be used for state-of-the-art semantic segmentation, [Fig 2] from [26]. . . . . 64

3.17 ‘Semantic units’ contained within the regions of images that participants have labelled as causing them to successfully remember that image. . . . 65

3.18 These objects frequently appear together inside the memorable regions of an image, of that category. Limited to three objects; higher amounts of object co-occurrences can be examined. . . . . 66

3.19 Structure of the Decoder. . . . . 67

3.20 Reconstruction accuracy for various image categories. . . . . 69

3.21 Comparison of the memorability results for a set of image categories between the VISCHEMA2 and VISCHEMA datasets. . . . . 71

3.22 VISCHEMA2 Latent Space Embedding. Green represents memorability and red represents false memorability, normalised between 0 and 1. Clustering of both memorable, and falsely memorable images is evident. Features that lead to the generation of memorable VMS maps are placed near each other, as are features that lead to the generation of VMS maps that indicate the scene is not so memorable. . . . . 72

3.23 Set of three images from VISCHEMA2 dataset and their predicted true VMS and false VMS on second and third lines. We find empirically that false schemas are often subsets of the true schema of the image that carries less information. For example, an image is memorable due to the presence of a man feeding a calf, yet the presence of just a man may lead to the false remembering of a scene. . . . . 73

## LIST OF FIGURES

3.24	VMS maps showing memorable (green) and falsely memorable (red) regions, for the images from the first column, are shown in the second column, and their corresponding predictions on the third column. . . . .	74
3.25	Predicted VMS maps for the given scene images. Ground-truth maps come from human data. Some human VMS maps contain false schemas (red), for visualisation purposes in this figure we only show predicted true (memorable) schemas. The best performing DF-VMS variant employs a Resnet backbone, self-attention, multiscale-information, and dual-feedback. VGG16 backbones do not capture the full spread of memorability; instead focusing strongly on semantic regions. ResNet backbones, with their richer feature extraction, perform better at VMS map prediction.	77
3.26	Examples 1, 9, and additional exemplars with predicted false memorability maps. As consistent with [83], false schemas are often a subset of the true schema, and are more difficult to predict. . . . .	79
4.1	Histogram showing the correlation between per image category consistency for Vischema 1 and 2 datasets and human observers' memory. Similar pattern of correlations between datasets indicates the reliability of using Visual Memory Schemas. . . . .	88
4.2	Generated images when fixing $\mathbf{Z}$ , where the sequence of generated images is displayed from left to right, while the memorability $\mathbf{M}$ is varied from low to high. Shown categories include kitchens, cathedrals, and living rooms. . . . .	89
4.3	Predicted low and high memorability for different memorability weighting factors for $\alpha$ , considering $\alpha = 0$ (in Equation 4.4) as baseline. When increasing $\alpha$ , all generated images have a higher memorability than the baseline. The most memorable images overall are obtained with $\alpha = 25$ , but the best pairwise effect is achieved with $\alpha = 10$ . . . . .	90
4.4	Differences in predicted memorability for low and highly memorable images generated with W-MEMGAN. . . . .	91
4.5	Memorable, shown within green boundaries and non-memorable, shown within red boundaries generated image pairs. Foils are shown within blue boundaries. . . . .	93

## LIST OF FIGURES

4.6	Generated high-memorability images (left) and their low-memorability pairs (right). VMS maps for each image are shown on the bottom row. . .	94
4.7	Difference in memorability (HR and VMS Intensity) for generated image populations. Degree 3 polynomial fitted for visualisation. . . . .	95
4.8	Memorability-constrained image generation model architecture. Pixel-Norm and Minibatch Standard Deviation layers omitted for clarity. . . .	103
4.9	Progressive generator with per-resolution memorability estimation. . . .	104
4.10	Memorability experiment structure. . . . .	108
5.1	Example of clutter algorithm working on a perceptually simple image and a more complex scene, as rated by humans. . . . .	113
5.2	A set of scenes sorted into ascending complexity, as rated by a group of human observers. The images below the arrow reveal the regions that humans labelled as complex (in blue) or simple (in red). Regions labelled as simple often contain textural variation (e.g, grass in image 1, or the sky/clouds in image 3), yet are labelled simple nonetheless. . . . .	116
5.3	The distributions of human decided complexity scores for the VISC-C dataset. . . . .	118
5.4	Relationship between annotation coverage, intensity, and complexity for scenes. As coverage and intensity of the complex channel increases; so does the human complexity score ratings, and vice-versa for simplicity. . .	118
5.5	Relationship between inverted scene complexity and 2d annotation metrics.	121
5.6	Complex/simple annotation coverage for upright (VISC-C, top) and inverted (VISC-CI, bottom) scenes. Coverage shows that much more of the image is indicated as complex or simple when inverted; despite low-level textural properties remaining the same. . . . .	121
5.7	Basic Complexity Prediction Architecture, with optional complexity map prediction head. . . . .	126
5.8	Examples of predicted complexity maps and their ground-truth counterparts from the test set. . . . .	127
5.9	Effect of network depth on complexity score prediction performance. Performance peaks in the penultimate processing block of each model, then plateaus. . . . .	128

## LIST OF FIGURES

5.10	Correlation with human ratings for both scores and complexity maps for different base network architectures. . . . .	128
5.11	Images which activate a sample of neurons from the final layer of a complexity prediction network. The network appears to learn both low-level and semantic features. . . . .	129
6.1	Pearson’s correlation between ground-truth human complexity ratings and ground-truth human memorability scores. . . . .	135
6.2	Pearson’s correlation between human complexity ratings and scene hit rates (left) and false alarm rates (right) . . . . .	136
6.3	Relationship between scene memorability (hit rate, left, and false alarm rate, right) and complex channel annotation intensity. Pearsons correlation.	137
6.4	Relationship between scene memorability (hit rate, left, and false alarm rate, right) and complex channel annotation coverage. Pearsons correlation.	137
6.5	Relationship between scene memorability (hit rate, left, and false alarm rate, right) and simple channel annotation intensity. Pearsons correlation.	138
6.6	Relationship between scene memorability (hit rate, left, and false alarm rate, right) and simple channel annotation coverage. Pearsons correlation.	138
6.7	Relationship between clutter (left) and symmetry (right) computational metrics and scene DPrime. All correlations significant. . . . .	142
A.1	These three objects frequently appear together inside the memorable regions of an image, of that category. VISHEMA Categories: Kitchen, Living Room, Work/Home, Big . . . . .	155
A.2	These three objects frequently appear together inside the memorable regions of an image, of that category. VISHEMA Categories: Small, Isolated, Populated, Public Entertainment . . . . .	155
A.3	Five object co-occurrences, all categories. . . . .	156
A.4	Seven object co-occurrences. Kitchen, Living Room, Work/Home, Big . . . . .	157
A.5	Seven object co-occurrences. Small, Isolated, Populated, Public Entertainment . . . . .	158

## LIST OF FIGURES

B.1	Structure of generator and discriminator blocks, showing interpolation and convolutional filter sizes. Similar structure to that of the standard progressive GAN. All convolutional layers employ Leaky ReLU activation and weight-scaling [70]. We find adding a hyperbolic tangent activation to the output of the interpolation layer to improve training speed and stability. . . . .	159
B.2	Flattened network diagram showing resolution and channels of each architecture block. Every resolution block output is passed through the memorability estimator. . . . .	160
B.3	Additional exemplars of generated highly-memorable images. . . . .	160
B.4	Additional exemplars of generated low-memorability images. Low-memorability images appear simpler, and often contain more “closed” perspectives, with less variation across the image. . . . .	161
C.1	Ten most complex upright scene images, as rated by humans. Complexity maps for these images are placed in the same location as their corresponding image in Figure C.2 . . . . .	163
C.2	Complexity maps for the ten most complex upright scenes, showing complex regions (blue) and simple (red) regions, as described by humans. Complexity score in upper left-hand corner. . . . .	163
C.3	Ten most complex inverted scene images, as rated by humans. Complexity maps for these images are placed in the same location as their corresponding image in Figure C.4 . . . . .	164
C.4	Complexity maps for ten high-complexity inverted scene images. Note increased annotation coverage compared to Figure C.2 . . . . .	164
C.5	Ten least complex upright scene images, as rated by humans. Complexity maps for these images are placed in the same location as their corresponding image in Figure C.6 . . . . .	165
C.6	Complexity maps for simplest upright scenes. Note prevalence of annotated ‘simple’ regions, matching low overall score. . . . .	165
C.7	Ten least complex inverted scene images, as rated by humans. Complexity maps for these images are placed in the same location as their corresponding image in Figure C.8 . . . . .	166



## LIST OF FIGURES

C.8 Complexity maps for least complex inverted scene images. Evidence for loss of localisation ability compared to upright low-complexity scenes (Figure C.6) . . . . .	166
C.9 Additional prediction results for complexity maps. . . . .	167

## List of Tables

2.1	The most common feature extractors shared among memorability prediction and analysis models. . . . .	24
3.1	Reconstruction accuracy for three deep learning architectures. . . . .	68
3.2	Comparison with Prior Work . . . . .	70
3.3	Prediction results for the VMS memorability channel. SH: single-headed output. KL: Kullback-Leibler Diver. . . . .	75
3.4	VMS false memorability channel prediction results. . . . .	76
3.5	VMS reconstruction results. True & False refer to memorable and falsely memorable schemas (green/red in images). $P^{2d}$ is the Pearsons 2D correlation [2, 83]. LaMem performance measured by Spearmans Correlation ( $\rho$ ). xA indicates no attention, xDF no dual-feedback, xM, no multi-scale information, xVMS, score prediction only. A dash in the table indicates the network does not compute that output. We include results for both a modern backbone, ResNet50, and for a fair comparison with prior work, a VGG16 backbone. A comparison with state-of-the-art is given against the current best model; vms-VAE from [83]. . . . .	80
4.1	Vischema 1 and Vischema 2 consistency, per category. Certain categories of images, such as kitchens or scenes involving public entertainment (playgrounds, theme parks) are more consistent than others, such as the isolated category. Higher consistency implies participants agreed on specific features that made the image memorable. . . . .	86

LIST OF TABLES

4.2 D-Prime analysis of human memory for each category in the Vischema 1 and Vischema 2 datasets. High values clearly indicate that the memory signal for the given image category is strong and thus image memorability for human observers is high. Certain categories have stronger signals than others, possibly due to easier or more available encoding schemas for that category among the human participants. . . . . 87

4.3 Comparison between MEMGAN and GANALYZE. . . . . 97

5.1 Results of a hierarchical regression analysis showing the contribution of each potential complexity factor towards explaining variance (coefficient of determination,  $R^2$ ) in complexity ratings for our VISC-C dataset. Together, clutter and symmetry explain 36% of human complexity (disjoint sets of human ratings explain 51% of each others variance). Entries in bold indicate significant increase in variance explained. Standard error of each linear model (Lm. Std.) and residual sum of squares (RSS) are reported for completeness, and is already incorporated into reported  $R^2$  120

5.2 Results of a hierarchical regression analysis run on human complexity ratings from the VISC-CI dataset (inverted scene images). The main contributors are clutter and symmetry (38%), with minor contribution from openness. Entries in bold indicate significant difference in variance explained. Std. Error is reported for completeness, and is already incorporated into given  $R^2$  . . . . . 123

5.3 Hierarchical Regression Results for the BOLD5000 dataset. Best explanatory model uses all factors, likely an effect of the more varied dataset, explaining 11.32% of variance in complexity ratings. These factors come close to human consistency over the dataset (one split of human ratings explains 12.67% of variance of the other split on average). . . . . 125

6.1 Comparing correlation of two-dimensional regions between memorability data and complexity data. All values significant. . . . . 139

LIST OF TABLES

6.2 Results of multiple linear regression, with Complex and Simple coverage removed to avoid multicollinearity concerns. Coefficients for each variable are shown, as is the coefficient of multiple regression (R) and variance explained (R-squared), as well as the variance explained when including all factors (af-Adjusted). All regressions are significant. Complexity can explain a small, but significant portion of variance inherent in memorability data for DPrime, hit rate, and false alarm rate. Significant values shown in bold,  $p < 0.001$ : \*\*\*, 0.05: \* . . . . . 141

D.1 Results of multiple linear regression. Coefficients for each variable are shown, as is the coefficient of multiple regression (R) and variance explained (R-squared). All regressions are significant. Complexity can explain a small, but significant portion of variance inherent in memorability data for DPrime, hit rate, and false alarm rate. Significant values shown in bold,  $p < 0.001$ : \*\*\*, 0.001: \*\*, 0.05: \* . . . . . 168

# Introduction

## 1.1 Motivation

Psychologists have long explored the characteristics of visual memory, investigating capacity and level of detail alongside accuracy and fallibility. However, until relatively recently two significant questions remained without clear answers:

- What makes an image memorable - and why are some images more memorable than others?
- Can we predict how well a human will recall having seen a given image, either with classical or machine-learned models?

Solving this problem in its entirety would require not only an understanding of the mechanisms of human visual memory, but also a method to understand and represent an image such that memorable factors can be evaluated. It would further understanding of why certain things are memorable, and not others, and provide insights into how the brain processes memory and what it prioritises. It is only with the advent of recent computational techniques that this problem has begun to become tractable. While there has been clear progress towards predicting memorability, deciphering exactly what causes an image to be memorable remains somewhat unclear; and so far defining an image in the terms of its exact "memorable components" remains difficult.

While analysing and predicting memorability remains in the realm of research, potential practical applications for this topic are numerous. Once memorability can be predicted,

this goes a long way to allowing memorability to be altered. The most obvious application is commercial, allowing for the memorability and hence efficacy of adverts to be assessed rapidly and automatically. Educational aids can be evaluated to determine how likely they are to be remembered, as can important public information and infographics. There are also medical applications - a baseline memorability score for a set of images could be used to track the decline of patients with cognitive diseases.

### 1.1.1 A Computational Approach

Recent advances in machine learning have led to techniques that allow computers to replicate certain human cognitive abilities. In certain cases, such techniques provide results indistinguishable from that of a human addressing the same cognitive task. It is this replicative ability that is of interest in the area of memorability prediction. Without computational assistance, predicting the memorability of an image is a nearly intractable problem. Not only are humans incapable of predicting which images are memorable, large scale human prediction of hundreds of thousands of images would be an exceedingly expensive, time consuming task. If a computer can be taught to emulate the function of human memory, these issues vanish. Computational power allowing, rapid image memorability prediction becomes possible. In the case of more complex models, determining why the model arrives at a given output can reveal hints about how human visual memory functions. These complex computational models are not without their drawbacks. They can provide stunningly accurate predictions, and even match human-level performance, but their interior logic can be obscure and difficult to interpret.

Computational memorability prediction and analysis is a field just under a decade old. In this relatively short time, memorability prediction has made progress in leaps and bounds (the progress in *analysis* remains harder to quantify). Generally, memorability prediction is framed as a regression task, the goal of which is to output a score, between 0 and 1, that indicates how likely that image is to be remembered by an ‘average person’. Psychological experiments gather data on a set of images, determining how well humans remember an image (often this takes the form of a repeat-recognition task). The predicted values and the ground truth values are then compared to determine the degree of consistency between them. If humans find one image generally more memorable than another, the computational model of memory should reflect this. Like most research that involves machine learning, this field rapidly grew to leverage the power of deep neural

## CHAPTER 1. INTRODUCTION

networks, the best of which are currently able to predict memorability scores with a consistency close to that which groups of humans share with each other. These models have allowed the relationship between memorability and other psychological image properties such as saliency, aesthetics, and interestingness to be evaluated. Neither of these properties are capable of explaining the variance inherent in image memorability, and in many cases, there is no relationship at all.

This research has shown that there is a high degree of consistency ( $\rho = 0.75$ ) between participants memory for images; in general, people will remember the same memorable image, and forget the same non-memorable image. Low-level image features, such as colour, intensity, or object counts do not correlate strongly with image memorability [69, 68]. Instead, high level semantic attributes such as image category, the contents of the image [20], the objects present [40, 141], and scene dynamics and category [94] appear to better correlate with image memorability scores. Features relevant to memorability can be extracted through deep learning mechanisms [9] in order to predict memorability scores [8, 44, 125] for images, with recent deep models reaching human-level performance.

### 1.1.2 Beyond Single-Score Metrics

Compared to overall memorability score prediction, there has been less research into examining memorability *across* an image, rather than with a single summative score. Probabilistic models have been created but lack a ground truth dataset to compare with. The effect of the memorability of individual objects in an image has been examined, but remains a much more difficult task due to need for segmentation of the image, a notoriously difficult problem. Recent work on this topic moves away from memorability score prediction towards a more complete model of visual memory. The Visual Memory Schema (VMS) maps gathered via the VISHEMA experiment define the regions of an image that causes that image to be either remembered - or falsely remembered. VMS maps are highly consistent (correlation histogram mean of 0.7) [2], indicative that participants agree on which regions cause a given scene image to be remembered. This work combines cognitive theories of visual memory with machine learning, and introduces the concept of visual schemas. Visual schemas are mental structures that enable an image to be remembered. These schemas allow generalisation about memorability across different images in different categories, and across individuals, providing significantly more information than a single score metric that describes memorability. With these

schemas, an image can be defined in terms of its ‘memorable regions’. The elements contained in these regions directly relate to that images’ memorability; that is, the structure, objects, and semantic units contained within this region aid in that image being remembered. However, predicting these schemas is more difficult than predicting a score alone. Predicting a one-dimensional metric is easier than predicting a three-dimensional schema which varies both spatially, and in intensity. Predictive efforts are made more difficult by the lack of available training data: the only currently existing dataset of Visual Memory Schemas and their corresponding scene images has only 800 images; a significantly lower amount compared to single-score memorability datasets (which number in the tens of thousands of images).

With memorability score prediction now having a close correlation with inter-human memorability scores, there is now the opportunity to start looking beyond score prediction towards a finer-grained understanding of memorability. The VISHEMA experiment represents an initial step in this direction. However, there remains a long way to go before the consistency between predictions about visual schemas and ground-truth scores reach the same level as memorability score prediction. However, further investigating visual schema generation for images could lead to models that better represent human visual memory and hence improve overall understanding about memory. Another relatively unexplored avenue opened by computational memorability prediction is examining the effects of attempting to modify images to improve or reduce their memorability. This moves beyond asking *what* makes an image memorable, to actively employing what we already know to create images that cause a direct change inside the human visual long-term memory system. Much existing work in this field focuses on either modifying existing images, or on those of face images which is a sample set distinct from those of the natural scenes. Recent advances in machine learning and image generation, combined with maps that define which regions of images cause an image to be remembered, open a path towards the generation of memorable images from scratch, rather than modifying already existing images. Attacking this problem is an important first step towards real-world applications of research into memorability.

### 1.1.3 Scene Complexity

The correlation between image memorability and several other perceptual characteristics has already been examined, and few of these characteristics are capable of explaining



## CHAPTER 1. INTRODUCTION

scene memorability. This holds for aesthetics, ‘interestingness’, and colour properties. However, the relationship between how complex a scene is and the memorability of that scene remains relatively unexplored. Understanding scene complexity, and how humans process and evaluate said complexity is a worthwhile endeavour by itself, leading us towards a better understanding of the brain and its vision processing systems. Much like image memorability, scene complexity suffers the same limiting factor of requiring humans "in-the-loop" to extract data for any given scene; predictive models offer the chance to evaluate complexity for any scene image, whether human data exists for that image or not. Perceptual complexity itself has prior theoretical grounding, which defines complexity as the intricacy or detail present in a line drawing [123], as the degree of difficulty involved in generating a verbal description of a texture [60], or evaluates complexity in context of aesthetics [35]. However, these measures do not specifically target scene perception; with initial research on scenes [105] finding evidence that clutter and mirror symmetry play a key role in visual complexity, along with openness and object organisation [103].

There are direct applications of scene complexity understanding, from marketing applications (e.g; perhaps you want your advert to be easier to visually process and comprehend and thus less complex), to potential impacts for psychological experiments (you may want all your visual stimuli to be of similar complexity to exclude a confounding factor) to healthcare applications (the evaluation of cognitive processing disorders; how easily a patient can process an image of known complexity). However, through the lens of image memorability, scene complexity affords us an additional metric that may help explain why some images are more memorable than others. A visual memory schema captures the overall memorable semantics of the image; but complexity may offer the ability to investigate how the overall *detail* present in the scene affects the memorability of that scene. Two problems face this line of inquiry. One is the lack of existing data, as there are very few scene datasets that exist with both memorability and complexity annotations; and none suitable for large-scale machine learning. The other is that factors that explain the complexity of scenes are not well understood, and that currently all complexity measures remain firmly in the single-score domain; with minimal exploration of which *regions* in an image contribute to its perceived complexity.

### 1.1.4 Research Direction

The research in this thesis focuses in three main directions, that each build on the concept of Visual Memory Schemas. These directions are as follows:

1. VMS maps define the areas of an image that cause the image to be remembered or falsely remembered. However, prediction of these maps is a difficult task and so far has only been accomplished in a limited capacity. The current limitations of generating these maps are examined, and improved methods of generating them explored. This involves the development and application of more advanced machine learning techniques to the problem of VMS map generation, as well as the creation of novel architectures/approaches. As the existing dataset is of limited size, posing issues for existing machine learning methods, larger-scale VMS datasets will need to be gathered. This further exploration of visual memory schemas with modern deep learning techniques offers the potential to better understand what makes images memorable.
2. It has recently become possible to create highly realistic images using generative models. Such generated images do not exist in the dataset used to train the models, and can be considered ‘new’ images. Combining VMS map models with generative models could lead towards the generation of *memorable* or *non-memorable* images. The generated images would be evaluated via human memorability experiments in order to evaluate how well the model learned to generate memorable/non-memorable features in images. Successfully accomplishing this further validates the VMS model of image memorability. The results of these experiments would have interesting implications for the future of the applications of memorability manipulation.
3. The relationship between memorability and complexity in scene images is not well understood, and neither are the elements that contribute to the perception of a scene’s complexity. Visual memory schemas offer the opportunity to investigate how the memorable regions of an image relate to that image’s complexity, and increase our understanding of both memorability *and* complexity. However, as no dataset currently exists that contains both scene complexity scores, labelled regions, and memorability information, this data must first be acquired. Such a dataset would ideally be large enough to afford the chance to develop neural

## CHAPTER 1. INTRODUCTION

network models capable of predicting scene complexity, and evaluation of this model could lead to a better understanding of how humans perceive complexity.

### 1.2 Thesis Structure

The remainder of this thesis is organised into six chapters, and is intended to both provide the reader with a background in computational memorability and complexity prediction, and of the advances made during this research project. The chapters are structured as follows:

**Chapter 2** is intended to provide the reader with a general background in memory, other perceptual image characteristics, and deep neural networks.

**Chapter 3** first introduces the concept of a Visual Memory Schema in greater detail, describes the existing dataset, and details the experiments conducted to gather further data and better understand what that data reveals about scene memorability. Secondly, the chapter explores progress made in developing neural network models of visual memory schemas, and evaluates several differing techniques as applied to image memorability prediction.

**Chapter 4** presents a novel neural network model that combines work on predicting VMS maps with that of generative models in an attempt to synthesise memorable or non-memorable images. The chapter also details the design and results of a repeat-recognition experiment to understand the efficacy of the model on human memory.

**Chapter 5** describes a perceptual complexity experiment that gathers two-dimensional complexity information from humans for a scene dataset, and operationalises several psychologically grounded factors that explain the complexity ratings given by humans. A neural model, combined with said factors reaches human-level performance for the dataset. The influence of semantics is explored through examining the complexity of inverted scenes.

**Chapter 6** analyses the relationship between scene complexity and scene memorability.

And finally, **Chapter 7** reviews the work as a whole, summarises the contributions of the work, and discusses the potential future directions this research could be taken in.

## 1.3 Publications & Presentations

The following papers have been published as a result of this work:

- C. Kyle-Davidson, A.G. Bors and K.K. Evans. ‘Predicting Visual Memory Schemas with Variational Autoencoders’. In: *Proc. British Machine Vision Conference (BMVC)*. 2019
- Cameron Kyle-Davidson, Adrian G Bors and Karla K Evans. ‘Modulating human memory for complex scenes with artificially generated images’. In: *Scientific Reports* 12.1 (2022), pp. 1–15

The following parts of the work have been presented as conference posters:

- Using Visual Memory Schemas for Modulation of Image Memorability (VSS 2021)
- Characterizing and Dissecting Human Perception of Scene Complexity (ECVP 2020, Psychonomics 2020)

The following parts of the work have been submitted for publication:

- Cameron Kyle-Davidson, Elizabeth Y. Zhou, Dirk B. Walther, Adrian G. Bors, Karla K. Evans. ‘Characterizing and Dissecting Human Perception of Scene Complexity’ *under consideration*

# Background

In this chapter we aim to give an overview of the fundamental concepts that this work employs. This covers both the basics of human memory, foundational machine learning, and perceptual image characteristics. These topics are vast, and cannot be covered in their entirety; instead we focus on areas directly relevant to the work presented in this thesis. From a psychological perspective, we briefly cover memory as a whole; then focus explicitly on Visual Long-Term Memory (VLTM). On the computational side of things, basic neural networks components, and common architectures used later in this work are defined. This chapter is intended to serve as an overview; more detailed literature sections that relate directly to discussed work are available at the start of each chapter, where relevant.

## 2.1 A Brief Overview of Memory

It is well accepted that memory can be effectively modelled as a combination of two different high level subsystems; that of *semantic* memory and that of *episodic* memory [41]. Semantic memory contains things implicitly known, such as how to read, speak, and perform arithmetic. In general, learned skills are recorded in semantic memory. Contrasting this, episodic memory records the autobiographical events of our lives. When events and items from our past are recalled, this utilises the episodic memory store. As disparate events can be separated by either (and both) time and space, this entails that there is a degree of temporal-spatial tagging to information stored in episodic memory; we can usually recall both the time and location of an event, and can additionally recall

## 2.1. A BRIEF OVERVIEW OF MEMORY

the temporal-spatial relationship between one event and others, for example, recalling that you walked into the kitchen prior to walking into the living room. [129] adopts the term ‘engram’ to refer to information encoded in specifically in episodic memory.

Semantic memory, however does not record events and experiences, instead storing rules, symbols, concepts, and the relationship between them; thus providing the foundational elements for storing implicit knowledge. Further differences arise when considering the loss of information from either system, as well as the consequences of information retrieval. Episodic events appear far more readily lost than semantic knowledge ("one does not forget how to ride a bike"), and understanding of what causes loss of semantic knowledge lags behind the understanding of which conditions lead to loss of episodic knowledge. No matter which system information is recalled from, the actual act of recalling is often entered into the episodic store (you remember remembering), providing an interesting form of feedback between the two systems.

While these subsystems are often considered separately, Tulving [41] originally hypothesised some degree of interdependence between them beyond that mentioned above. While not all episodic encodings require an intervention of the semantic memory system, some experiences may benefit from semantic store assistance; for example, a mathematician may better remember a seminar talk than a lay-person due to semantic knowledge of the presented formulas. The act of recalling was hypothesised to combine engram information with semantic store information in order to *reconstruct* the memory.

Since Tulving’s theory was written, much research has supported the distinction between the two types of memory; and most convincing is neurological studies that find clear evidence that the semantic and episodic stores can be damaged independently of each other. However, Greenberg finds that semantic and episodic memory are in fact reasonably intertwined, and damage to one system impedes the other[52], especially with regards to the learning of new information. As Tulving hypothesised, the episodic store is instrumental in fast learning of semantic knowledge; and when this store is damaged, the ability to learn new skills diminishes. In turn, when the semantic store is disrupted, the ability to encode new episodic memories is similarly harmed; both memory stores appear to support the other. While it is still possible to remember experiences with a damaged semantic store, and to learn skills with a damaged episodic store, the ability to do so is greatly below normal.

## CHAPTER 2. BACKGROUND

This culminates in the conclusion that not only does the semantic store facilitate the encoding of episodic memories, but that episodic memory also aids in the encoding of semantic knowledge. That is, known skills help you to recall events, and recorded events help you encode new skills. This entanglement of memory subsystems holds at retrieval as well, with episodic memory providing a fast pathway for the efficient retrieval of semantic knowledge; when episodic memory is impaired, semantic recall falters. When the semantic store declines in functionality, while episodic memories can still be recalled, these memories lack specific detail. Further decline leads to worse autobiographical recall in general.

It has been known for a long time that memory is not perfect; and is not an exact, lossless recording of data. Instead, episodic memory is widely considered to be reconstructive, rather than reproductive. Tulving frames this as a ‘recoding’ of stored engrams; a set of operations that takes place on the engram once it has already been encoded into memory. Thus, remembered events are not reproduced exactly as they occurred. Instead, memories may be pieced together from recorded fragments. Schacter [116] hypothesised that the reason for this is that a constructive memory system can be re-purposed to imagine future events, and that lack of a rote-recording system is a positive, rather than a negative, and is in fact representative of an *adaptive* recording system. The past does not repeat verbatim in the future, but it does echo, and being able to draw upon multiple prior experiences aids in constructing adequate responses to future situations. Indeed, similar brain regions activate when imagining the future versus remembering the past. [116, 117]. We further examine the reliability of memory, and visual long-term memory in general, in Section 2.2. In part, in this thesis we probe and model visual long-term memory, a subset of episodic memory, by investigating the memorability of scene images that are perceived and stored for longer than a few seconds.

### 2.2 Remembering Images

The human capacity for recognising images that we have seen at an early time appears very large. Standing [126] evaluated the capacity of visual memory through a recognition task with increasingly large amounts of images. In the largest experiment conducted, 10,000 images were shown to each participant. Standing found a linear relationship; as the number of images shown increased, both the number of images remembered

## 2.2. REMEMBERING IMAGES

**and** the number of images forgotten increased, leading Standing to hypothesise that the capacity of visual memory is ‘practically limitless’. It is generally assumed that as memories move from working memory, to short term memory, to long term memory, that the detail in the memory fades, leaving only a general gist of the image; such as the category of the image, and general scene elements. While the capacity of visual long-term memory appears very large, this could be an illusion - storing just the gist of an image requires much less information than storing a detailed representation of the image. This stored gist trace would be sufficient to determine which image you have seen before when presented with multiple possible options, but recognition performance would start to degrade when those options are semantically similar to each other. Brady conducted a study to determine the capacity of *detailed* visual long term memory via a 2-Alternative Forced Choice methodology paired with three options - category distinct foil, same category, different item foil, and same category, same item, different state foil. 2,500 real world objects were used in the images shown. Despite the difficulty, recognition performance remained high, dropping to only 87% in the most difficult case [15]. Repeats of images were also tested, and identified correctly 96% of the time. It follows from this that visual memory not only has a large capacity for images, but that representations stored in visual memory are highly detailed. Brady places the information-theoretic capacity of VLTm at approximately 228000 unique codes.

Cunningham, however, while agreeing on capacity, concludes that long-term memory remains highly dependent on gist [33], and that the difference in memory performance is often due to differences in testing techniques. Two Alternative Forced Choice, a common choice for memory studies, where two images are compared and one must be chosen, may not accurately reflect the memory stored in the brain, and it is unclear what influence familiarity vs recollection has on the choice. Cunningham makes use of an ONR (old/new recognition) test to reduce the effect of noisy or incomplete recollections leading to a correct result regardless of quality of the memory stored. The Brady experiment was replicated, finding that ONR performance degraded where 2AFC performance did not. It appears that while the capacity of VLTm is large; this capacity is in fact highly dependent upon stored gist traces as well as detailed representations. So far, these experiments have examined the memorability of objects; it is natural to assume that remembering scenes is more difficult, and that there is likely to be a fall in performance when tasked to recall complex scene images. However, Konkle *et al.* [79] show that



## CHAPTER 2. BACKGROUND

even tasked to remember 2,800 scene images, and shown same category-distractors, recognition accuracy for scenes is high. This implies that scene memorability is high-fidelity, and stores enough detail about a scene image for it be selected against other, similar scenes. That is, more than just the scene category is preserved in visual long-term memory. As memory for abstract images is very poor [80], it appears that some kind of preexisting mental structure is required for this level of memory performance.

It has been found that VLTm is also subject to the level of processing (LOP) effect [5], where deeper processing leads to better recognition. Here, level of processing refers to the amount of additional processing undertaken when viewing a stimulus; for example, judging the ‘intelligence’ of a face shown in a photograph. The LOP effect has primarily been studied in faces; with somewhat mixed results. Generally, making some form of judgement on a face enhances recognition of that face; though it is unclear exactly what depth of processing is necessary to cause the effect. In some cases, tasks thought to require deeper processing show less of an effect than shallower tasks. Recently, this effect has been investigated beyond that of faces; examining the effect on images of doors. The LOP effect was consistent for this image set, though modest, whereas in the contrasting verbal processing experiment, effect size varied widely. Baddeley *et al.* [5] relates this to the concept of affordance, where the relation of some concept to an organism *affords* some possible action, such as a chair being capable of being used for sitting, or as a potential weapon. Baddeley notes that rich encoding does not necessarily lead to good recognition unless the coding is sufficiently complex enough to defeat similar distractors. Hence, one reason the LOP effect appears relatively small for visual stimuli is that discriminative features in the stimuli set used were not powerful enough to defeat similar distractors present in that dataset. Verbal stimuli lend themselves more easily to semantic elaboration when being deeply encoded, as they afford a rich tapestry of related words and meanings. Door images, however, afford little to the observer, and hence their encoding depends more upon perceptual features present in the image.

It would be incorrect to assume that an images memorability is a binary property. While a single person may recall or forget an image, over a population, that image’s memorability exists on a continuum; between ‘most likely to be recalled’ and ‘least likely to be recalled’. Le-Hoa Vo [134] shows that these differences in intrinsic image memorability appear rapidly after presentation of the image, and the longer the lag between presentation and test, the greater the divergence between memorable and less-

## 2.2. REMEMBERING IMAGES

memorable images. Vo defines memorability as a function of the hit rate of the image, where a hit corresponds to a previously shown image being recognised. The target image could be repeated at any one of four ‘lags’ after that image was shown, which correspond to how many images are shown in-between repetitions of the target image. The shortest lag on average was 20 seconds, as each image was shown for two seconds with a 500 millisecond fixation target in between, while the longest lag corresponds to over ten minutes. Poorly memorable images show a decrease in recognition of 20% after 20 seconds. After ten minutes, this has decreased to 32% compared to the drop from 97% to 78% for highly memorable images. Vo also tracked pupillary response and blink rate to gain an understanding of how cognitive load differs when recalling memorable or non-memorable images. In this context, pupillary response refers to the change in size of the pupil, and blink rate refers to how many blinks occur per measured time period. Blink rates tend to decrease under high cognitive loads, while pupils dilate more in response to ‘seen’ items vs new items. Poorly memorable images correspond with increased pupillary responses and decreased blink rates. Vo states that the increased pupillary responses mirror the greater cognitive load required for recollective processes, and hence that poorly memorable images are more difficult to retrieve than memorable images.

While the neural correlates of memory are still not well-understood, it does appear that there exists a distinct processing stream associated with memory, that ‘tags’ viewed stimuli for later encoding. Bainbridge *et al.* conducted a study employing fMRI imaging paired with a task that involved dividing stimulus into male/female (for faces) or indoor/outdoor (for scenes) [7]. No mention of memory was made to the participants. After the scanning task, participants are tasked with a memory test that they were not aware was coming. There was evidence of significant sensitivity in the ventral visual stream and the medial temporal lobe to the memorability of viewed image. Forgotten images when viewed again caused a similar stimulus to arise in the memorability-sensitive brain regions as the first time the image was viewed. These brain regions are the same brain regions that activate during first time viewing of the image. This processing stream is termed ‘memorability’, as it appears responsible for determining whether a given stimulus should be remembered. Memorability occurs beyond low-level perception (no sensitivity in early visual cortex), and may ‘reflect the statistical distinctiveness of a stimulus along a multidimensional set of axes’, and hence be used to tag stimuli for

## CHAPTER 2. BACKGROUND

later memory encoding by the medial temporal lobe. Later work reinforces the idea of a ‘perceptual trace’ of memorability, finding that signals associated with highly memorable images propagate across several brain regions associated with high-level visual processing [97]. The brain appears to be able to subject memorable stimuli to a deeper level of visual processing than comparative low-memorability stimuli.

It is natural to assume that intending to remember an image improves how well that image is remembered. Given a task, it makes sense that exerting effort at that task will lead to better performance. Previous studies have shown little to no effect from intending to remember images (although there is a significant effect when the stimuli is verbal). Block *et al.* suggests this may be due to other effects overshadowing the effect, such as the level of processing effect combined with rehearsal strategies [13]. To determine whether an effect exists when these confounding variables are excluded pictorial stimuli are shown rapidly after one another, preventing either deep analysis or rehearsal. Block found a significant intent to remember effect vs incidental remembering when participants were tested with briefly presented, unrehearsable pictorial stimuli of faces. This appears to indicate that the intent-to-remember effect only arises in the most difficult cases. Evans and Baddeley [43] test this further, employing visual stimuli that have distinctive detail removed. In the relevant case where an intent to remember effect appeared, participants were tasked with remembering scenes of doors that had potential diagnostic features removed. It may be the intent to learn helps in selection of diagnostic features, or simply increases the amount of features encoded, and in most cases is not required. Only in the most difficult cases is a conscious effort beneficial for visual memory.

### 2.3 The Fallibility of Visual Memory

As we established in Section 2.1, episodic memory, and by way of inheritance, visual long-term memory, is reconstructive, rather than reproductive. To reiterate, while visual long-term memory clearly has a large, detailed capacity, it is by no means perfect. Errors often occur, some due to the reconstructive nature of episodic memory, and others due to perceptual errors that occur at sensory input. While this reconstructive ability is almost certainly an evolutionary advantage, it does lead to an interesting defect; that of false remembering. While it is obvious that during an image memorability experiment some

### 2.3. THE FALLIBILITY OF VISUAL MEMORY

images will almost certainly be forgotten, somewhat more surprising is that some images will be marked as ‘remembered’ even if the participant has never seen that image before. In fact, many of the previously examined studies show these ‘false recognition’ events. In addition to this apparent tricking of the reconstructive visual memory system, there also exists perceptual effects that alter stimuli almost as soon as they are received as input. Most notable to that of image perception is the boundary extension effect; where participants remember more of a scene than in fact they actually saw - constructing an ‘artificial’ boundary beyond the edges of the image.

While it is certainly interesting to learn about remembering, false remembering offers the equally valuable opportunity to gain a better understanding of exactly how visual long-term memory operates. Koutstaal *et al.* tested the recognition performance of older and younger adults for detailed coloured pictures of objects, looking specifically at false recognition [81]. While both older and younger adults showed significant false recognition for each image category, older adults showed reduced recognition of unrelated targets (targets not similar to the overall image category theme being tested), indicating they relied more on conceptual/perceptual similarity, which Koutstaal believes is indicative that only the gist trace of the image is being retained; thus making it easier to ‘false alarm’ due to similar gist traces from similar images. Specifically, for within category lures, older adults had a higher false alarm rate vs younger adults. This may indicate that correct recognition of images is due to specific, detailed traces, but in the case of false recognition, recognition defaults to a gist trace, sensitive to general semantics present during the initial encoding of the viewed image.

While not strictly related to visual memory, episodic memory itself is vulnerable to ‘misinformation’ where a memory is affected by post-encoding information. Loftus demonstrates that being warned about misinformation does not necessarily avoid the damaging effect of this new information [91]. Loftus goes on to show that it is possible, over several weeks, to construct rich, detailed, and entirely false memories in participants. This is demonstrated by holding a series of interviews with a participant about an event that never occurred. As the interviews progress over the weeks, the false memory becomes increasingly detailed. Given that entire false experiences can be implanted by a researcher, it is not surprising that this occurs in the much more limited case of believing to have seen an image. Interestingly, these false memories contain less detail [118], which matches nicely with the later work of Koutstaal *et al.* [80]

## CHAPTER 2. BACKGROUND

Other errors occur not during encoding, but at perception. The boundary extension effect is where, upon viewing an image of a scene, and then later being asked to identify if that same image is identical or zoomed in, participants commonly choose the ‘zoomed in’ option, suggesting that people construct the scene mentally beyond the actual boundaries shown. Intraub shows that this effect takes place in only 1/20th of a second, and hypothesises that this effect is an integral part of the perceptual system. Our senses exist to let us construct the world around us, and the boundary extension effect appears to ‘pre-empt’ parts of the world that may be likely to be looked at shortly [66]. Spano et al later examined these boundaries further, and find that they persist even among people with impaired hippocampuses [124], indicating that it is a brain-wide phenomena, and not localised to one area associated with memory.

Much remains to be understood about false remembering from the perspective of visual long-term memory. Just as it not entirely clear what causes certain images to be remembered over others, it is equally unclear as to what causes the false remembering of certain images. The only thing that is clear is that human memory is certainly not infallible, and that a complete model of visual memory would be capable of explaining both why an image is memorable, and why an image might cause false remembering. As we explore later, computational models have allowed great progress in the former; the latter remains relatively unexplored. However, memorability is not the only characteristic attributable to images, and the next section explores other perceptual measures that *may* associate with memory.

### 2.4 Interestingness and Aesthetics

We have established so far that images can be remembered, and that how memorable a given image is can vary. But memorability is not the only intrinsic characteristic common to images. Images can be judged along multiple different perceptual axes, all of which depend upon the content of the image. Most studied are those of interestingness and aesthetics; both image properties consistent among observers, and which have been shown to be capable of being modelled. While we will study image memorability itself in more detail in Chapter 3, in this section we will briefly discuss these other perceptual metrics, and whether how interesting, or pretty, an image is, has anything to do with how well it is remembered.

### 2.4.1 Interestingness

What causes images to be perceived as interesting? It could be their degree of aesthetic attractiveness; asking whether ‘prettier’ images are more interesting. Turner *et al.* compares the idea that interestingness relies on appraisal of a high degree of pleasantness against the idea that stimuli can be interesting and unpleasant [1]. Participants were asked to view paintings and rate them for emotional and cognitive responses on a bipolar Likert scale. Turner found that ratings of pleasantness and ratings of interest were essentially independent. Disturbing paintings tended to be appraised as more interesting, though less enjoyable. At least in the case of paintings, it appears that interestingness lacks a relationship with aesthetics, and that visual stimuli can be both unpleasant to observe, yet interesting, and alternatively pleasant, but boring. Instead, it appears that interestingness is a function of image content and composition. Dhar *et al.* examines how well interestingness can be predicted by several high level attributes, including compositional, content, and sky-illumination attributes [37]. The work takes advantage of measures of photographic quality commonly used among photographers, including opposing colors, low depth of field, and the two-thirds rule. These metrics, when used to train a Support Vector Machine (see Section 2.5), performs extremely well at predicting interestingness. Generally, more interesting images appear to be clearer depictions of their category; with less interesting images being less clear or more cluttered.

But does the degree of interestingness of an image have anything to do with the memorability of that image? While at a glance the assumption that interesting images are more likely to be remembered makes sense, in practice, there is little relationship between the two properties. Isola *et al.*, in one of the first papers on computational memorability [68] (see Chapter 3) briefly examines interestingness and memorability, and finds no relation. Gygli *et al.* also finds that interesting images are not necessarily memorable [55], and instead finds a negative correlation between how interesting an image is, and the memorability score of the image. This may be due to an artifact of the dataset; memorable but dull images appear to contain singular objects; whereas interesting images contain more detail, which may make them more difficult to recall. Interestingness does correlate with "assumed memorability" ratings from the human participants, which suggests that when estimating how memorable an image is likely to be, humans employ interestingness as a metric; even though this does not actually predict memorability. In contrast, aesthetics and interestingness are correlated with each other, contradicting

## CHAPTER 2. BACKGROUND

Turners original findings. This may be because the dataset of images used by Gygli and Isola lacked images specifically designed to be unpleasant. In order to facilitate computational prediction of interestingness, Gygli introduces a measure of ‘unusualness’, which defines how different an image is from its neighbours, and uses a similar metric to determine how unusual selected patches in the image are in relation to each other, hypothesising that a key aspect of how interesting an image is unusual features across an image. This predictive measure, combined with aesthetics estimation and several other metrics, including a complexity estimator, can predict interestingness with a strong correlation to ground-truth scores. They find the most unusual images tend to be the most interesting.

### 2.4.2 Aesthetics

Aesthetics in images generally refers to how ‘pleasing’ that image is to perceive. In this context, we might consider a landscape photograph of rolling hills, lit by a deep orange sunset highly aesthetic, and yet an image of a decaying garbage heap much less so. Being capable of determining the aesthetics of images has several real-world applications, such as image retrieval (‘find the best looking image in a database’), or as a teaching aid for novice photographers. However, it is also of interest to cognitive scientists; determining what causes images to be perceived as aesthetic, and which factors relate to aesthetics, helps to reveal how the brain processes visual stimuli.

Dhar *et al.* uses their set of describable image attributes (previously used for interestingness evaluation) and a dataset of 16,000 images with aesthetic ratings to train a support vector machine. They find that the same attributes that lead to good estimations of interestingness also lead to good aesthetic predictions, reinforcing the findings of Gygli *et al.* that for photographs, aesthetics and interestingness are interrelated. Much like interestingness, the predictive power of these high-level attributes is much greater than previously studied low-level metrics such as contrast or brightness [71]. Murray and Perronnin note a need for a large, diverse dataset for aesthetics reserach, and hence introduce AVA: A large scale database for aesthetics [98]. AVA contains over 250,000 images, combined with a variety of accompanying meta-data, including aesthetic scores, semantic labels, and photographic style-related labels. They show that generic models trained on a large-scale dataset outperform small-scale models that employ hand-picked features, tested over the same dataset. They additionally find that images with

## 2.4. INTERESTINGNESS AND AESTHETICS

the greatest variance in aesthetic ratings tended to be non-conventional photographs; i.e, those more open to individual differences when it comes to personal interpretation of the image.

Aside from these unconventional images, in general, real world scenes have consistent aesthetic preference scores between individuals. However, abstract images are much more specific to individual tastes. Vessel and Rubin compare scene preferences, abstract image preferences, and abstract vs scene preferences [132]. They suggest that visual preferences are driven by semantic content of stimuli, and shared semantic interpretations lead to shared preferences. They confirm this by de-emphasising the semantic content of real world scenes by intermixing real images with abstract in the image streams shown to participants. In this condition, individual preferences arise, which they hypothesise is caused by direct comparisons between previously viewed abstract images and real-world images. Abstract images themselves may not span the same semantic context as real world images, leading to less preference correlation; there is less ‘shared meaning’. The lack of agreement between the preferences of abstract images means computationally predicting aesthetic preference from a general dataset is likely to be more difficult than predicting aesthetics of scene images; and that the performance of artificial prediction of abstract art preference is unlikely to ever match that of real-world scene performance.

### 2.4.3 Relationship to Memorability

We have already seen above, in the work of both Isola *et al.* [68] and Gygli *et al.* [55] that memorability has either no relation with interestingness, or a weak inverse correlation. That is, how interesting an image is has apparently little to do with how well it will be remembered. Khosla *et al.* introduces LaMem [75], a large image dataset composed of several image subsets with various perceptual ratings. These subsets include ratings for image popularity, ranked as the view-count of images drawn from Flickr, saliency, taken from an image eye-fixation dataset, and emotions from the affective image dataset. Ratings for aesthetics are taken from the AVA dataset. All subsets have memorability ratings. Khosla *et al.* finds that these attributes, despite being high-level, have relatively little to do with memorability.

Highly memorable images appear to be more popular, but aside from the most memorable case, the difference disappears. The reason for this difference is not discussed,



## CHAPTER 2. BACKGROUND

and appears unknown, though it is reasonable to consider that it may be due to context effects (discussed further in chapter 3). Saliency shows a minor effect, with a difference between the most and least memorable images; more memorable images contain more consistent fixations between participants. This implies that memorable images contain singular specific items to focus on (i.e an object), though this is may be an effect of the dataset; remembering a single object is likely to be easier than remembering a scene image.

The emotional content of the image appears related to its memorability, with strong negative emotions (disgust, anger, fear) being more easily remembered than other emotions. The least memorable emotions seem to be pleasanter, such as awe, and contentment. This matches findings in [68]. Interestingly, images rated as ‘amusing’ appear statistically similar in memorability to those with ‘disgust’ ratings. However, the images with these ratings were drawn from a dataset specifically designed to contain images with affective content; a dataset of scenes or objects are unlikely to have a strong emotional component to their memorability.

Khosla *et al.* [73] find no relationship at all between aesthetics and memory; not even the weak negative correlation that might be implied by aesthetics strong relationship with interestingness. We can say with confidence that across a large and diverse dataset, that how ‘pretty’ an image is has little to do with whether that image will be remembered. In general, while image memorability has some limited relationships with other image properties, none of these properties are capable of fully explaining the memorability of an image, and certainly are not capable of describing memorability in all cases. Humans are certainly capable of remembering images even if they do not cause a consistent emotional response, or contain single objects to focus upon. This implies that memorability is a distinct image property; intrinsic to the image itself. Later work on image memorability takes advantage of relatively recent developments in both classical and neural-network based machine-learning methods; the next section describes some of these techniques, and their application to memorability prediction is discussed throughout this work.

### 2.5 Machine Learning

This section is intended to give an overview of some of the more common machine learning techniques that are often applied to the problem of memorability prediction.

This ranges from support vector machines (SVM), which used to rank among the most commonly used, to convolutional neural networks and other, more esoteric networks that are in use today.

### 2.5.1 Support Vector Machines

The technique that developed into what we today call a ‘Support Vector Machine’ was originally proposed in 1992 by Boser *et al.* [14] as a method for finding the maximal margin between ‘training patterns’ and a decision boundary. In other words, a support vector machine (SVM) is a machine learning algorithm that attempts to learn a decision boundary that divides a set of datapoints into different classes [32]. The ‘decision boundary’ is the optimal hyperplane that can best divide the training data into their separate classes. This hyperplane, by definition, has the greatest margin between the differing classes in the data, and is constructed by the support vectors of the data. These support vectors are those closest to the hyperplane (hence a subset of the input data), and as such are instrumental in describing the direction and placement of the plane. So far, this description works for *binary* classification; one hyperplane to divide two classes of data. In the case of multi-class classification, the problem is broken down into multiple binary classification problems; either finding multiple hyperplanes across the data, or determining a hyperplane capable of separating one class from all others. Most memorability studies that make use of support vector machines in fact use them for support vector regression (SVR). SVR operates in a very similar manner to a SVM, the only difference being that the hyperplane is used for regression rather than classification.

Naturally, the use of a hyperplane implies that the data is linearly separable; that is, it already exists in ‘clusters’ that between which, a straight line (in the two-dimensional case) can be drawn. In practice however, complex real-world data is unlikely to exist in this structured form. To solve this issue, SVMs employ the ‘kernel trick’ method [63]. This approach transforms the data (in whichever dimension it currently exists) into a representation in a higher dimension. This allows a non-linear lower-dimensional feature space to be restructured into a higher-dimensional linear feature space, which allows an optimal hyperplane to be found.

### 2.5.2 Feature Extraction

For support vector machines to function (and indeed, all machine-learning techniques) they need to be provided with some form of input. For SVMs, this is rarely the original data; it is far more efficient to find some relevant features that capture whichever characteristic is being predicted, extract them from the data, and use those instead. Memorability prediction models tend to fall into two camps; those that make use of more classical features, and those that make use of neural networks, either for final classification, or as an entirely contained system. Older work, before the general rise of deep neural networks tends towards classical features combined with support vector machines. Newer models either use neural networks to extract features and then classify based upon those features with an SVM, or are self-contained systems that use neural networks for both feature extraction and prediction. It is hence worth providing an overview of what these classical features actually *are*, and how they are produced, before we discuss computational memorability prediction in detail in Chapter 3.

Table 2.1 shows some of the most common types of classical features employed for memorability prediction. Though not an exhaustive list, these features in some combination show up in most classical models.

Graph-based visual saliency (GBVS) [56] is a computational model of human saliency - i.e, how likely is it that an image feature will capture our attention, developed by Harel *et al.* This model captures human behaviour with a receiver operating characteristic (ROC) of 98% against human ROC; making GBVS a highly accurate model of human attention. GBVS works by calculating the dissimilarity between a given region of the image with other neighbouring regions. These regions are used as nodes in a graph, where the transition cost between each node is based upon the similarity of each region. Areas of the image are then highlighted with an intensity based upon the amount of time a random walker would spend at a node before continuing. As crossing from a similar region to a disjoint similar region is less likely, these areas are walked less versus nodes that correspond to similar regions.

Histograms of oriented gradients (HOG) [34] calculates the direction of edges in an image, which in turn captures the shapes present within the image. HOG is invariant to geometric transforms aside from object orientation, and is generally used for object detection. The structural similarity index (SSIM) compares two images and produces

## 2.5. MACHINE LEARNING

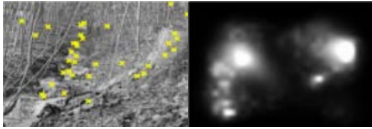
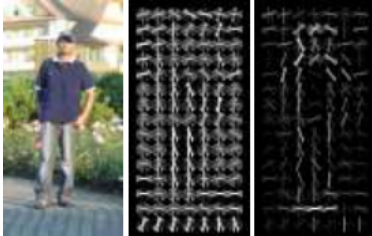
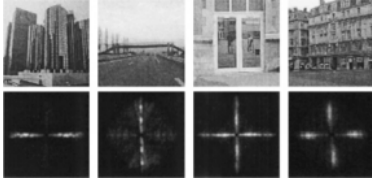
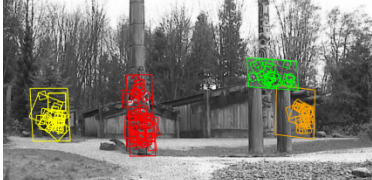
Acronym	Full Name	Graphic	Reference
GBVS	Graph Based Visual Saliency		Harel 2007[56]
HOG	Histogram of Oriented Gradients		Dalal 2005[34]
SSIM	Structural Similarity Index		Wang 2004[137]
GIST	N/A		Oliva 2001[103]
SIFT	Scale Invariant Feature Transform		Lowe 2004[92]

Table 2.1: The most common feature extractors shared among memorability prediction and analysis models.

a metric of how similar they are to one another. SSIM works upon various windows extracted from the two images, and compares the averages, variances, covariances, and dynamic ranges of three metrics; luminance, contrast, and structure (local intensity patterns).

The GIST model [103] provides a computational representation of the ‘spatial envelope’ of an image; where the spatial envelope is a low dimensional representation of that scene. Each dimension is based upon a perceptual feature that captures the spatial arrangement and textural makeup present in the scene. This model is loosely related to ‘scene gist’ a low-level, relatively un-detailed representation of a scene that can be extracted in less than 100ms. Rapidly determining the category of an image seen for very short amounts of time; or being able to describe a few objects present in the scene and their surrounding

## CHAPTER 2. BACKGROUND

context both involve the use of gist. The GIST model represents the scene as a set of spectra along the various dimensions of naturalness, openness, roughness, expansion, and ruggedness (pictured in Table 2.1). These spectra are generally similar for images in the same category, and diverge for images in different categories.

Scale invariant feature transform (SIFT) [92] is a feature detection algorithm that can be used to generate feature vectors that describes an image. SIFT functions by extracting many scale-invariant keypoints from an image. Each keypoint is tagged with an assumed orientation, which helps SIFT remain rotation invariant. The feature vector consisting of extracted keypoints can then be used to match objects in a separate image to the same object in the original image (as the keypoints will match, even if the object is rotated or present at different scale). For computing memorability scores, the generated SIFT keypoint vector is passed directly into a support vector machine without further processing, and simply serves as description of the image.

### 2.5.3 Neural Networks

Neural networks have a history that stretches back over fifty years, and to cover every variation and evolution of the basic concept would be impractical. Nonetheless, modern neural networks share several defining features, and in this section we will briefly review these shared functions upon which the majority of neural networks rely on to operate. A neural network is a set of connected artificial neurons. These neurons are generally structured into distinct layers, with each layer receiving the output of the preceding layer. As an input flows through a trained network from start to end, the input signal is transformed by the weights and biases of the artificial neurons into an output signal that represents some learned metric of the input, for example, the class of the input sample. The simplest possible ‘neural network’ is the perceptron; essentially a single artificial neuron [110]. This artificial neuron has a set of inputs  $x$ , a set of weights  $\mathbf{W}$  and a set of biases  $\mathbf{b}$ . The output,  $y$  is computed through the relation  $y = \sigma(\mathbf{W}x + \mathbf{b})$ .  $\sigma$  represents an *activation function*; a function that non-linearly transforms the output. This, much like an SVM, allows the perceptron to learn a hyperplane that separates the data. In the case of linearly separable data, perceptrons are guaranteed to converge. However, as with SVMs, most real-world data is not linearly separable, and in this case, perceptrons will never converge.

However, this relatively significant issue can be solved by stacking perceptrons into layers, and feeding the output of the previous layer forward to the next. The aptly named *Multi-Layer* Perceptron (MLP) represented the first step toward modern deep-learning, and allows the additional layers of the network to learn non-linear transformations of the data; much as the SVM kernel trick allows for non-separable low-dimensional data to be projected as points in a separable higher-dimensional space. Multi-Layer Perceptrons are powerful machines, and can be considered universal function approximators, indicating that the calculation performed via the weights of a trained MLP can represent a large assortment of functions. However, this does not imply that those weights can be learned; simply that it is possible that some assortment of weights can *exist* to approximate a given function.

In order for a neural network to be useful it must be *trained*. The weights and biases are initially set to random values; and are thus highly unlikely to solve whichever task the network is intended to solve. The process of updating these weights to values that allow the network to solve an arbitrary task is known as *training* the network, and requires two key components; a loss function, and an optimiser.

### Loss Functions

The loss function of the network does nothing more than compare the current output of the neural network to the desired output of the network, and return some metric that indicates the difference between these two values. The choice of loss function is relatively critical to the performance of the network; and there is no guarantee that the loss function that works well for one task can be readily applied to a different task. There is a wide variety of loss functions, far more than could be listed here. However, a few are both common enough, and robust enough to different problems that they are worth mentioning here. For classification problems, the categorical cross-entropy loss  $\lambda(y, \hat{y}) = - \sum_{n=1}^N \sum_{c=1}^C y_n^c \cdot \log(\hat{y}_n^c)$  has seen extensive use. For categorical cross-entropy, the label  $n$  with class  $c$ ,  $y_n^c$  is a binary value (0 or 1) that determines whether the label is a member of class  $c$ . The prediction  $\hat{y}_n^c$  is a probability between 0 and 1 that label  $n$  is a member of class  $c$ . The goal is to minimise the difference between the actual class label and the predicted class label. The log function penalises large errors more than small errors; which helps to prevent confident, yet incorrect predictions.

## CHAPTER 2. BACKGROUND

For regression problems there are no class labels, so cross-entropy loss cannot be used. The natural choice in this case is often Mean Squared Error (MSE), which penalises the squared difference between the output of the network and the actual labels. The MSE loss is defined as follows:  $\lambda = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ . The MSE loss can easily be extended to multiple dimensions, and such can be used to allow neural networks to find the error in two-dimensional data (for example, learning to generate saliency maps). There are many more possible loss functions, such as the variational loss found in autoencoder models of the same name, and the Wasserstein and hinge losses found in GANs, which will be discussed later.

### Optimisers

Optimisers are algorithms responsible for the actual learning of the network; they update the weights of each neuron based on the error provided by the loss function. However, the optimiser needs to know which direction the weights should be changed in to minimise the error. This is accomplished via the *backpropagation* algorithm. Backpropagation itself is not an optimiser; instead, it is an algorithm that can calculate the gradient of the loss function with respect to the current weights of the network. The optimiser uses these gradients to ‘step’ along the weight gradients. Exactly how the optimiser uses the provided gradients differs depending on the optimiser used; as with loss functions, there’s a wide variety of possible choices for optimisation algorithms, some of which may work better than others for certain problems. Generally however, they are all variations on **Stochastic Gradient Descent** (SGD), an algorithm which calculates the loss gradient for a random subset of the input data and updates the weights by an amount defined by the step size (or *learning rate*). This has the effect of traversing the multidimensional landscape defined by the loss function, with the goal of coming to rest in a ‘global minimum’ - the set of weights with the lowest possible loss.

SGD and has mostly been superseded by more modern algorithms designed to reduce the amount of iterations required by the network to converge. These include the RMSProp algorithm, based upon the AdaDelta algorithm, which maintains a different, adaptive learning rate for each parameter in the network. In the AdaDelta [145] algorithm, the learning rate is based on continually accumulating gradients, which eventually results in the learning rate shrinking to nothing. RMSProp solves this by keeping a decaying moving average of the calculated squared gradients; allowing it to focus on newer calculated

gradients and avoid the learning rate diminishing. The Adam optimiser [77] likewise maintains a decaying average of both previous gradients, and those gradients squared. These additional parameters serve to simulate momentum; allowing the optimiser to ‘skip’ over local increases in loss if the loss has been decreasing up to that point. Adam has seen successful use in many types of neural network architectures, from CNNs to generative models; though it does come with its own hyperparameters that occasionally need to be tuned to the problem at hand, complicating training.

#### 2.5.4 Convolutional Neural Networks

Convolutional neural networks (CNN) are among the most common types of neural networks seen today, and are likely responsible for the popularity of deep-learning. Before CNNs, machine learning problems involving images required carefully handcrafted features, or were constrained to very small images. With CNNs, it suddenly became possible to train a neural network to classify a wide variety of images with remarkable accuracy. This developed into CNNs becoming the primary choice for object recognition, image segmentation, and more. While usually known for their effectiveness at image classification, they also often appear in any complex model where learning features about spatially sensitive data is useful. In more recent studies examining memorability prediction in particular, CNNs tend to dominate due to their ability to extract useful features from image data.

A standard Multi-layer Perceptron network involves connections between each neuron of one layer and all the neurons of the preceding layer. While this allows the network to integrate global features (‘fully connected’ layers are still used for this purpose), applying this model to high-dimensional data such as images, results in a network with an enormous amount of connections. As all these connections need their weights updating, the model quickly becomes computationally bound. Additionally, while images tend to have spatial constraints on their structure, fully connected models have no concept of spatial locality - each input is treated independently even when some spatial relationship exists. E.g, an MLP can not learn to detect the "eye structure" in faces invariant of where in the image the eye actually appears.

CNNs solve this problem by connecting each neuron with only a small window of the input at a time, termed the ‘receptive field’ (See Fig 2.1) of the neuron. This window



## CHAPTER 2. BACKGROUND

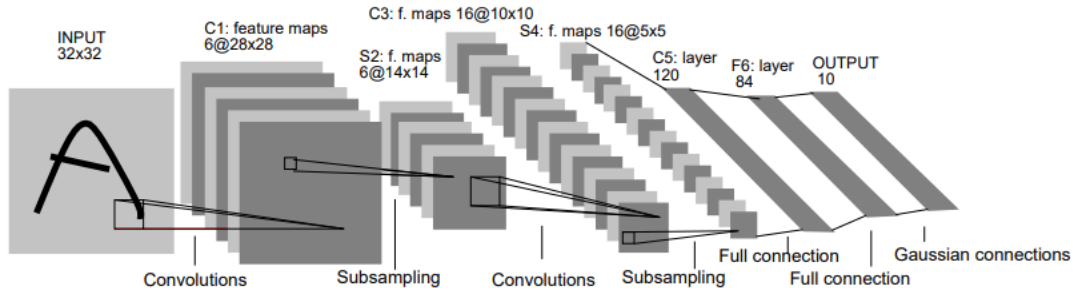


Figure 2.1: LeCunn et al's LeNet - one of the first convolutional neural network architectures. The feature maps mentioned are analogous to filters. Dimensionality reduction is accomplished through subsampling. [Fig. 2 in [87]]

'slides' over the image. This sparse connectivity means fewer weights are required to learn the patterns in the data, as the receptive field does not have to cover the whole image. This spatial constraint allows each neuron to learn to detect spatially local features. In the case of the above (simplified) example, an "eye detector" may arise in a neuron; that will activate whenever its receptive field encounters an eye in the input. This detector is structured as a set of 'filters' that activate in response to given patterns in the previous layers. As the input flows through the CNN network, these filters become more powerful. Filters in early layers may detect lines, edges, or corners, while later layers detect arrangements of these components - objects. Much in the same way one might employ convolution with an edge-detection kernel to find edges in an image, a convolutional neural network uses convolution with its filters to detect whichever feature the network filters have learned it's advantageous to detect. These convolution operations give the network its name.

To help keep the number of computations required under control, and to force the network to learn the most relevant features in the data, as the data is passed through the network it usually undergoes some form of dimensionality reduction. The features that survive the reduction process are considered to be the most important to solving whichever task the network is being trained to perform; and generally, the deeper the network, the more powerful, and more abstract, the representation of the input data becomes. There are many forms of dimensionality reduction, but usually this takes the form of some kind of pooling; often used is MaxPooling, which preserves the maximum

value in the pooled regions, or AvgPooling which preserves the average value of the pooled region.

As with multilayer perceptrons, convolutional neural networks generally have some form of activation after each layer. Historically, both the logistic function and hyperbolic tangent functions have been used, with the hyperbolic tangent ( $\tanh$ ) outperforming the older logistic function. Unfortunately, both these activations tend to *saturate* where large activations become locked to 1, and small activations locked to 0 (or -1). As the error backpropagated through the network depends upon the derivative of these functions, the lack of sensitivity leads to the *vanishing gradient* problem, where the gradient needed to update the weights and keep the network learning tends towards zero. Once this happens, training collapses. The Rectified Linear Unit (ReLU), defined as  $RELU(z) = \max(0, z)$  helps solve this issue by essentially being unbounded in the positive direction. This helps prevent the vanishing gradient problem, while remaining a non-linear function. However, if the output of a pre-activation neuron becomes negative despite the input, it will be clamped to zero, and stay that way; never learning (a ‘dead’ neuron). An incremental update to the ReLU activation, the Leaky ReLU helps prevent this by allowing small negative values to pass through the activation function; keeping the advantages of ReLU without the issue of neurons becoming unable to learn.

### 2.5.5 Recurrent Neural Networks

Recurrent neural networks (RNN) are a form of neural network that maintains a ‘memory’ of the data that it has seen previously. This allows it to process information based not only upon the current input, but also upon the previous input that it has seen at an earlier time. This kind of neural network works best with data that has some natural ordering, usually along a ‘time’ axis. This includes data such as text (recurrent networks frequently form the backbone of natural language processing architectures) and audio.

A basic RNN architecture contains a hidden state that propagates forward, but has no control over what this hidden state should contain. A more complex form of RNN, the Long Short-Term Memory (LSTM) network [62] introduces the concept of learnable ‘gates’ that allow the network to learn which information should be persisted, and which information can be safely forgotten. The ‘first’ gate in the LSTM learns what information to accept into the LSTM core state, the second two gates decide what information to

## CHAPTER 2. BACKGROUND

update that state with, and the final gate decides the output of the neuron. This allows the network greater flexibility in deciding how best to adapt to the data it is being shown. While standard RNNs are not widely used, LSTMs remain popular.

### 2.5.6 Additional Network Types

There are of course many more network types than just CNNs and RNNs, some of which we discuss later in this thesis, such as the Variational Autoencoder (VAEs) (Chapter 3.2) and Generative Adversarial Networks (GANs) (Chapter 4). While the building blocks of these networks are mostly the same basic methods discussed in the preceding chapters, these types of neural network are more specialised, and benefit from a contextual discussion rather than a general overview. In general, modern neural network architectures, whatever their application, tend to make use of either CNN or RNN components. If a neural network is involved in processing images in any form, it is a safe assumption that convolutional layers will be involved, and if the network needs to modify its output based on previous input, recurrent layers are likely to make an appearance. In the next section we will discuss transfer learning, and a few common neural network architectures that see widespread use.

### 2.5.7 Transfer Learning & Common Architectures

When considering computational image memorability prediction, and indeed many other domains, it rarely makes sense to develop and train a network completely from scratch. After all, neural networks are expensive and time consuming to train. A sensible alternative is re-tasking and extending an architecture that's already been developed, and is known to work well. If the pre-trained weights for that architecture are also available, this could easily save a vast amount of computation time; there is no point re-training a network for object detection for the purposes of memorability estimation, if there already exists a network that performs object detection for the purpose of image classification! All that is necessary is the network be re-tasked from classifying images to predicting memorability. This is, in essence, *transfer learning*, where a network originally trained for one task is re-trained for another.

As we have seen above, neural networks learn feature detectors of increasing power throughout their layers. These detectors are abstract, and in a sufficiently complex image classification network trained on a diverse dataset, could be responsible for detecting

objects, animals, or people. It is only in the last few layers of the network that these detectors are recruited for the purpose of determining image class (e.g, an airplane, a boat, a beach). There is no reason these features cannot be employed for an entirely different purpose; and there are several ways to accomplish this. In the simplest case, the weights of the network can be frozen (not updated during training), with only the last few layers left free to update. The network can then be trained on a different dataset, and the layers responsible for integrating the features into a prediction will learn to classify on the new dataset. In more complex cases, those layers can be entirely removed and replaced with a stack of layers more suitable for the new task. In doing so, as these new layers are the only layers that need to be trained, the network can be trained much faster, and with less training data, on the new task than on its original task. This has significant benefits when working in problem domains that benefit from basic functionality (e.g the ability to detect objects), but do not have sufficient training data to train object detectors from scratch in a deep neural network. We will now briefly examine two popular convolutional neural network architectures that have seen widespread use, and often form the "backbone" for applications that involve transfer learning.

## VGG

The VGG ("Visual Geometry Group") architecture [122] was influential enough that it is still commonly used today, despite being nearly seven years old at time of writing. The VGG network achieved state-of-the-art performance on the ILSVRC-2014, a large-scale classification challenge based on the ImageNet dataset [36]. The challenge involves classifying images into one of a thousand possible categories. The most common variant, VGG16, uses 16 layers with trainable weights, and this notation holds for all other VGG-type architectures. The convolutional layers use convolutions with a receptive field size of 3x3, and a stride of 1, which provides the ability to capture spatial directions in the input; but are relatively computationally cheap. The last three layers of the network are fully connected, with the final layer containing one thousand neurons which match to the one thousand classes of the ILSVRC challenge. All layers use the Rectified Linear Unit non-linearity. A trained VGG network contains deep features that lend themselves well to other applications; in the original paper the network was transfer-learned across several different datasets, showing good performance on each. Since then, any application that

## CHAPTER 2. BACKGROUND

could make use of deep features that describe the semantic content of images has often made use of the VGG network.

### **ResNet**

While the VGG architecture was highly successful, it was constrained by an upper limit on number of layers that the network could have before it became both too time-consuming, and too unstable to train. Indeed, adding layer upon layer eventually leads to the network eventually destabilising, with each layer added causing a decrease in final training accuracy. This problem was eventually solved by an architecture we know now as ‘ResNet’ - a network that contains residual connections [58]. These connections allow much deeper networks to be easily trained; a well known variant has 152 trainable layers. The residual connections themselves are implemented by way of shortcut connections which ‘skip’ a subset of layers, adding the original input back to the processed output. The hypothesis in the original paper suggests these skip connections, which act as an identity operation of the original input, allow the network to choose to either add deeper representation of the problem, or to continue with the original input if deeper representation causes a greater loss. Residual networks have set state-of-the-art performance on a wide variety of image classification/object detection datasets, and have become widespread. If a transfer-learning problem requires a more powerful representation than the VGG architecture can provide, often a residual backbone is used instead. However, even now, exactly why residual connections improve performance is not well-understood. Interestingly, as the number of layers in a residual network increase, the individual response of that layer decreases, suggesting there is some form of learnt normalisation across the whole network that prevents destabilisation. While there have been improvements to the architecture over the years since ResNets’ introduction, the basic principle, that residual connections improve performance, remains the same.

## **2.6 A General Overview of Complexity**

It is readily apparent that humans are capable of determining the complexity of a given image; shown a blank canvas and an abstract painting, it is easy to identify the more complex of the two. However, it is less clear how humans perceive the everyday complexity in which they are immersed; that of the natural scene. Like memorability, image complexity has also seen increasing focus on applying computational techniques in or-

## 2.6. A GENERAL OVERVIEW OF COMPLEXITY

der to model and understand how humans perceive the complexity of a given image. However, before modern machine-learning based approaches, psychologists have long attempted to understand which factors cause an image to be regarded as complex or non-complex [12]. While the majority of these efforts are not scene focused, they nonetheless reveal some clues on how the brain processes complexity in general. As discussed earlier, image memorability appears to have both a gist trace and a detail trace. While gist is relatively well understood [102, 104, 85], which features of the image contribute to the detail trace in memory has had less examination. Could image complexity serve as an analogue for the ‘detail’ level of the image, and in some manner interact with visual memory? This is especially interesting to consider in the context of natural scene images; perhaps the complexity of the scene affects how well that scene can be recalled. While the relationship of aesthetics, interestingness, and other intrinsic image properties with memorability is well understood, how image *complexity* relates to image memorability is much less clear.

The origins of complexity theory and its relation to other image properties can be traced back to the early 20th century, where G. D. Birkhoff defined an aesthetic measure [12] as a ratio between the order and the complexity of an image, where complexity relates to the count of elements and order relates to the count of regularities present. A few decades later complexity is examined in the context of aesthetics [35], finding that for very simple images (polygons) symmetry was a key determinant of rated complexity, while later still complexity was redefined either as the detail present in a line drawing of an image [123] or as the degree of difficulty in providing a verbal description of a texture in by Heaps & Handel [60]. Heaps & Handel find complexity to be correlated significantly with the structure, orientation, and repetitiveness of the texture. In this case, structure refers to how much organisation vs randomness exist in the lines and parts of the texture.

In 2004, Oliva *et al.* [105], in the first study to examine scenes explicitly, hypothesises that complexity perception is affected both by the variety of objects in the scene, and the variety of surface textures present. In an experiment to determine which factors affect perception of complexity, they find that complexity could be modelled along two main dimensions for interior scenes; that of mirror symmetry and that of clutter. A year on, in 2005, Rigau *et al.* [108] proposes an information-theoretic framework for modelling complexity, based upon Birkhoffs’ aesthetic measure, which partitions an image

## CHAPTER 2. BACKGROUND

into several homogeneous regions then calculates the mutual information between these regions. How this relates to human perceptual complexity was not explored, nor was it explored in later work [109], which uses Kolmogorov complexity [78] as a potential measure for image complexity. Kolmogorov complexity can most easily be understood as the length of the shortest program to compute a given output on a universal computer; i.e, the most compressed that output can be. Kolmogorov complexity is uncomputable, but can be estimated by compression algorithms [108]. Random structures are difficult to compress, so it has been suggested that the complexity of an image is related to the structures in an image that lie somewhere on an axis between trivial regularity (the 'order' of Birkhoff) and meaningless randomness; that is complete order and complete randomness are similarly lacking in complexity [39].

To investigate the overlap between machine methods and human perception, Cardaci *et al.* frame image complexity as a fuzzy process [21], and conduct a trial to evaluate whether their computational method matches reported complexity values from human observers for paintings. They extract local image features and build an entropy-based distance function to determine how far a given image is from the simplest image in the set. However, they define human perceptual complexity as related to the perceived time to observe an image; and while they find a relation between their fuzzy measure and perceived time, it is unclear what visual processes drive this. Yu *et al.* [143] instead examine spatial information measures (such as edge magnitude) and find they correlate with compression-based complexity measures [45]. However, these measures are often tested on line drawings, polygons, or icon images; all of which are a long way from a rich natural scene. Even paintings, while of interest for aesthetics perception, do not reveal much about how *scene* complexity is perceived. Finally, in [95] it was found that computational measures of complexity correlated with ratings of visual complexity, and ratings of visual complexity correlated with measures of affect, but computational complexity did not correlate with affect. That is, complex pictures tend to be rated more 'pleasant' and 'arousing' than non-complex images, yet existing computational techniques do not indicate this relationship exists. There appear to be minimal studies that directly relate image complexity to image memorability, though one study suggests that high complexity images may be more memorable than medium complexity images [123].

## 2.7 Summary

It is evident there is much yet to be understood about complexity perception; and even more to understand about scene complexity perception. Later in this work, in Chapter 5 we discuss modern approaches to analysing and predicting image complexity, including various methods that make use of deep learning. However, it is still unknown exactly how image complexity and image memorability relate; and for scene images, there is even less data. Later in the same chapter we explore the gathering of complexity data purely for scenes, and discuss how this relates to those scenes memorability. In the next chapter however, we will discuss how the memorability data for those scenes was obtained and what this reveals about image memorability in general via a new approach: the Visual Memory Schema.



## Visual Memory Schemas

While there have been several approaches to predicting the memorability of images through computational means, until now these approaches have been limited to a single ‘score’ that defines how memorable an image is; without explaining why the image has that score. In contrast, VMS Maps identify the regions in an image that cause a human to be able to remember that image. For the first time, there exists a dataset that contains two-dimensional human memorability data. However, due to the resources required to gather this sort of data, the original dataset is limited, consisting of only 800 images; making prediction of VMS maps difficult. In this chapter, we present the relevant background needed to understand computational image memorability prediction, and our efforts to expand available VMS datasets and VMS prediction methods. We explore the differences in two-dimensional memorability information across categories, and for the first time employ computational methods to quantify the semantic ‘units’ that make up a Visual Memory Schema; that is, we find which arrangement of elements and objects in a scene cause that scene to be remembered.

### 3.1 Background

Studies of human visual memory in psychology stretch back decades, but research employing computational methods to understand image memorability are relatively recent. With an increase in computational power and an advancement in image processing techniques, computational investigation into perceptual image properties became possible. For the first time, large-scale crowd-sourced image memorability datasets could be ac-

quired, and based on this data, early machine learning techniques could be employed to learn from this data; and predict the memorability of a given image. As classical techniques gave-way to deep learning, prediction accuracy has only increased; eventually reaching human-level performance. However, these methods (and the data on which they are trained) give a single score for the entire image; and do not reveal what it is about the image that *causes* it to be remembered. Later work introduces Visual Memory Schemas, two dimensional memorability maps. Training computational models to learn which parts of an image are memorable and which is not is significantly more complex than single-score regression; and there is still a way to go before reaching human level performance.

### 3.1.1 Computational Memorability

The first study to introduce the notion of large-scale computational memorability prediction is that of Isola *et al.* in 2011. Isola developed a ‘memory game’[69] in which workers on the Amazon Mechanical Turk platform were presented a series of images, displayed for one second, with a 1.4 second gap in between. The workers were asked to press the space bar when a repeated image was shown (a variant of the old/new recognition test). Each series was 120 images long, which constituted a ‘level’ in the game, and each participant could complete up to thirty levels. 665 participants played the game. Of the images shown to the participants, 2222 images were targets, and 8220 images were fillers, which were not repeated. Participants were not shown ahead of time which images were targets, and each target was repeated only once. Similarly, each filler was only shown once. Each image was scored by an average of 78 participants, with the mean memorability score lying around 67.5% (defined as the percentage of correct recognitions), with a false alarm rate of 10.7%. Isola also found that when humans are asked to predict if an image is likely to be memorable or not, the results were actually weakly negatively correlated with memorability[68], indicating that humans are actually very bad at determining whether an image is likely to be remembered or forgotten.

Critically, the Spearman’s rank over 25 randomly split memorability trials is 0.75, indicating a high degree of consistency between participants. Because of this high degree of consistency, Isola hypothesised that there is in fact some intrinsic component to memorability in images. If memorability had more to do with the participants viewing the image than the image itself, then the overall consistency of memorability scores would

## CHAPTER 3. VISUAL MEMORY SCHEMAS

be expected to be much lower (or perhaps nonexistent). Because of this consistency, indicative that it is more to do with the image than the viewers, the memorability score of an image should be able to be predicted by an algorithm. Isola developed a predictive algorithm based upon classical global image features that included pixel histograms, GIST, SIFT, HoG and SSIM (detailed in Chapter 2.5). Together with a Support Vector Regression (SVR) machine, the predicted scores had a rank correlation of 0.54 with the ground-truth human scores. Object statistics were also examined, finding that simple statistics such as mean class coverage or the count of an object are not predictive of memorability, though scene category did appear to summarise much of what made an image memorable.

The first, and until much later, only investigation into memorability beyond a single score was Khosla *et al.*'s [73] work into predicting memorable regions of a given image. By examining selected regions, and predicting how likely said region is to be forgotten or hallucinated, then pooling these feature maps into one overall map, the general memorability of the image can be predicted. Khosla *et al.* use a probabilistic model to simulate a 'noisy memory process', and hypothesise that the likelihood of an image being remembered is the distance between the actual image and a noisy degraded internal representation. In the model, they define this distance as the inverse of that image's memorability score. Multiple descriptors are used for each feature region: gradient, color, texture, saliency, shape, semantic. They achieve a Spearman's rank correlation of 0.5 between ground truth and predicted, though they do not predict a true memorability score, only a ranking between images. Different regions have different memorability scores, but how accurate these are for the region itself cannot be calculated, as no ground truth region-memorability dataset existed at this time, and Khosla did not create one. This means the accuracy of these 'memorability maps' cannot be verified against human data.

Khosla later introduces the LaMem database [75], a dataset of 60000 images and a memorability score for each image. This database is used to train a CNN to predict memorability, reaching a rank correlation of 0.64. Rather than train the model from scratch, Khosla used a pre-trained model and retrained it over the LaMem dataset, terming the resulting model 'MemNet'. The model was first pre-trained on both the Places dataset (made up of over seven million labelled scenes) and the Imagenet dataset, and used an AlexNet [3] backbone (which consists of eight layers, five convolutional

### 3.1. BACKGROUND

and three pooling). Interestingly, the network finds that faces and bodies correlate strongly with memorability, while regions of natural scenes seem not to. This is the first major application of a large-scale convolutional neural network to the problem of memorability prediction, and achieves a rank correlation outperforming every hand-picked method presented previously. Following this success, most image memorability prediction methods involved neural networks to some degree; even if just as feature extractors.

Dubey *et al.* [40] used a CNN pretrained on the Imagenet dataset for feature extraction purposes. The features were then passed into an SVR machine for memorability score prediction. Imagenet [36] is a vast database of images that is commonly used for pre-training of neural networks, under the hypothesis that the same learned features that work well for *classifying* images will also work well for predicting other image-related characteristics. Dubey’s CNN model achieved a correlation of 0.7 with ground truth human scores, though this score may be artificially high due to the small amount of images used (850). Dubey also combined this model with a semantic segmentation technique that could extract individual objects from images, using it to predict the memorability of these objects. This network had a much lower correlation of 0.39. However, as the CNN alone performed well, the error in the prediction is more likely to do with errors in the segmentation technique (a notoriously difficult problem) than with the prediction network.

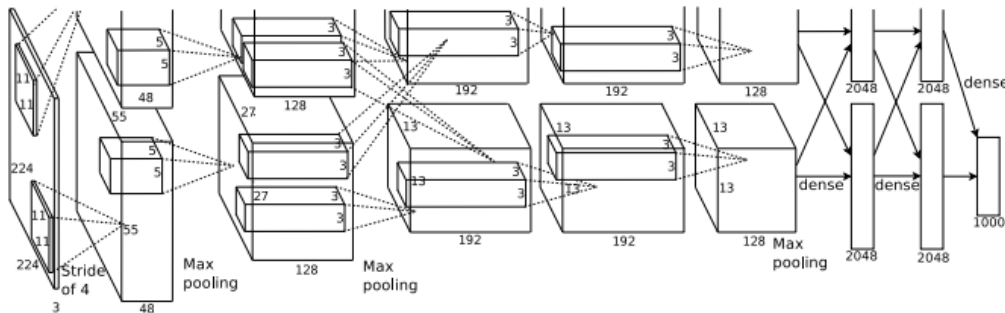


Figure 3.1: The AlexNet Architecture [3]. The second GPU processing stream is truncated, but follows the same architecture as the first.

Lukavsky [94] explores the effect of an image being different from its neighbours, as well as the effect of being in a different category to its neighbours. They accomplish

### CHAPTER 3. VISUAL MEMORY SCHEMAS

this with a convolutional neural network known as ‘Places-CNN’, which is built upon the AlexNet architecture and trained on the Places dataset [148] to classify images into scene categories (compared to Imagenet, which tends to be object categories). First, feature vectors are extracted by truncating the network prior to the final classification step. Secondly, the L2 norm is used to compute the difference between extracted features from two images. A smaller L2 norm means the semantic content of the image is closer to each other, and thus the images are "in-context" with each other. This method was used to computationally determine which images are out of context with others in the dataset, and hence to explore the effects of context on memorability.

Yoon *et al.* [141] later examines the effect of spatial relationships upon memorability by using both a neural network to segment objects from the image, and a neural network to predict the memorability of those segmented areas. For memorability prediction, Yoon used Khosla’s MemNet [75] to extract memorability-relevant features, and then employed DilatedNet [149] to extract a segmented map of the scene. These features were combined via a support vector regression machine, which was trained to infer the memorability score. This model achieved a correlation of 0.66 with the ground truth, close to the LaMem human split correlation of 0.68, and significantly more accurate than Dubey’s earlier model.

Most early works were not "end-to-end" neural networks; rather than having a single neural network that computes memorability scores, they instead used neural networks for feature extraction, but passed these features into other algorithms to compute memorability scores. In contrast, Squalli-Houssani *et al.* [11] develop an end-to-end neural network that incorporates both CNN features and features from an image-captioning system, intending to capture the powerful memorability-descriptive effects of semantics. Squalli-Houssani accomplishes this by making use of an LSTM-CNN combination. The CNN extract relevant images features, and the LSTM network uses these features to infer likely combinations of words that describe the image - essentially, captions that describe the semantic content of the image. These features are merged with standard CNN features extracted using the VGG16 architecture. By combining these features together, they achieve a final Spearman’s correlation of 0.72 over the LaMem dataset. However, to accomplish this images are divided into four distinct classes, from low to high memorability, based upon their LaMem scores, turning memorability prediction into a classification, rather than regression problem. It is generally easier to compute

### 3.1. BACKGROUND

aggregated classes rather than a direct regression score, which explains the unusually high performance.

Fajtl *et al.* also investigates the pairing of recurrent neural networks with convolutional neural networks for the purposes of memorability prediction and develops a neural network[44] that uses an iteratively generated attention based metric. This network uses an LSTM to generate attention maps (three iterations provide the best results). It should be noted that these attention maps are unrelated to what is typically known as ‘attention’ in the cognitive science world; they are not models of saliency or eye fixation. They instead determine which parts of the image the neural network should process. To create these maps, a method known as ‘soft attention’ is used, which assigns a probability weight to every informational element in the feature maps of the network, rather than only ‘attending’ to elements that are greater than some arbitrary boundary. These probabilities determine how much the neural network weight the network should assign to a given element, which in turn weights the future generation of attention maps. Fajtl achieved a 0.677 Spearman’s rank, very close to the 0.68 of human consistency.

The base of the network is a pretrained residual network. Features are extracted from the penultimate layers of the network, and fed into the attention predictor and LSTM network, whose iterative attention maps are then summed together. The LSTM hidden state is mapped to the normalised memorability score of the input image to regularise the final output. The loss function:

$$L = (\hat{y} - y)^2 + \lambda L_\alpha, \quad (3.1)$$

is standard root mean squared error combined with a penalty  $\lambda L_\alpha$  that encourages the model to explore all image regions over the LSTM iterations, and prevent the algorithm becoming ‘stuck’ in one spatial region. This penalty is a function of all activations over the attention maps. While this approach certainly provides good overall results, the improvement is relatively minor compared to the same network with attention features disabled (0.663), indicating that learned deep features in a sufficiently complex network remain the best predictor for memorability, and that likely the most critical elements for memorability prediction is 1.) a sufficiently deep network and 2.) a sufficiently large dataset.

**What can computational approaches tell us about memorability?**

Image saliency defines regions of an image that draw an observers attention. Do high levels of saliency influence how memorable an image is? Mancas found that eye fixation durations were longer the more memorable the image [96], and that congruency between fixations is also higher for more memorable images vs less memorable images. This is indicative of some link between attention and memorability. Mancas built upon this by constructing a classifier that made use of attention-based features (saliency), combined with Isola’s original image attributes, and found they improve image memorability prediction. Even when 1512 of Isola’s feature dimensions were replaced with 17 attention based dimensions, overall Spearman’s rank consistency between predicted and ground truth was still higher than Isola’s model alone (0.479). Celikkale [23] later explored combinations of semantic features (the scene category label), object features (annotations on the image describing it’s object content) and dense visual features, such as colour histograms, GIST, HOG, and SSIM (described in more detail in Section 2.5.2, with a method that pools together salient regions in the image. Object level saliency and bottom-up saliency maps are obtained and used for attention-based pooling of image regions, that generates the final feature vector, and allowed Celikkale to achieve a Spearman’s rank correlation of 0.52 with the ground truth [23]. While it is clear that saliency has a relation to memorability, it certainly does not fully explain it, as techniques that predict using saliency remain far from the human level consistency.

In general, exactly how image content relates to the ‘memorability’ of that image is still not well understood, and multiple feature dimensions (extracted from the image) are required to provide a reasonable explanation of single-score image memorability [113]. The best descriptors of image memorability appear to be high level scene semantics, those that deal with emotion, scene dynamics, actions, and demographics. Isola [67] found that these attributes alone outperform all other tested feature extractors, with a rank correlation of 0.51 with ground-truth scores. Combining these attributes with other semantic predictors such as objects present and scene category boosts the final performance of the predictor to 0.54, which isn’t surpassed until neural network based methods are developed. Taking a more fine-grained approach, Dubey *et al.* [40] found that individual objects present in an image have varying degrees of memorability. Ground truth memorability values for objects in images were obtained in similar fashion to Isola *et al.* Participants played a ‘memory game’, though in this case shown images were masked,

### 3.1. BACKGROUND

leaving only individual objects available for encoding. Object memorability has a high consistency (Spearman's correlation of 0.76), which suggests that individual objects have a level of memorability intrinsic to their structure.

Natural, outdoor scene images are some of the most difficult images to predict memorability for, with machine techniques falling short of efforts to predict memorability for both indoor scenes and object-focused images. Lu *et al.* [93] find that certain Hue-Saturation-Value (HSV) values correlate with human memorability, and develop a dataset that contains only natural scene images (such as forests and deserts). The HSV feature contributes to memorability prediction of the natural scene images to a larger degree than low-level predictors and the model overall outperforms Isola *et al.*'s model, with the HSV feature resulting in an increase of 7.3% to the Spearman's rank correlation between ground truth and predicted values. However, Lu's dataset is very small, consisting of only 258 images, and colour has been shown previously to be only weakly predictive of memorability. The effect seen here is likely a result of the small dataset combined with the difficulty of the task, but does indicate that in the absence of rich descriptive features colour does play a small part in memorability.

Bylinskii *et al.* [20] find that images distinct to their context are remembered. For example, in a dataset of deserts, a forest image may be better remembered as it stands out against the context in which it has been presented. Bylinskii also finds that certain image categories are more memorable than others, and that memorability rankings of scene categories have a Spearman's rank of 0.68 over 25 half splits. Similarly, Isola found that scene category alone had a correlation of 0.37 with ground truth memorability. One possible explanation for variation in category memorability is that scene categories with a greater amount of contextually distinct images appear to be more memorable; it is potentially variety throughout the category that improves the memorability of that category. An alternative explanation for variances in memorability could be due to the perceived depth and motion of that image. Basavaraju *et al.* compute the depth and motion of an image with optical flow and depth estimation methods. A set of convolutional models were trained to predict memorability based on depth, motion, or both. Neither the model based on depth or the model based on motion outperformed the original MemNet CNN introduced by Khosla *et al.* However, when these features are combined, this model slightly outperforms MemNet (0.64 vs 0.655) [9]. The issue here may arise from lack of an accurate baseline. Both motion and depth of the images were



## CHAPTER 3. VISUAL MEMORY SCHEMAS

found computationally, not drawn from the (unavailable) ground truth. If the calculation of these factors is not accurate, it would be difficult to draw any conclusions about memorability. Nonetheless, it appears even with potentially inaccurate depth/motion estimations there is enough additional information to improve memorability prediction performance.

### 3.1.2 Visual Memory Schemas

In cognitive science, a schema is a mental construct that facilitates the encoding of a scene. For example, the average person may maintain a ‘kitchen’ schema that consists of arrangements of common elements typically found in a kitchen. Viewed scenes that better match this schema are therefore better encoded and retrieved. Visual Memory Schemas represent a way of operationalising this idea of a ‘schema’ and extracting which scene elements directly correspond to the mental structures that enable remembering of the scene. Visual Memory Schemas were introduced recently in the work of Akagunduz *et al.*[2] via the VISHEMA Experiment, culminating in the creation and analysis of an 800 image scene dataset paired with 800 ‘Visual Memory Schema’ (VMS) maps. These VMS Maps capture the regions in the scene images that cause a person to remember, or falsely remember, that scene. In turn, these regions are thought to contain elements that match the cognitive schema for that scene. The images and VMS maps have a resolution of 700 pixels by 700 pixels and are full colour. For this dataset, images widely regarded in the literature as being ‘highly memorable’ are excluded, by purposefully removing images with recognisable landmarks, attention-drawing text, and people looking directly at the camera. This results in a more stable dataset, as the memorability data for each scene is more likely to be effected by scene semantic content rather than known memorable features. In the VISHEMA experiment, participants ( $n = 90$ ) are asked to memorise 400 images drawn from the dataset, and then tested on another set of 400 images (of which 200 are repeats and the other 200 are fillers) to determine how well those images are remembered. Participants are asked to select on a scale between 0 and 100 how confident they are that they have seen that particular image before. Over a certain threshold (30) participants are asked to draw boxes on the image over the regions of the image that they believe has caused them to remember that image. These maps are highly consistent, with a Pearsons 2D correlation of 0.7 - participants agree on the areas that caused the image to be recognised. VMS Maps hence are two dimensional probability

### 3.1. BACKGROUND

distribution maps that indicate how likely a region in an image is to cause that image to be remembered or falsely remembered (Fig. 3.2). These maps represent cognitive elements, shared among the participants that took part in the experiment ( $n = 90$ ) that influence the memorability of an image. Because of this spatial element, Visual Memory Schemas allow analysis of which regions in an image causes a human to remember, or falsely remember, that image.



Figure 3.2: An example from the VISHEMA dataset produced in [2]

True VMS Maps, which indicate the areas that cause an image to be correctly remembered have a high level of consistency between randomised equal splits of the participants, while False VMS maps, which indicate areas that cause an image to be falsely remembered, have a lower level of consistency. From these data it appears that while it is relatively easy to agree on what is memorable, regions that cause humans to believe

## CHAPTER 3. VISUAL MEMORY SCHEMAS

they have seen an image, when in fact they have not, are more subjective. A VMS Map of an image has a Spearman's correlation with the computed saliency (GBVS) of that image (0.581), much lower than the correlation between participant VMS Maps (0.7), indicative that saliency cannot fully explain what makes an image memorable. Additionally, VMS Maps were compared with eye fixation data gathered at the same time the experiment was conducted, and no significant correlation between eye fixations and VMS maps was found. This clearly indicates VMS Maps clearly capture information about memorability beyond that of simple attention-based metrics. True VMS maps are more consistent than false VMS maps across observers, which is hypothesised to be because the encoding of more easily remembered images relies upon more established mental schemas, whereas falsely remembered images are more due to reliance on individual experience.

Akagunduz *et al.* [2] employ transfer learning and five different neural network architectures in order to determine the best combination for predicting combined VMS maps. The five different pretrained networks were MemNet, and four VGG variations: VGG-S, VGG-M, VGG16, and VGG19. The original classification layers of the networks are removed, and new layers attached consisting of 3 256 neuron hidden layers and a 400 neuron output layer. Twenty-one variations on these architectures are tested, dependent upon which final layer of the pretrained networks the new output network is appended to. The final output of each network is a  $20 \times 20$  pixel combined VMS map. Each network is trained for each possible 80/20 split of the training data, and considering two possible loss functions (The L1 and L2 norms). This results in 210 total different experiments. They find that deeper layers in the neural architecture perform better at reconstructing VMS maps, though interestingly the deepest layer in the network perform more poorly compared to previous layers. This is hypothesised as being caused by the deepest layer being fine tuned for image classification rather than VMS map reconstruction. The best performing network is VGG19, and the best reconstructed category is 'work-home' with a Pearson's 2D correlation of 0.677 with ground truth data.

### 3.2 Methodology

Predicting the memorability score for an image representing how likely a given image is to be remembered by a human during a recognition test, is a difficult task - memorability

## 3.2. METHODOLOGY

has been shown to be associated with the semantic content of the image, a complex dimension to extract. With the advent of large memorability datasets that contain tens of thousands of images paired with ground truth memorability scores, recent deep learning models have succeeded in achieving close-to-human performance in predicting how likely an image is to be remembered. Previous work in the arena of memorability prediction has been engineered with the goal of predicting memorability scores for a given image. Few research studies explored creating models capable of predicting the regions of an image that contribute the most to an image’s memorability. These models’ predictions of memorable regions lack a clear relation to the ground truth, as until very recently no dataset of the regions that cause *humans* to find a given image memorable, existed. In this section, we present the methodology of several approaches to predicting visual memory schemas. We start with a variational autoencoder based approach, trained on the original VISHEMA dataset of 800 images. We then explore additional computational techniques which may aid in the prediction of visual memory schemas, and finally we develop a novel architecture that incorporates these techniques, and makes use of existing single-score memorability datasets. Results for all proposed approaches can be found in Section 3.3.

### 3.2.1 Predicting Visual Memory Schemas with Variational Autoencoders

Autoencoders (AE) attempt to learn efficient latent-space encodings of the input data that would allow its reconstruction from such an encoding. A variational autoencoder (VAE) [76] is an extension of the AE, which has the training aim to maximise the lower bound of the marginal log-likelihood of the data following encoding and reconstruction. This means minimising the KL divergence between the posterior and *a priori* data distributions during the training. Rather than just learning a compressed encoding of the data, a VAE learns a probability distribution that is an approximation of the true probability distribution of the underlying data. This allows a VAE to be used as a generative model based on sampling in the latent space.

VAEs are made up of two components - an *encoder* which converts input data  $x$  into a latent space representation  $z$ , and a *decoder* that converts a latent space variable  $z$  back into data  $x'$  akin to the input  $x$ . Convolutional neural networks (CNNs) are used for implementing both the encoder and the decoder. The encoder is defined as

## CHAPTER 3. VISUAL MEMORY SCHEMAS

a probabilistic machine  $q_\theta(z|x)$  that extracts a specific latent space code  $z$  where  $\theta$  represents the parameters of the encoder's network. Meanwhile, the decoder maps the information in a probabilistic sense defined by  $p_\phi(x|z)$  in the opposite way from the code  $z$  back to the data space  $x$ , where  $\phi$  defined the parameters of the decoder network. The encoder and decoder are related through the loss function which consists of two components:

$$L(\theta, \phi) = -E_{z \sim q_\theta(z|x)}[\log p_\phi(x|z)] + KL(q_\theta(z|x)||p(z)) \quad (3.2)$$

where  $KL(\cdot)$  represents the Kullback-Liebler divergence between the *a priori* distribution of the latent space  $q_\theta(z|x_i)$  and its estimated distribution  $p(z)$ . The first term from equation (3.2) represents the reconstruction loss and the second term regularises the learnt distribution. The latter term helps the VAE to learn to group conceptually similar data in the same regions of the latent space.

Here, we are aiming to develop a generative method for Visual Memory Schemas (VMS), for a given input image (specifically, those of scenes). In our approach we aim to generate both true and false VMSs, simultaneously. This is defined as an image-to-image translation problem by making use of an VAE consisting of two CNNs, with the first one, the encoder designed to learn a mapping from an image to a latent code, while the decoder to learn the mapping from that latent code to a VMS. Previous work [94, 46] has shown that CNNs work well at extracting high-level image features that also allow for the prediction of memorability [11]. CNNs such as VGG-16 network have also been shown to be capable of learning to reconstruct VMS maps at some degree for certain image categories [2]. We propose using VAE models which have good ability to learn data classification in the latent space, as exemplified in Fig. 3.3. This model would allow a good separation of the false and positive VMS encoding spaces and then for the generation of dual channel VMS maps for generic scene input images corresponding to true and false VMS structures in which given random memorable images produce latent codes similar to those indicated experimentally by humans in memorable images. Moreover, the learned latent space modelled by VAEs can be easily inspected in order to find relations between the memorability and false memorability of images; and to determine whether extracted deep neural network features are separable into those that define high-memorability VMS maps, and those that define low memorability VMS maps.

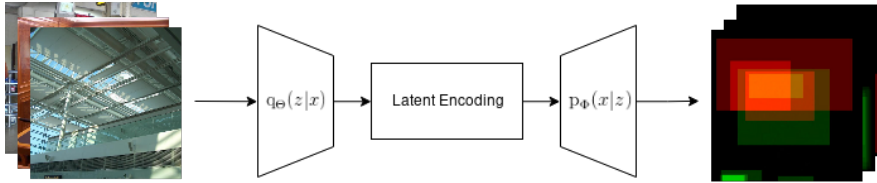


Figure 3.3: Predicting VAEs in images using an autoencoder.

For the training we use a pre-trained VGG network architecture [122] for the encoder after truncating the network before the classification step and using only the convolutional layers. The final output of the VGG architecture will be connected to a dense layer in order to compress the representation further, followed by the latent encoding. In CNNs the deep features that would emerge capture structures of the objects in the scene [147] and semantic structures [50] present in the input image.

The decoder benefits from being able to be simpler than the encoder. Whereas the input of the encoder consists of real world scenes, the output of the decoder is a VMS map, which consists of only two channels representing the spatial density of how likely a given image region is to cause that image to be remembered. There is no benefit in using a very deep architecture for the decoder, as we do not need to recreate any meaningful semantic features in the output. Additionally, a simpler architecture keeps the number of trainable parameters low, which is important when considering the low amount of available training data.

The loss function for this model is similar to the standard VAE loss function from (3.2), with the exception that in the reconstruction term, instead of reconstructing the *original* image data, aims to reconstruct associated information, such as VMSs. If  $X$  is the set of scene images and  $Y$  the set of associated VMS maps, with  $x \in X$  and  $y \in Y$  representing corresponding images drawn from these sets, our loss function is:

$$L = -E_{z \sim q_\theta(z|x)}[\log p_\Psi(y|z)] + KL(q_\theta(z|x)||p(z)) \quad (3.3)$$

where  $\Psi$  represents the parameters associated with the VAE reconstructing the VMSs data  $y$  at the end of the encoder. We additionally investigate replacing the reconstruction term with the l1-norm as in [2].

### 3.2.2 Exploring Visual Memory Schema Prediction with Multi-Scale Information, Depth, and Self-Attention

We have shown that an artificial learning model, such as a Variational Autoencoder (VAE) [76], can predict VMS maps for scene images (see results in Section 3.3). However, the family of models capable of specifically indicating regions from images which are responsible for their memorisation, have not been studied in depth compared to their single-score counterparts. Here we propose multiple different approaches to memorability map prediction, examining the effects of multi-scale information, non-local self-attention, the inclusion of depth information, and various combinations of these factors. We also draw on evaluation metrics from visual saliency prediction in order to set a new, comprehensive baseline for VMS map prediction.

To accomplish this, we developed a series of models capable of predicting visual memory schemas for scenes, testing the influence of depth, self-attention, and multi-scale information. We examine both the impact of latent-space dimension on our variational architectures, as well as develop standard deconvolutional models, and for each network where feasible we test the effect of introducing self-attention and depth information. Our goal is to discover both which techniques are applicable to VMS prediction, and to set a variety of comprehensive baselines for future work.

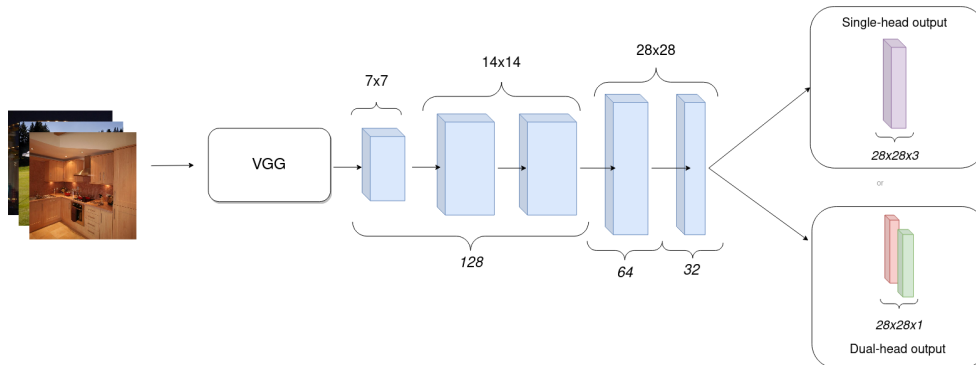


Figure 3.4: End-to-end deconvolutional network showing single and dual headed outputs. The height and width of the convolution filters is given above, while the channels are given below the diagram. The dimensions of the output is given below each output.

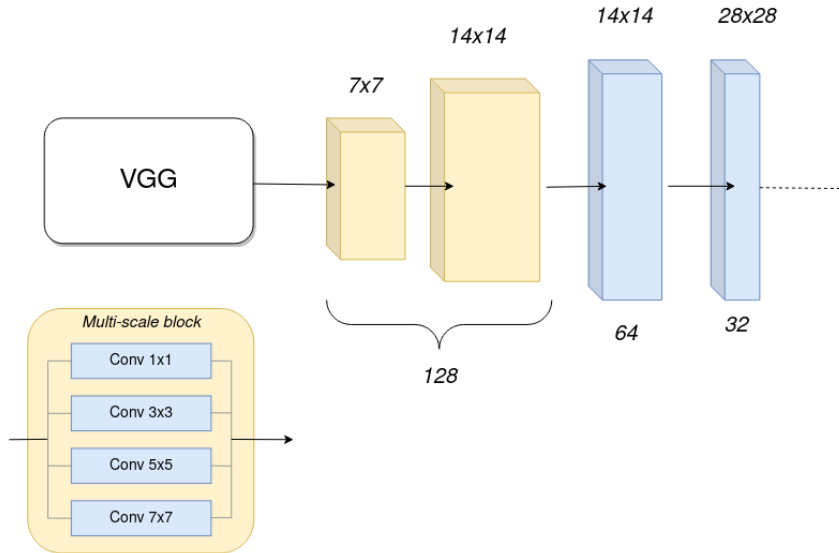


Figure 3.5: Multi-scale VMS predictor with multi-scale blocks (MSB) from [64].

### Deep learning architectures

We choose three architectures as potential baselines against which to evaluate further developments to our VMS predictor models. First, we choose a straightforward end-to-end deconvolutional (CNN-deconv) architecture similar to that used in [2]. A pretrained VGG16 network feeds features into five convolutional blocks, with upscaling at specified intervals, as in the architecture shown in Figure 3.4. The output of the network is represented by one (single-headed) or two (dual-headed) memorability maps. The former generates a two-channel memorability map, while the latter generates both memorable and falsely memorable maps as distinct outputs. All convolutional blocks use a filter size of  $3 \times 3$  aside from the final outputs, which are  $1 \times 1$ .

Structures that influence image memorability arise at various scales in the image. Given the recent success of multi-scale information in finding conditional image correspondences for image-retrieval [64], we employ a similar methodology for enabling a deep learning architecture with multiple scale analysis and assess its efficiency for visual memory schema prediction. This multi-scale architecture replaces the three starting convolutional blocks with two multi-scale blocks (MSB) in the architecture from Figure 3.5. Finally, given the capabilities of image generation by VAEs [76], we also consider our





(shown to be effective for saliency map prediction [82]), and the ELBO loss (3.4). Additionally, we expand the research undertaken in [83] by varying the size of the latent space as  $|\mathbf{z}| = \{8, 32, 64, 128\}$ , where  $|\cdot|$  denotes the cardinality.

### Studying the influence of depth in the scene

Previous research indicated the importance of depth in the scene for influencing the memorability *score* prediction performance, according to Basavaraju *et al.*, [8]. However, whether this effect holds for visual memory schemas has not been explored. In the experiments undertaken in this study we generate depth maps for our dataset using MiDaS [86], a state of the art monocular depth estimation model. We concatenate features learnt from depth images with the features from the original image with the same dimension as shown in Figure 3.6 .

### Introducing self-attention mechanisms

Cognitive structures that lend themselves to remembering are rarely single objects in an image. Frequently, memorable regions are scattered throughout an image, or indicate an arrangement of objects (such as for example that of a table surrounded by chairs in an indoor scene) rather than a single object (a glass of water). A structural or semantical representation of the scene can indicate additional memorisation clues [141]. Non-local blocks [136] are designed to capture long-range dependencies by allowing the network to determine which features should be attended to, across the entire input. In the following we integrate the ‘Embedded Gaussian’ variant from [136] in order to determine whether long-range modelling aids VMS map prediction.

Given the embedding spaces  $W_\phi \mathbf{x}_i$  for the given input  $\mathbf{x}$ , and the learnable weighting hyperparameter  $\lambda$  and the re-introduction of original feature maps given in [146], the self-attention output is given by:

$$\mathbf{y} = \lambda \operatorname{softmax}(\mathbf{x}^T W_\theta^T W_\phi \mathbf{x}) g(\mathbf{x}) + \mathbf{x}, \quad (3.6)$$

where  $g(\mathbf{x})$  is a linear function of the input.

We combine the non-local blocks with our memorability predictors in the following manners: in multi-scale architectures, after the first multi-scale block, after the second

## CHAPTER 3. VISUAL MEMORY SCHEMAS

multiscale block + convolutional layer, and prior to the output (including in the architectures where we also consider the depth). In our variational architectures, we include the non-local layer in the decoder, two layers before the output.

### 3.2.3 A Dual-Feedback Approach to Visual Memory Schema Prediction

While we have expanded available visual memory schema datasets from just 800 images to over 4000 (details in Section 3.3.1), compared to single score datasets, this is still a relatively small amount of data. The LaMem dataset [75] contains 60,000 images paired with single-score memorability data. Although these images are not scene-focused (and may consist of objects, faces, or even animals), it would be advantageous if this data could be taken advantage of from the perspective of *two-dimensional* memorability. To that end, we design a new architecture that can be trained both on visual memory schema and scene data, while also containing an auxiliary loss we can train on the LaMem dataset, in the hope that the network can learn additional memorable features. These features can then be re-used for identifying which regions of a scene cause that scene to be remembered (or falsely remembered). In this section we describe the architecture and loss function for a Dual-Feedback VMS Prediction Network.

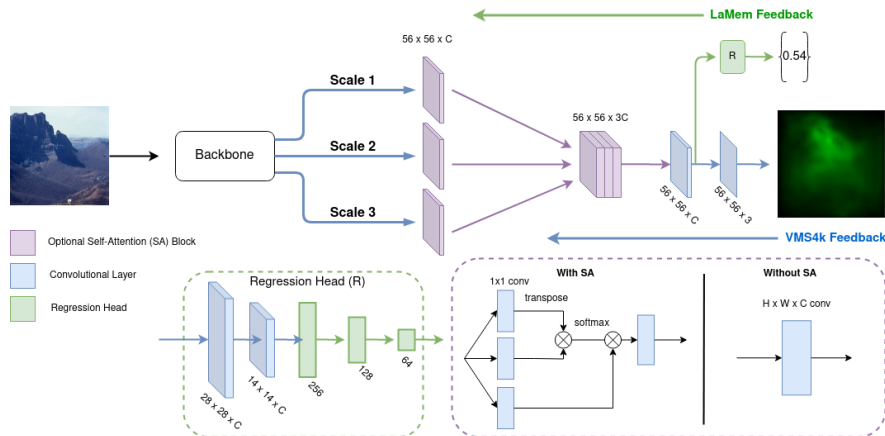


Figure 3.7: Architecture of proposed Visual Memory Schema predictor with Dual Memorability Feedback. Colors refer to layer types and are given in the legend.

Drawing on our results from Section 3.3.3, our architecture for visual memory schema

## 3.2. METHODOLOGY

prediction employs transfer learning, self-attention, and multi-scale information. To take advantage of existing memorability datasets, we additionally employ a dual feedback mechanism and condition the network to predict both memorability maps and memorability scores for input images. The architecture for the network is shown in Fig. 3.7. The network first extracts features from multiple scales in the backbone architecture, optionally computes attention maps for these features, and finally combines these multi-scale attention maps to predict the output map.

**Multi-scale Feature Extraction** We consider two backbone architectures: VGG16 [122] and RESNET50 [58], and employ these to extract semantic features from the input images. As memorability information occurs at multiple different scales throughout an image, we extract the semantic features at three different scales corresponding to processing blocks in the backbone architecture. Given an input image  $I_n \in \mathbb{R}^{224 \times 224 \times 3}$ , for each backbone we extract feature maps at  $S_1 \in \mathbb{R}^{56 \times 56 \times 256}$ ,  $S_2 \in \mathbb{R}^{28 \times 28 \times 512}$ , and  $S_3 \in \mathbb{R}^{14 \times 14 \times 512}$ , where  $S_1$ ,  $S_2$ , and  $S_3$  we call Scale 1, Scale 2, and Scale 3 respectively. All scale images are passed through a  $1 \times 1$  convolution for dimensionality reduction resulting in  $S_1, S_2, S_3 \in \mathbb{R}^{C \times H_s \times W_s}$  where  $C$  is hyperparameter defining the number of desired feature maps for each scale, and  $H_s$  and  $W_s$  define the height and width of the feature map at that scale.

**Optional Self Attention** Self attention has shown promise in single-score memorability predictors [44]. We examine whether self-attention offers any benefit for memorability map prediction. Given the embedding spaces  $W_\phi \mathbf{x}_i$  for the given input  $\mathbf{x} \in S_1, S_2, S_3$  [136], and the learnable weighting hyperparameter  $\lambda$  and the re-introduction of original feature maps given in [146], the self-attention output is given by:

$$\mathbf{y} = \lambda \text{ softmax } (\mathbf{x}^T W_\theta^T W_\phi \mathbf{x}) g(\mathbf{x}) + \mathbf{x}, \quad (3.7)$$

where  $g(\mathbf{x})$  is a linear function of the input. We compute self-attention maps for each scale. Each embedding space is parameterised by a  $1 \times 1$  convolution. If self-attention is disabled, each block is replaced by a  $3 \times 3$  convolution with  $C$  channels.

**Feature Concatenation & Dual Feedback** Whether self-attention is enabled or not, the multiscale feature maps are combined via channel-wise concatenation, giving a singular weight matrix representing memorable features at the three scales. With  $S_1, S_2, S_3 \in \mathbb{R}^{C \times 56 \times 56}$ ,  $S_m = [S_1, S_2, S_3]$ ,  $S_m \in \mathbb{R}^{3C \times 56 \times 56}$ . This is followed by two

## CHAPTER 3. VISUAL MEMORY SCHEMAS

output heads. The primary output consists of a  $3 \times 3$  convolution followed by a  $1 \times 1$  convolution that produces VMS map  $V$  for input image  $i$ ,  $V_i \in \mathbb{R}^{56 \times 56 \times 3}$ . The auxiliary head consists of two stacked  $3 \times 3$  convolution + max pooling blocks, followed by channel-wise global average pooling [88], and the output score  $L_i \in (0, 1) \subset \mathbb{R}$  is given by four stacked fully connected layers with  $\{F, \frac{F}{2}, \frac{F}{4}, 1\}$  neurons respectively. We choose  $F$  to be 256 and  $C$  to be 16, balanced for available compute budget, dataset size, and empirical studies (a greater value for  $C$  did not lead to additional performance gains).

### Loss Function

We train our predictor via the loss function given in Equation 3.8.

$$Loss(V, L) = \frac{1}{v} \sum_{i=1}^v (V_i - \hat{V}_i)^2 + \alpha \frac{1}{k} \sum_{i=1}^k (L_i - \hat{L}_i)^2 \quad (3.8)$$

The first term represents the loss over the samples of ground truth and predicted memorability maps, with  $V$  representing a predicted visual memory schema and  $\hat{V}$  representing a ground-truth map. The second term contains the loss over ground truth and predicted memorability scores,  $L$  and  $\hat{L}$  respectively.  $v$  and  $k$  represent sample populations of training data.  $\alpha$  is a weighting hyperparameter that controls the contribution of memorability score feedback when training to predict visual memory schemas. This can be set to 0 to disable dual feedback, and train on visual memory schema data alone.

## 3.3 Experimental Results

In this section we present the results for the proposed approaches given above. We start with a detailed description of all the datasets used in this work, from the initial VISHEMA dataset, to those that we have developed over the course of this project. We additionally quantify the elements that actually ‘make up’ a schema; providing a human-readable description of mental schemas that aid in the remembering of scene images. We then give results for the initial VAE-based model, over the original 800-image VISHEMA dataset, before examining potential model improvements on our expanded 1600 image dataset. Finally, we show the results for our current deep learning model over a new 4000+ image dataset of scenes and visual memory schemas.

### 3.3.1 Visual Memory Schema Datasets

Over the course of the work presented in this thesis two new datasets consisting of images and their corresponding VMS maps have been developed. In this section we will describe these datasets, how they were gathered, and their nomenclature. We will also explore the additional information we gain by taking a two-dimensional view of scene memorability compared to a single-score approach. The datasets used in this work are as follows:

- VISCHEMA
- VISCHEMA 2
- VISCHEMA PLUS
- VMS4k

VISCHEMA is the original dataset from the Akagunduz *et al.* experiment described above. VISCHEMA 2 is a replica of that experiment, consisting of 800 new images in the same categories as VISCHEMA, and with the same pre-processing paradigm applied (images that contained obvious text, people looking at the camera, and obvious landmarks were removed). VISCHEMA PLUS refers to these two datasets combined into a single 1600 image/VMS Map dataset, representing a 100% increase in available visual memory schema data. The available data was then further increased via the VMS4k experiment, resulting in over 4000 total scenes with paired VMS maps.

#### 3.3.1.1 VMS4k

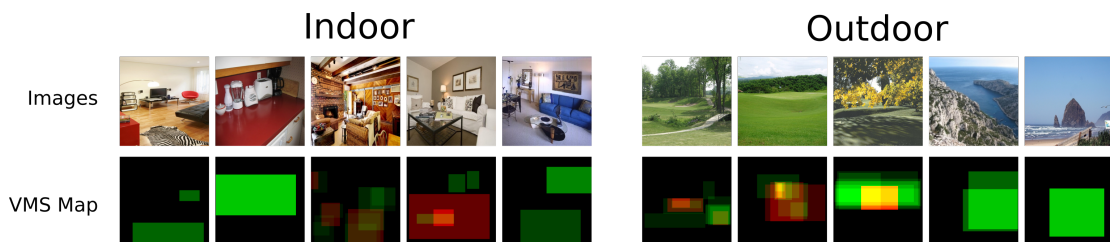


Figure 3.8: Examples from the VMS4k Dataset. Green areas indicate that region caused the image to be remembered, red areas indicate regions that caused an image to be falsely remembered; indicated as seen despite never being shown to a participant.

## CHAPTER 3. VISUAL MEMORY SCHEMAS

The full VMS4k dataset consists of 4000 images. These images are divided into two categories: indoor scenes, and outdoor scenes. The images themselves are drawn from the SUN dataset [140]. The indoor category is made up of 2000 images, the majority of which are extracted from the SUN kitchen and living room categories, with additional images from the conference room and airport terminal categories. These images provide a general collection of commonly encountered indoor environments, with a focus on environments encountered day-to-day. The outdoor category is more varied, and contains 2000 images extracted from the house, skyscraper, amusement park, playground, pasture, golf course, mountain, badlands, coast, and hill SUN categories. As environments encountered outdoors tend to be more varied than those indoors, a wider variety of images were collected for the purposes of the outdoor category.

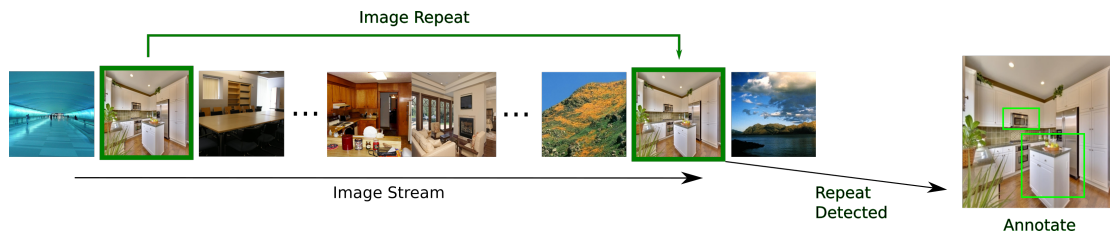


Figure 3.9: Repeat-recognition experiment structure

While the original VISHEMA experiment [2] used a two-phase in-person study/test paradigm, we instead design a continuous image-stream experiment, similar to [75]. This allows us to employ cloud-based experimentation platforms. Our dataset was divided into image sequences of 600 images, consisting of 200 targets, 200 fillers (i.e. images that were not repeated), and 200 repeats of the targets, yielding 20 distinct image sequences, each seen by human observers. Target repeats were distributed throughout the sequence such that there was an average of 300 images between the first showing of a target and its repeats. Each image was shown to the participant for three seconds. Once an image in the stream was indicated by the participant to have been remembered, they were asked to annotate the image with the region(s) of the image that they believed caused them to remember the image (Fig. 3.9). Participants were allowed to annotate multiple regions in the images. In total, 93 participants undertook the experiment. Participants show good memory performance for the images shown during the image sequences, (Fig. 3.10) with the majority of participants showing a  $d'$ -prime of over 2.0, indicating suitable performance.

### 3.3. EXPERIMENTAL RESULTS

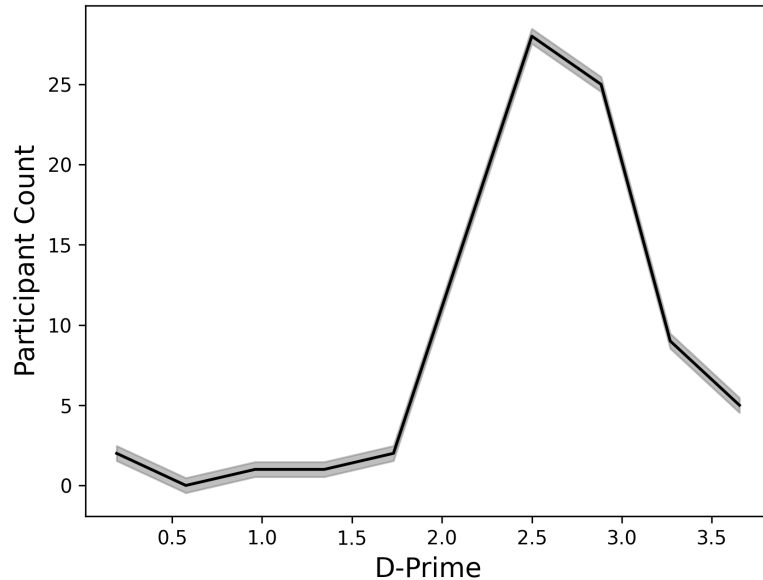


Figure 3.10: Participant D-Primes reveal good memory performance for the shown images. No participants were excluded from the analysis.

Of these 4000 shown images, not every image in the sequence was either (1) recognised as a repeat or (2) falsely recognised as a repeat. These images lack annotations, and for the purposes of this dataset, can be safely ignored. After this process, this leaves 3,461 images with corresponding annotations indicating the regions that caused the participants to remember that image. Examples from both the indoor and outdoor categories with corresponding memorability maps are shown in Fig. 3.8. The VMS map images consist of two channels; one containing regions labelled as memorable, and one containing regions that are ‘falsely memorable’; i.e, regions that caused the participant to false alarm on the image. In this work, we focus primarily on *memorability*, and concern ourselves with the memorability channel of the visual memory schemas. However, the dataset does contain false-memorability information that could be utilised in future work. We are able to safely combine this dataset with existing VMS datasets for a total of 4,261 image/VMS pairs (3,461 novel).



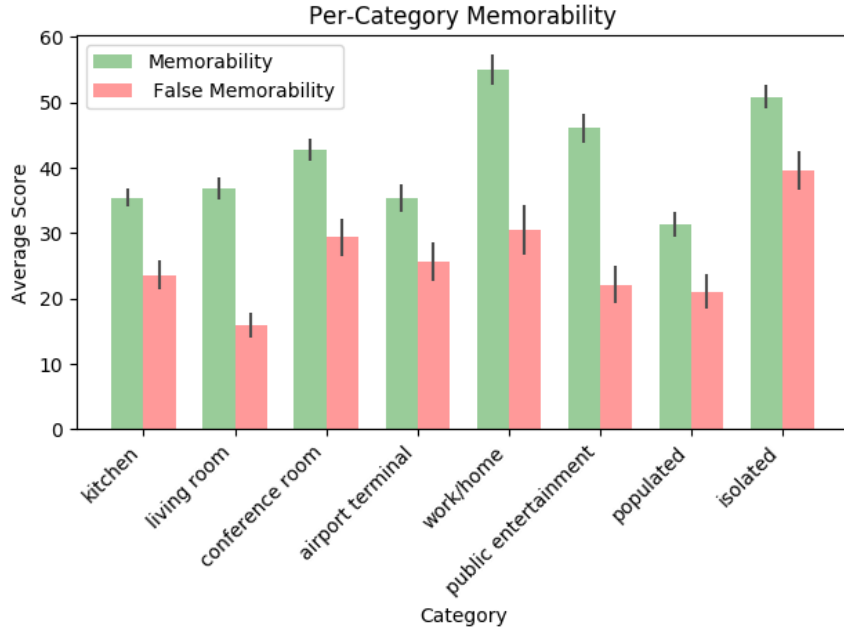


Figure 3.11: True and false memorability for the VISCHEMA image set.

### 3.3.1.2 Category Differences

It is well known that memorability varies across categories; some are by default more memorable than others. For the VISCHEMA dataset, the memorability (and false memorability) of each category is shown in Fig. 3.11. In this case, we condense the visual memory maps down to a single score based on the average value of each channel (either memorability, or false memorability) of the map. This represents how consistent participants were when annotating regions of the image as memorable, or falsely memorable. However, in condensing this information down to a single score, the two-dimensional aspect of the data is lost.

Two-dimensional memorability annotations allow us to understand not just which images are memorable, but the differences between images that, on the surface, appear to have the same level of memorability. Such a difference is obscured if a single-score perspective is taken. We investigate the difference between the two categories of scene images (Fig. 3.12), and find no significant difference ( $p > 0.05$ , one-way independent ANOVA) in memorability, defined by per-image hitrate (correct detection of the target). There

### 3.3. EXPERIMENTAL RESULTS

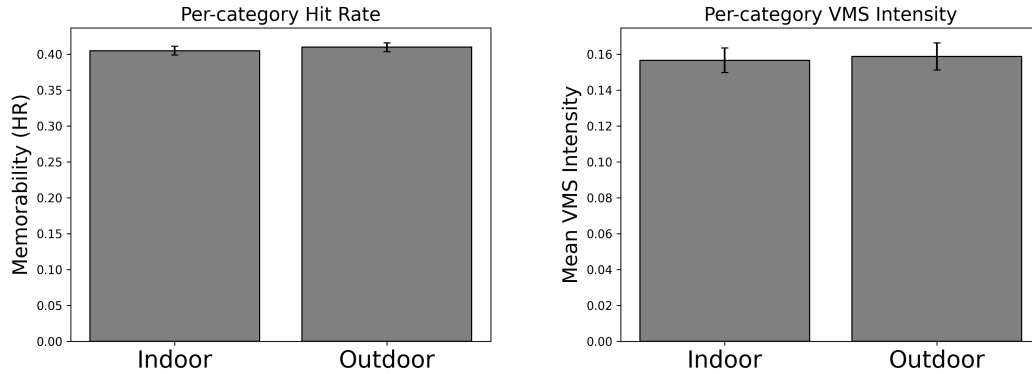


Figure 3.12: There is no difference in memorability performance between categories as measured by hit-rate or VMS intensity (an analogue for participant consistency).

is also no significant difference ( $p > 0.05$ , one-way independent ANOVA) in participant consistency for indoor or outdoor scene categories, defined by the average intensity of the VMS memorability channel (0.157, indoor vs 0.159, outdoor). While prior work finds memorability differences across categories (and indeed we show this for the VISHEMA dataset), in those cases the categories were significantly more fine-grained compared to the coarseness of "indoor" or "outdoor". In this case, both categories can be considered 'identically memorable' - at least if just a single-score rating of memorability is considered.

However, the two-dimensional annotations reveal more differences between the categories (Fig. 3.13, Fig. 3.14) than are immediately apparent from examining single-score metrics. Indoor scenes had significantly more annotations ( $p < 0.05$ , Kruskal-Wallis) per-image than outdoor scenes, which show a clear bias towards lower counts of annotations; that is, participants believe fewer regions of the image caused them to remember that image compared to indoor scenes (Fig. 3.15). This suggests that despite the similar overall memorability between the two categories, the memorability of outdoor scenes is related to fewer semantic structures within the scene, whereas for indoor images, multiple regions spread spatially across the scene together cause that scene to be remembered.

Beyond number of labelled memorable regions, we also find a significant difference ( $p < 0.05$ , Kruskal-Wallis) in the sizes of the memorable regions (Fig. 3.15) between the two categories, with memorable regions in indoor scenes being significantly smaller,

### CHAPTER 3. VISUAL MEMORY SCHEMAS

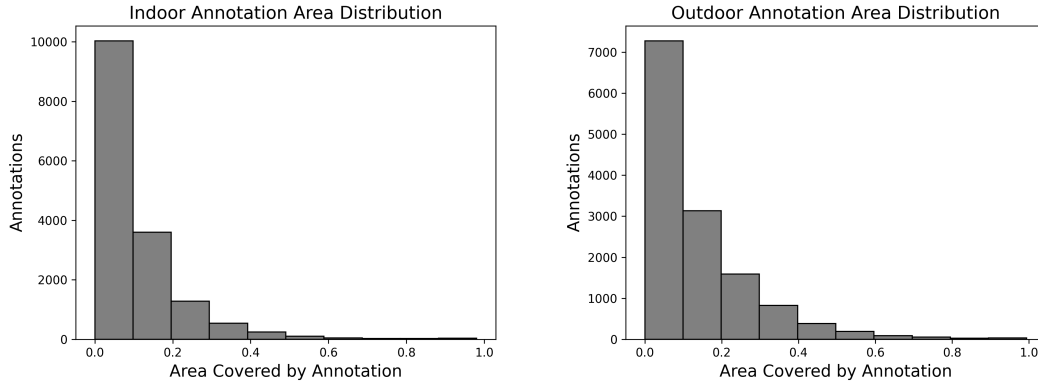


Figure 3.13: Outdoor scenes (right) show bias towards larger annotation areas compared to indoor scene images (left).

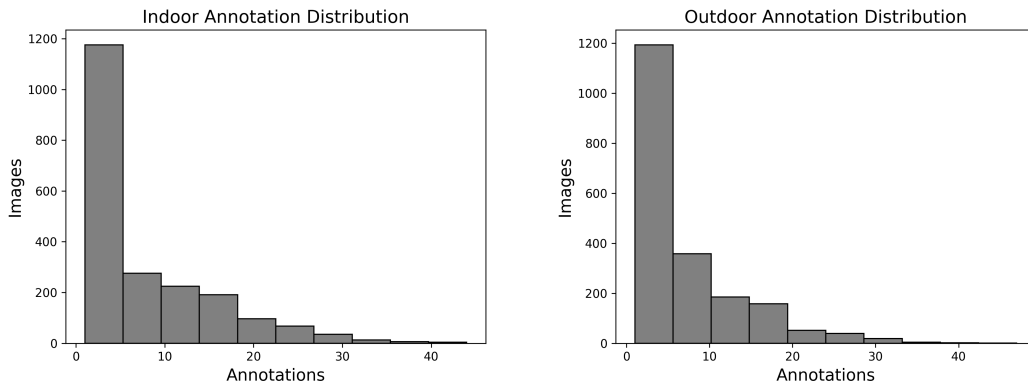


Figure 3.14: Outdoor scenes (right) show a greater bias towards fewer per-image annotations than indoor scenes (left).

(size defined by percentage of image covered with annotation) than those from outdoor scenes. Intuitively, this makes sense; outdoor scenes often portray grander vistas than indoor scenes (a coastline, vs a kitchen) and as such have appropriately sized memorable semantic structures. Hence, memorable indoor scene images appear memorable due to multiple smaller regions (a combination and arrangement of multiple objects, e.g tables, chairs, couches), while outdoor scene images are memorable due to larger, more singular regions (a mountain; a coastline). These details are lost when VMS maps, and image memorability in general, is treated as a singular score.

### 3.3. EXPERIMENTAL RESULTS

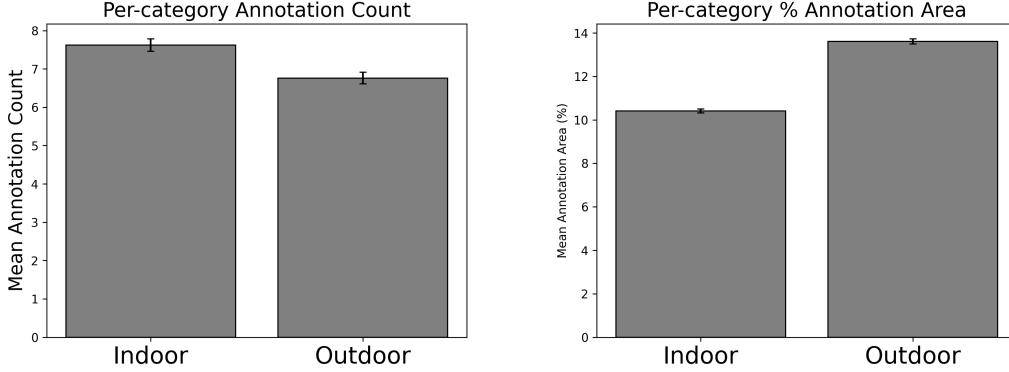


Figure 3.15: The average annotations per-image is significantly greater for indoor, than outdoor scenes (left), and there is a significant difference between the sizes of annotations between indoor and outdoor regions (right).

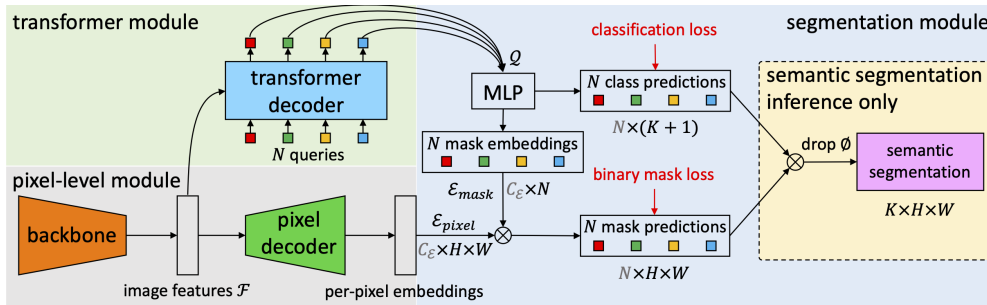


Figure 3.16: MaskFormer architecture; a neural network that can be used for state-of-the-art semantic segmentation, [Fig 2] from [26].

#### 3.3.1.3 Quantifying the Schema

While visual memory schemas reveal the regions that drive scene memorability, and hence represent the schema elements used to encode that image, it is difficult to go from a VMS map to a human-understandable description of the schema. A person can easily determine the objects and arrangement of elements contained within a memorable region; but to do this over the entire VMS4k dataset would be intractable, both timewise and financially. Instead, we would like to be able to computationally gather the scene elements that have been captured within a memorable region. This is not an easy task; the ground-truth images in VMS4k come with no pixel-level labels that reveal which objects and semantic units (walls, skylines, floors, fields, etc) are contained in

## CHAPTER 3. VISUAL MEMORY SCHEMAS

any given image. Extracting which objects and semantic units have caused an image to be memorable, and generalising this over our VMS categories would allow us to extract what is being contained within every memorable region in the dataset; revealing the actual schemas being used to encode our scene images.

To do this, we employ the MaskFormer architecture [26] (Fig. 3.16). MaskFormer is a semantic segmentation network. While an object-detection network may be tasked to identify every object in a scene, and be able to delineate said objects with bounding boxes, the goal of a semantic segmentation algorithm is to decompose an image into a set of pixel-level labels, that identify exactly which object, or semantic unit, that pixel belongs. MaskFormer takes a slightly different approach, instead attempting to generate and classify binary masks, each of which segments out one part of the image. A transformer component (‘transformer decoder’) [131] generates sets of class predictions and mask embeddings via a multilayer perceptron (MLP). The pixel decoder extracts per-pixel embeddings, which are combined with the output of the transformer decoder to compute both a binary mask. The output of the MLP is used directly to generate class predictions. The mask and class predictions are then combined via matrix multiplication in the final module of the network. The network that we use for extracting the content of memorable regions is pre-trained upon the ADE20k-Full dataset, with 847 classes. At the time of writing, MaskFormer is both more efficient and more accurate than other segmentation models.

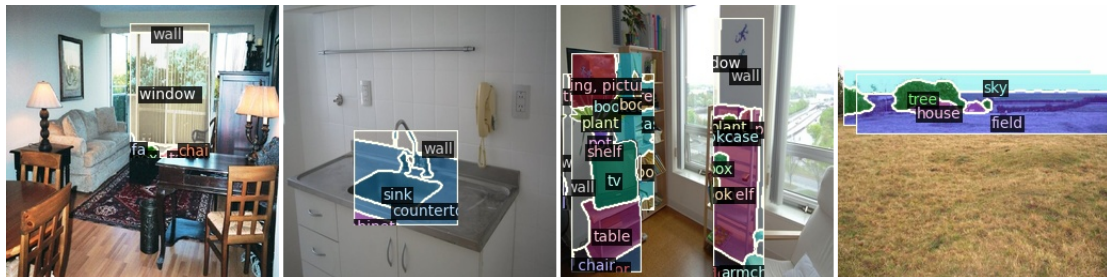


Figure 3.17: ‘Semantic units’ contained within the regions of images that participants have labelled as causing them to successfully remember that image.

Some examples of content found and labelled inside memorable regions of the VMS4k dataset is shown in Fig. 3.17. While the predictions are not always perfectly accurate; they are accurate enough that a reasonable picture of the schema for each image can be seen. For example, in the image of a field; it is obvious that not one single element

### 3.3. EXPERIMENTAL RESULTS

contributed to that image being remembered. Instead, it is the arrangement of the house, with the trees, placed in a field with the sky as a background. These are the scene elements that have matched with the mental schema held in the participants which labelled this image, and aided in encoding of the scene. To extract a general schema for each category, we ask which scene components commonly occur with each other *inside* memorable regions; that is, which arrangement of elements most frequently leads to a region of the image being labelled as causing recognition of the scene. We do this by calculating the number of times each extracted element co-occurs with other element(s) across all memorable regions in that image.

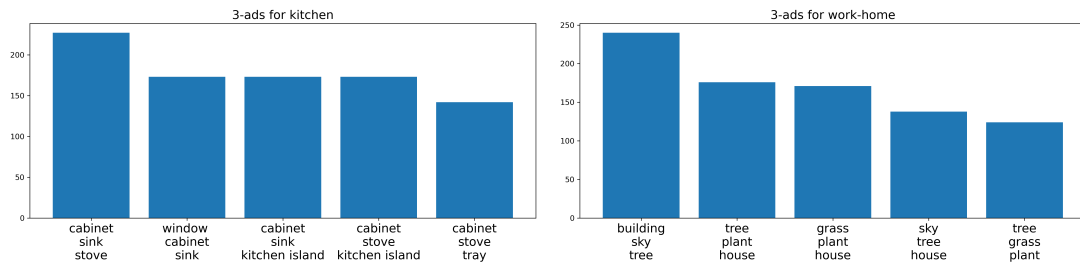


Figure 3.18: These objects frequently appear together inside the memorable regions of an image, of that category. Limited to three objects; higher amounts of object co-occurrences can be examined.

In Fig. 3.18 we show some examples of this procedure, for the kitchen and work-home (pictures of houses, or office buildings) category. We limit this analysis to co-occurrences of just three objects; higher amounts of objects can also be examined (see Appendix A). Likewise, we only show the top five most frequent ‘schemas’. From this we can determine that the most likely cause of encoding of a kitchen image is the presence of cabinets, sinks, and stoves (greater than other arrangements of memorable kitchen semantic units; e.g the presence of cabinets, stoves, and trays). For the work-home category most frequent is buildings, skylines, and trees; whereas arrangements of trees, grass, and plants appear to occur less frequently inside the regions that have caused recollection of that image. These elements, appearing together, appear to capture the ‘schema’ used to encode scene images for a given category; we have gone from a mental schema, to two-dimensional maps, and finally to human-understandable descriptions of those schemas for each VISHEMA category. While we have hypothesised that some scenes are remembered better due to their content; and because they better match a held

schema in a human observer, through quantifying that schema we can see that this does in fact appear to be the case. Some arrangements of objects are labelled more frequently as "causing the remembering of that image" than other arrangements of objects; across entire categories of similar scenes.

### 3.3.2 Variational Autoencoder Approach

We train our variational network over the VISHEMA dataset, and we also use the images from the VISHEMA2 dataset (ground-truth scores were not available at the time of this study) for evaluating the model. We employed the LaMem dataset to evaluate the relationship between VMS maps and single-score ratings. For the encoder we use a pretrained VGG-16 network to extract a  $7 \times 7 \times 512$  representation of an image, then compress this further using an  $n$  dimensional dense layer, which leads to a latent space with a dimension of  $m$ . All parameters of the VGG network are frozen, by considering learning rates set to 0 during training, to avoid damaging the deep features while training on a small dataset such as ours. We employ data augmentation for training due to the small size of the training set. Data augmentation involves various realistic image manipulations, such as for example shifting the image either horizontally or vertically by 0.1 of the total image width, zooming the image, and horizontal flipping, which artificially increases the training data, and helps to reduce network overfitting.

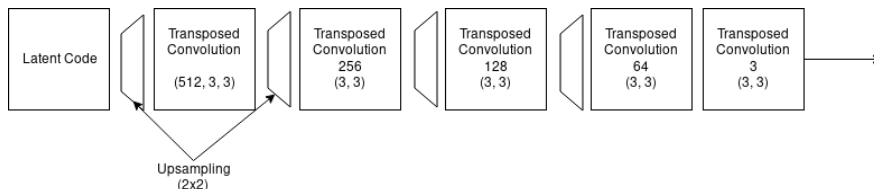


Figure 3.19: Structure of the Decoder.

The decoder consists of a five layer upsampling network, shown in Fig. 3.19, that implements transposed convolutions in order to convert the  $m$ -dimensional latent variable space back into an image. We apply batch normalisation after every convolution and employ l2 kernel regularisation [31],  $\lambda = 0.02$ , and a learning rate of 0.0001. We use a batch size of 32 and train the network for 250 epochs with 20 steps per epoch. In the experiments we evaluate three different architectures considering: 1)  $n = 64$  and  $m=8$ ; 2)  $n = 64$  and  $m=8$  with an  $l1$  reconstruction loss; 3)  $n = 128$  and  $m=32$ . The input and output of the entire architecture is a  $224 \times 224$  image. The model is implemented

### 3.3. EXPERIMENTAL RESULTS

in Keras<sup>1</sup>.

We evaluate reconstruction results of the original VISCHEMA dataset using both standard mean squared error (MSE) over all test images and the two dimensional Pearson product-moment correlation coefficients  $\rho^{2D}$ . We average the results on all true VMSs, and false VMSs, separately. True VMSs represent the VMS map regions indicated by participants in the experiments that represent what made them remember that image, while false VMSs represent regions from images, falsely indicated by people that made them remember those images. Actually those images have not been shown to them before. We obtain this metrics for all visual schemas and then evaluate the relation between this metric and the more standard ‘memorability score’ provided in the LaMem dataset [44]. The relationship between visual memory schemas and computational saliency is also explored. Computational saliency maps for the VISCHEMA datasets are generated via the Graph Based Visual Saliency (GBVS) algorithm [56].

Finally, we employ a single-score memorability prediction network and evaluate the relation between the VISCHEMA datasets memorability scores of the predicted VMS and the VMSs corresponding to the choices made by people, for both datasets, VISCHEMA and the VISCHEMA2. For all evaluations of our memorability metrics and standard memorability scores we follow prior work from [69], [75] and use Spearmans rank correlation.

Latent Space Dimension (m)	VMS	$\rho^{2D}$	MSE
32	True	0.545	92.54
	False	0.369	70.526
	All	<b>0.57</b>	85.379
8	True	0.513	90.812
	False	0.333	64.228
	All	0.53	83.472
8 and L1 norm in (3.3)	True	0.543	72.348
	False	0.168	25.131
	All	0.559	72.052

Table 3.1: Reconstruction accuracy for three deep learning architectures.

Table 3.1 shows the reconstruction results in terms of both MSE and Spearmans rank

<sup>1</sup><https://keras.io>



### CHAPTER 3. VISUAL MEMORY SCHEMAS

correlation,  $\rho^{2D}$ . The network with an  $m=8$  dimensional latent space and an  $l_1$ -norm component to its loss function has the overall best MSE, while the network with the overall best Pearsons correlation with the ground truth is the network with a  $m=32$  dimensional latent space. Our overall  $\rho^{2D}$  results are slightly worse than those presented in [2], though it should be noted that we generate both the true and false maps simultaneously. This allows us to investigate how well the individual true and false VMS are reconstructed. In general, false VMS maps are more difficult to accurately reconstruct than true VMS maps. This is likely due to the overall lower consistency between human observers for false VMS maps. While what is memorable tends to be well agreed on among people, what causes false remembering of an image is more varied, and this effect crosses over to generative models. Interestingly, we find that a higher dimensional latent space has the best effect on reconstruction accuracy, rather than the use of an  $l_1$ -norm in the loss term. This is due to the effect of the second term in the loss function from equation (3.3) and indicates that higher dimensional spaces are better at capturing ‘memorability’. For the rest of this section we evaluate the results of the network with a  $m = 32$  dimensional latent space, given that this architecture performs the best as measured by the  $\rho^{2D}$  metric.

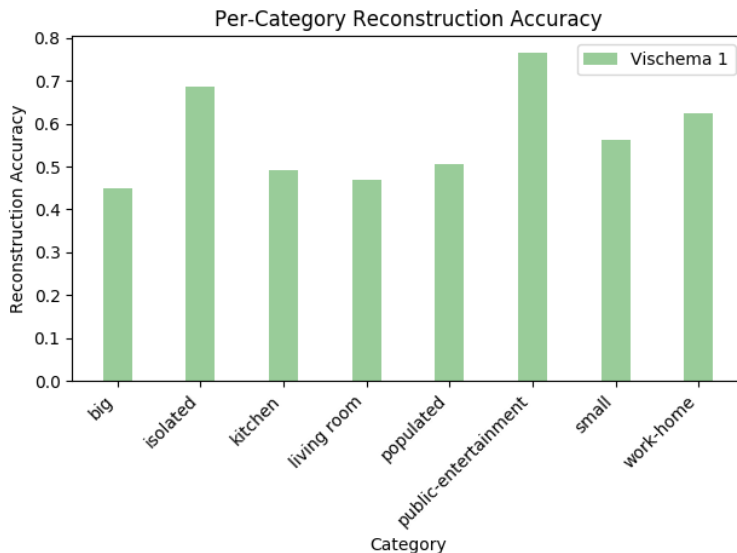


Figure 3.20: Reconstruction accuracy for various image categories.

Figure 3.20 shows the reconstruction accuracy measured by  $\rho^{2D}$  for each category in the VISCHEMA dataset, over the 160 image test set. We find that the best performing

### 3.3. EXPERIMENTAL RESULTS

category is that of Public Entertainment, with a correlation of 0.766, which is better than the results from [2] which found that the Work-Home image category had the best performance with a correlation of 0.677. A comparison with prior work is shown in Table 3.2.

Work	Best Category	$\rho^{2D}$	Worst Category	$\rho^{2D}$	Overall $\rho^{2D}$
Previous Method	Work/Home	0.677	Living Room	0.506	0.588
Our Method	Public Entertainment	0.766	Big	0.449	0.57

Table 3.2: Comparison with Prior Work

The worst performing category for VMS reconstruction is the "Big" which contains images of airport terminals with a correlation of 0.449. In general, we find that categories that have high overall memorability tend to reconstruct better than the categories with low overall memorability. Differences from prior work may also be due to generating higher resolution images, which captures more detail in some categories yet causes more divergence in categories with less available memorability information. We found that the correlation between predicted VMS maps and saliency maps, provided by the Graph Based Visual Saliency (GBVS) algorithm [56], to be 0.69 which agrees with other results on the relationship between memorability and saliency [40, 2]. GBVS is a well used saliency measure, but VMS maps offer more than saliency alone. When averaging on all image categories and comparing with saliency, we found that false VMS maps have a correlation of 0.625 while true VMS maps have a correlation of 0.704.

#### Memorability Results

We generate 800 predicted VMS maps for the 800 images in the VISHEMA2 dataset and find that the distribution of memorability and false memorability agrees with that of the original ground truth dataset, according to the results from Fig. 3.21 with Spearman's ranks of 0.929 and 1.0, respectively for  $p < 0.01$ . This is due to the similarity of the datasets, but it also shows that the proposed model has successfully learned to generate VMSs that agree on a category-wide scale despite being trained with no category labels. Additionally, we find that in general the higher the memorability of an image, the higher its own false memorability, as we can observe from the similarity of the clusters of the latent space embeddings of the Memorability and those corresponding to False Memorability, shown in Fig. 3.22a and 3.22b, respectively. Images that tend to be

### CHAPTER 3. VISUAL MEMORY SCHEMAS

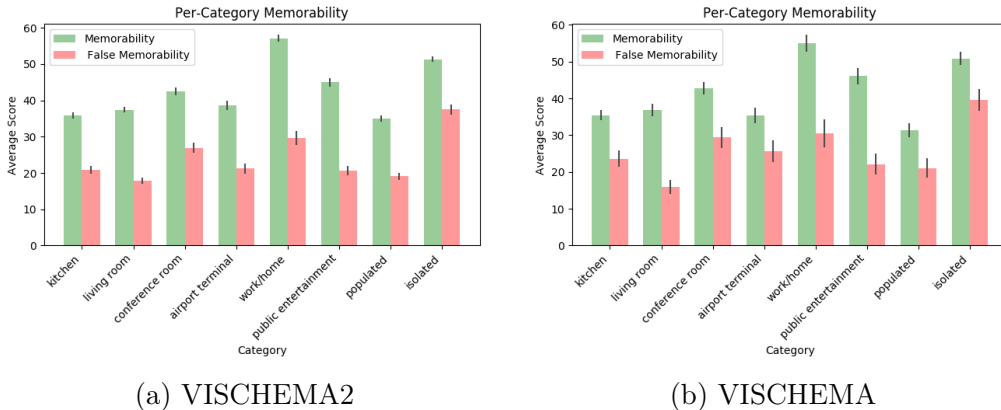


Figure 3.21: Comparison of the memorability results for a set of image categories between the VISCHEMA2 and VISCHEMA datasets.

highly memorable also tend to be highly falsely memorable. In Fig. 3.23, three images from VISCHEMA2 are shown on first line and their corresponding true and false VMSs are shown on second and third line, respectively.

Predicted memorability scores for both VISCHEMA 1 and 2 datasets were obtained by employing the AMNet network [44]. These scores were then compared to the memorability metric used for evaluating visual schemas. No significant relationship was found between the per-category memorability metrics and the predicted category memorability scores aside from VISCHEMA2’s "Populated" category which had a Spearmans rank correlation with the AMNet scores of 0.203 with  $p < 0.01$ . It appears that VMSs, even predicted schemas, do not directly relate to predicted memorability scores for the same images, and that unlike our VMS prediction model, predicted memorability scores may not take fully into account what humans find memorable. It has been shown that deep neural networks take the simplest approach possible to solving a problem [17], and it is possible that memorability prediction models are working on factors that do not necessarily align directly with memorability if some other learned metric provides a ‘good enough’ approximation. This could explain why predicted scores do not align with VMS maps.

We also examine the relationship between the ground truth memorability scores and our metric by predicting VMSs for a 10,000 image subset of the LaMem dataset, used in [44], and estimating only the true memorability score for them. We then use the Spearmans

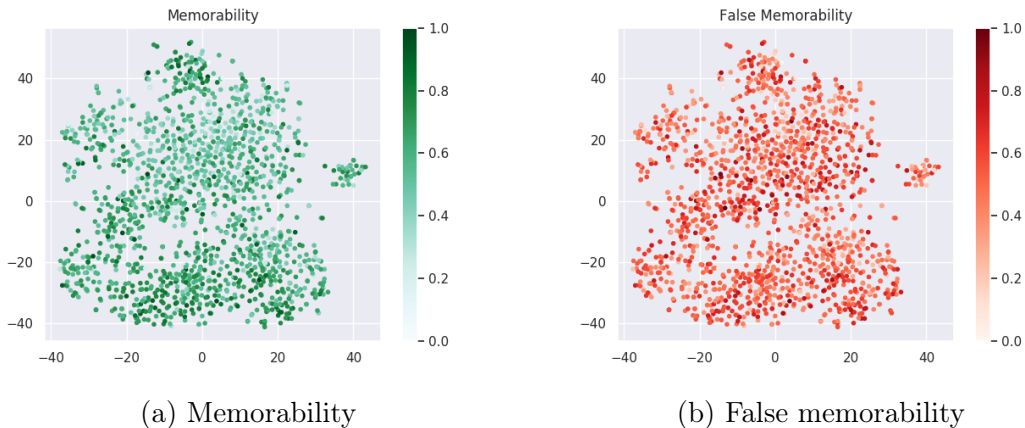


Figure 3.22: VISCEMA2 Latent Space Embedding. Green represents memorability and red represents false memorability, normalised between 0 and 1. Clustering of both memorable, and falsely memorable images is evident. Features that lead to the generation of memorable VMS maps are placed near each other, as are features that lead to the generation of VMS maps that indicate the scene is not so memorable.

rank to compare the ground-truth score and our metric. We find a rank correlation of 0.147 with  $p < 0.01$ , indicating that VMS maps and experimentally-based memorability scores are weakly, but significantly, related.

### 3.3.3 Multi-Scale Information, Depth, and Self-Attention

We use the **VISCEMA PLUS** dataset, with 1600 scene images and 1600 corresponding memorability maps. We divide this dataset using a standard split of 70% training set, 20% validation set, and a 10% test set which we use for analysis. Two images from this dataset can be seen in the first column from Figure 3.24.

Prior work, including our own, evaluates the efficacy of VMS predictors with two distinct measures; the Pearson 2D correlation [2], and the mean squared error (MSE) [83]. We choose three additional probabilistic measures as evaluation measures in order to evaluate our VMS predictors: Kullback-Leibler Divergence (KLD), Earth Mover Distance (EMD), and Histogram Similarity (SIM) [19], metrics commonly used to evaluate saliency map models. We also employ the pixelwise Spearman rank correlation,  $S^{2D}$ , as the measure commonly used to evaluate memorability score predictors. The ‘best’

### CHAPTER 3. VISUAL MEMORY SCHEMAS



Figure 3.23: Set of three images from VISHEMA2 dataset and their predicted true VMS and false VMS on second and third lines. We find empirically that false schemas are often subsets of the true schema of the image that carries less information. For example, an image is memorable due to the presence of a man feeding a calf, yet the presence of just a man may lead to the false remembering of a scene.

metric depends on application; some applications may value a small mean squared error distance, others a model that displays statistically similar behaviour to human ground truth, even at the cost of a greater MSE. By selecting a variety of metrics, we offer future work a comprehensive analysis of VMS prediction models, and the chance to build on a model best suited to whichever future application is necessary.

The deconvolutional networks are trained for 100 epochs (after which there is no improvement against the validation set), and optimised via RMSProp using a learning rate of  $\eta = 0.0001$ . Each deconvolutional network outputs a  $28 \times 28$  pixel VMS map for a given input image, as VMS maps are robust to rescaling. The VAEs are trained for 500 epochs, and output a VMS map at the same resolution as the input image. Features from the pre-trained VGG16 network were L2 normalised before reaching the trainable

### 3.3. EXPERIMENTAL RESULTS

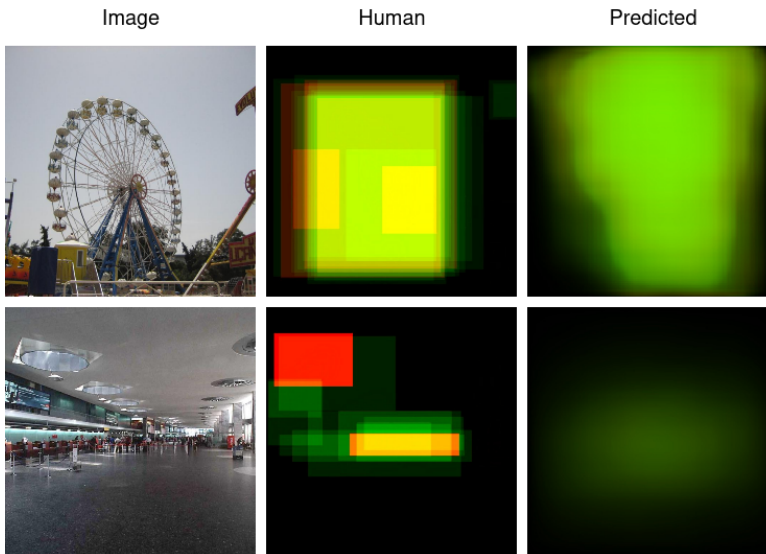


Figure 3.24: VMS maps showing memorable (green) and falsely memorable (red) regions, for the images from the first column, are shown in the second column, and their corresponding predictions on the third column.

layers. All networks were trained on a single NVIDIA 1080 Ti GPU.

The prediction results for the VMS memorability channel are provided in Table 3.3 and those for the false memorability are shown in Table 3.4. In these tables, we denote by MSB when considering multiscale blocks, attention (or att) where we use non-local neural blocks, as described in Section 3.2.2, and ‘Depth,’ when using depth maps according to Section 3.2.2. VAE latent spaces are denoted by L + the latent dimension  $|z|$ .

For memorability, the best performing straight deconvolutional networks were trained with the KL Divergence loss from (3.4), which provides the best MSE performance from all tested architectures. For the false memorability, a simple MSB-based network sets the record for MSE, although attention-based MSB networks come close. These results for MSE outperform prior work by a significant margin. The superior performance of the KL-loss may explain why VAEs remain the best overall approach. With limited data, it is not surprising that VAEs with a smaller latent-space ‘bottleneck’ perform better. By combining the ability of VAEs to extract low-dimensional memorability/false-memorability features with non-local neural networks long-range dependency capture, the VAE+Att L8 Model sets the state of the art results for four memorability metrics and three false-memorability metrics. The baselines (VAE aside) performed poorly at both

### CHAPTER 3. VISUAL MEMORY SCHEMAS

Table 3.3: Prediction results for the VMS memorability channel. SH: single-headed output. KL: Kullback-Leibler Diver.

Model	MSE↓	$P^{2D}$ ↑	$S^{2D}$ ↑	KLD↓	EMD↓	SIM↑
CNN-deconv	70.09	-0.03	0.03	2.1	159.67	0.4
MSB	86.79	-0.01	-0.06	1.31	142.4	0.41
CNN-deconv SH	61.99	0.02	0.04	2.86	147.6	0.4
MSB SH	69.84	0.14	0.21	1.04	197.44	0.44
VAE (from 3.2.1) [83]	87.23	0.46	0.51	1.06	36.01	0.52
MSB-Attention	58.83	0.1	0.19	1.29	191.42	0.44
MSB-Depth	76.24	0.22	0.29	1.32	151.67	0.45
MSB-Depth+Att	70.99	0.24	0.37	<b>0.99</b>	186.75	0.46
MSB-Attention SH	69.63	0.31	0.32	3.01	80.8	0.46
MSB-Depth SH	77.36	0.13	0.2	1.88	141.46	0.42
MSB-Depth+Att SH	67.98	0.24	0.4	1	187.83	0.46
MSB-Attention KL	<b>53.78</b>	0.22	0.29	-	179.93	0.46
MSB-Depth KL	67.3	0.31	0.44	-	157.02	0.48
MSB-Depth+Att KL	79.2	0.34	0.41	-	106.1	0.49
VAE L8	92.44	0.48	0.52	-	36.3	<b>0.53</b>
VAE L64	83.57	0.47	0.52	-	35.06	0.51
VAE L128	96.13	0.43	0.47	-	47.22	0.49
<b>VAE+Att L8</b>	87.65	<b>0.49</b>	<b>0.53</b>	-	<b>34.17</b>	<b>0.53</b>
VAE+Att L32	87.88	0.46	0.51	-	36.88	0.52
VAE+Att L64	84.4	0.46	0.51	-	36.53	0.51
VAE+Att L128	91.31	0.44	0.48	-	42.91	0.49

true memorability and false memorability prediction, as it can be seen from Tables 3.3 and 3.4.

The poorest performing architecture is the straight deconvolutional network. The initial introduction of multi-scale blocks improves performance slightly, and producing a single output improves performance significantly. Both the introduction of self-attention and depth information improves memorability prediction, though depth information alone causes significantly poorer performance when predicting false memorability. Depth and attention modules combined exceed the performance of either one alone. With this additional information, there is minimal difference between single-headed or dual-headed approaches. As with prior work, the prediction of the false-memorability channel remains significantly more difficult than that of memorability prediction. This is because false memorability maps are more varied and less consistent than positive memorability maps.

### 3.3. EXPERIMENTAL RESULTS

Table 3.4: VMS false memorability channel prediction results.

Model	MSE↓	$P^{2D}$ ↑	$S^{2D}$ ↑	KLD↓	EMD↓	SIM↑
CNN-deconv	39.9	-0.05	-0.09	8.73	33.3	0.12
MSB	<b>35.96</b>	-0.12	-0.16	9.5	23.92	0.05
CNN-deconv SH	39.94	-0.13	-0.19	9.98	37.29	0.08
MSB SH	38.54	-0.03	-0.03	8.12	<b>22.64</b>	0.11
VAE (from 3.2.1) [83]	75.66	0.34	<b>0.37</b>	1.85	36.38	0.36
MSB-Attention	38.53	0.12	0.15	2.17	186.03	0.29
MSB-Depth	69.7	-0.07	-0.17	6.58	35.52	0.15
MSB-Depth+Att	63.29	0.09	0.09	4.61	69.9	0.25
MSB-Attention SH	47.15	0.09	0.09	5.77	63.22	0.24
MSB-Depth SH	57.89	-0.2	-0.32	9.5	32.09	0.07
MSB-Depth+Att SH	66.6	0.17	0.17	3.28	67.42	0.28
MSB-Attention KL	38.62	0.23	0.26	-	122.24	0.33
MSB-Depth KL	48.33	0.17	0.25	-	159.54	0.3
MSB-Depth+Att KL	57.76	0.07	0.08	-	114.91	0.26
VAE L8	83.27	0.35	<b>0.37</b>	-	30.77	0.36
VAE L64	62.53	0.31	0.33	-	36.67	0.34
VAE L128	86.33	0.29	0.33	-	72.98	0.33
<b>VAE+Att L8</b>	74.66	<b>0.36</b>	<b>0.37</b>	-	29.73	<b>0.37</b>
VAE+Att L32	73.41	0.34	<b>0.37</b>	-	35.61	0.36
VAE+Att L64	67.86	0.33	0.36	-	47.24	0.36
VAE+Att L128	73.57	0.3	0.32	-	54.68	0.33

Nonetheless, incorporating self-attention mechanisms provides a significant improvement to the performance, likely due to being able to capture longer-range dependencies. We achieve a  $P^{2D}$  of 0.49 for the positive memorability and 0.36 for false memorability respectively, which exceeds all previous models tested on VISHEMA dataset. While single-score models have matched human-level consistency, with a baseline for human VMS consistency of 0.69, VMS prediction still has a way to go before reaching the level of single-score predictors.

#### 3.3.4 A Dual-Feedback Approach

Here we discuss the implementation details required to train the dual-feedback network and present prediction results over the VMS4k dataset. The Dual-Feedback VMS (DF-VMS) Network is trained using the Adam optimiser [77] with a learning rate of  $5 \times 10^{-5}$



### CHAPTER 3. VISUAL MEMORY SCHEMAS

with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . Each model is trained for 250 epochs on an Nvidia V100 GPU. The network is trained on two datasets. The first dataset is VMS4k, divided into a train/validation/test split of 85%/5%/10%. Each input consists of a random batch of scene images and their corresponding human annotated (ground truth) Visual Memory Schemas. The second dataset is LaMem [75], with each training example consisting of an input image (not necessarily a scene image) and its corresponding one-dimensional memorability score. We train the network in a ‘tick-tock’ fashion, first on the LaMem training set, then on the VMS4k training set, repeating each epoch until training is complete. For our backbone we use either VGG16 or RESNET50, pre-trained on the imagenet dataset. The weights of the backbone architecture are not updated during training. We empirically choose  $\alpha = 40^{-1}$  which helps prevent the network focusing on predicting scores over our primary objective; the VMS maps. The network takes approximately 18 hours to train on a single V100 GPU. We evaluate our architecture on VMS4k and use LaMem as an optional auxiliary feedback mechanism. There is no two-dimensional memorability data associated with the LaMem dataset.

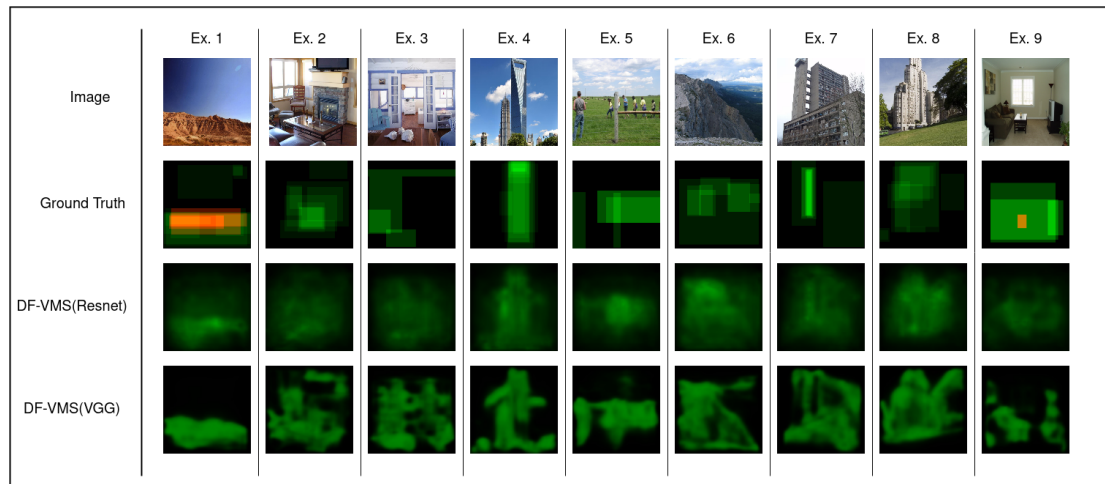


Figure 3.25: Predicted VMS maps for the given scene images. Ground-truth maps come from human data. Some human VMS maps contain false schemas (red), for visualisation purposes in this figure we only show predicted true (memorable) schemas. The best performing DF-VMS variant employs a Resnet backbone, self-attention, multiscale-information, and dual-feedback. VGG16 backbones do not capture the full spread of memorability; instead focusing strongly on semantic regions. ResNet backbones, with their richer feature extraction, perform better at VMS map prediction.

Results for reconstruction accuracy on VMS4k are shown in Table 3.5. While we obtain

### 3.3. EXPERIMENTAL RESULTS

the best results with the ResNet50 backbone feature extractor, we include results using the older VGG16 architecture for the purposes of comparison to prior work. In the table, we show results for the previous best performing, variational autoencoder based model, vms-VAE [83], on both the original VISCHEMA dataset, and on VMS4k. Our DF-VMS model outperforms this model, even when vms-VAE is trained on our VMS4k dataset (and thus benefits from the additional data) while using the older VGG16 as a backbone for DF-VMS. Our analysis reveals that prior models are not capable of taking advantage of our larger dataset. This drawback does not affect the proposed DF-VMS approach. Our qualitative results indicate that DF-VMS models that use the VGG16 backbone tend to give overconfident memorability predictions over the semantic content of the image, but do not capture the ‘spread’ of memorability across the image. To verify that the model was not simply learning to activate on strong edges, we include results for a baseline canny edge detector based approach. We find that this results in poor performance compared to any of our networks; indicating that all models are learning to detect ‘memorable regions’ rather than areas of strong edges.

Through our DF-VMS model we boost visual memory schema prediction performance by 11.8% for true (memorable) schemas and by 8.6% for falsely memorable schemas compared to prior work. In Fig. 3.25 we show a set of predicted examples for a variety of both indoor and outdoor scene images along with their ground-truth human VMS maps. Some human VMS maps (Ex. 1, Ex. 9) contain red areas that indicate regions that lead to false remembering. See Fig. 3.26 for these examples with predicted false memorability maps. While the VGG-backbone generates confident and clear predictions; in practice, these fail to capture less memorable regions of the image, and overall a deeper backbone leads to superior performance by offering features that capture regions which do not contain the strongest memorable signal. For completeness (we do not focus on score prediction), we include results for the LaMem test set from our auxiliary output. We achieve reasonable results for this despite significant differences between the VMS4k dataset (scene memorability) and the LaMem dataset (generic image memorability i.e. frame-filling objects, faces, or people).

**Ablation Testing** To evaluate the impact of the various optional model improvements (attention, dual feedback, multi-scale information) we train the best performing model a further three times with a given aspect *excluded* ( $\mathbf{x}$ ) from the model, and show the results in Table 3.5. In the table,  $-\mathbf{xA}$  indicates attention excluded (i.e, not present in

### CHAPTER 3. VISUAL MEMORY SCHEMAS



Figure 3.26: Examples 1, 9, and additional exemplars with predicted false memorability maps. As consistent with [83], false schemas are often a subset of the true schema, and are more difficult to predict.

the model),  $-\mathbf{xDF}$ , dual-feedback excluded, and  $-\mathbf{xM}$ , multiscale information excluded. Additionally, we test performance purely on the auxiliary memorability loss by disabling visual memory schema feedback ( $-\mathbf{xVMS}$ ). All ablation models were trained for the same number of epochs as the original model. We find that in general disabling any of these factors leads to poorer model performance, with the most drastic decrease occurring when dual feedback is disabled, except for in the case of the relatively shallow VGG16 backbone; in this case disabling dual feedback leads to an even greater performance increase over current SoTA. This is because the features extracted by the VGG16 network are not rich enough to support the additional constraints on learning imposed by the LaMem feedback, and leads to an overall destabilising effect. However, in either case both models still exceed current SoTA, and the deeper ResNet network does not suffer from this destabilisation. The LaMem feedback appears to improve results in one of two ways: 1.) by better predicting human ground truth in the memorable regions of the image (DF leads to the network better understanding how semantic image features relate to memorability) and 2.) by reducing erroneous predictions for regions of the image that are unlabelled; neither memorable nor falsely-memorable. Hence, by employing existing large single-score memorability datasets as an auxiliary loss, an increase (1.5%) in performance can be gained on sufficiently deep networks when predicting visual memory schemas, without gathering more VMS data (a time consuming and expensive task). Despite the differences between the VMS4k and LaMem dataset, the model has learned additional features that relate to the memorable regions of scene images despite the LaMem dataset not being scene-focused. Interestingly, disabling training on

### 3.4. CONCLUSION

Backbone	Method	Dataset	True P2d	False P2d	LaMem ( $\rho$ )
None	Edge Detection	VMS4k	0.234	0.216	-
VGG16	vms-VAE	VMS4k	0.395	0.357	-
	DF-VMS	VMS4k + LaMem	0.425	0.374	0.552
ResNet50	DF-VMS-R	VMS4k + LaMem	<b>0.513</b>	<b>0.443</b>	0.466
	DF-VMS-R-xA	VMS4k + LaMem	0.497	0.435	0.444
	DF-VMS-R-xDF	VMS4k	0.488	0.423	-
	DF-VMS-R-xM	VMS4k + LaMem	0.497	0.418	0.446
	DF-VMS-R-xVMS	LaMem	-	-	0.28

Table 3.5: VMS reconstruction results. True & False refer to memorable and falsely memorable schemas (green/red in images).  $P^{2d}$  is the Pearsons 2D correlation [2, 83]. LaMem performance measured by Spearmans Correlation ( $\rho$ ). xA indicates no attention, xDF no dual-feedback, xM, no multi-scale information, xVMS, score prediction only. A dash in the table indicates the network does not compute that output. We include results for both a modern backbone, ResNet50, and for a fair comparison with prior work, a VGG16 backbone. A comparison with state-of-the-art is given against the current best model; vms-VAE from [83].

VMS4k leads to worse single-score performance; indicating that spatial memorability maps gathered from humans could be applied in future work to boost single-score prediction performance.

## 3.4 Conclusion

In this chapter, we have developed both new visual memory schema datasets, and new approaches for predicting visual memory schemas for arbitrary scene images. We show that our initial VAE model is capable of predicting Visual Memory Schemas for a given input image, and can generate both true and false VMS maps simultaneously at over ten times the resolution of previous approaches. Moreover, we find a very close correlation between the ground truth per-category metrics and the predicted per-category metrics, and finally show that current single-score memorability prediction does not appear to correlate with ground truth or predicted VMS metrics, and that these metrics do have a significant, but weak, positive correlation with ground truth memorability scores from the LaMem dataset. This indicates that VMSs can provide additional information about image memorability which is not traditionally captured by other memorability prediction methods.

## CHAPTER 3. VISUAL MEMORY SCHEMAS

Following on from this, we explored several different approaches for Visual Memory Schema map prediction. We examined the effect of depth in the scene, self-attention mechanisms, multi-scale blocks, and when varying the size of VAE latent-spaces for generating VMSes corresponding to both positive and false memory. We consider various performance metrics for all models in order to set a baseline for future work. We achieve state-of-the-art results for VMS prediction for deep learning architectures, such as VAEs and CNNs, when considering non-local self attention. Finally, we develop DF-VMS, a novel dual-feedback based Visual Memory Schema prediction model. DF-VMS model is trained both on VMS4k, a scene dataset with two-dimensional memorability information, and on an a single-score dataset, LaMem. Through ablation tests, we show that prediction of VMS maps is significantly improved by allowing the model to learn from existing single-score datasets, and additionally through the inclusion of self-attention and multiscale information. Interestingly, we also find that disabling memorability map feedback is highly detrimental to single-score prediction performance. Our model achieves state of the art performance, exceeding all our previous approaches, when predicting memorable or falsely memorable regions of a scene image, on a large memorability dataset of over 4000 scenes and VMS maps.

However, our contributions do not lay solely in deep learning VMS prediction models. By starting from an initial seed of 800 scenes, we first double this to 1600 images paired with visual memory schemas that follow the paradigm of the original experiment. We then develop a new continuous paradigm suitable for online experimentation that allows us to gather VMS maps in much greater quantities, developing a dataset of over 4000 scenes and VMS maps. We find from this data that category differences that are not immediately apparent from single score metrics appear when considering two-dimensional metrics, and that through modern segmentation techniques, we identify a human-readable "schema" for each category. That is, we extract the objects that, when appearing together, make a scene memorable. Our VAE model allows us to inspect the models learnt latent space and reason about whether scene features cause memorable images to group together, while exploring various theoretically promising techniques shows that we can boost VMS prediction in a significant fashion by considering state of the art computational methods. These techniques come together in our dual-feedback VMS model, which, allow us to set a new state-of-the-art for visual memory schema prediction, by taking advantage of existing large-scale single score memorability datasets. We

### 3.4. CONCLUSION

now consider whether we can use these predictors for more than *just* predicting VMS maps - can we instead use them to modulate human memory itself?

# Modulating Human Memory

## 4.1 Introduction

As we have seen in Chapter 2, cognitive science research of human visual episodic memory over the last few decades reveals both large storage capacity and a surprising ability to retain detail [15, 16]. Recent work at the intersection between the fields of machine learning and cognitive psychology have exposed another property of visual memory for images: consistency between observers [69]. Showing a set of images to a human population sample, most members of that sample will remember roughly the same subset of images. This implies that to a certain extent, image memorability (*i.e.* how likely the average person is to remember a given image) is an implicit property of the image itself. Image memorability does not correlate strongly with simple image characteristics such as colour, intensity, the number of objects present in the scene [69], or with attention, and is robust to overt cognitive influence [6]. Rather, high-level scene attributes help explain the memorability of images [68], such as the content of the image (for example the presence of "a person") or the dynamics occurring in the captured scene ("throwing a ball"). While memorability is affected very weakly by certain global features, such as average image hue and contrast, semantic context plays a stronger role, explored in [73]. These features are related to, but not completely explained by, objects present in the image [40], and specifically, their location and size [8]. These findings have led to attempts to predict image memorability using computational tools, which find the best predictor to be high-level semantics, such as the image scene category [68]. Later work established the influence of scene category and contextual distinctiveness on memorab-

ility [20], and current state-of-the-art models employ automatic deep feature extraction via convolutional neural networks (CNN) [49, 44]. The field of image memorability prediction has advanced to the point where CNN-based models can predict how likely an image is to be remembered with human-level consistency (Spearman rank correlation coefficient of  $\rho = 0.67$ ) [49].

Initially the majority of research studies framed the problem of image memorability prediction as regression to a one-dimensional score. Recent research results develop an understanding of memorability as a two-dimensional property that varies across an image, resulting in the extraction and analysis of cognitive relational patterns that capture the regions of scene images human observers deem memorable. These relational patterns, known as Visual Memory Schemas (VMS) [2], capture the cognitive representations and structures that humans use to organise and encode a given image into memory, have high consistency between humans ( $\rho = 0.70$ ), and a limited relation with one dimensional single score predictors for memorability. VMS internal consistency (measured via Pearsons 2D correlation) is higher than both VMS correlation with eye fixations ( $P^{2D} = 0.50$ ) or saliency ( $P^{2D} = 0.58$ ) [83]. Compared to image memorability prediction, fewer works tackle the task of modifying the memorability of images, or that of generating images that are intended to be less or more memorable. Modifying the memorability of face images was explored in [74], where it was found that active appearance models [28] could be employed to adjust various facial features associated with memorability. Deep generative models have also shown some success in modifying image memorability, from face generation [121], to employing style transfer [120], to transformer-based network capable of modifying the memorability of a seed image [49].

In this chapter, we present a generative model we call ‘MEMGAN’, capable of synthesizing completely new photo-realistic scene images by using two-dimensional maps of memorability. These maps are based upon cognitive relational patterns, which reveal the mechanisms humans employ to encode scene images in memory. We validate this approach by performing a repeat-recognition human experiment, and find that our generated images significantly modulate the memory performance of human observers. When designing our approach, we set out to verify that visual memory schemas (VMS) capture information that is memorable in an image to a sufficient degree to constrain a generative model that can synthesise completely new scenes which in turn are able to modulate human memory. We start with the analyses of the per-category consistency



and the per-category memorability signal (measured by D-Prime) for the VISCHEMA image datasets [133], and explore the relationship between consistency and memorability in order to verify that VMS maps are suitable descriptors of memorability. We then consider two deep learning generative adversarial techniques for generating memorable images : based upon the Wasserstein loss metric [4], and a progressively growing network. This approach allowed us to investigate what effect modifying the visual memory schemas the scenes were based on, has on the generated images. The generative neural network requires feedback on the memorability of its synthesised scenes during training time. The network generates hundreds of thousands of images during its training, and memorability feedback is necessary for every generated image. To deal with this constraint, we train a VMS prediction model based directly upon human data that can produce VMS feedback for arbitrary scene images. The predictor learns which features (from indoor scenes) make up a visual schema for our experimental kitchen scenes. It is this feedback that constrains the generative model. We evaluate our generated scenes via a human observer memory experiment, testing if our newly generated more memorable images are remembered better than the generated low memorability images. These findings allow us to acquire new insights into the efficacy of modulating the performance of human memory via images generated to activate specific visual memory schemas in human observers.

Developing the capability to generate memorable scene images without requiring an initial image seed has clear practical and theoretical applications. Such a technique could be applied to create highly effective memorable advertisements, improve educational tools, and there is also the potential for medical applications, such as tracking the decline in memory of patients with advancing cognitive deficits by providing a targeted baseline of memorability. A completely data driven approach such as this would provide significant advances to the methods used in cognitive science for the study of mental structures for the organisation of thought and behaviour employed by humans.

## 4.2 Results

### 4.2.1 VMS Consistency and Memorability

VMS maps capture spatial and relational components of episodic memory, and hence contain additional information compared to single-score based image memorability meth-

ods. In order to evaluate the validity of the data driven VMS maps as image memorability predictors we examine their image category consistency. We employ a method from signal detection theory to extract a global ‘memorability’ signal for each category for human observers and then evaluate the correlation between this global signal and VMS map consistency. The evaluation is performed on both the VISHEMA 1 dataset [2], which consists of 800 images and their corresponding 800 2D memorability maps, and the VISHEMA 2 dataset [83], an expansion to VISHEMA 1 which consists of another 800 images and memorability maps.

Category	Consistency	
	VISHEMA 1	VISHEMA 2
Isolated	0.556	0.447
Populated	0.624	0.562
Public Ent.	0.706	0.661
Work/Home	0.674	0.57
Kitchen	0.628	0.479
Living Room	0.568	0.446
Small	0.611	0.525
Big	0.637	0.595

Table 4.1: Vischema 1 and Vischema 2 consistency, per category. Certain categories of images, such as kitchens or scenes involving public entertainment (playgrounds, theme parks) are more consistent than others, such as the isolated category. Higher consistency implies participants agreed on specific features that made the image memorable.

The VISHEMA 1 and 2 datasets contain a variety of images, grouped in the following categories : Isolated, Populated, Public, Entertainment, Work/Home, Kitchen, Living Room, Small and Big. The consistency of the VMS maps, on a category-by-category bases for both VISHEMA 1 and 2 is presented in Table 4.1. The consistency is calculated by taking 25 splits of the data (one split creating two VMS maps for each image, each built from an equal division of human annotation data) and correlating the resulting VMS maps against each other, using the Pearson’s Correlation Coefficient. For all image categories the correlation is positive, and in many cases, strongly positive as is the case for the “entertainment” category, composed of images of fairgrounds and playgrounds. Observers tend to agree with each other on which regions allowed them to remember the image in the categories that show strong consistency signal.

CHAPTER 4. MODULATING HUMAN MEMORY

Category	D-Prime	
	VISCHEMA 1	VISCHEMA 2
Isolated	1.008	0.692
Populated	1.47	1.197
Public Ent.	2.037	1.813
Work/Home	1.896	1.38
Kitchen	1.602	1.257
Living Room	1.725	1.252
Small	1.52	1.4
Big	1.741	1.7

Table 4.2: D-Prime analysis of human memory for each category in the Vischema 1 and Vischema 2 datasets. High values clearly indicate that the memory signal for the given image category is strong and thus image memorability for human observers is high. Certain categories have stronger signals than others, possibly due to easier or more available encoding schemas for that category among the human participants.

$$D' = z(HR) - z(FAR) \tag{4.1}$$

In order to test that visual memory schemas can capture image memorability we calculate the signal strength of the observers' memory for the given images by using the sensitivity index, also known as the  $D'$  (D-Prime) measure. The sensitivity index,  $D'$  is a measure from signal detection theory that represents the strength of a given signal, in our case characterising the human observers ability to remember the given image. The equation is shown in Equation 4.1, where  $z$  is the z-transform. The results for the  $D'$  scores are provided in Table 4.2 and similar to the consistency of memorable regions show that not all image categories are equally memorable. Strong overall positive correlation between image memorability measured with  $D'$  and per category consistency of VMS maps for both VISCHEMA 1 ( $\rho = 0.83$ ,  $p < 0.05$ ), and VISCHEMA 2 ( $\rho = 0.76$ ,  $p < 0.05$ ) suggest a robust relationship between the two measures. When comparing this correlation for each image in each category, we also see a positive correlation, shown in Fig 4.1. The overall high VMS consistency and positive correlation with the image memorability signal (measured by  $D'$ ) indicates that VMS maps are a good descriptor of image memorability. In the following we refer to the combined VISCHEMA 1 and 2

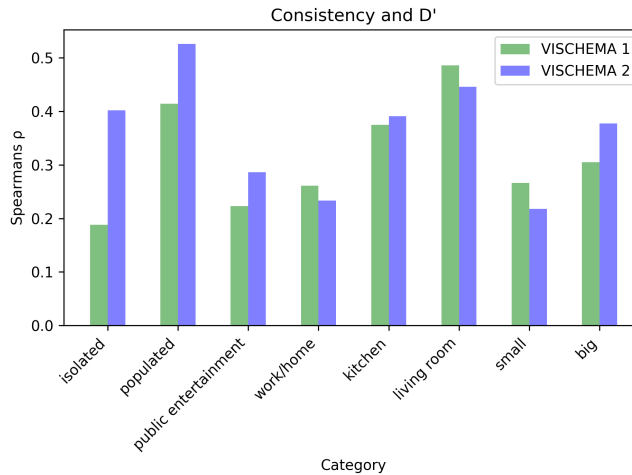


Figure 4.1: Histogram showing the correlation between per image category consistency for Vischema 1 and 2 datasets and human observers’ memory. Similar pattern of correlations between datasets indicates the reliability of using Visual Memory Schemas.

datasets as the VISCHEMA PLUS dataset.

#### 4.2.2 Generating Memorable Images Based on VMS Maps

In our study we start by developing two different deep learning network architectures for generating memorable complex scenes, one based upon the Wasserstein GAN [4] capable of producing images up to  $128 \times 128$  pixels and the ProGAN architecture [70] architecture capable of producing images of up to  $256 \times 256$  pixels resolution. In order to generate images of varying level of memorability we explore incorporating the data driven VMS maps into the deep learning training and generation algorithm in two different ways. The first is by considering it as a single score while the second is as a spatial map constraint in the loss function used for training the deep learning models. We evaluated these two different constraints in the Wasserstein GAN architecture by assessing whether the newly generated images can produce a differential score when applying a computational single-score artificial memorability predictor. Our ProGAN-derived architecture is capable of generating images of a sufficient quality and resolution for human observer experiments, with which we validate our approach.



Figure 4.2: Generated images when fixing  $\mathbf{Z}$ , where the sequence of generated images is displayed from left to right, while the memorability  $\mathbf{M}$  is varied from low to high. Shown categories include kitchens, cathedrals, and living rooms.

### Single-score constraint

The first implementation of VMS maps in a Wasserstein GAN architecture is in a form of a single score that is based on the average intensity of the VMS map (*i.e.* observer consistency) and is used to modify the memorability of the generated image. We hence refer to our Wasserstein-based memorability generation network as W-MEMGAN. We generate a range of images characterised by various levels of memorability, from low to high, by fixing the generators latent code  $\mathbf{Z}$ , which controls the semantic content of the generated images, and varying the memorability input  $\mathbf{M}$  to control the memorability of the generated images.

The newly generated images are created in ascending memorability in order to examine the variation space between exemplars of non-memorable and memorable images of a given image. Figures 4.2a and 4.2b show the generated images for different examples of scene from different scene categories, obtained by fixing  $\mathbf{Z}$  while varying  $\mathbf{M}$  from low memorability to high memorability. Just from visual evaluation of the images it is evident that clear differences emerge between images when increasing the memorability constraint. We can observe in all the scenes from Figure 4.2a, that as memorability increases, semantic details and a more realistic ‘kitchen-like’ appearance emerges. The low memorability cases appear to display semantic ‘noise’ representing a collection of mismatched features with loose spatial relations. The less memorable images may dis-

## 4.2. RESULTS

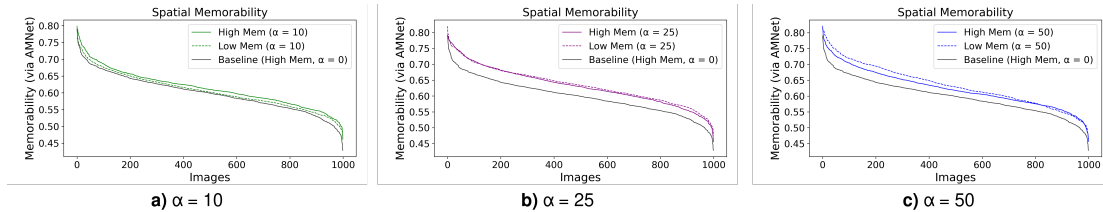


Figure 4.3: Predicted low and high memorability for different memorability weighting factors for  $\alpha$ , considering  $\alpha = 0$  (in Equation 4.4) as baseline. When increasing  $\alpha$ , all generated images have a higher memorability than the baseline. The most memorable images overall are obtained with  $\alpha = 25$ , but the best pairwise effect is achieved with  $\alpha = 10$ .

play the typical elements of a kitchen, but lack structure, or rather the correct spatial relationship between the elements. It appears that by defining visual memory schemas as constraints of memorability results not only in the appearance of memorable semantic details, but also enforces spatial relationships between these details. This lends evidence that VMS maps capture semantic details and structures which match learned schemas held in human cognition. From Figure 4.2b we can observe that when increasing the memorability, this results in a better image structure, clarity, and detail, resulting in images that better match human cognitive schemas.

In order to evaluate the newly generated images in a more quantitative fashion we generate 2000 images by setting the memorability constraint  $\mathbf{M}$  either to very low or to very high. This results in the generation of pairs of images where only the memorability information varies between the two generated images while having the same random seed  $\mathbf{Z}$ . These images are then evaluated using AMNet [44], an independent memorability prediction network. AMNet predicts the memorability of images on a scale between 0 and 1.0, allowing us to calculate the difference between our population of intended memorable and non-memorable generated images, while also allowing us to inspect the difference between the newly generated paired images. The results in Figure 4.4a show a statistically significant difference in memorability ( $p < 0.01$ ) between the two populations. Images generated to be memorable clearly show a trend to be predicted as more memorable compared to the baseline population of low-memorability images. To note is that not all the highly memorable generated images are themselves equally memorable independent from the memorability modulation as we have seen when looking at memorability across different scene categories. Thus, when examining the pairs of

## CHAPTER 4. MODULATING HUMAN MEMORY

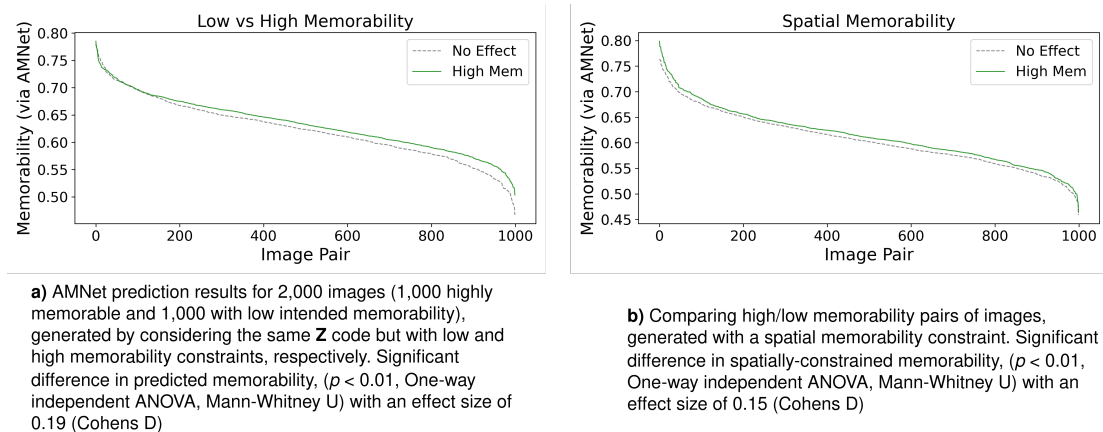


Figure 4.4: Differences in predicted memorability for low and highly memorable images generated with W-MEMGAN.

our generated images, we find that as overall image memorability decreases, it becomes more difficult to influence the memorability of certain scene image categories which are already not particularly memorable. When the image generated to be memorable has a predicted memorability above 0.65, then 79.5% of the pairs of memorable and non-memorable images have a positive difference in memorability. When memorability falls below 0.65, only 40.7% of the pairs have a positive difference in memorability, where a ‘positive difference in memorability’ indicates that the image generated to be memorable is predicted as more memorable than the image generated to be non-memorable.

### Spatial map constraint

A single score for an entire image does not capture spatial information about the memorability in the scene. As VMS maps reveal, not all regions of the image are equally memorable and in many cases memorability is concentrated on certain structures inside the image. We hypothesise that these carry semantic information that matches corresponding cognitive structures (schemas) used by the observers to encode and then retrieve information from long-term memory. It is highly unlikely that a single score represents the entirety of the memorability of an image. There are most likely multiple characteristics within an image associated and encoded with an episode of encountering that image. Instead, we hypothesised based on numerous findings from Cognitive Sci-

ence that how closely a viewed scene corresponds to a cognitive schema plays a role in image memorability or rather how much is an image memorable to a human observer. While single score methods might be able to predict image memorability, they do not reveal anything about why the image is memorable for a human observer, or which elements in it cause that image to be remembered. We instead base this constraint on the concept of a visual memory schema represented in two dimensions; an organisational map of semantic elements shared amongst human observers that enable the encoding and recognition of scenes.

This method naturally lends itself to a two-dimensional representation of image memorability; the regions captured inside a visual memory schema map are thought to directly represent the semantic elements that lead to that image’s encoding and recognition. These elements correspond with schemas held in the brain; cognitive structures that represent the typical elements (and arrangement of elements) of a scene. A human, through life long experience and acquired knowledge, may construct a schema of a kitchen, learning that a kitchen may contain countertops, an oven, and kitchen appliances (this is an example; real schemas are likely more complex and flexible). Scene images that better match this mental schema in both arrangement and semantic presence have an encoding advantage against kitchen scenes that lack these elements or arrangements. Computational measures that employ visual memory schemas can be thought of as learning a method to replicate human scene memory that more closely mirrors the method the human brain uses to encode scene images; the visual schema.

Hence, to take advantage of the 2D characteristics of VMS maps, we modify W-MEMGAN to take as input a  $10 \times 10$  pixel map describing the intended spatial memorability of the generated image. The provided input are artificial VMS maps created using a deep learning method trained on VISHEMA 1 and 2 (VISHEMA PLUS), similar to those obtained from human observers. As with single-score VMS constrained memorability, employing artificial 2D VMS maps to alter the memorability of generated images also results in a statistically significant difference between populations of 1,000 generated memorable and 1,000 generated non-memorable images, shown in Figure 4.4b. These findings indicate that both single-score and spatial constraints extracted from the VMS maps incorporated into our W-MEMGAN architecture are capable of modulating the memorability of newly generated images evaluated by an artificial memorability predictor such as AMNet. The non-spatial single-score implementation of the VMS results in a



## CHAPTER 4. MODULATING HUMAN MEMORY



Figure 4.5: Memorable, shown within green boundaries and non-memorable, shown within red boundaries generated image pairs. Foils are shown within blue boundaries.

greater effect size of the difference in memorability (0.19 *vs* 0.15, Cohens D) compared to the spatial method. We postulate that this could be the result of additional difficulty of integrating a spatial constraint compared to calculating a single score constraint for the entire image.

To examine the effect of the feedback strength of the memorability feedback mechanism we tested several different values for  $\alpha$ , the hyper-parameter which controls the ‘strength’ of the mechanism and defines how strongly we intend memorability to affect our generated images. The effect of four different values of  $\alpha$ : 0, 10, 25, and 50 on the prediction of low and high memorability in generated images is shown in Figure 4.3). For  $\alpha = 0$ , the memorability predictor provides no feedback to the network, disabling the influence of VMS maps and hence is used as a baseline. Best results are achieved for  $\alpha = 10$ , resulting in the clearest difference between high and low-memorability images and noticeably above those of the baseline, Fig. 4.3a. Using an  $\alpha = 25$  resulted in generation of images with high memorability scores but with reduced ability to discriminate between the low and high memorable exemplars. Higher values for  $\alpha$  prevent the W-MEMGAN from distinguishing between high and low memorability images, instead just raising the memorability of every image generated by the network, as shown by the results from Figures 4.3b and 4.3c.

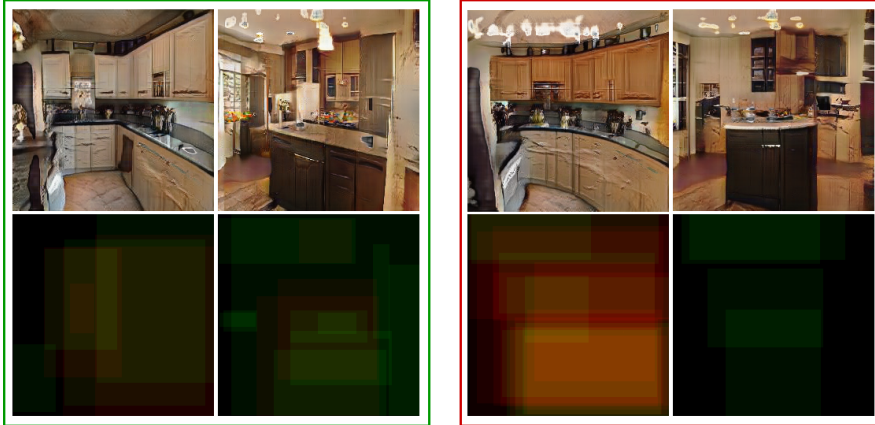


Figure 4.6: Generated high-memorability images (left) and their low-memorability pairs (right). VMS maps for each image are shown on the bottom row.

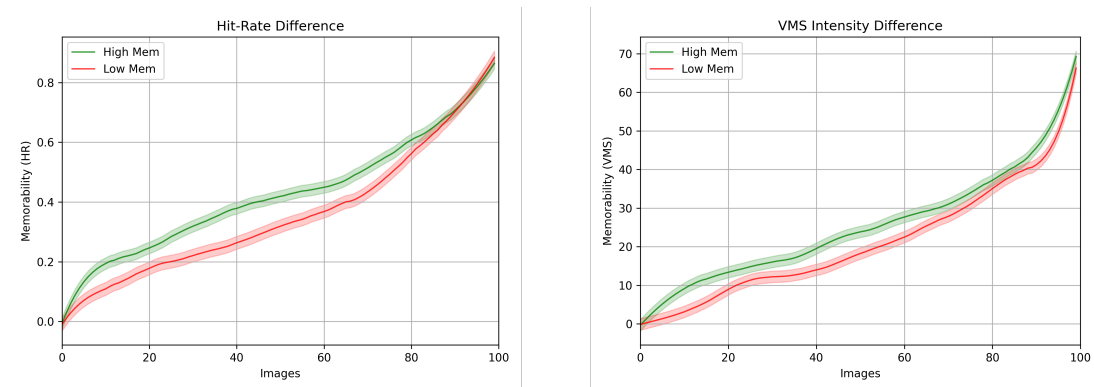
### 4.2.3 Human memory performance for generated images

We test the feasibility of directly modulating image memorability using VMS maps by conducting a visual memory experiment with human observers. The images used in the experiment were generated by our second architecture; based upon ProGAN [70] combined with memorability feedback, which we term ‘MEMGAN’. MEMGAN enables the creation of higher resolution images at a much higher quality than the W-MEMGAN architecture. Based on our previous results, we weight the memorability constraint to a value of  $\alpha = 10$ , which gives the best partition between memorable and non-memorable generated images. Examples of images generated are shown in Figure 4.5. In the experiment, human observers were asked to view a stream of generated images presented for 3 second each one at a time. Participants were asked to recognize images they recognized as repeats and indicate upon identifying a repeat the areas in the image that made them remember the image. This allowed us to evaluate the memorability of generated scenes through the hit rate of the images (how often an image was successfully recognized as a repeat) and the consistency of the VMS maps across observers (regions in the image indicated as memorable areas, see the examples from Figure 4.6).

To determine how consistent our participants were with each other, we take 25 equal splits of our visual memory schema map and hit data for each split, and then compare them against one another. We find a hit-rate consistency of 0.3 (Spearman’s  $\rho$ ,  $p < .0001$  for all splits) and an overall VMS map consistency of 0.38 (Pearson Linear Correlation

## CHAPTER 4. MODULATING HUMAN MEMORY

Coefficient). The VMS map consistency is lower than the 0.67 presented in [2], but this is expected given that our task contains only two different categories of images, and is thus a very homogeneous stimulus set. Nonetheless, there exists a clear consistency between participants.



**a)** Comparison between the hit-rates HR of images generated to be high-memorability against those generated to be low-memorability. Images sorted by ascending memorability. Significant difference in hit-rates ( $p < 0.05$ , One-way ANOVA, Kruskal-Wallis, Mann-Whitney U)

**b)** Comparison between the VMS intensity of images generated to be high-memorability against those generated to be low-memorability. Images sorted by ascending memorability. VMS intensity correlates with overlapping labelled memorable regions, and hence indicates more memorable regions. There is a significant difference in VMS memorability ( $p < 0.05$ , One-way ANOVA, Kruskal-Wallis, Mann-Whitney U)

Figure 4.7: Difference in memorability (HR and VMS Intensity) for generated image populations. Degree 3 polynomial fitted for visualisation.

### Differences in Observed Memorability for Generated Images

The evaluation of the hit rates for the generated high and low memorable images, (Figure 4.7a, shows that high memorable images result in both average higher hit rates (0.45 for high and 0.39 for low memorable) and average higher false alarms rate (0.20 for high and 0.16 for low memorable). However there is a statistically significant difference ( $p < 0.05$ ) for the hit rates but not for the false alarms rates between generated images as highly memorable and those with low memorability. This pattern of results, of a robust difference in hit rates and a lower difference in false alarms, is expected [83, 38] given that the same structures that enable easier encoding of a scene, also make it more likely for a human to believe they have seen that scene. Indeed, within the VISHEMA image set with which the memorability evaluator was trained, a rise in memorability corresponds with a rise in false memorability ( $\rho = 0.19$ ,  $p < 0.001$ ).

### Differences in Observed Visual Memory Schemas for Generated Images

Comparison of differences in Visual Memory Schemas between high and low memorability images required that we first condense each VMS map down to an average intensity. Outliers beyond two standard deviations of the mean were excluded. This gives us separate values for both images that were correctly recognised as seen before (true memorability) and those misrecognised (false memorability). There is a statistically significant difference ( $p < 0.05$ , effect size 0.36 Cohens D) in the VMS memorability channel for highly memorable images *vs* low memorable images, with a robust Bayes factor  $\ln(BF) = 1.117$  indicating substantial evidence for the effect of modulating memorability (see Figure 4.7b). As before, there is no statistical difference for false memorability between image categories. We also compared predicted VMS maps from the generator network to those outlined by the human observers and found a Pearsons Correlation of 0.49 ( $p < 0.05$ ), a Spearman rank correlation of 0.5 ( $p < 0.05$ ), with a population average mean-squared error of 58 between predicted and human-gathered VMS maps. These results indicate that the relational memorability patterns used to generate the images are effective in defining visual memory schemas in the same images, which correlates positively with those indicated by human observers.

### Image Pair Analysis

The results for both the hit rate and observed VMS’s are encouraging, and clearly show a statistical difference between the overall populations of highly memorable and low memorability generated images. However, the images were generated in pairs, with high and low memorable versions of the same scene image, as defined by a fixed latent code, and by modulated memorability, and thus can be compared as such. This requires the evaluation of the pair-wise difference, between an image with the same latent code but modulated memorability. Results indicate a statistically significant difference ( $p < 0.02$ , paired T-test, Wilcoxon signed-rank test) for both hit rates and VMS memorability.

### Comparing our results with an independent computational predictor of memorability

We also compare the results obtained on our human observer study with those of a recent state-of-the-art memorability predictor [44], trained on the same dataset (LSUN, [142]) from which we drew the training data for our MEMGAN models. We find no significant

## CHAPTER 4. MODULATING HUMAN MEMORY

Table 4.3: Comparison between MEMGAN and GANALYZE.

Method	Primary Focus	Log-odds increase	Constraint	Seed Image Required?	Pretrained Generator Required?
GANALYZE [49]	Objects/Animals	0.19 / step	Single-score	Yes	Yes
MEMGAN	Indoor Scenes	0.31	2D Map	<b>No</b>	<b>No</b>

correlation between the memorability scores calculated by the independent memorability predictor for our images with the experimentally obtained hit rates or VMS map intensity of our images. This finding suggests that memorability predictors based only on single-score models of memorability are missing important characteristics of human visual memory, and that artificial predictors fail to predict human memory performance for generated images. However, we do see a significant effect ( $p < 0.05$ , Paired T-test, Wilcoxon signed-rank, Mann-Whitney U), when comparing paired predicted scores for our generated high and low memorability images, which suggests computational predictors can differentiate between populations of generated images.

### Comparison with prior work

Comparison with prior work is made difficult due to both inter-experiment paradigm differences and differences in the datasets employed by previous work compared to ours. We cannot compare with work that examines face memorability, as memory for faces employs a different mechanism than that of scenes [115]. The most sensible comparison of our work is with that of Lore *et al.* [49], where a transformer is employed to shift the memorability of images within the latent space of a BigGAN [18] network. However, our experiment is more difficult than that of [49] for several reasons. Our stimulus set is indoor scene focused, rather than consisting of objects and animals, and hence is more homogeneous compared to [49], which makes remembering our images more difficult [20]. Secondly, the gap between target and repeat is, on average, longer than the longest gap in the memory experiment of [49], again making the task more difficult. Thirdly, we must manipulate the semantic content of entire scenes, whereas in [49] the differences in object memorability are due to changes in object size, brightness, object centeredness, and object shape. Most of these factors cannot be manipulated to make *scenes* more memorable. Henceforth this comparison should be taken with these key paradigm differences in mind.

We employ the same method as in [49] to calculate the log-odds difference between our

low and high memorability image sets for our scenes. In Table 4.3, we show the log-odds increase, which captures how much more likely a "high memorability" image is to be remembered by a human than a "low memorability" image. We also indicate the memorability constraint type, whether an initial seed image is required to be modified, and if a pretrained generator is necessary. We find the log-odds of remembering an image in the "high" category increase by 0.31 compared to those from the "low" category. Despite our harder memorability task, our results are comparable to that of [49], while being able to train in only 14 days on 4x Nvidia 1080 Ti, compared to BigGAN requiring 15 days on 8x Nvidia V100s (significantly more powerful GPUs). Our approach also does not require an initial "seed image" to modify. While the memorability of scenes and objects cannot necessarily be directly compared, the log-odds increase being comparable between both approaches is additional evidence that Visual Memory Schemas are good descriptors of memorability.

### 4.3 Discussion

In this chapter we presented and evaluated a method of generating scene images constrained by a construct from cognitive science: visual memory schemas, and tested its validity to modulate human episodic memory of images. The modelling of the VMSs is data driven and based on human memory study. We directly manipulate the visual schemas of images in a generative deep learning model (MEMGAN) and hence influence the final memorability of generated images. To our knowledge this is the first example of a generative model specifically trained from scratch to generate memorable scene images employing two-dimensional memorability data gathered from human observer experiments. Moreover, we double the size of an existing two-dimensional memorability dataset, and for the first time investigate the relationship between VMS map consistency and image memorability, along with presenting per-category consistency data for VIS-CHEMA categories. Encouragingly, consistency values remain high for both the original VIS-CHEMA dataset and our second replicated experiment, confirming the validity of this approach of gathering two-dimensional memorability maps.

There is currently a limited number of existing approaches to the problem of modification of image memorability. Sidorov *et al.* [121] examine various methods for altering the memorability of images, from basic-photo editing techniques such as adjusting the sat-

## CHAPTER 4. MODULATING HUMAN MEMORY

uration of the image, to the employment of an attention-based Generative Adversarial Network (GAN) for generating memorable face photographs. The memorability data used as input for training the GAN was drawn from artificial memorability predictors. They find that both their altered and generated images produce changes to the artificially predicted memorability score of images but do not have any data on human observers. This approach is similar to that from [120], in which a deep style-transfer model was trained to automatically apply ‘filters’ such as sepia tones or saturation boosters to images in order to boost said images memorability. In [49], Lore *et al.* develop a transformer network that can be attached to an existing generator in order to adjust the memorability of generated images. While this approach does leverage existing trained networks to generate photorealistic images, this is dependent upon the chosen generator, and additionally requires a generated ‘seed’ image for the network to adjust. As a feedback mechanism they employ single-score memorability predictors. They show through human recognition trials that the images adjusted to be more memorable tend to be empirically more memorable.

Prior approaches to memorability modification require a starting image (real or generated), and it is the memorability of this image that is then modified. We instead desire to create an approach that can generate images without requiring this initial seed image, and can instead synthesise recognisable scene images given only a latent code and a desired memorability. As proof of concept that cognitive relational patterns can serve as the basis for a generative network for memorable scene images we develop and train an architecture that can synthesise low-resolution memorability-constrained images. We evaluated the potential of both single-score VMS based memorability and spatial memorability as a driving mechanism for scene image generation. The generated images, despite low resolution and relatively poor quality, are capable of causing a significant effect in a state-of-the-art third-party memorability prediction network that had never previously seen the generated images. Further, by modifying the strength of the memorability feedback mechanism our memorability constrained images can be made to display both higher *and lower* image memorability compared to a baseline of generated images where the memorability feedback network is disabled. Interestingly, placing too much emphasis on the feedback network causes the network to lose discriminative power, becoming unable to correctly generate images with high vs low memorability, yet generating images that were predicted to be much more memorable overall. This provides

### 4.3. DISCUSSION

evidence that we can manipulate the memorability of generated images in a meaningful way. Image memorability based on VMS maps appears to control both the emergence of semantic details as well as the spatial relationships created between these details.

The final test of visual memory schemas as viable mechanisms for modulating image memorability is whether our approach could functionally work with actual human observers and not only with computational memorability predictors. By integrating our memorability evaluator and loss component with a more advanced generator allowed us to influence the memorability of relatively high resolution, high-detail images. While we lack the resources to generate high resolution photo-realistic images, the images we do generate show clear structure, detail, and are certainly recognisable as belonging to their intended category. From our results (and given that we based our model constraint on visual memory schemas), we hypothesise that more memorable scenes better match the cognitive schema of that scene contained within the human mind. In this work, we observe that making a scene more memorable results in changes to the structure and content of the generated scene compared to the same scene generated to be of a lower memorability. We hypothesise these differences cause the image to become closer or further from the mental representation (i.e, the schema) of that scene which is stored within the human brain. However, it is unclear whether the differences in schema between high and low memorability images also result in greater difficulties visually recognising the image as an exemplar of its class. To test this, we employed a scene recognition deep neural network (ResNet152 [58]) trained on the Places365 dataset [148] to categorize every synthesised image. We find that in general the majority of images are classified as their class (or a highly similar, related class, e.g, galley vs kitchen). For highly memorable generated images, 96% of images are correctly classified and for low memorability generated images, 95% of them are correctly classified as kitchens by the scene recognition network. There appears to be little difference in how visually recognisable the generated images are as members of their class; and the demonstrated memorability effect appears independent of visual recognizability. Given that human ability to categorize scene images generally exceeds that of neural networks; the recognizability scores shown are best viewed as a lower bound. Additionally, as human memory is not contingent upon resolution [127, 48, 139] and perfect photo-realism, the images we generate serve well for their intended purpose. We show through human observer memory experiment that the images we generate to be more memorable are more likely to be detected



## CHAPTER 4. MODULATING HUMAN MEMORY

correctly as repeated images by humans. Additionally, we find that the false-alarm rate of said images also increases, an encouraging sign that our images are truly modulated by visual memory schemas, as the exact same effect appears in the VISHEMA dataset of real images. The cognitive schemas that aid the remembering of scenes also lead to false remembering when presented with a memorable image modelled on the schema, even if that image has never been seen before. Critically, we are able to generate memorable images without requiring a seed image, such as the approach employed in [49], and verify that two-dimensional maps of memorability can be employed to modulate memorability, rather than relying on single-score approaches.

In summary, our results indicate we were able to both fool computational memorability predictors, and manipulate human visual long-term memory via artificially generated images, constrained with a two-dimensional visual memorability schema concept borrowed from cognitive psychology, for which there are neural correlates [128]. It may appear circular that we have constrained a model with visual memory schemas, (which indicate memorable regions) and find that our generated images are indeed memorable. However, this only appears this way because the data shows an effect on human memory; there was no guarantee that this was possible to accomplish. There is additionally no guarantee that the generative model would be able to be constrained by the visual memory schemas. There is little work in this area (and none that examines the visual schemas of generated images); and in essence the model is the test - investigating whether it is, or it is not possible to use visual memory schemas to synthesise scenes that can modulate human memory. We find that by employing VMS maps we are able to generate completely new artificial scenes that cause a desired modulation of human memory as tested by a human observer memory experiment. This has interesting implications for the future study of image memorability, as well as real-world applications for memorability research.

We have designed a neural network that appears to understand visual memory schemas to a sufficient enough degree to use them to visibly change the output of a generated scene, based upon a brand new, extrapolated or invented schema (of controllable memorability), that we want the scene to match. The generative network is constrained by an artificial VMS map predictor that can produce two-dimensional memorability maps for arbitrary scene images; the greater the difference between the predicted VMS map and the target VMS map for a synthesised scene, the more the network is penalised. As we have shown,

when we constrain a synthesis network with a VMS predictor, we find that we are able to generate scenes that affect human memory for those scenes, or rather, affect their the performance on a memory test. We learn from this that visual memory schemas appear a strong enough descriptor of what information humans encode into memory to enact visible changes on the synthesised images based upon the input schema.

## 4.4 Methods

### 4.4.1 Memorability Estimation Feedback Network

The assessment of image memorability is performed by employing a Visual Memory Schema prediction model developed in [83], which is based on the Variational Autoencoder (VAE) [76] learning model. A VAE is made up of two convolutional networks: the encoder aiming to extract a latent space representing the data, and the decoder which aims to reconstruct the given data. Following training, given an image, the VAE is used to predict its corresponding VMS map. We train this model on the VISHEMA PLUS dataset containing 1,600 image/VMS pairs. The output of this model is a two-dimensional VMS map. This predicted VMS map is based upon the latent space of the VAE, which corresponds to a learnt mapping of image features to memorability based upon multiple human observations for the input image. We only consider the ‘memorability’ channel of the VMS maps (true schemas), and do not make use of the ‘false memorability’ (false schemas) information. For the given VMS data ( $\mathbf{x}$ , the encoder of the VAE infers a latent space  $\mathbf{z}$ , by using the following loss function :

$$L(\theta, \phi) = -E_{\mathbf{z} \sim q_{\theta}(\mathbf{z}|\mathbf{x})}[\log p_{\phi}(\mathbf{x}|\mathbf{z})] + KL(q_{\theta}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \quad (4.2)$$

where the former term represents the log-likelihood of VMS reconstruction by using the decoder network and the latter represents the Kullback-Leibler (KL) divergence between the variational distribution  $q_{\theta}(\mathbf{z}|\mathbf{x})$  and the prior  $p(\mathbf{z})$  aiming to assess the image reconstruction ability of the network.  $\theta$  and  $\phi$  represent the parameters of the VAE’s encoder and decoder networks, respectively.

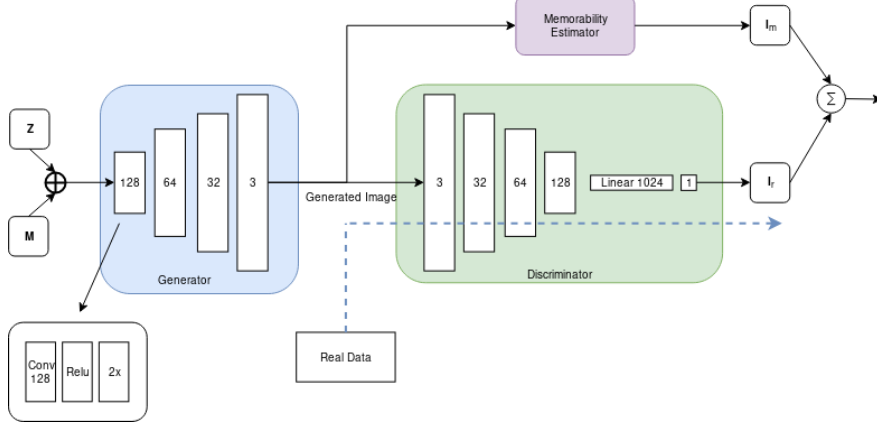


Figure 4.8: Memorability-constrained image generation model architecture. PixelNorm and Minibatch Standard Deviation layers omitted for clarity.

#### 4.4.2 W-MEMGAN Architecture & Training

The diagram of the deep learning architecture used for generating memorable images is shown in Figure 4.8. It consists of a Generator  $G$ , a Discriminator  $D$ , and the memorability feedback network  $\mathcal{M}$ . While the generator creates memorable images, the discriminator evaluates the ‘realness’ of the generated images, and the auxiliary memorability network evaluates whether the memorability of the generated image matches the memorability defined by a memorability constraint  $\mathbf{M}$ .  $\mathbf{M}$  in this case may either be a two-dimensional target VMS map, or a single target memorability score. The image generation network  $G$ , corresponding to the generator from WGAN, aims to synthesise an image  $\hat{\mathbf{I}}$  using random variables  $\mathbf{Z}$  as inputs, which defines the latent space of the MEMGAN, while  $\mathbf{M}$  acts as the memorability constraint :

$$\hat{\mathbf{I}} = G(\mathbf{Z}, \mathbf{M}). \quad (4.3)$$

The output of the generator is a generated image  $\hat{\mathbf{I}}$ , whose memorability score is as close to  $\mathbf{M}$  as possible. Both  $\mathbf{Z}$  and  $\mathbf{M}$  are drawn from Gaussian distributions. The generator is constrained by both the discriminator  $D$  and by the memorability feedback network  $\mathcal{M}$ , which estimates the memorability map  $\hat{\mathbf{I}}_m = \mathcal{M}(\hat{\mathbf{I}})$ . The discriminator  $D$  is implemented as an improved Wasserstein GAN model [54] which employs a penalty term on the discriminator loss yielding better performance and stability when compared

to the classical GAN [51].

During the training,  $\mathbf{Z}$  is sampled randomly from a Gaussian distribution, and  $\mathbf{M}$  is either sampled from the Gaussian distribution or  $\mathbb{P}_t$ , the distribution of possible target VMS maps, depending whether the network is being trained for spatial memorability or single-score memorability. When training the discriminator  $D$ ,  $\mathbf{M}$  is discarded, as it is only necessary for training the generator, where it is used to calculate the memorability loss. This has the effect of penalising the generator if the generated images are not of a similar memorability to that defined by  $\mathbf{M}$ . For example, if the image was intended to be memorable while actually it is not memorable, the generator loss will increase. Each training epoch consisted of 60,000 kitchen images drawn from the LSUN database [142], and the network was trained for 500 epochs, which took approximately 8 days on  $4 \times$  Nvidia 1080 Ti GPUs.

#### 4.4.3 MEMGAN Architecture & Training

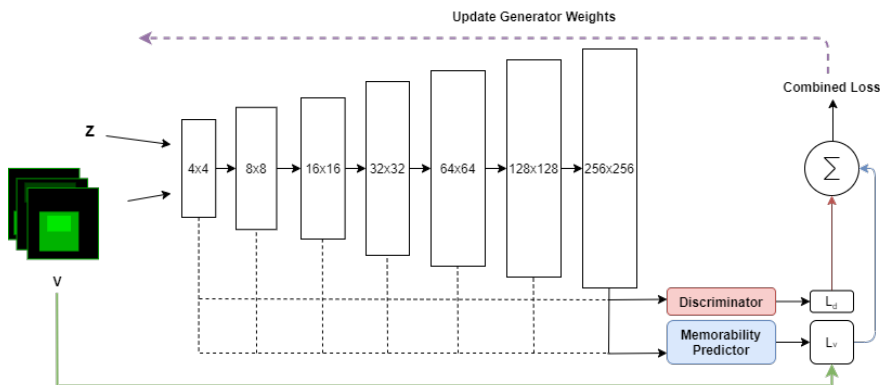


Figure 4.9: Progressive generator with per-resolution memorability estimation.

The Wasserstein GAN based network does not generate memorable images at a sufficiently high resolution and quality for human trials. Given the latent code  $\mathbf{Z}$  and an artificially generated target 2D visual memory schema (VMS) map  $\mathbf{V}$ , the goal was to generate a sufficiently realistic  $256 \times 256$  pixel image from  $\mathbf{Z}$ , whose VMS is close to that of  $\mathbf{V}$ . We hence combine the memorability feedback network with a more suitable generator architecture, that of the progressive GAN [70]. We draw  $\mathbf{V}$  from  $\mathbb{P}_t$ , the possible target VMS maps.

## CHAPTER 4. MODULATING HUMAN MEMORY

While aiming to obtain photo-realism is preferable, it is not a strong requirement for our architecture and subsequent human observer experiment. As long as the image generated is recognisably as a member of its target category, a human observer will employ the correct visual schema when encoding the image into memory. This allows us to reduce the capacity of the network compared to the original progressive GAN [70], which results in an accelerated training time on the available hardware. A simplified (for visualisation purposes) architecture is shown in Figure 4.9. The MEMGAN architecture we develop bears superficial similarities to both ACGAN [101] and InfoGAN [25], though rather than predicting discrete class labels or extracting interpretable dimensions in an unsupervised fashion, it generates memorable images, without a prerequisite seed image (such as those used in [49]), while being supervised by human observer-based cognitive structures. The generative network architecture has specific processing blocks for each image resolution, as can be observed in Figure 4.9. The output image of each resolution block is passed through the memorability predictor as the network generates more accurate images of increasing resolution. As each resolution block takes over the information produced by the previous layer of processing blocks, the connection of those blocks to the memorability predictor is dropped. This allows the memorability signal to affect all resolutions of the generator during training. We only generate up to a resolution of  $256 \times 256$  to limit the computation time, which is ever increasing when attempting to generate images of higher resolutions. The training time is reduced at the cost of losing some detail by reducing the capacity of the  $256 \times 256$  and  $128 \times 128$  resolution blocks by half. Finally we add a *tanh* activation function at the output, before merging different resolution blocks, which aids stability.

We trained two deep generative networks in order to generate images for our human memory experiment, one with a memorability constraint (MEMGAN) and another without, whose purpose was to generate foil images. Both networks were trained for 200 epochs. Each resolution block was slowly introduced to the network over a duration of ten epochs, and then trained for an additional ten epochs before the next resolution block was introduced. Each of the first five resolution blocks of  $128 \times 128$  pixels from Figure 4.9, was shown a total of 4,800,000 images. The final block of  $128 \times 128$  was shown 2,400,000 images. These images were drawn from a dataset of 240,000 kitchen scene images, and the same number of living room scene images, both drawn from the LSUN database [142]. This allowed a suitable balance between resolution, quality, and

required total training time. We follow the example set in [77] with the following parameters:  $Lr = 0.0015$ ,  $\beta_1 = 0$ ,  $\beta_2 = 0.99$ ,  $e = 1 \times 10^{-8}$ . Each of the two networks was trained for 14 days on  $4 \times$  Nvidia 1080 Ti GPUs.

#### 4.4.4 Loss functions

Both our memorable image generators are designed to use the same loss function, the Wasserstein metric combined with a component which calculates the difference between the desired and generated memorability for a given image. This training mechanism works for both single-score and VMS map memorability training examples.

$$L = \mathbb{E}_{\hat{\mathbf{z}} \sim \mathbb{P}_z, \mathbf{v} \sim \mathbb{P}_t} [D(G(\hat{\mathbf{z}}, \mathbf{v}))] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_r} [D(\mathbf{x})] + \lambda Loss_{gp} + \alpha \mathbb{E}_{\hat{\mathbf{x}} \sim \mathbb{P}_g, \mathbf{v} \sim \mathbb{P}_t} [(\mathcal{M}(\hat{\mathbf{x}}) - \mathbf{v})^2] \quad (4.4)$$

The loss function is designed to embed a memorability predictor and contains the following components: a generator network  $G$ , a discriminator  $D$  and memorability predictor network  $\mathcal{M}$ . Considering the latent code distribution  $\mathbb{P}_z$ , target VMS distribution  $\mathbb{P}_t$ , real image distribution  $\mathbb{P}_r$ , predicted VMS distribution  $\mathbb{P}_v$ , and generated image distribution  $\mathbb{P}_g$  based upon the latent code  $\hat{\mathbf{z}}$  and  $\hat{\mathbf{v}}$  we define the loss function in Eq. (4.4). The latent code  $\hat{\mathbf{z}}$  is drawn from a Gaussian distribution and  $\mathbf{v}$  from a distribution of target VMS maps, where height, width, and intensity of VMS regions is drawn from a uniform distribution.  $\lambda Loss_{gp}$  refers to the gradient penalty loss in [138].  $\alpha$  controls the strength of the memorability loss.  $\mathbb{P}_g$  represents the probability of the generated data and  $\mathbb{P}_r$  is the probability of the real data. The additional term controlled by the hyperparameter  $\lambda$  prevents the gradients inside the discriminator from violating Lipschitz continuity, whereas the first two terms evaluate the Earth-Mover distance between the generated and real distributions. The additional memorability loss, combined with the Wasserstein loss, constrains the image generation by both ‘realness’ and memorability simultaneously.

#### 4.4.5 Generating Images for Human Observer Experiments

We generated low-memorability and high-memorability kitchen images with our MEMGAN. To avoid making the task too difficult, we also generate memorability-unconstrained im-

## CHAPTER 4. MODULATING HUMAN MEMORY

ages of another interior scene category to act as foils in the memory experiment, living rooms. Our target memorability-modulated images are generated in pairs, with a fixed latent code  $Z$  per pair, varying the desired target VMS map between low and high memorability (modulating memorability constraint  $\mathbf{M}$ ); for each highly-memorable image there is a non-memorable image defined by the same latent code. We generated several hundred pairs and additionally memorability-unconstrained images as foils. Foil images and target image pairs suffering from extreme distortion were excluded from this study to avoid differences in memorability being caused by drastic quality differences between categories. In order to avoid any bias in the selection of images, if one image of a target pair is of acceptable quality to be included in a human trial, then the other image of the pair is automatically included as well. What is more there is little chance of biasing the images for memorability one way or another, as it has been shown in prior studies that humans cannot intrinsically predict the memorability of any given image [68].

We selected 100 pairs of high and low-memorability generated images, for 200 memorability-constrained images overall. We additionally selected 200 generated living room scene foils of suitable quality. The resulting 400 images were used as a stimulus set for the human observer memory experiment that tested the validity of our memory modulation. We quantified the image quality differences by employing the Fréchet Inception Distance (FID) [61] and note minimal differences between categories. The high-memorability images have a FID of 108 and the low-memorability images a FID score of 104, while the non-constrained images (foils) had a FID score of 88. Based on these minimal differences, it is highly unlikely that differences in image quality are affecting the memorability of our images. The VMS training datasets can be found at <https://www.cs.york.ac.uk/vischema/>

### 4.4.6 Human Memory Experiment

With our stimulus set of generated images, we conduct a human recognition memory experiment with 119 participants. Each participant saw one of ten unique sequences of 150 images, with 40 targets (*i.e.* repeats of once presented images in the sequence) per sequence. Both foils and target images could be repeated. Each image was shown on-screen for 3 seconds. Where an image was repeated, we ensured a minimum of at least 30 images between first showing and repeat (see Figure 4.10). Each sequence was viewed on average by 12 different participants. Participants were asked to indicate by pressing

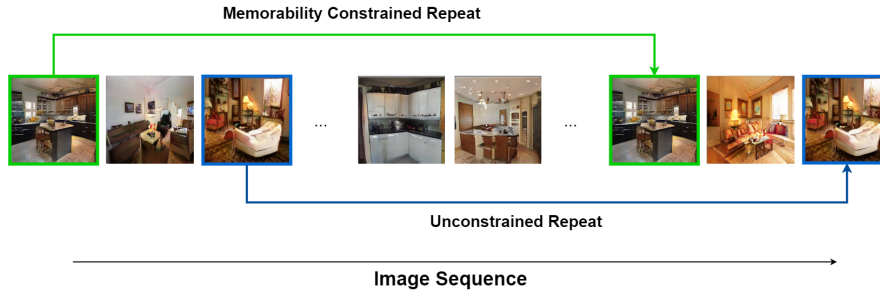


Figure 4.10: Memorability experiment structure.

a button when an image they were viewing was a repeat. If they correctly indicated a repeat it was considered a hit and if not, then a miss. When participants indicated that they recognised a repeat, they were asked to annotate the regions in the image they believe caused them to remember the image. This allowed us to gather two-dimensional memorability maps for our generated images. Each sequence took approximately 9 minutes on average for a participant to complete. We then analyse these results through several statistical tests, primarily a one-way independent Analysis of variance (ANOVA). We also employ Mann-Whitney U tests to verify that our effect occurs in the intended direction, and Kruskal-Wallis and Wilcoxon signed-rank tests to verify that our results hold if distribution assumptions are relaxed. Critically, no single participant was shown both the high-memorability and the low-memorability image of a given pair in the same image stream. This prevented recognition of images by previously viewing the same image with a different memorability value, rather than remembering a repeat of the target. Participants were paid at a rate of \$7.02 per hour using the crowdsourcing platform Prolific, and prescreened such that all participants were between 18 - 65 years of age and fluent in English. This experiment was approved by the Departmental Ethics Committee of the Dept. of Psychology, University of York, UK, and follows relevant guidelines given by that committee. Informed consent was given by participants, and they were free to withdraw at any time.

#### 4.4.7 Evaluating Scene Recognition Differences

To determine whether there was any difference in recognizability between high and low memorability images that has arisen due to differences in their structure, we employ a deep neural network. We select a ResNet152 model [58], which has been pretrained on



## CHAPTER 4. MODULATING HUMAN MEMORY

the Places365 [148] dataset, such that it can classify images into one of 365 different scenes. If the network predicts the image is a kitchen (or kitchen related) we record this as a successful recognition of the scene. As the network predictions can be specific to the *type* of kitchen (for example, the network is capable of differentiating a ‘galley’ style kitchen from a ‘wet-bar’) we select a set of categories that closely relate to kitchens; and assume any prediction in this set is a correct recognition that the image shown is a kitchen. This subset consists of: ‘kitchen’, ‘wet\_bar’, ‘galley’, ‘restaurant\_kitchen’, and ‘sushi\_bar’. We then accumulate predictions by running all images from both the low memorability category and high memorability category through this network, and record the predictions. We find that the low memorability category has a recognition rate (correct predictions) of 95%, and for high memorability, a recognition rate of 96%.

### 4.5 Summary

In this chapter we have presented and evaluated MEMGAN, a method of synthesising scene images constrained by visual memory schemas. This application of both generative networks and cognitive science allows us to directly manipulate the visual schemas of generated images and hence influence their resulting memorability. We evaluate the outputs of our model by conducting a memory experiment on human observers, finding that scene images generated with high-memorability visual memory schemas result in superior memory performance from the human participants, while low-memorability visual schemas result in more forgettable images. Additionally, we show a high degree of correlation between the predicted visual memory schemas of our generated images and the real-world obtained visual memory schemas of human observers, indicating we were able to manipulate human mental schemas towards those of our target schemas. This has interesting implications for the future study of image memorability, as well as real-world applications for image memorability research, and further validates the Visual Memory Schema approach for the purposes of characterising human long-term visual memory. In the next chapter, we turn to investigating another perceptual image characteristic; scene complexity, which may relate to the memorability of that scene. We investigate how we can extract, understand, and model human perception of scene complexity; asking which elements contribute to a human perceiving a given scene as ‘complex’ or ‘simple’. Later, in Chapter 6 we analyse the relationship between the memorability of a scene, and that scenes’ complexity.

# Perceptual Scene Complexity

## 5.1 Introduction

To eventually investigate whether complexity and memorability relate, we first need to investigate how humans perceive complexity itself. It is obvious that humans can rapidly evaluate the complexity of their surroundings; it is not difficult to determine whether our surroundings are relatively simple, or contain some inherent level of complexity. However, it remains relatively unknown which mechanisms underlie this perception; determining these may lead to a better understanding of how the human visual system operates, and how it processes scene complexity. In addition to theoretical advancements, there are also numerous practical applications for the study and measure of perceptual complexity. These include marketing applications (e.g; perhaps you want your advert to be easier to visually process and comprehend, and thus less complex), impacts for psychological experiments (you may want all your visual stimuli to be of similar complexity to exclude a confounding factor) and healthcare applications (the evaluation of cognitive image processing disorders; how easily a patient can process an image of known complexity). The study of scene complexity may also help inform the development of virtual reality environments; a simulated world desiring realism should be capable of matching the complexity of real environments, without appearing too simplistic, or overly complex. The development of complexity models allows the extraction of complexity values from scenes to take place automatically without requiring a human-in-the-loop for each application. Without these models, the majority of practical applications become significantly more difficult; requiring costly human intervention for

## CHAPTER 5. PERCEPTUAL SCENE COMPLEXITY

every instance of the application.

As discussed in Chapter 2, the first apparent quantification of what humans might perceive as complexity appears in the early 20th century [12], defined as the count of elements in an image. Later work redefines complexity as the intricacy or detail present in a line drawing [123], or as the degree of difficulty involved in generating a verbal description of a texture [60], or evaluates complexity in the context of aesthetics [35]. However, these measures do not specifically target scene perception; with initial research on complexity perception in scenes [105] finding evidence that clutter and mirror symmetry play a key role in visual complexity, along with openness and object organisation (e.g. factors based upon scene gist research [103], where gist represents the general semantic content of the scene). As computing systems became more powerful, and the field of information science evolved; so too have definitions of complexity and techniques for calculating it. We have already encountered some of these techniques in Chapter 2; those which employ Shannon entropy of the image [144, 22], under the hypothesis that more complex images have a greater level of entropy (or disorganization), and simpler images contain more redundant information (and hence, lower entropy). Also discussed was Kolmogorov complexity [78], another information theoretic measure. These entropy-based measures appear to be one method of operationalising visual clutter [111], as the more cluttered the image, the more disorganised the image, hence the greater entropy. Naturally, information-theoretic measures are somewhat divorced from human perception. An image of random, coloured noise is high-entropy, yet meaningless to a human.

More recent research has turned to finding combinations of metrics that predict visual complexity [30, 99]; some information theoretic, some more grounded in human perception. These models are capable of predicting human complexity scores with an accuracy greater than any single predictor alone. The most recent work focuses on developing neural models of perceptual image complexity, finding that visual complexity information arises within the feature maps of deep convolutional networks [114], and similarly that multiple regions across the brain are involved with the representation of the complexity inherent in naturalistic stimuli [53]. Progress in understanding human perception of visual complexity, especially in the area of natural scene perception [30], is made more difficult by a lack of high-quality, varied scene datasets [99]. Existing datasets are either small (sub-200 images) [29], or are object-focused, which leads participants to evaluate the complexity of the object that fills the frame rather than the image as a whole. While

## 5.1. INTRODUCTION

object complexity likely contributes to the overall perception of complexity in a given scene, in order to understand scene complexity these objects must be placed in the wider context of their surroundings. Finally, drawing from image memorability research, it is becoming more apparent that perceptual image characteristics, while often represented as a single score for a given image, are better represented as two-dimensional properties that vary across an image [2]. Currently, available datasets indicate that the complexity rating a human may give is based on the entire image, which ignores the local properties of complexity within that image.

Our aim is to address previous shortcomings by developing human observer based, high quality, two-dimensional scene complexity datasets, and computationally operationalizing psychological measures of perceptual complexity. We choose four different metrics: clutter, symmetry, entropy, and openness, each hypothesised or evidenced to have some relation to complexity in prior work. We employ these measures to develop an understanding of exactly which perceptual factors account for human perception of visual complexity, ‘factorising’ out the degree to which each metric helps to explain human variance in complexity ratings. Our primary dataset, which we call ‘Vischema-Complexity’ (VISC-C), is based upon a categorical scene dataset [2], and consists of 800 images with 800 complexity scores; giving a rating for each image, obtained from a human observer study. In addition, critically, it contains 800 ‘complexity maps’ that capture the image regions that participants find simple or complex; and for the first time reveal the image areas that contribute to perceptions of scene complexity. We also introduce VISC-CI, a complexity dataset of complexity scores and complexity maps from human observer study of vertically flipped variants of our scene images. Vertical inversion results in destroying or damaging the semantic content present in an image [135, 100, 72, 42], when perceived by a human, thus allowing the quantification of the effect of scene semantics on perceptions of image complexity. Further we generalized our analysis to an existing image set, BOLD5000 [24]. Finally, we develop and evaluate a neural network model capable of simultaneously predicting complexity scores and two-dimensional complexity maps. We examine which features these “black box” neural models have learned to associate with perceptual complexity by dissecting the network and examining individual artificial neurons.

## 5.2 Factorising Complexity

Upon review of studies investigating human perception of visual complexity it is evident that multiple factors contribute to this perception, and in part some of these factors can be operationalised with computational measures. However, It is difficult to ground complex information-theoretic measures to human perception. As the first step in our investigation, we instead define a set of four possible complexity measures (entropy, clutter, symmetry and openness), chosen for both simplicity and their existing grounding in cognitive psychology. We evaluate their success in explaining the variance inherent in human complexity perception obtained from human observer experiments and recorded in the VISC-C, VISC-CI and BOLD5000 datasets. As color has been found to show contradicting results both as relating to complexity [30] and to not relating to complexity [27], we err on the side of caution and include color as integral part of the factors we examine where appropriate (clutter, symmetry, openness).

### 5.2.1 Clutter

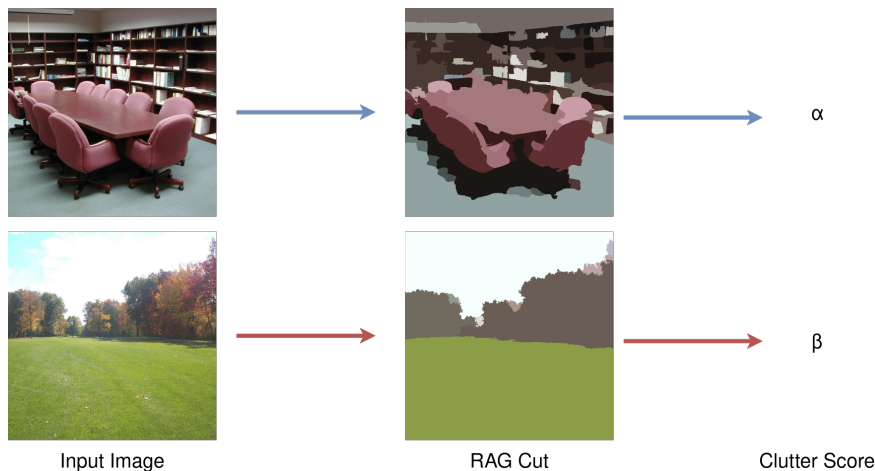


Figure 5.1: Example of clutter algorithm working on a perceptually simple image and a more complex scene, as rated by humans.

It is intuitive that the level of variation across an image would, in some fashion, be related to the complexity of that image. Prior research has revealed that human perception of clutter is one of the components that correlates with scene complexity [105]. There have been various attempts to characterise clutter, primarily through information-theoretic

measures [111]. Here, instead of an information-theoretic entropy-based approach, we characterise clutter as the number of separable regions computed by a normalised graph-cut of the region-adjacency graph of an image [119]. The normalised graph cut here divides an image into a number of ‘perceptually distinct’ regions. This has the effect of grouping similar parts of the image together into one average-color region. Our hypothesis here is that images that are perceived to be more complex would be decomposed into a greater number of distinct and separable regions, whereas simpler scenes are segmented into less regions; as overall they contain more ‘perceptually similar’ parts. The cost of dividing a graph into two disjoint regions is the summed weights of the edges removed to cause the bisection. The optimal bisection of this graph is the bisection with the lowest cost (i.e, that optimally separates two perceptually distinct regions). The normalised cut of graph  $\mathbf{G} = (\mathbf{V}, \mathbf{E})$  into distinct sets  $A, B$  is given in Equation 5.1.  $Cut(A, B)$  computes the sum of edge weights removed, and  $Assoc(A, V)$  is the sum of edge weights from  $A$  to all vertices in the region-adjacency graph.

$$Ncut(A, B) = \frac{Cut(A, B)}{Assoc(A, V)} + \frac{Cut(A, B)}{Assoc(B, V)} \quad (5.1)$$

### 5.2.2 Patch-based Symmetry

Much of the research into how symmetry affects the perception of images is conducted with “global symmetry”. This is defined as the difference between two regions of an image generated by bisecting the image either horizontally or vertically; equivalent to folding an image in half and determining how well each half matches the other. When participants are asked to give rankings on the complexity of images, this global symmetry of the image has been found to be a significant component of those rankings [105], and evidently relates to complexity in some manner. Computationally, most symmetry extraction methods focus on detecting the axis of symmetry of objects, or for determining where rotational symmetry appears in an image [90, 57, 106]. These methods are object-focused, and hence provide less information about the general symmetry present in a scene. Instead, we focus on extracting the symmetry of patches across the image, a compromise between computationally intensive methods that identify the symmetry of objects, and simpler methods that evaluate global bilateral symmetry. Our approach in particular works well for scenes; whose main semantic

details are often aligned in a horizontal plane. We hence define local patch symmetry as the mean of the horizontal and vertical symmetry contained within arbitrary-sized patches across the scene image. Given an image patch  $N_{ij}^{h \times w \times c}$ , at location  $(i, j)$ , we bisect the patch vertically giving  $(A^{h \times \frac{w}{2} \times c}, B^{h \times \frac{w}{2} \times c})$ , where  $A_{ij} = N_{i, 0 < j < \frac{w}{2}}$  and  $B_{ij} = N_{i, \frac{w}{2} < j < w}$ , defining  $F_h(A)$  as the horizontal flip of  $A$ , the horizontal symmetry of the patch is simply  $\text{sym}_h(N) = \sqrt{(f_h(A) - B)^2}$ . The vertical case is similarly defined. Hence,  $\text{sym}(N) = \frac{H_n^{\text{sym}} + V_n^{\text{sym}}}{2}$ , and the overall symmetry of image  $I$  given by  $\text{sym}(I) = \frac{1}{|K|} \sum_{i=0}^{I_{\text{cols}}/s-1} \sum_{j=0}^{I_{\text{rows}}/s-1} \text{sym}(N_{i \cdot s, j \cdot s}^{h \times w \times c})$  where  $K$  is the set of patches extracted, and  $s$  the stride.

### 5.2.3 Entropy

It is common in the literature on complexity to examine measures of entropy and the relationship between entropy and complexity. For the sake of completeness, we consider the Shannon entropy of the image histogram  $H = -\sum_k p_k \log_2(p_k)$  with  $p_k$  representing the probability of finding a pixel of  $k$  intensity over the image. Intrinsically, and certainly for simplistic images, it's generally expected that an increase in entropy corresponds with an increase in perceived complexity.

### 5.2.4 Openness

Despite psychological evidence for the influence of scene openness [105] on complexity, this factor remains relatively unexamined in computational approaches to perceptual complexity. Images with clear horizon lines and lack of boundaries are said to be 'open' (e.g, a field), and scenes that lack these, to be closed (e.g, a photograph of a kitchen taken perpendicular to a flat surface). We compute openness following the methodology from [112], and predict openness scores for every image in our dataset.

## 5.3 Experiment 1 - Two Dimensional Complexity

To evaluate which factors relate to human perception of complexity, we conducted an experiment on human observers, and tested our predefined computational measures against human complexity ratings. The experiment was designed to capture both complexity scores and two-dimensional annotations across a series of scene images. By evaluating our computational measures against the same images, we can determine which factors

### 5.3. EXPERIMENT 1 - TWO DIMENSIONAL COMPLEXITY

explain human perception of complexity. We term the dataset resulting from this experiment “VISC-C” for “VISHEMA-COMPLEXITY”.

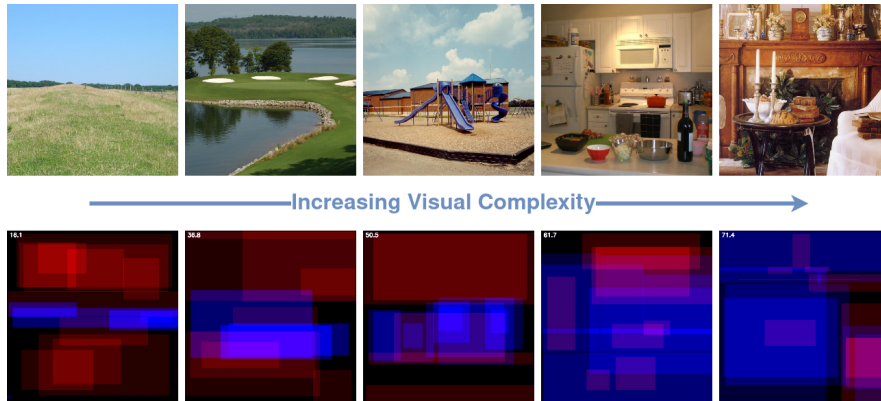


Figure 5.2: A set of scenes sorted into ascending complexity, as rated by a group of human observers. The images below the arrow reveal the regions that humans labelled as complex (in blue) or simple (in red). Regions labelled as simple often contain textural variation (e.g, grass in image 1, or the sky/clouds in image 3), yet are labelled simple nonetheless.

#### 5.3.1 Participants

A total of forty participants aged between 18 and 65, and fluent in English participated in the experiment. There were no other preconditions. Participants were paid for their participation and no personally identifiable information about participants was gathered or stored by the authors. Participants were free to withdraw from the study at any time. The experiment was approved by the ethics board of University of York, UK.

#### 5.3.2 Materials

The stimuli used were images from the VISHEMA: a categorical scene dataset initially gathered for the purposes of image memorability experiments [2]. The dataset consists of 800 images with a resolution of 700 x 700 pixels. The image-set is divided into eight classes of 100 images each, with each class corresponding to a commonly encountered scene category. Available classes are: kitchen, living-room, conference-room, airport-terminal, work/home (containing images of houses/office buildings), public entertainment (amusement parks/playgrounds), populated outdoor scenes (pastures/golf



courses), and isolated outdoor scenes (mountains/badlands). Example images are shown in Figure 5.2.

### 5.3.3 Procedure

The experiment was conducted online via Prolific [107], an online experimentation platform. Participants were shown a continuous stream of 200 scene images and completed the task at their own pace. For each image in the stream, they were first asked to rate the complexity of the image on a scale between 0 (least complex) and 100 (most complex). Once participants gave a rating, they were then asked to annotate the image. Each participant was asked to annotate either complex regions or simplistic/simple regions in the image. In no case was any participant asked to annotate both the simple and complex regions of the same scene image. In this manner we acquire independent annotations of both simple and complex regions for each of the images in the dataset.

Rather than employ a 2-Alternative Forced Choice paradigm, we designed a continuous complexity experiment. In 2-AFC experiments, two images are compared, and the most complex image selected. This both runs the risk of inducing comparative bias [130], and additionally implies that comparisons must eventually be converted (via one of several possible transformations) into a single rating. Even with a continuous paradigm, participants may begin to reference prior images as a baseline for future images they view. A participant that views a sequence of simplistic images and then a slightly more complex image may over-rate the complexity of that image, and vice-versa. In our experiment, every image stream shown to a participant was first randomised to minimise this effect and avoid potential biasing issues [47] that may arise in 2-AFC style complexity experiments; hence no two participants saw the same stream of images, and the average complexity score and annotations for the image can be considered independent of the context of the other images in the stream. We obtained 10 score ratings, and 10 annotations for each of the 800 images (five complex annotations and five simple annotations). A participant had to label at least one, and at most three, rectangular regions in an image before continuing on to the next image.

### 5.3.4 Data Analysis

We employed a hierarchical regression analysis (HRA) to analyse the contribution of each potential computational factor to perceived complexity, considering the contribution of

### 5.3. EXPERIMENT 1 - TWO DIMENSIONAL COMPLEXITY

the previous factors. We based our initial ordering of the factors on the order of their singular degree of correlation with human complexity ratings. Manipulating the order in which the factors were entered into the HRA, did not have any significant effect on the result. Notably, we tested whether entropy or clutter as the first factor results in decreased explanatory power of whichever factor is added second; and find that this does not change the outcome of the analysis. Hence, we start with clutter, then in turn add entropy, patch-based symmetry, and openness. The complexity score of any given image is defined as an average of scores from participants who saw that image. We concatenated all the per-image annotations into a singular two-channel ‘complexity map’, which captures complexity in one channel, and simplicity in the other.

#### 5.3.5 Results

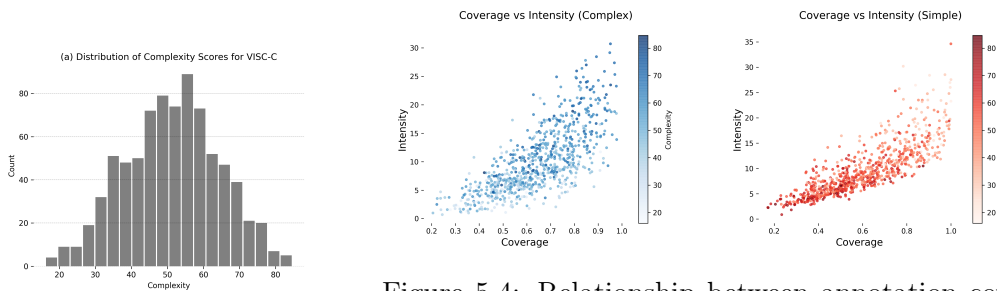


Figure 5.3: The distributions of human decided complexity scores for the VISC-C dataset.

Figure 5.4: Relationship between annotation coverage, intensity, and complexity for scenes. As coverage and intensity of the complex channel increases; so does the human complexity score ratings, and vice-versa for simplicity.

Figure 5.3 shows that human complexity ratings follow a Gamma distribution across images, a property reasonably expected for a scene dataset. Most images are unlikely to be either minimally or maximally complex. The mean complexity score for the images was 51.25, and the standard deviation was 13.14. We know from prior work that complexity ratings given by humans for images are consistent. However, there is little data on the consistency of complexity ratings purely for scene images, and whether participants agree that the same regions of the scene are simple or complex. We evaluated both the consistency of participant scores; and the consistency of our two-dimensional complexity annotations. Participant consistency was measured by dividing the participant data into two splits, and computing both complexity maps and scores from each half of the

## CHAPTER 5. PERCEPTUAL SCENE COMPLEXITY

data. We compared the scores from each split via the Spearman’s correlation, and the two-dimensional maps via the Pearson 2D Correlation (P2D), following prior literature [2]. We evaluated 100 splits for the scores, and 25 splits for the complexity maps. Participants show a strong agreement in their complexity scores ( $r = 0.72$ ). They also saw a good agreement on the complex regions of an image ( $P2D = 0.41$ ), and to a lesser extent, on the simple regions of the image ( $P2D = 0.27$ ). From the score consistency data, we can say that, on average, a random symmetrical split of human complexity ratings can explain 51% of the variance of the other splits ratings; the other 49% is surmised to be due to individual differences between participants.

To evaluate the two-dimensional annotations, we considered two properties; annotation coverage, which we quantify as the percentage of the image covered by simple or complex annotations, and the average intensity of the complex and simple channels. Intuitively, we assumed that a more complex image should contain more complex annotations, and a simple image should contain more simple annotations. The more intense these annotations in the complexity map, the more agreement exists between participants that the indicated region is of consequence, and the more complex (or simple) the region. We find that both annotation coverage and annotation intensity are strongly related to the complexity scores given by the participants. Annotation coverage and intensity is predictive of complexity score (multiple linear regression,  $R^2 = 0.6$ , Figure 5.4) and is indicative that the participants are labelling the images in-line with their scores. These results indicate that our two-dimensional annotation maps are indeed capturing both complexity and simplicity, and are strongly associated with “single-score” measures of complexity.

The results of a hierarchical regression analysis are provided in Table 5.1. Our computational complexity factors explain approximately 36% of the variance inherent in human complexity ratings. This is encouraging given that two disjoint sets of human complexity ratings explain 51% of the variance in each set. Generally, we can say that any measure that approaches or exceeds this ‘target score’ of 51% captures complexity to the same degree as the human visual system. Lastly, the results indicate that human complexity ratings are well explained by both clutter, and patch-based symmetry, and that entropy and openness contribute little. Visual clutter explains the most variance in complexity scores, followed by local symmetry. It is intuitive that the more cluttered the scene, the more complex the scene. Conversely, the more locally symmetrical features exist in

## 5.4. EXPERIMENT 2 - THE EFFECT OF SEMANTICS

Table 5.1: Results of a hierarchical regression analysis showing the contribution of each potential complexity factor towards explaining variance (coefficient of determination,  $R^2$ ) in complexity ratings for our VISC-C dataset. Together, clutter and symmetry explain 36% of human complexity (disjoint sets of human ratings explain 51% of each others variance). Entries in bold indicate significant increase in variance explained. Standard error of each linear model (Lm. Std.) and residual sum of squares (RSS) are reported for completeness, and is already incorporated into reported  $R^2$

Model	RSS	Adjusted $R^2$	$\Delta R^2$	Lm. Std. Error	Significance ( $p$ )
(constant)	29.35	-	-	0.19	-
<b>Clutter (C)</b>	<b>20.57</b>	<b>0.2983</b>	<b>0.2983</b>	<b>0.16</b>	<b>&lt;0.001</b>
C, Entropy (E)	20.55	0.2983	0	0.16	>0.05
<b>C, E, Symmetry (S)</b>	<b>18.84</b>	<b>0.3557</b>	<b>0.0574</b>	<b>0.15</b>	<b>&lt;0.001</b>
C, E, S, Openness	18.82	0.3557	0	0.15	>0.05

the scene, the less complex the scene is rated; there is less locally novel information to be processed. Entropy appears to have minimal explanatory power for perceptual scene complexity, as does openness.

### 5.4 Experiment 2 - The Effect of Semantics

The aim of our second experiment was to investigate the role scene semantics play in perception of scene complexity; we ask to what degree the participants' complexity ratings depend upon the semantic content of the scene. In order to investigate this, we rotationally invert our image dataset, disrupting the processing of semantic content for human observers. As with Experiment 1, we evaluated how our computational factors explain the perception of complexity of inverted scenes by human observers. These factors do not extract any semantic information from the scene. If they explain a considerable amount of variance inherent in inverted complexity scores, then perceived complexity for inverted images is very likely to be bottom-up driven and independent of semantic meanings. We term the dataset resulting from this experiment "VISC-CI" for "VISHEMA-COMPLEXITY INVERTED".

### 5.4.1 Participants

A new group of 40 participants, aged between 18 and 65 years of age were recruited for the second experiment. Participants were made aware they would be viewing inverted images and their consent to participate was obtained prior to completing the experiment. This experiment was approved by the ethics board of the University of York.

### 5.4.2 Materials & Procedure

The images and the procedure in this experiment was identical to Experiment 1 except that the presented images were rotationally flipped, producing an inverted variant of the scene. The data analysis employed in the experiment was the same as reported in Experiment 1

### 5.4.3 Results

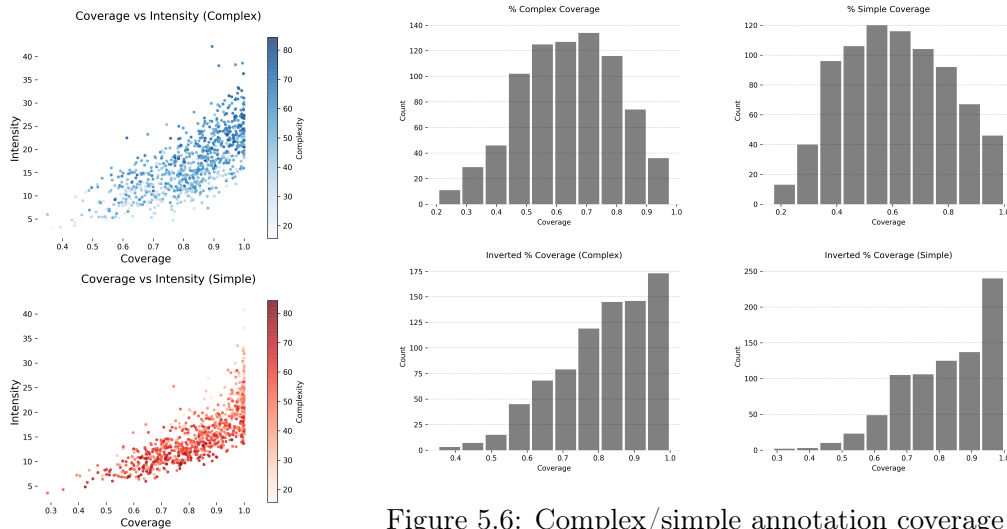


Figure 5.5: Relationship between inverted scene complexity and 2d annotation metrics.

Figure 5.6: Complex/simple annotation coverage for upright (VISC-C, top) and inverted (VISC-CI, bottom) scenes. Coverage shows that much more of the image is indicated as complex or simple when inverted; despite low-level textural properties remaining the same.

Complexity scores for inverted images show a mild skew towards being rated as more complex (mean = 53.20, standard deviation = 13.31) compared to upright images. There

#### 5.4. EXPERIMENT 2 - THE EFFECT OF SEMANTICS

was also a lower degree of agreement in complexity scores among observers compared to upright images ( $r = 0.60$ ). Despite these variations, complexity scores between upright scenes and inverted scenes correlate strongly together ( $r = 0.77$ ), which suggests even when inverted (semantic structure disrupted) participants are still able to determine the complexity of the image (though with a lower degree of inter-participant agreement).

While participants agreed to the same degree on the complex regions of inverted images as they did for upright ( $P2D = 0.39$ ), there was a lower degree of agreement between participants for the simplistic regions ( $P2D = 0.17$ ). This reflects the increased difficulty of the task, and is an initial indication that destruction of semantic structure affects complexity perception; especially in the case of determining what is simple. Participant consistency is decreased compared to Experiment 1, with a split of human data explaining 35% of the variance of its corresponding half (participant consistency of  $r = 0.59$ ).

The two-dimensional annotation properties of the delineated regions (annotation size and coverage) correlate strongly with given complexity scores ( $r = 0.66$ ), but show a significant skew towards a larger total annotation area (Figure 5.5) compared to upright scenes. Interestingly, we find that by inverting the scene images we caused a significant change in annotation coverage (Figure 5.6), with a greater percentage of the image being indicated as complex or simple. This suggests that participants find it more difficult to localise exactly what within the image is complex, or simple; defaulting to a global view of complexity for the entire image. These results imply that for images lacking semantic information, humans fall back to lower-level, global features when perceiving complexity, but do make use of semantic content where it is present.

Our complexity factors explain 38% of the variance in complexity scores (Table 5.2); exceeding the average human consistency of 35%. In this case; low-level classical features appear to explain all the variance in the human ratings. Given that inverting the scene damages the semantic information present in the image, we can hypothesise that the remaining 15% of variance not captured in the case of upright scenes we observed in Experiment 1 is due to the semantic structure of the scene images shown.

Table 5.2: Results of a hierarchical regression analysis run on human complexity ratings from the VISC-CI dataset (inverted scene images). The main contributors are clutter and symmetry (38%), with minor contribution from openness. Entries in bold indicate significant difference in variance explained. Std. Error is reported for completeness, and is already incorporated into given  $R^2$

Model	RSS	Adjusted $R^2$	$\Delta R^2$	Lm. Std. Error	Significance ( $p$ )
(constant)	30.05	-	-	0.19	-
<b>Clutter (C)</b>	<b>20.59</b>	<b>0.314</b>	<b>0.314</b>	<b>0.16</b>	<b>&lt;0.001</b>
C, Entropy (E)	20.59	0.313	-0.001	0.16	>0.05
<b>C, E, Symmetry (S)</b>	<b>18.80</b>	<b>0.372</b>	<b>0.059</b>	<b>0.15</b>	<b>&lt;0.001</b>
<b>C, E, S, Openness</b>	<b>18.633</b>	<b>0.378</b>	<b>0.006</b>	<b>0.15</b>	<b>&lt;0.01</b>

## 5.5 Experiment 3 - Generalizing to a Different Dataset

Experiment 3 examines how well our computational factors generalize to another existing image set, BOLD5000. BOLD 5000 is a dataset of 4914 images for which there is accompanying neuroimaging data primarily used for training and testing computer vision models [24]. Available for this dataset is a set of complexity ratings, gathered by researchers at the University of Toronto. This data is not yet publicly available. The images in the dataset is an amalgamation of images from other different image sets as follows: COCO, depicting objects [89], ImageNet, depicting diverse content of objects and scenes [36], and scene images based on categories from SUN [140].

### 5.5.1 Participants

The data was gathered via 1118 participants from Amazon Mechanical Turk, who were compensated for their participation. The participants were only recruited if they lived in either in Canada or the USA, and had approval rates greater than or equal to 75%. The experiment was approved by the University of Toronto Research Ethics Board.

### 5.5.2 Materials

For the purpose of the experiment, 4914 images from BOLD5000 were used. Each image in the dataset has a resolution of 375 x 375 pixels. These images consist of 1999 images from COCO, 1915 images from ImageNet, and 1000 scene images based on categories from SUN. Images from COCO were collected to depict objects and images

## 5.5. EXPERIMENT 3 - GENERALIZING TO A DIFFERENT DATASET

from ImageNet either depict objects or scenes.

### 5.5.3 Procedure

The experiment collected complexity ratings from participants for selected images from the BOLD5000 image set. The experiment ran on each participant’s computer using the Inquisit [65] software. The images were pseudo-randomly assigned into groups such that each image received ratings from 50 participants. The images were presented sequentially in a random order to each participant and each participant viewed and rated 252 images. Participants provide three different ratings for each image on a 5-point Likert scale. Question 1 was: “How symmetric do you think this image is?”; Question 2: “How simple or complex is this image?”; the response options were 1 = “very simple”, 2 = “simple”, 3 = “neutral”, 4 = “complex” and 5 = “very complex”. Lastly, Question 3: “How much do you enjoy looking at this image?”. Participants needed to respond to all three ratings in sequence before the next image appeared.

### 5.5.4 Data Cleaning

Exclusion criteria were established to ensure high data quality. To detect participants always giving the same response, the variance of responses for each participant in a 15-rating sliding window were computed. Participants with a variance less than or equal to 0.2 on average were excluded. Also excluded were participants with average variance between 0.2 and 0.5 and mean reaction time shorter than or equal to 250 ms. Data from participants who did not finish the entire experiment were also discarded. These criteria resulted in the exclusion of the data of 143 participants (12.8%). Data collection continued until there were 50 valid ratings per image. For this study, only complexity ratings for each of the images were considered. Ratings for complexity were converted to z-scores separately for each participant by subtracting the mean of their responses and dividing by the standard deviation. Z-scored ratings for each image were averaged over participants for further use. We used this dataset, and conducted a hierarchical regression analysis similar to that in Experiment 1 and 2, with complexity ratings as the dependent variable and our computational factors as independent variables.



## CHAPTER 5. PERCEPTUAL SCENE COMPLEXITY

Table 5.3: Hierarchical Regression Results for the BOLD5000 dataset. Best explanatory model uses all factors, likely an effect of the more varied dataset, explaining 11.32% of variance in complexity ratings. These factors come close to human consistency over the dataset (one split of human ratings explains 12.67% of variance of the other split on average).

Model	RSS	Adjusted $R^2$	$\Delta R^2$	Lm. Std. Error	Significance ( $p$ )
(constant)	90.994	-	-	0.14	-
Clutter (C)	88.674	0.0253	0.0253	0.13	<0.001
C, Entropy (E)	84.238	0.0739	0.0486	0.13	<0.001
C, E, Symmetry (S)	82.176	0.0964	0.0225	0.13	<0.001
C, E, S, Openness	80.625	0.1132	0.0168	0.13	<0.001

### 5.5.5 Results

First, we evaluated the consistency in human complexity ratings over the BOLD5000 dataset. On average, a random split explains 11% of the variance in the other split after normalization within each participant. The consistency in human ratings is lower in this experiment than in Experiment 1 or 2. This is most likely caused by both the high diversity in the image set, and due to the dataset being primarily object-focused; all of which might result in lower consistency across participants compared to a scene dataset that consists of commonly encountered natural scenes. A post-hoc analysis shows that the rating consistency is higher in a subset of images that consists of scenes only. On average, a random split in COCO images explains 6.6% of the variance in the other split; a random split in ImageNet images explain 8.9% of the variance in the other split; and a random split in scene images based on SUN explains 11% of the variance in the other split.

The results of the hierarchical regression analysis are shown in Table 5.3. Our computational complexity factors explain approximately 11% of the variance inherent in human complexity ratings, which is close to the rating consistency across participants. The hierarchical regression analysis shows that all four computational factors contribute to explaining variance in human ratings. Compared with experiment 1 and 2, more of the involvement of entropy and openness for ratings in this image set might be explained by the higher diversity of images from BOLD5000. Nonetheless, the results here confirm results in experiment 1, that our computational factors are able to explain a

large proportion of variance in human complexity ratings. It indicates that our analysis is generalizable to a larger and more diverged set of images. The results also suggest that while for scenes semantics appear to play a part, for diverse object-focused images, low-level computational methods appear sufficient to explain human ratings.

## 5.6 Modelling Complexity

In Experiment 1 and 2 we established that annotation statistics for simple and complex regions extracted by human observers are strongly associated with overall global image complexity score, and that low-level computational measures explain a large proportion of variance inherent in complexity ratings. We now examine the efficacy of employing deep neural networks to predict both scene complexity scores and complexity maps. Further we ask whether neural networks are capable of capturing the semantic component of image complexity. Of interest is discovering whether deep neural networks learn features which can be used in conjunction with classical clutter and symmetry features in explaining human perception of image complexity. We develop a neural complexity model that can predict 2D complexity maps and scores simultaneously.

### 5.6.1 Predicting Complexity Scores & Maps

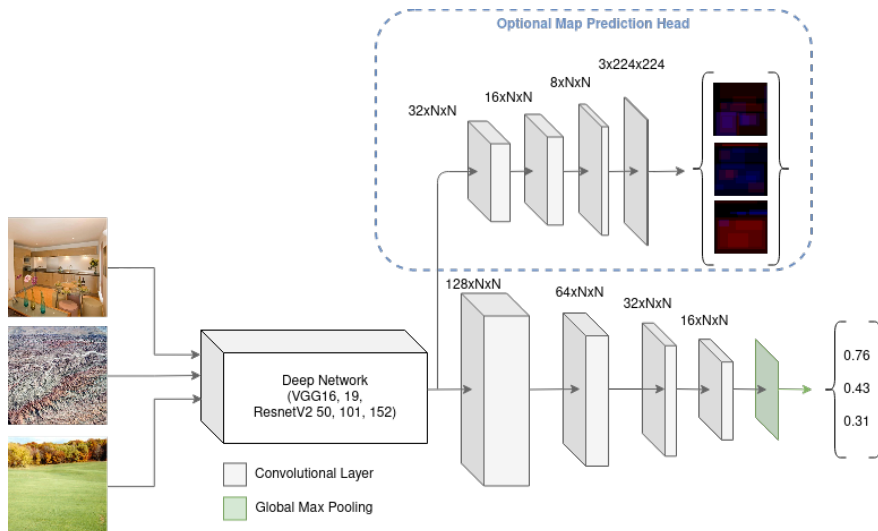


Figure 5.7: Basic Complexity Prediction Architecture, with optional complexity map prediction head.

## CHAPTER 5. PERCEPTUAL SCENE COMPLEXITY

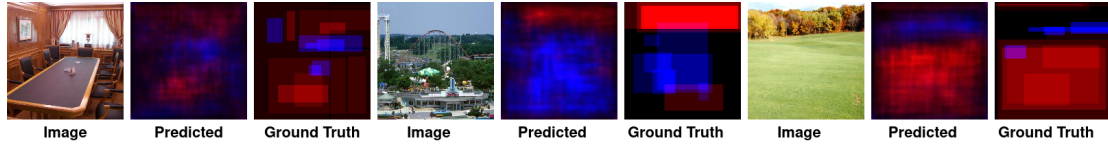


Figure 5.8: Examples of predicted complexity maps and their ground-truth counterparts from the test set.

We performed transfer learning upon five architectures: VGG16, VGG19 [122], and ResnetV2 [59, 58] with 50, 101 and 152 layers, to develop a different variants of complexity prediction network; which we term ‘ComplexityNet’. Each network has its classification head removed, and a four-layer convolutional regression head attached at a selected point in the network (as shown in Figure 5.7). While there has been work towards the artificial prediction of memorability maps [83], this remains unexplored for complexity prediction. To resolve this, we include an optional fully-convolutional complexity map prediction head, tasked to generate complexity maps for the input images. To evaluate the effect of network depth on complexity prediction, we systematically attach the regression head after each major processing block in each target network (results shown in Figure 5.9). Each ComplexityNet variant is then trained for 100 epochs with RMSProp (learning rate: 0.0001), and cross-validated on 8 splits of the data. From this cross-validation, we obtain predictions for every image in the VISC-C dataset. When the complexity map prediction head is enabled, the network is trained simultaneously with both scores and maps as inputs. We use the standard mean squared error for both score and map regression, and use ReLU activation functions throughout the network, aside from each output, which terminates with a sigmoid activation. The training process takes approximately six hours on a single NVidia Tesla V100.

Our complexity prediction model performs well at predicting complexity scores for scene images. When considering complexity map prediction, ComplexityNet achieves good performance for both scores and maps, with the best-performing model (when considering both scores and maps) achieving a Spearmans correlation of  $\rho = 0.67$  with human scores, and generating complexity maps that correlate with human complexity maps (complex annotations:  $P^{2D} = 0.54$ , simple annotations:  $P^{2D} = 0.49$ ). Samples of predicted complexity maps and their human observer-based maps can be seen in Figure 5.8, and prediction results from all tested architectures in Figure 5.10. More prediction examples are available in the Appendix.

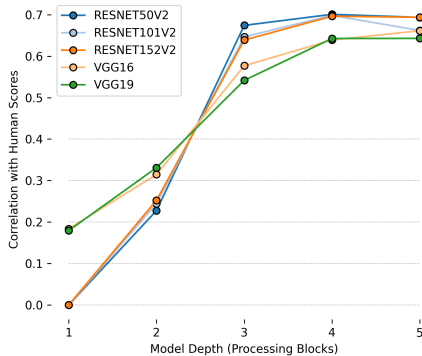


Figure 5.9: Effect of network depth on complexity score prediction performance. Performance peaks in the penultimate processing block of each model, then plateaus.

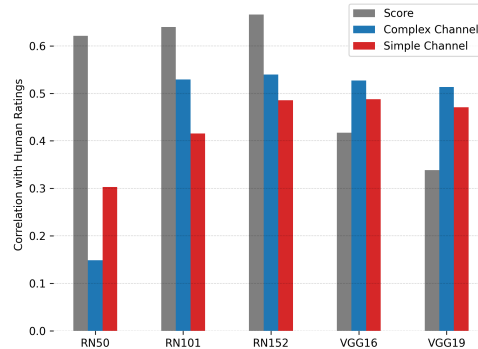


Figure 5.10: Correlation with human ratings for both scores and complexity maps for different base network architectures.

### 5.6.2 What Neural Networks Learn about Complexity

Do neural networks learn mostly low-level features, or do they extract semantic features with relations to complexity? To investigate whether neural networks learn features orthogonal to low-level computational measures, we combine the previous results of our hierarchical regression analyses with the predicted score outputs from our best performing ComplexityNet (based on RESNETv2-152). If the neural network adds little additional variance explained, we can assume that the neural prediction is based primarily on low level features. On the other hand, if the network can explain more variance inherent in complexity in addition to low-level features, this implies the network is learning semantic features that relate to complexity. To investigate this further, we employ network dissection [10] to ‘take apart’ our neural model. This allows us to determine which image features are important for complexity prediction, and to examine which image features the network is considering when predicting complexity scores for scene images. We dissect our best-performing ComplexityNet model, and examine each neuron from the final convolutional layer of the complexity prediction head; 16 neurons in total. Each neuron is assigned a set of images that best activate that neuron.

This neural network dissection reveals the images that activate each trained neuron (shown in Figure 5.11). With this dissection we make our “black-box model” transparent,

## CHAPTER 5. PERCEPTUAL SCENE COMPLEXITY

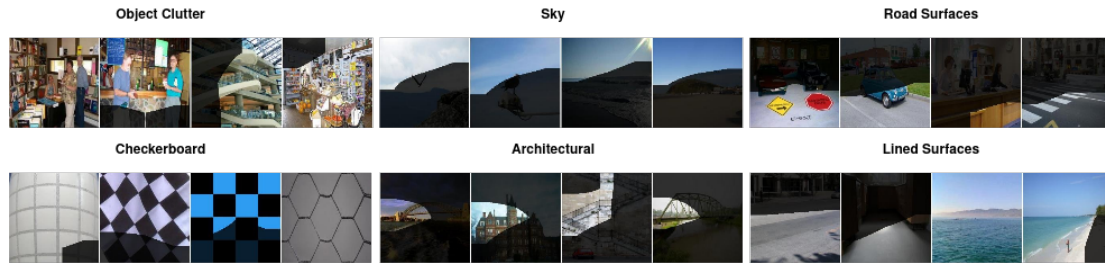


Figure 5.11: Images which activate a sample of neurons from the final layer of a complexity prediction network. The network appears to learn both low-level and semantic features.

and can analyse the features that each neuron in the output layer of the network searches for in the input scene image. We find in the network an emergence of both neurons that can detect low-level repeated features, such as checkerboard patterns or lined surfaces; as well as neurons focused on semantic structures such as skies, architectural elements, and road surfaces. Also interesting is the development of neurons that appear to detect clutter and activate strongly for images which contain large amounts of assorted objects. This is indicative of the importance of semantic information in complexity perception and reinforces prior literature in finding visual clutter influences perception of complexity; even inside neural networks modelled after human perception.

By combining our computational factors with the predicted score outputs from our best performing ComplexityNet (based on RESNETv2-152) in a hierarchical regression analysis, we can explain an additional 17% of complexity score variance, orthogonal to global image features such as clutter and symmetry. In total, this explains a total of 52% of the variance inherent in human complexity, matching the variance that can be explained in one set of human ratings by another randomly chosen set of human ratings. This suggests that complexity perception functions as a combination of both global image features (clutter, symmetry) and semantic information (architectural details, presence of roads, or skies); and that to predict complexity accurately, both are necessary.

### 5.7 Discussion

In prior work it has been shown participants in general are capable of evaluating the complexity of images in general. However, most complexity datasets are either small

[29], consist of simple images designed to investigate low-level processing (polygons, line drawings) [123, 60], or contain a mixture of images with only a small scene component [114]. It remains unclear exactly which features contribute to human perception of the complexity of scenes. Additionally, prior work tends to treat image complexity as a single rating for the whole image; which may obscure details on how complexity varies across a scene. In our studies, we specifically set out to both investigate scene complexity as a property that may not be constant across a scene, and to determine how scene complexity itself can be explained; including evaluating the effect of semantics. Here, we define ‘semantics’ in the context of scene complexity as referring to the collection of elements that give the meaning to the scene. Semantics provide a sense of context to the scene being viewed. We develop two scene datasets, VISC-C (Experiment 1) and VISC-CI (Experiment 2), for human perceptual scene complexity prediction and understanding. Compared to prior scene datasets, ours consist of high-resolution, high quality scenes images, are more varied, and include two-dimensional human annotations.

In Experiment 1 we find that the complexity ratings given by participants for our scene images are highly consistent, with one human split of complexity ratings being able to explain 51% of the variance in the other split. We also find that the two-dimensional annotations given by participants correlate strongly with the participants ratings. From this we can infer that the annotations (our complexity maps) are indeed capturing the complexity in the image, and reveal that different regions of scenes do vary in their perceived complexity. This two-dimensional dataset allows us to see that even in simple scene images, there are complex regions; and in complex scenes, simple regions. To determine whether we can explain the complexity of these scene images, we develop explainable, psychologically grounded computational measures for image complexity analysis. These computational measures are based in prior work that either hypothesises, or has shown, that these measures may play a part in complexity perception, even if this prior work does not directly involve scenes. Whereas in prior work with simpler images, entropy appears to play a role, for our scene images, we find no significant relation with complexity scores. Instead, we find we can describe a majority portion of human variance (36%) with our measures that are less information-theoretic; the clutter and symmetry factors.

From the results of Experiment 1 we can conclude that some portion of human complexity perception does appear to make use of lower level, global image properties. We

## CHAPTER 5. PERCEPTUAL SCENE COMPLEXITY

hypothesise the remaining, unexplained variance is in fact the effect of the high-level semantics of the scene. To investigate this, we conducted Experiment 2, a replica of Experiment 1, except all images are rotationally inverted, known to damage the semantics present in the image. Our results show that we can explain all human variance present in complexity for inverted scenes with our global image measures that do not employ semantics. This implies that when a scene is lacking in semantics, human perception of complexity falls back to global image features. The fact that the annotations for the inverted scenes also shows a lack of precise ‘localisation’ - that is, the annotations are spread out over the entire image, also appears to support this. We then ask how we can extract this semantic content, and use it for complexity prediction. To accomplish this, we develop our ComplexityNet neural network model, trained to extract the semantic features from images and re-task these features to predict both complexity scores, and complexity maps.

To determine whether our results are generalizable, in Experiment 3 we compute our global image properties for a large dataset of varied images (BOLD5000), only a small proportion of which are scenes. While this dataset lacks two-dimensional annotations, it does have complexity ratings for each image. Our results mirror that from Experiment 1, showing that global image features work well for explaining the variance in human complexity ratings across this dataset. However, overall participant consistency is lower than that of our scene dataset; in part due to the high variety of types of images, and it’s primarily object-focused nature. The BOLD5000 dataset, rather than consisting of varied scene images, also includes a vast array of different object photographs. Given this variety in image type and content compared to our scene-focused datasets, it is not surprising that we find all four of our computational factors become significant. Semantics appear to play less of a role when considering the complexity of object-focused images; as in general humans agree less on how complex objects are, compared to scene images.

To fully understand whether semantic information aids in complexity perception, we combined the output of ComplexityNet with our global image features, and find that we can capture all of the human variance (52%) inherent in single-score complexity ratings. However, neural networks are often ‘black boxes’; it is difficult to understand exactly which image features the network is using to give its’ complexity prediction. By employing a ‘network dissection’ technique [10] to discover these features, we find that both

low-level (checkerboards, lined surfaces, clutter) and semantic feature (sky, architectural details) extractors arise in the neurons of such a model. These results are consistent with what we observe from Experiment 1 and 2, that both global image features and semantic structure appears necessary to model human complexity perception.

## Conclusion

In this study we have developed three new datasets for the purposes of understanding how humans perceive the complexity of scene images. Two of these datasets focus entirely on scenes, and contain two-dimensional annotations that indicate the complex or simple regions in these scenes. The other dataset is both large; and diverse. Through state-of-the-art computational techniques we characterise human complexity as being in part explained with ‘global image properties’ (clutter, symmetry, entropy, openness) and by a ‘semantic’ part; which we capture with a neural network.

Global image properties explain a large proportion of human variance, and indeed all variance of scenes where those have had their semantic content disrupted. With a neural model and these properties combined, we can explain all the variance of human complexity perception in our dataset of natural scene images. Through network dissection, we have found that the network learns both global image features features and semantic features that relate to scene complexity; and both are necessary to explain why humans find some scenes complex, and some simple.



## Complexity & Memorability

### 6.1 Introduction

As discussed previously in Chapter 2, visual long-term memory appears to encode both a general ‘gist’ of the scene as well as specific details that enable a previously viewed scene to be selected from an array of similar scenes [79]. The gist trace has been the focus of much research [102, 104, 85] and captures rapidly extracted information (e.g, scene category), providing a general, undetailed ‘overview’ of the scene. However, how scene *detail* influences memory remains relatively unexplored. This may be due to the difficulty in extracting detail itself; while the gist of a scene can be identified with rapid serial visual presentation, no singular method exists for finding the detail trace. Recently, a study by Evans & Baddeley [43] proposes a two-level processing model for scene memory, while also employing visual complexity as an analogue for scene detail. The initial processing stage of the two-level model is based on gist, extracting nothing more than general image features, whereas the second stage facilitates encoding of idiosyncratic scene elements. This work reveals that differences in image *complexity* appear to affect how well a given image is remembered. Images of man-made scenes (indoor scenes etc, assumed high complexity) are better remembered than natural scenes (outdoors, low complexity). Additionally, in the case of door images with and without detail, those doors with detail are better remembered than those without. In essence, the level of detail present in a scene appears to directly affect the memorability of that scene. This is somewhat reinforced by the work of Saraee *et al.* [114] who find that a computational estimate of complexity is positively correlated with the average hit-rate of categories

## 6.2. COMPLEXITY RATINGS & MEMORABILITY

drawn from the FIGRIM dataset [20].

In this chapter we endeavour to explore the relationship between scene memorability and scene detail more comprehensively than prior work. As has been shown in both earlier work and this work, Visual Memory Schemas provide a good framework for understanding the memorability of a scene. This allows us a strong, two-dimensional baseline from which to investigate this relationship; detail and memorability can be examined across an image. Prior work has operationalised ‘detail’ as corresponding to the level of perceived complexity present in a scene; our work to understand perceptual complexity has given us a robust two-dimensional dataset of scene complexity. Unlike Saraee *et al.* [114] we will not be limited to computational estimations of scene complexity, and we can go beyond the binary categorisations of ‘simple’ and ‘complex’ of Evans & Baddeley [43], as our dataset has complexity information from human observers for every image. We also have the opportunity to explore the relationship in greater detail than just hit rate, as examined by Saraee *et al.*. Here, we examine whether complexity scores given by humans corresponds with the memorability of scene images from the VISHEMA dataset, where we use the d-prime measure to encode memorability. We then break this apart, examining the relationship between the two-dimensional complexity annotation statistics we explore in Chapter 5 and both hit rate and false alarm rate of our scene images. Finally, we explore how the complexity of image regions affect the memorability of that region, and briefly investigate whether our computational metrics, that help explain perceptual complexity also relate to scene memorability.

### 6.2 Complexity Ratings & Memorability

We first examine the relationship between our human complexity single-score ratings, and compare against memorability scores also obtained from humans for the VISHEMA dataset. Each scene image in the 800 image dataset now has a two-dimensional Visual Memory Schema, a two-dimensional Complexity map, single-score memorability information (Hit rate, False alarm rate, and D-prime) and single-score complexity information (a rating between 0 and 100). We begin by comparing the complexity ratings for each image with the corresponding d-prime score for that image, and find a significant positive correlation (Figure 6.1). That is, as the participants complexity ratings increase, so too do the memorability ratings for those scenes.

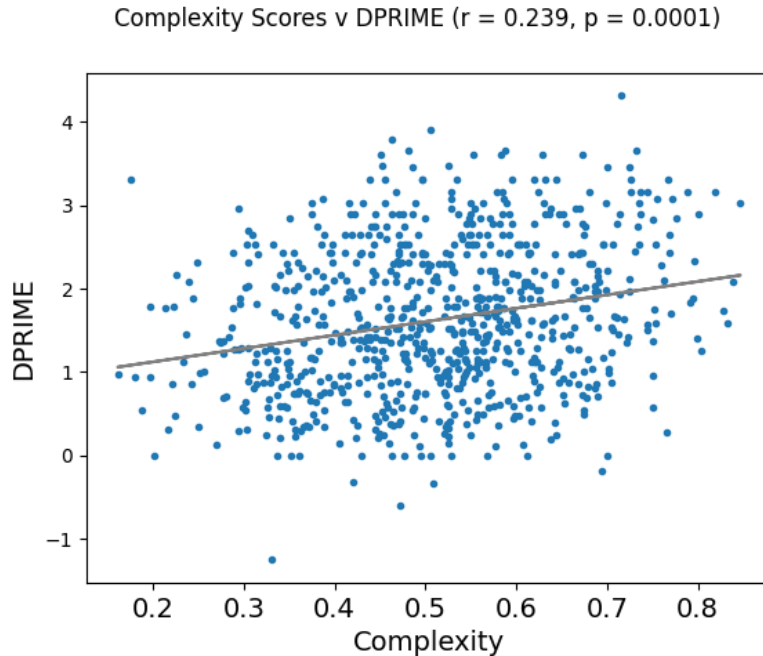


Figure 6.1: Pearson’s correlation between ground-truth human complexity ratings and ground-truth human memorability scores.

However, D-prime alone does not necessarily tell the whole story; is the rise in memorability as complexity ratings increase carried by hit rate, false alarm rate, or both? To determine this we compare complexity ratings separately with both the hit-rate and false alarm rate for each image, shown in Figure 6.2. Both the relationship between complexity ratings and hit rate, and complexity ratings and false alarm rate, is significant. A rise in complexity ratings corresponds with a rise in hit rate (indicating the image is more likely to be correctly recognised) as well as with a decrease in false alarm ratings (indicating that the image is less likely to be falsely recognised, despite never being shown). This suggests that complexity, and by inference, detail, has a role in both increasing the likelihood of recognising an image, and decreasing incidences of false recognition. A greater level of detail prevents a viewed scene being confused with a previously encoded scene, as the detail provides more idiosyncratic information that can be encoded. This helps to separate the encoded image from the viewed image and helps reduce false recognition. The detail present may also facilitate correct recognition by providing more features that can be encoded and later recalled.

## 6.3. TWO-DIMENSIONAL STATISTICS

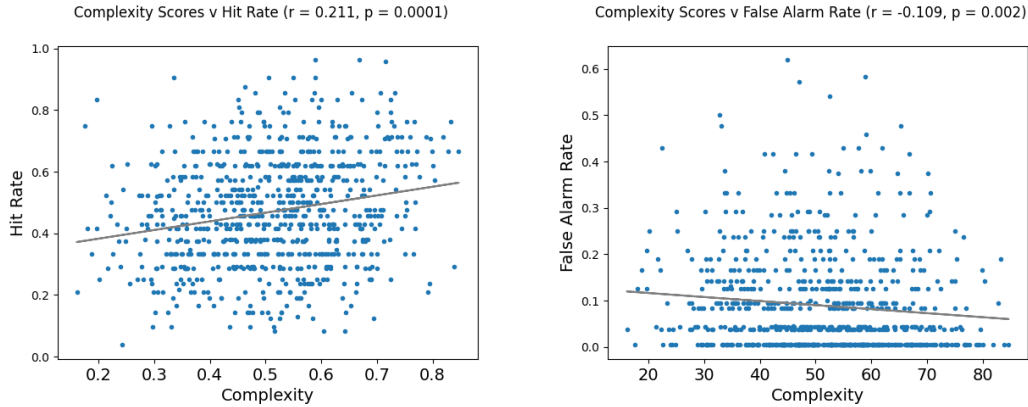


Figure 6.2: Pearson's correlation between human complexity ratings and scene hit rates (left) and false alarm rates (right)

### 6.3 Two-dimensional Statistics

To gain further understanding we can break down our singular complexity score by examining the two-dimensional annotations for both complex image regions and simple image regions. This allows us to determine which complexity-related characteristics are responsible for a scene being remembered better by a human. For example, we can ask if a scene is more memorable because it contains more agreement on complex regions, larger complex regions, or if perhaps it's related to the *simple* areas of the image instead. We employ the same metrics to evaluate two-dimensional complexity annotations as in Chapter 5, that is 1.) the average intensity of the complexity map (for either complexity or simplicity) representative of consistency, and 2.) the amount of image covered in annotations as a percentage. Generally, the greater the intensity and coverage of the complex channel of the complexity map, the more complex the image, and vice-versa for the simple channel. We first examine what these metrics tell us about the memorability of scene images, then secondly we directly compare complexity annotations to memorability annotations.

#### 6.3.1 Complexity Annotations

In Figure 6.3 we show the relationship between scene memorability (both hit rate, and false alarm rate) and the average intensity of the complexity channel of each image's complexity map. This intensity is analogous to participant agreement; the more an-

## CHAPTER 6. COMPLEXITY & MEMORABILITY

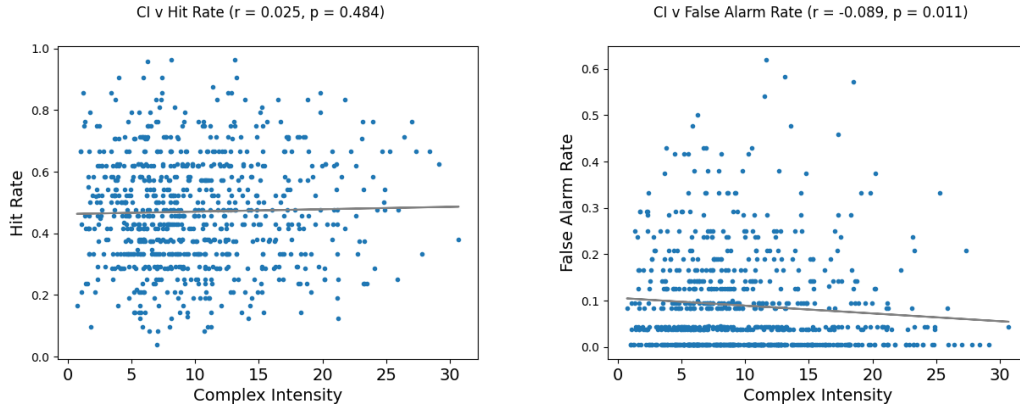


Figure 6.3: Relationship between scene memorability (hit rate, left, and false alarm rate, right) and complex channel annotation intensity. Pearsons correlation.

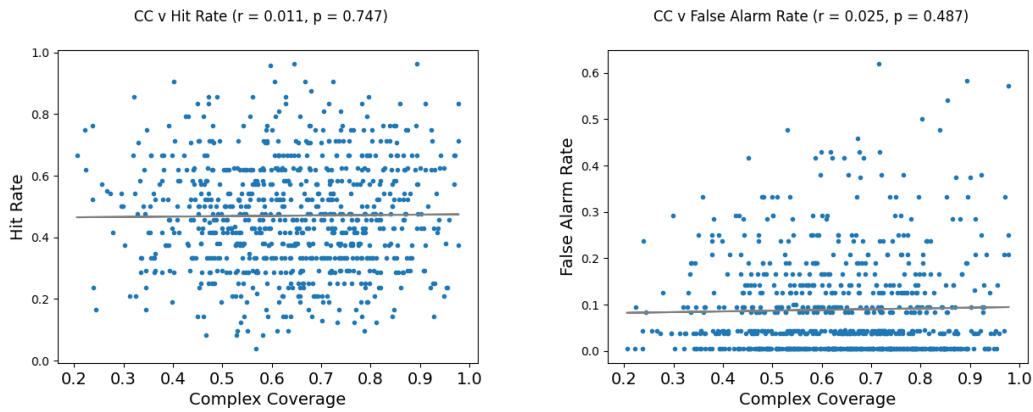


Figure 6.4: Relationship between scene memorability (hit rate, left, and false alarm rate, right) and complex channel annotation coverage. Pearsons correlation.

notations overlapping on a single region, the greater the intensity of the channel. We find that participant agreement on what in the image is complex has no significant relationship with the hit rates of the image. However, there is a weak, but significant negative correlation between complex channel intensity and false alarm rates. Turning to annotation coverage (Figure 6.4), we find that there is no relationship between complex channel coverage and image memorability. In general, there is little evidence that metrics that describe complex annotations specifically have much to do with scene memorability; aside from possibly reducing instances of false recognition.

## 6.3.2 Simplicity Annotations

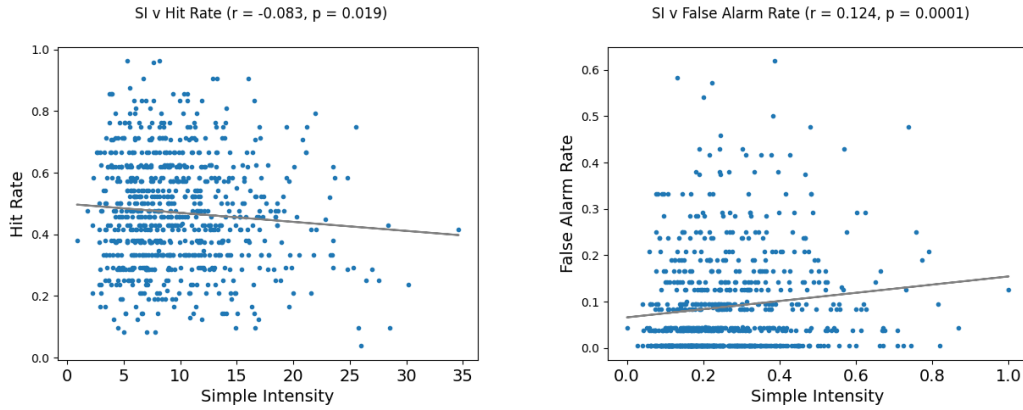


Figure 6.5: Relationship between scene memorability (hit rate, left, and false alarm rate, right) and simple channel annotation intensity. Pearsons correlation.

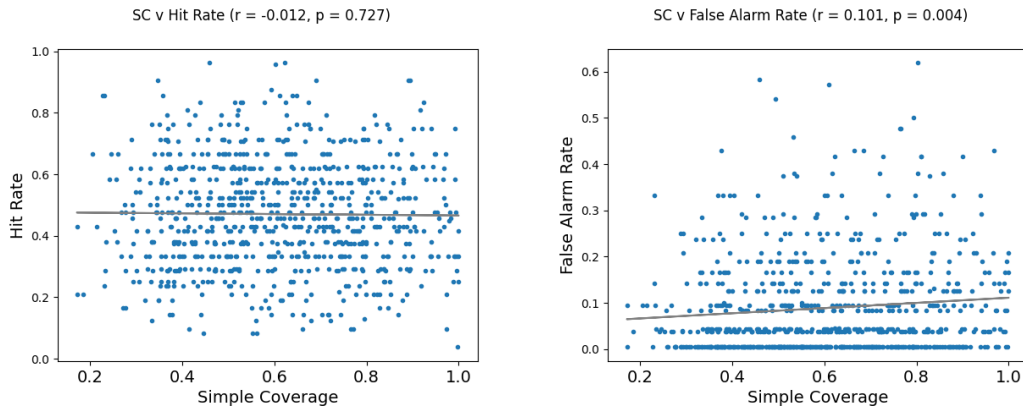


Figure 6.6: Relationship between scene memorability (hit rate, left, and false alarm rate, right) and simple channel annotation coverage. Pearsons correlation.

We now show the relationship between simple annotations (annotations given by participants to demarcate perceptually simple areas of the scene) and scene memorability. In Figure 6.5 as above we show the relationship between memorability and average intensity of the simple channel, while in Figure 6.6 we show the relationship between memorability and simple annotation coverage of the scene. Unlike the two-dimensional complexity, simple annotations show a significant correlation between average intensity (participant agreement) and both hit rate and false alarm rate. Generally, the more agreed-on simple regions in the scene, the lower the hit-rate, and the greater the false-alarm rate. For

annotation coverage, while there is no relationship with hit-rate (i.e, more simple regions across an image does not effect likelihood of correct recognition), there is a significant relationship with false alarm rate.

### 6.3.3 Region Memorability

While comparing annotation statistics with single-score memorability data reveals some interesting relationships between complexity and memorability, this still requires a condensing of two-dimensional information into one dimension. However, because our data for both memorability and complexity is two-dimensional, we can directly compare the two sets of maps (complexity/memorability). The memorability map data (Visual Memory Schemas, Chapter 3) contains both a ‘true schema’ channel; indicating regions that caused the scene to be correctly remembered, and a ‘false schema’ channel, indicating regions that cause false remembering. The complexity data (Complexity Maps, Chapter 5) contains both a ‘complex’ channel indicating complex regions, and a ‘simple’ channel indicating simple regions. Here, we directly correlate these different two-dimensional maps together using the Pearsons 2D correlation [2]. The results are shown in Table 6.1. All correlations are significant.

Table 6.1: Comparing correlation of two-dimensional regions between memorability data and complexity data. All values significant.

	Memorability		False Memorability	
	Complexity	Simplicity	Complexity	Simplicity
$\rho$	0.5	-0.06	0.34	-0.05

We find that there is a strong correlation between regions labelled as perceptually complex, and both regions that caused participants to remember an image, and to falsely remember an image. Simple regions are weakly, though significantly, negatively correlated with memorability and false memorability. In general, it appears that for scene images, complex regions appear to be a driving force for both recognising the scene, and also falsely recognising the scene. Simple regions, however, do not lead to that region being indicated as having caused a recognition, or a false recognition.

## 6.4 Computational Methods

In this section, we briefly examine some computational methods to better understand the relationship between complexity and memorability. First, rather than examining each factor independently, we consider complexity as whole. Secondly, we evaluate how well computational measures that work well for complexity can explain memorability.

### 6.4.1 Multi-factor Analysis

It appears evident that the complexity of a scene image has an effect on how likely that scene image is to be remembered, forgotten, or even falsely remembered. To further investigate the relationship we conduct a series of multiple linear regression analyses considering one memorability metric (DPrime, hit rate, false alarm rate) with a series of complexity metrics (Complex/Simple channel intensity, Complex/Simple channel coverage, and human complexity ratings). For our analysis, we remove the complex/simple coverage factors due to multicollinearity concerns. While this will not affect the descriptive power of the model, it can make interpretation of results more difficult. If the purpose of the model is purely explanatory, we can capture a higher total variance explained by including all factors. However, to determine which of these factors are relevant, we employ a subset of our factors that maintain a good degree of independence. A full table of results that includes all factors is included in Appendix D. We present the results of this analysis in Table 6.2, and additionally show the effect of keeping all factors (af) in the row titled ‘af-Adjusted R-Squared’.

We find that for DPrime, hit rate, and false alarm rate, complexity is capable of explaining a small, yet significant portion of variance in memorability scores. For D-prime, we can explain 5.6% of the variance (6.8% when including coverage metrics), and for hit rate and false alarm rate, 4.9% and 1.5% respectively. Interestingly, we can explain much less of the variance in false alarms with complexity than in hit rates, likely a result of the greater degree of human variation in false alarms than in hit rates. For Dprimes and hit rates, the most significant predictor are the human complexity ratings; though hit rates benefit from complex channel intensity. For false alarm rates, the only significant predictor from complexity is that of the average simple channel intensity.



## CHAPTER 6. COMPLEXITY & MEMORABILITY

Table 6.2: Results of multiple linear regression, with Complex and Simple coverage removed to avoid multicollinearity concerns. Coefficients for each variable are shown, as is the coefficient of multiple regression (R) and variance explained (R-squared), as well as the variance explained when including all factors (af-Adjusted). All regressions are significant. Complexity can explain a small, but significant portion of variance inherent in memorability data for DPrime, hit rate, and false alarm rate. Significant values shown in bold,  $p < 0$ : \*\*\*, 0.05: \*

	D-Prime	Hit Rate	False Alarm Rate
Constant	1.000	0.3156	0.098
Complex Intensity	-0.051	<b>-0.096*</b>	-0.014
Simple Intensity	-0.360	0.01	<b>0.061*</b>
Complexity Scores	<b>1.431***</b>	<b>0.355***</b>	-0.042
R	0.244	0.23	0.136
R-squared	0.06	0.053	0.018
Adjusted R-Squared	0.056	0.049	0.015
af-Adjusted R-Squared	0.068	0.06	0.027
Observations	800		

### 6.4.2 Computational Metrics

In Chapter 5 we showed that two computational metrics, based on prior psychological studies, perform well at explaining human perception of complexity. Given that it appears complexity and memorability have some relation, we now investigate whether these computational metrics also show a relationship to scene memorability. We do this by calculating the Pearson’s correlation between these metrics (computed for each image) and the DPrime of each image. The results are shown in Figure 6.7.

We find clutter, which is positively correlated with complexity, is also positively correlated with memorability as measured by DPrime. Symmetry, negatively correlated with perceived complexity, is likewise also negatively correlated with memorability. Generally, the more clutter and less symmetry present in a scene, the more memorable that scene should be; and vice versa.

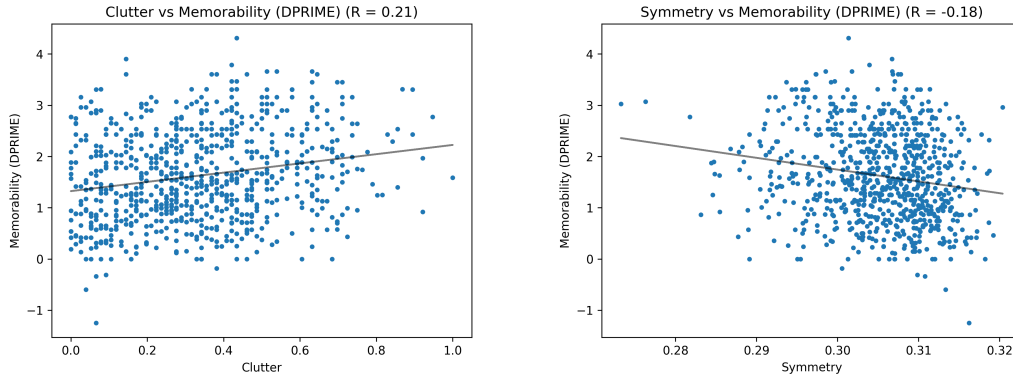


Figure 6.7: Relationship between clutter (left) and symmetry (right) computational metrics and scene DPrime. All correlations significant.

## 6.5 Discussion

In this chapter we have made use of our previous work to investigate the question of whether perceptual complexity could potentially account for the ‘detail’ trace in human memory. There is limited prior work that investigates the relationship between complexity and memorability. One such study finds a potential relationship considering an artificial predictor of complexity [114] with hit rate, while another shows both that man-made images appear to be remembered better than natural images, and that door images with detail removed are remembered worse. For this study, scenes with man-made features are considered complex, and scenes without, consisting of natural features are considered simple. In our work, having ground-truth complexity scores for scene images means we do not have to use an artificial predictor, nor divide images into two categories of man-made or simple. Instead, we can directly compare complexity data to memorability data *for the same scene image*.

Most importantly, we do indeed find a relationship between human complexity ratings and human memorability data. There is a mild (0.239) but highly significant ( $p < 0.001$ ) correlation between the single-score ratings given by participants regarding the scenes complexity, and the DPrime score for that scene. It is evident that generally, the more complex the scene, the more memorable that same scene. This relationship is carried both in the hit rate and the false alarm rate of a given scene; the more complex the image, the greater the likelihood of a correct recognition, and the lower the chance of

## CHAPTER 6. COMPLEXITY & MEMORABILITY

an incorrect recognition. A greater level of detail present in a scene may provide more potential features that can be encoded during the first time the image is viewed, which helps to both correctly identify a repeat of the scene, while also helping to filter out incorrect matches that lead to false recognition. Interestingly, prior work only finds that detail (considered between groups of manmade vs natural scenes) is only carried in false alarm scores [43]. By considering complexity at an image level, rather than a category level, we leverage a more fine-grained approach in which results in the data pattern shown here. Notably, we allow both for simple man-made images and complex natural images, a distinction that may be lost through grouping.

Breaking down complexity into either ‘complex’ or ‘simple’ metrics based upon the results of the two-dimensional complexity maps allows us to further investigate this relationship. When looking only at one factor at a time, the amount of image covered in complex annotations (indicating that a specific image region is a complex one) has no relation to either the hit rate or the false alarm rate of the scene. Participant agreement on complex regions likewise only appears indicative of reduced false alarm rate. In other words, when participants are more likely to indicate the same region as complex in a scene, that scene is also more likely to have reduced instances of incorrect recognition. For the simple channel, the data shows an inverse pattern; the more agreement on simple regions in the scene, the more likely that image is to be falsely recognised. This also carries a small, but significant reduction in hit-rate. Generally, it appears (when considered solo) that the simple regions in the image have more to do with the memorability of that image than the complex regions. However, this is not necessarily a complete picture. Examining the relationship between complexity-based annotations and memorability-based annotations, it is evident that the complexity of the region appears to drive the memorability of that same region. In essence, while simplicity appears to drive false-alarms up, when an image is correctly recognised, the regions that have caused this correct recognition are, in part, related to the complexity of that region. This pattern also appears for the ‘false schemas’; which also appear to be driven by region complexity. It appears that even in a simple image, which should have a greater incidence of false alarms, the effect that causes correct recognition of that image is still partially to do with the complex regions present inside that scene. This makes sense, given our prior data on complexity; even simple images can contain potentially complex regions, a detail visible in our two-dimensional data.

## 6.5. DISCUSSION

But how much of memorability can be explained with complexity-based measures? To answer this we conducted a multiple linear regression, and find that in general we can explain 5.6% of the variance of human memorability for scenes with factors selected to reduce multicollinearity. If we relax this constraint, we can explain up to 6.8% of the variance. This is not a large amount, but it is significant; from this data we can certainly say that complexity, and hence detail, has *something* to do with the memorability of scenes. It should be noted that complexity is not necessarily the be-all end-all of detail; there could certainly be elements of ‘detail’ that are not captured in single-score and two-dimensional metrics of ‘complexity’ - however, complexity appears to be a suitable enough analogue to explain some of the variance in memorability. The data from the multiple regression analysis supports that found previously; for memorability as defined by DPrime, the significant predictor is that of the human complexity score ratings. Breaking this apart into hit rate and false alarm rate reveals that for hit rate, the primary predictor remains human complexity ratings, with a weak negative relationship between complex agreement and scene hit rate. This may indicate that overly complex scenes could actually suffer a *decrease* in memorability compared to those that strike a middle-ground between complexity and simplicity. For false alarm rate, the only significant predictor is how much participants agree on which image regions are simple. This reinforces our earlier finding that more simplicity leads to more false alarms; though the variance explained for false alarm rates is low, likely due to the inherently larger variation in false alarms than in hit rates; the same problem that causes false schemas to be more difficult to predict than true schemas.

Finally, given that the data so far is strongly indicative that complexity plays some role in scene memorability, it would be unusual if metrics developed to explain complexity did not also have some degree of correlation with memorability. We test this, and find that our results for our clutter and symmetry metrics correlated against memorability data show the same pattern that you would expect if they were correlated against complexity data. Notably, that greater clutter increases scene memorability, and greater symmetry decreases it. In general, it does appear that complexity, serving as an analogue for detail in the scene, does indeed relate to memorability. It is clear that there is an element of detail processing employed during scene encoding, and that detail can both cause an effect in recognition, and in false recognition. Most convincing is the evidence that pairs human complexity ratings to image DPrimes; with higher complexity comes

## CHAPTER 6. COMPLEXITY & MEMORABILITY

higher DPrimes, and with lower complexity, lower DPrimes. Specifically, when a scene is remembered, the memorable regions of that scene and the complex regions of that scene appear related, whereas false recognitions appear to be linked to the overall simplicity of the image. This adds to the so far relatively sparse body of work that attempts to determine the effect of scene detail on human memory performance, and our results both reinforce prior work, and provide new details on how complexity and memorability relate to one another.

### 6.6 Summary

In this chapter we have drawn upon both Visual Memory Schemas and our prior work on human complexity perception to investigate whether complexity and memorability are related. We use complexity to operationalise ‘detail’, an element of scene images that is hypothesised to assist in the encoding of images into visual long-term memory. We compare memorability data and complexity data in several different methods, starting by examining single-score ratings of complexity against single-score ratings of memorability, before progressing to investigating what two-dimensional complexity information reveals about human memorability. We find that complexity and memorability do indeed relate, and that complexity can explain a small, but significant portion of the variance inherent in scene memorability. Interestingly, while complexity appears to drive hit rates, and simplicity, false alarm rates, there is some evidence that high agreement on complex scene regions is negatively correlated with correctly recognising that scene.

# Conclusion

In this chapter we conclude the thesis, starting by providing a summary of the main research conducted in Section 7.1. We then discuss the general conclusions for this research in Section 7.2, and examine possible directions this research could take in the future in Section 7.4.

## 7.1 Thesis Summary

At the start of this thesis, in Chapter 1, we outline the motivation for conducting research into scene memorability. Notably, we had three main aims for this work: first, to examine potential avenues for improving visual memory schema map generation, expanding VMS datasets, and to gain a greater understanding of what may make images memorable. Secondly, to explore whether visual memory schemas could be employed to synthesise scene images designed to be either memorable, or non-memorable, and to test this effect in humans. Thirdly, and finally, we aimed to explore how humans perceive the complexity of scene images, and then make use of complexity data as an analogue of the ‘detail’ trace in human memory; exploring how memorability and complexity relate.

In Chapter 2 we set the scene for this research, presenting to the reader an overview of human memory, exploring which image characteristics may or may not relate to image memorability, before turning to general machine learning techniques and ending with a general overview of complexity. To avoid overburdening the general background chapter, beyond this, each chapter contains a specific literature survey that examines relevant

## CHAPTER 7. CONCLUSION

prior work to that chapter. We identify several key areas and address them - 1) The small overall VMS map dataset size. 2) The lack of techniques designed to predict VMS maps. 3) Minimal prior work into attempting to control the memorability of scenes, and none that do not require a starting ‘seed image’ to modify. 4) No exploration of complexity perception as a property that varies across a scene. 5) Complexity has not been examined as potentially serving as the ‘detail’ trace of scene memorability; and in fact, the relationship between complexity and memorability has not been examined in the case of ground-truth data existing for both.

### 7.1.1 Visual Memory Schemas

For Chapter 3 we start by noting that a significant challenge facing research into visual memory schemas is the lack of suitable data. This hampers both investigative techniques and machine learning models of scene memorability. To solve this issue, we first replicate the original VISHEMA experiment, doubling available data, before designing a continuous paradigm that can be hosted online. This allowed us to eventually increase available VMS data from 800 scenes/vms maps, to over 4000 scenes/vms maps. We use this data to evaluate a variety of machine learning techniques, exploring a variational approach, the effect of various hypothetically grounded techniques, before developing a novel architecture that gives SoTA VMS map prediction performance. The data allowed us to explore differences in memorability that are lost when scene memorability is reduced to a single score.

### 7.1.2 Modulating Human Memory

In Chapter 4 we propose two GAN networks that integrate a VMS map predictor to serve as an auxiliary ‘memorability loss’ for synthesised images. The purpose of these networks are to investigate whether we can use VMS maps to create scene images of a targeted memorability. This serves both as an interesting question in its own right, and as a further validation of the VMS map approach. We find that we do appear to cause a significant difference in hit rate between scenes generated to be low memorability vs those generated to be high memorability. Post-hoc analysis finds little difference in either the quality or the recognizability of these scenes, suggesting it is memory being affected. Empirically, differences in structure between low and high memorability images can be observed. Compared to prior work, our approach requires no initial seed image (we do

not modify an existing image to be more or less memorable), no existing pre-trained generator, and focuses entirely on complex indoor scenes.

### 7.1.3 Factorising Scene Complexity

We study how humans perceive complexity in Chapter 5. Encouraged by results in two-dimensional memorability, we likewise go beyond single-score complexity and gather complexity maps that reveal both perceived complex and perceived simple regions across scene images. We find that metrics that describe these annotations perform well at describing the overall complexity rating given to the scene images we used as our source dataset. By employing (and computationally operationalising) psychologically grounded metrics of complexity we find we can describe a significant portion of human variance with scene clutter and scene symmetry. Through a trained neural network (for semantic extraction) we explain the remaining portion of complexity variance, and through results on an inverted scene dataset settle on a two-pronged model of complexity; perception relies on both global image characteristics and scene semantics. Through network dissection we find that the network has learned to detect both low-level patterns and semantic scene elements, reinforcing this theory. Notably, we gather our complexity data on the VISHEMA dataset; creating for the first time a dataset that has both two-dimensional memorability data *and* two-dimensional complexity data.

### 7.1.4 Complexity & Memorability

Finally, in Chapter 6 we make use of our prior work in scene complexity and ask whether we can use complexity as an analogue for the detail trace of human memory. Prior work has suggested that the degree of complexity in the scene may affect how memorable that scene is, but until now no dataset existed that had both complexity and memorability data for every data point. Prior work relies on either artificial predictors for complexity, or on the reasonable assumption that manmade scenes are more likely to be complex, and outdoor scenes more likely to be simple. In our work, however, we have ground-truth human data that can be directly compared, and can directly explore the relationship without needing to separate scenes into two different categories. We find that complexity plays a small but significant part in scene memorability, and our global image features that can explain part of complexity equally correlate with memorability.



## 7.2 Conclusions

Here we provide some general conclusions from the work conducted in this thesis.

- By expanding available VMS datasets from 800 images to over 4000 we allow future work to benefit from a significantly expanded amount of data, often which is the limiting factor for machine learning approaches. We find that there are several differences in scene memorability that only appear when considering two-dimensional data, that are occluded when only considering a single-score representation of memorability. Through state-of-the-art segmentation techniques we are able to extract the ‘schema’ that visual memory schemas capture, for the first time determining which elements appearing together in scene images correspond with recognition of that scene. Through computational techniques, we move from a mental schema, to an annotated scene image, to a human-readable description of that mental schema that is suitable for quantitative analyses.
- We build upon our initial VMS prediction model, considering various alternative techniques to improve VMS prediction. We finally develop upon a novel architecture that makes use of both two-dimensional VMS data, as well as single-score memorability data, of which there is a significantly larger amount. Combined with the best-performing techniques that we isolate in a series of comprehensive tests, this model sets the baseline for VMS prediction. Interestingly, while our model was never intended to predict memorability scores for the single-score dataset, we find that disabling two-dimensional memorability feedback significantly impairs the ability of the model to predict single-dimensional memorability scores. This serves as yet more evidence that the VMS approach is a suitable descriptor for scene memory.
- By retasking VMS predictors and combining these with generative models we endeavour to synthesise scene images that can modulate human memory. We find that our images generated to be more memorable appear to be so, and vice-versa with scenes generated to be less memorable. It appears that human memory is susceptible to memorability-modulated artificially generated scene images, and that visual memory schemas provide a powerful enough description of memory to enable this.

## 7.2. CONCLUSIONS

- Human perception of complexity appears two-fold, and how we perceive the complexity of scenes is based both upon global image characteristics as well and the semantics present in the image. We find we can explain a significant portion of variance with our global image metrics, and for inverted scene images with damaged semantics can explain *all* the variance. By using a neural network to capture semantics that our global measures miss, we are able to explain all variance inherent in human complexity perception. We develop a dataset of scenes that contain both two-dimensional memorability information, and two-dimensional complexity information that highlights which areas of scenes are perceived as complex, and which are perceived as simple. We find these annotations correspond with the overall ratings of complexity that humans give to these scene, and open the path to a two-dimensional consideration of scene complexity.
- Finally, we find that complexity and memorability are related. While prior work has shown that scenes thought to be simple are less memorable than complex scenes, and has proven this with simpler images (doors with detail removed vs the same doors unaltered), or investigated artificial metrics of complexity against memorability, we here show the relationship with ground-truth data for both complexity and memorability. We present a comprehensive investigation of complexity against memorability, hypothesising that complexity may capture (or partially capture) the ‘detail’ trace of human memory. We find we are able to explain a small, but significant portion of variance in memorability with complexity, and also find that image regions that lead to recognition are often complex. In essence, leveraging our prior work in complexity perception we can say that the level of complexity present in a scene is a significant element in determining whether that scene will be remembered. Notably, we find that higher complexity images, in general, show higher incidences of hit rates and low incidences of false alarm rates; and that detail likely assists both recognition and reduces false recognition. We finally show that computational metrics relevant for complexity are also relevant for memorability, with scene clutter positively correlated with remembering an image, and high amounts of local scene symmetry negatively correlated with remembering.

### 7.3 Limitations

There are some limitations to this work that should be considered. While we have made progress in expanding available visual memory schema data, the amount of data remains relatively small compared to other computer vision datasets. This makes training a neural network challenging, due to both potential lack of variability, as well as the risk of overfitting. While data augmentation can alleviate these concerns, it does not perfectly resolve the issue. Additionally, while we endeavoured to use a wide variety of scene types, there are natural limits to that breadth of the data we were able to gather. For the original VISHEMA dataset, the data is limited to eight categories; and for VMS4k, the categories were constrained simply to ‘indoor’ and ‘outdoor’ in order to ensure there was an equal amount of that type of scene in each category. The scene types used to make up these categories were balanced as much as was feasible, but perfect balancing was not possible, which may introduce some element of bias *inside* the indoor and outdoor categories. This could be addressed with additional data.

In our studies we make use of behavioural data. This has some potential drawbacks; it is expensive and time-consuming to gather, and has the potential to be influenced by inherent biases and participant strategies. While for memorability data this issue is less pronounced (recognition performance of a scene is easy to test) for complexity there is no obvious test for performance. While participants agree on what in a scene is complex or simple, and in the score rating given to that scene, it remains unknown exactly what is being done by the participants when asked to rate complexity. This would require further research, with different experimental paradigm.

When it comes to complexity, although participants are highly consistent; and their ratings correlate with computational metrics, complexity as a visual concept lacks a concrete definition. While memorability of a scene can be categorised as ‘how well human observers can remember that scene’, a similar definition for complexity remains elusive, with prior work suggesting that it is as simple as ‘the count of elements present’ or ‘the verbal difficulty in describing a texture’. For scenes, both these definitions fall short; scenes are more than just their objects, and contain many and varied textures. We know that humans can consistently rate complexity, that perception of complexity appears to rely on both semantic and low-level features, and that in some fashion it is related to human memory of scenes. Beyond this, there may be a rigorous definition that

captures the data; or it may be an intrinsic feature of perception, similar to aesthetics or interestingness.

## 7.4 Future Work

Thanks to the relatively broad nature of this research, there are multiple directions this research could be taken in, both computational and psychological. Here, we describe a few potential avenues relevant to future applications of this work.

- The most obvious extension is yet more visual memory schema data. Modern single-score memorability datasets contain tens of thousands of datapoints; our contribution still lags behind this. By gathering more data, either with the original paradigm or our crowdsourced alternative, offers the chance to improve existing models and gain further understanding around the nature of visual memory schemas.
- We do not expect our dual-feedback model to be the be-all and end-all of visual memory schema prediction. There is certainly a promising direction in taking advantage of existing memorability datasets, and future approaches may wish to consider not just one, but multiple existing datasets to aid in VMS prediction. We kept our networks relatively shallow to best use available computing power; deeper networks with additional data could result in a promising increase in VMS map predictive performance. Likewise, while we found self-attention and multi-scale information performed adequately, recent generic classification models have found high levels of success with transformer based techniques; none of which have been examined in the context of predicting the memorable regions of scenes.
- There is significantly more to explore when considering the modulation of human memory with synthesised scenes. We were limited to a single indoor scene; there is no reason this technique could not be applied to class-conditional GANs, and explore the effect of altering the memorability of multiple outdoor and indoor scenes, should compute resources be available. Additionally since we conducted our study, there has been much progress in the area of generative networks; future work may be able to approach photo-realism across a variety of scene categories.
- In our work we mostly explored memorability, or ‘true schemas’. However, we also

## CHAPTER 7. CONCLUSION

have significant data on ‘false schemas’, or the areas that lead to false recognition of a scene. False schemas are significantly harder to predict compared to false schemas, and less consistent amongst humans. It would also be interesting to develop models that can more accurately model false schemas; and compare these to models of true schemas to determine which factors are relevant in triggering a false recognition. The segmentation analysis of Chapter 3.2 could also be applied to false schemas to determine which scene elements lead to a false recognition compared to a true recognition.

- Finally, while we have examined complexity and memorability together, we have only been able to do this for 800 scene images. It would be interesting to expand available complexity data to match that of the two-dimensional memorability data, and as such gain an even greater insight into the relationship between detail and memory. There has been little investigation into the features contained within complex or simple regions; further investigation into this may help reveal which semantic elements being present are critical for a scene region to be perceived as complex or simple.

### 7.5 Summary

In this thesis we have explored the memorability of scene images, leveraging a technique known as a visual memory schema map, and developing new methods of computationally predicting these maps. We gain insight into what contributes to these schemas through computational techniques, and employ our predictors to find that we can in fact modulate human memory with visual memory schemas. We develop a two-dimensional complexity dataset and dissect human complexity into a variety of psychologically grounded metrics, before using our data to investigate the relationship between visual memory and scene complexity. In summary, we find that we can capture the detail trace of human memory through complexity, and that detail has a small but significant impact on how well scene images are remembered.

## Co-occurrences of Objects in Memorable Regions

Here we provide co-occurrence counts for  $n$ -objects that appear together inside the memorable regions of images. These objects can be thought of together as providing a human-readable representation of a schema that leads that scene to be remembered. Viewed per-category these reveal the most memorable combinations of objects in that category; and hence capture the ‘schema’ for an entire category at once. We provide graphs for every category for 3, 5 and 7 object co-occurrences, each providing a greater level of detail, but with more specificity.

## APPENDIX A. CO-OCCURRENCES OF OBJECTS IN MEMORABLE REGIONS

### A.1 3 Objects

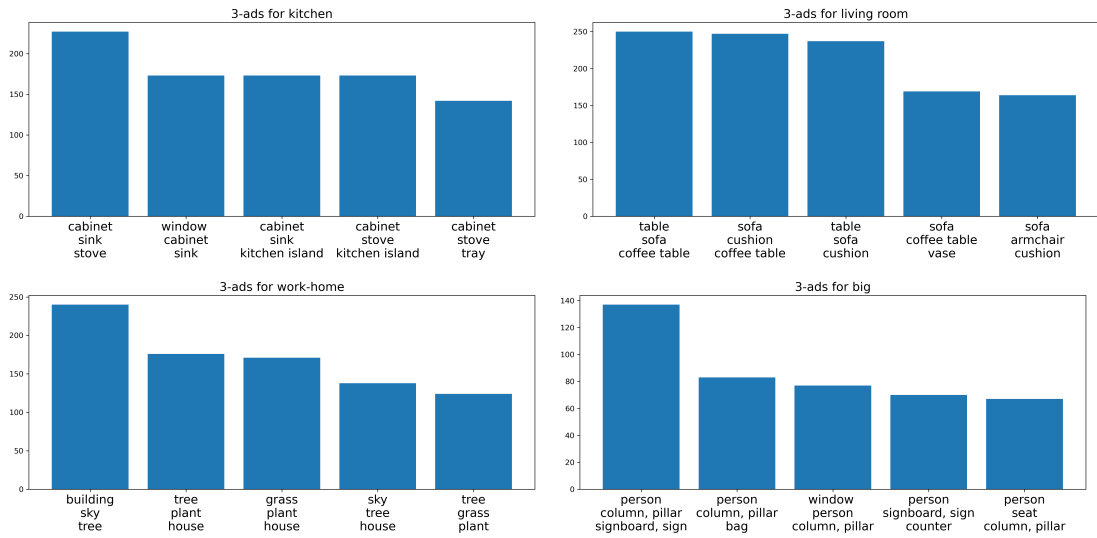


Figure A.1: These three objects frequently appear together inside the memorable regions of an image, of that category. VISCHEMA Categories: Kitchen, Living Room, Work/Home, Big

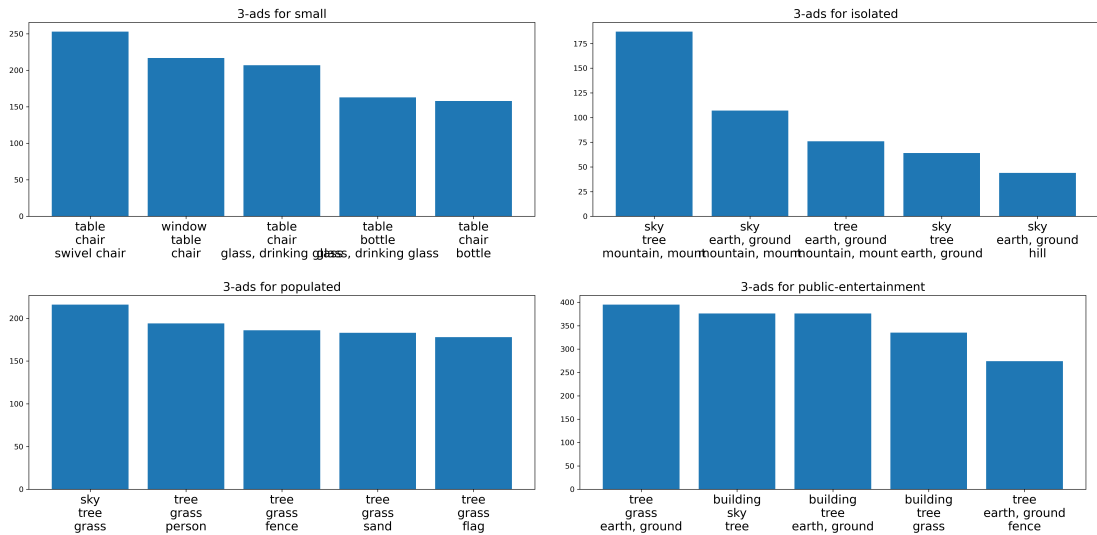


Figure A.2: These three objects frequently appear together inside the memorable regions of an image, of that category. VISCHEMA Categories: Small, Isolated, Populated, Public Entertainment

## A.2 5 Objects

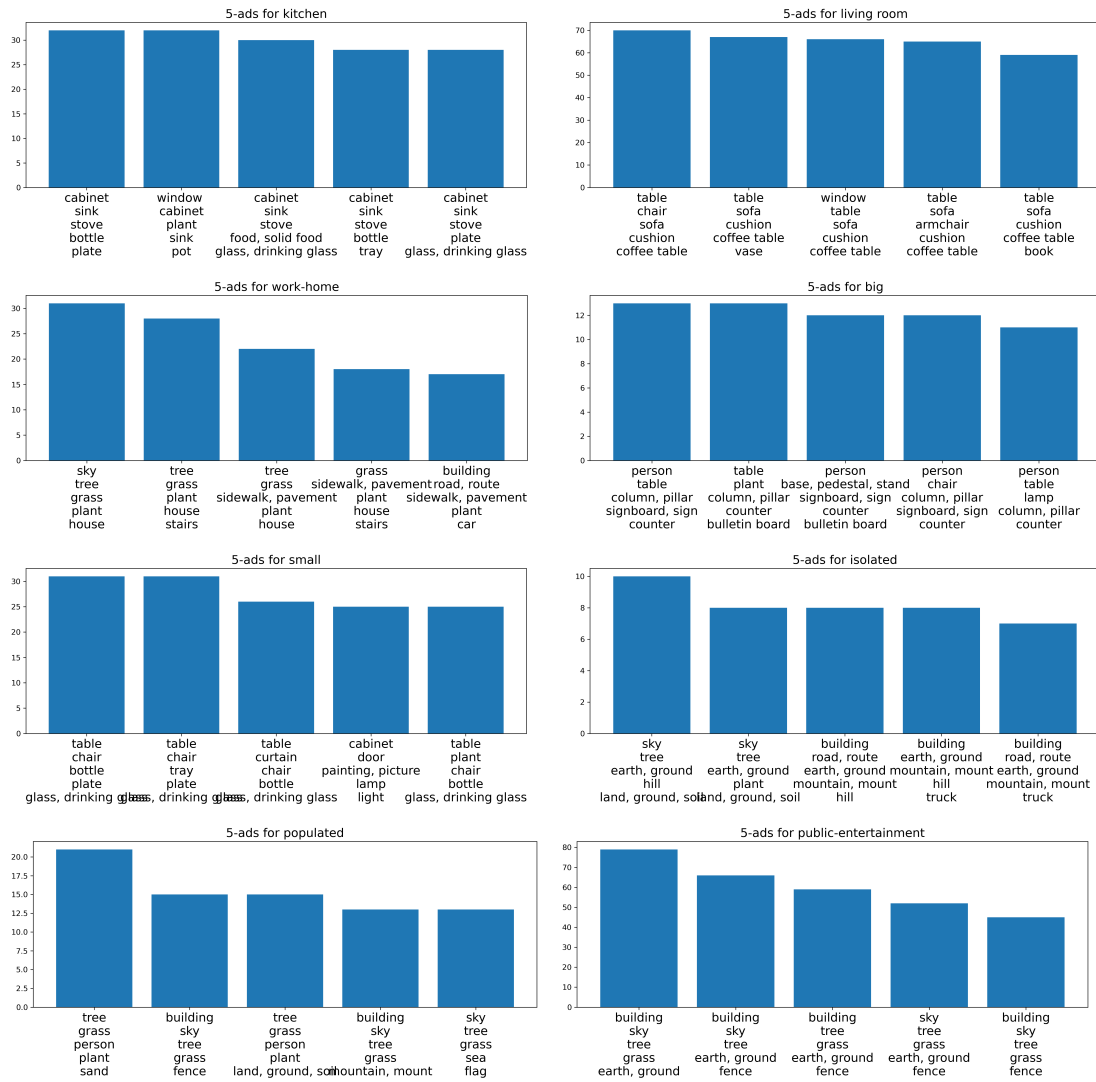


Figure A.3: Five object co-occurrences, all categories.



APPENDIX A. CO-OCCURRENCES OF OBJECTS IN MEMORABLE REGIONS

A.3 7 Objects

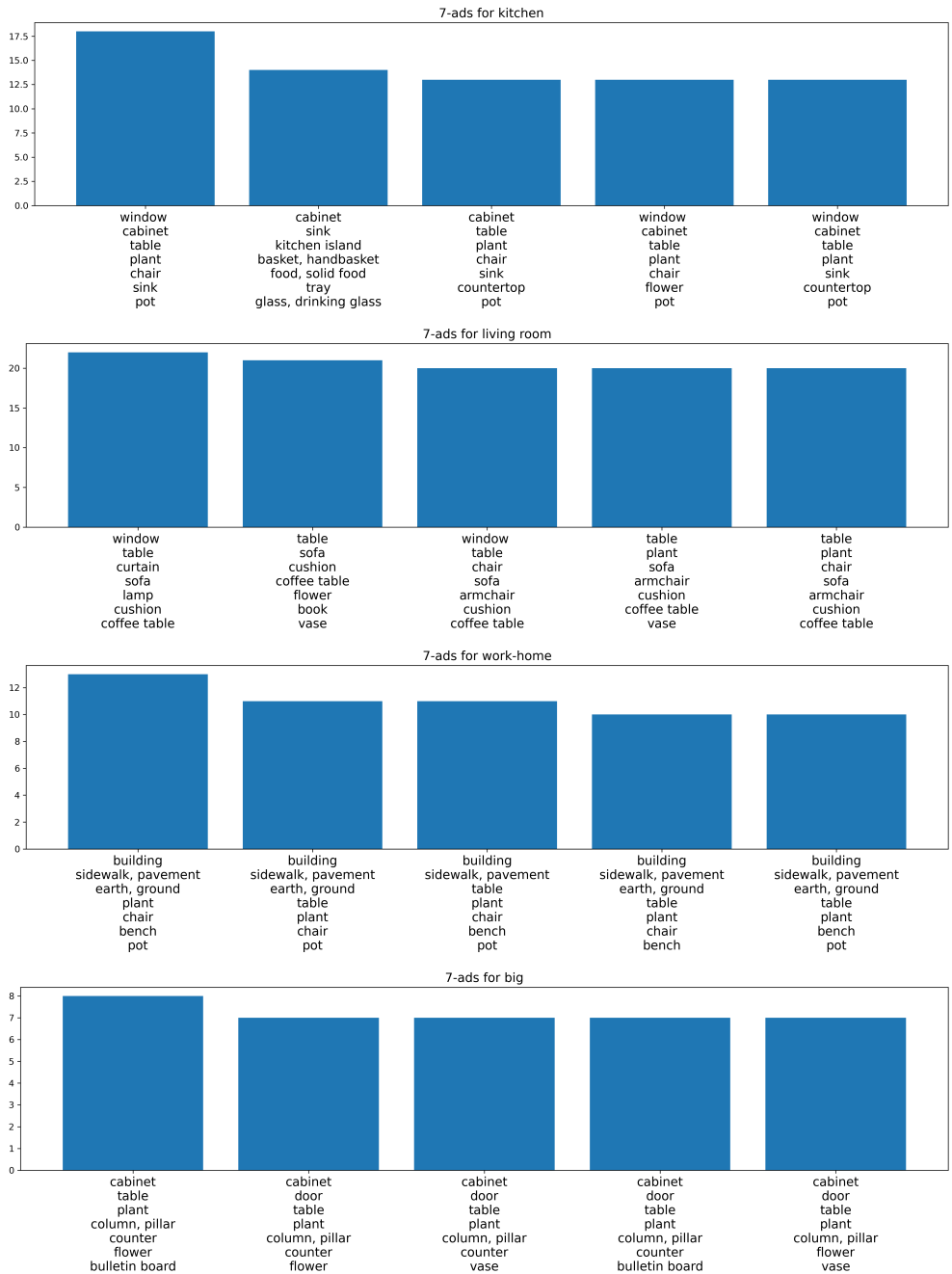


Figure A.4: Seven object co-occurrences. Kitchen, Living Room, Work/Home, Big

### A.3. 7 OBJECTS

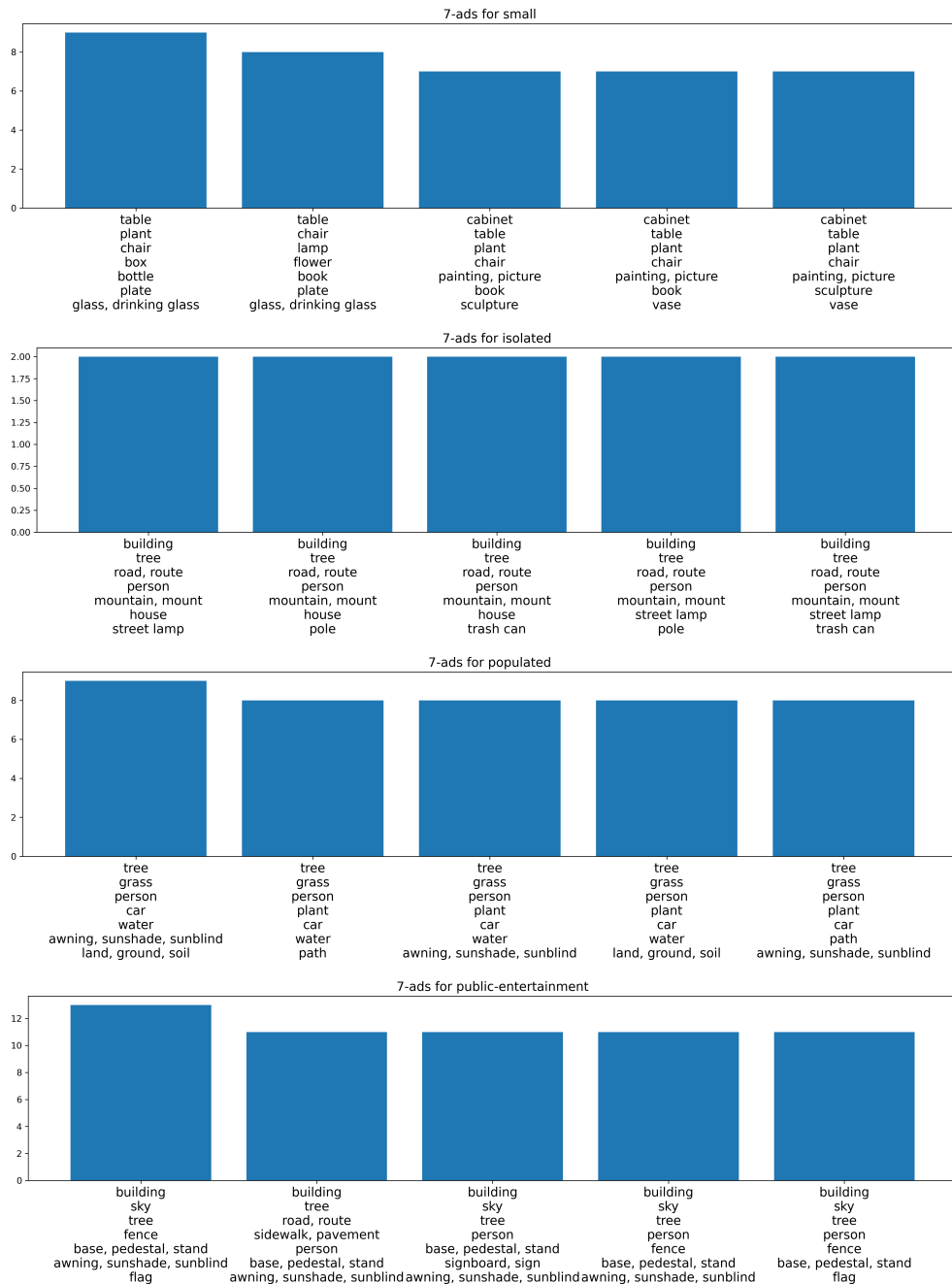


Figure A.5: Seven object co-occurrences. Small, Isolated, Populated, Public Entertainment

# Images with Modulated Memorability

## B.1 Additional MEMGAN Architecture Details

The following figures (figures B.1, B.2) are intended to support Chapter 4.

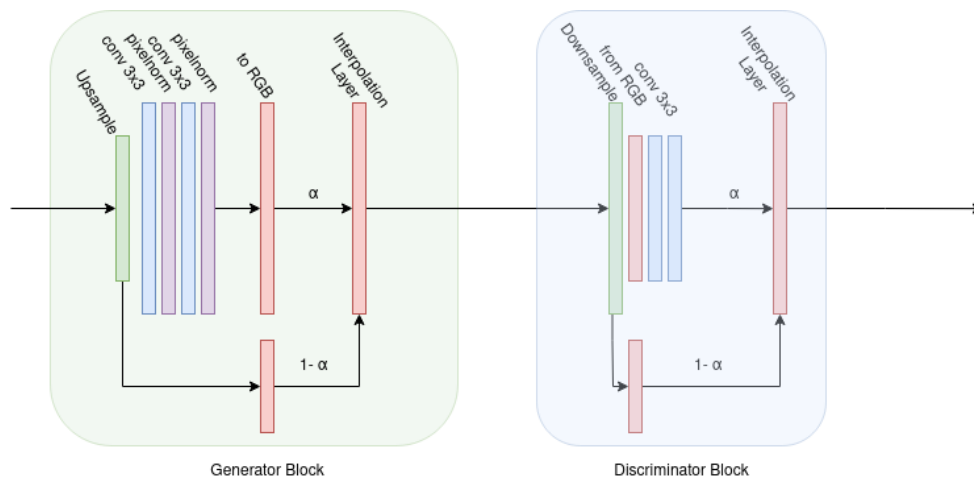


Figure B.1: Structure of generator and discriminator blocks, showing interpolation and convolutional filter sizes. Similar structure to that of the standard progressive GAN. All convolutional layers employ Leaky ReLU activation and weight-scaling [70]. We find adding a hyperbolic tangent activation to the output of the interpolation layer to improve training speed and stability.

## B.2. HIGH-MEMORABILITY IMAGES

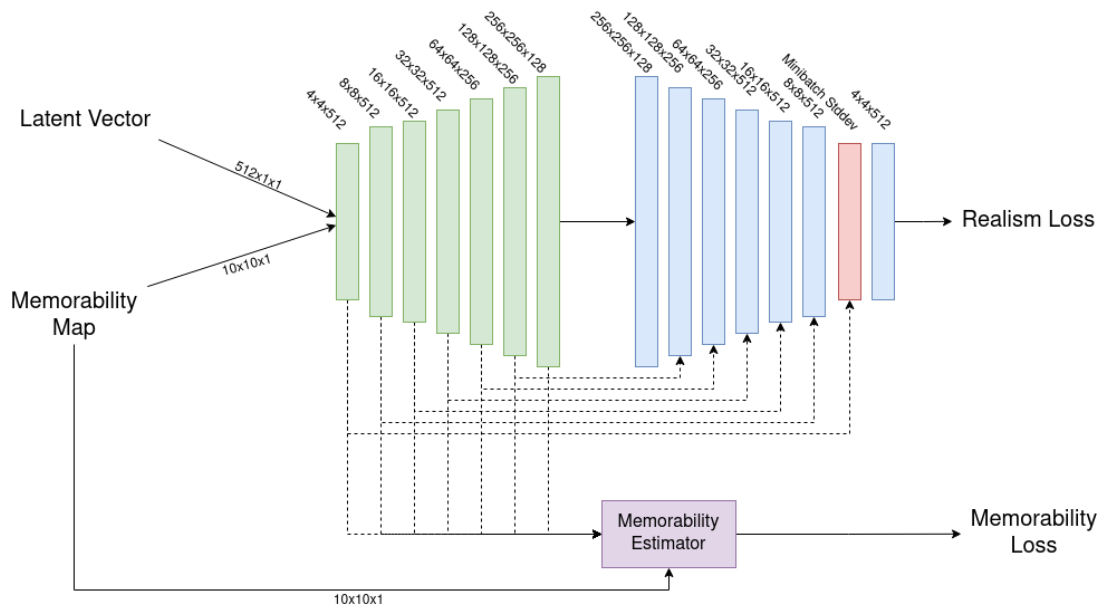


Figure B.2: Flattened network diagram showing resolution and channels of each architecture block. Every resolution block output is passed through the memorability estimator.

## B.2 High-Memorability Images



Figure B.3: Additional exemplars of generated highly-memorable images.

## APPENDIX B. IMAGES WITH MODULATED MEMORABILITY

### B.3 Low-Memorability Images



Figure B.4: Additional exemplars of generated low-memorability images. Low-memorability images appear simpler, and often contain more “closed” perspectives, with less variation across the image.

## Complex and Simple Images

The sections below contain additional examples from the complexity datasets gathered during as part of this research. The last sections contains additional prediction results for complexity maps from the complexity prediction neural network. The entire dataset is available online<sup>1</sup>.

### C.1 High Complexity Images

The following figures show the ten most complex scenes for both upright and inverted scene images, and are intended to support Chapter 5. In all complexity map images, blue areas represent complex regions, and red areas represent simple regions. All regions are gathered from human annotations. Additionally, in the upper left hand corner, each map contains the overall complexity *score* assigned to the scene.

---

<sup>1</sup><https://ccpl.hosted.york.ac.uk/research>

## APPENDIX C. COMPLEX AND SIMPLE IMAGES

### C.1.1 Upright Scenes



Figure C.1: Ten most complex upright scene images, as rated by humans. Complexity maps for these images are placed in the same location as their corresponding image in Figure C.2

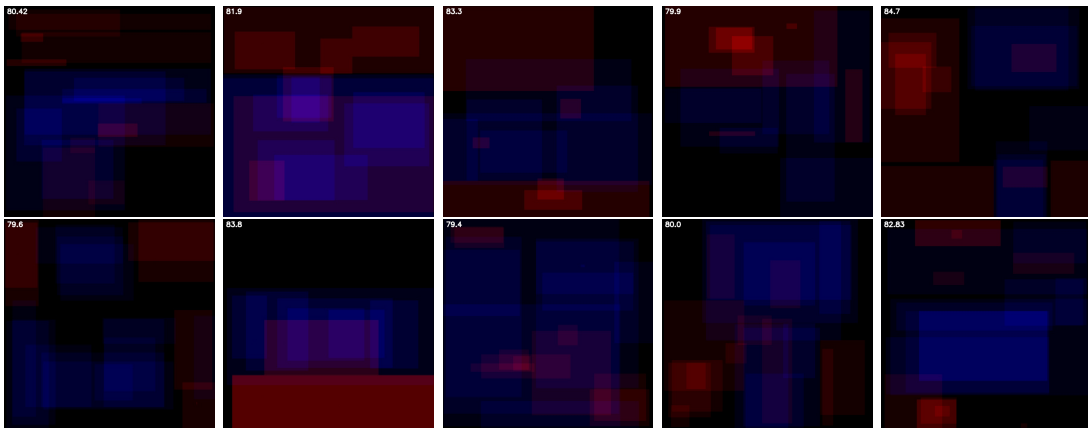


Figure C.2: Complexity maps for the ten most complex upright scenes, showing complex regions (blue) and simple (red) regions, as described by humans. Complexity score in upper left-hand corner.



## C.2. LOW COMPLEXITY IMAGES

### C.1.2 Inverted Scenes

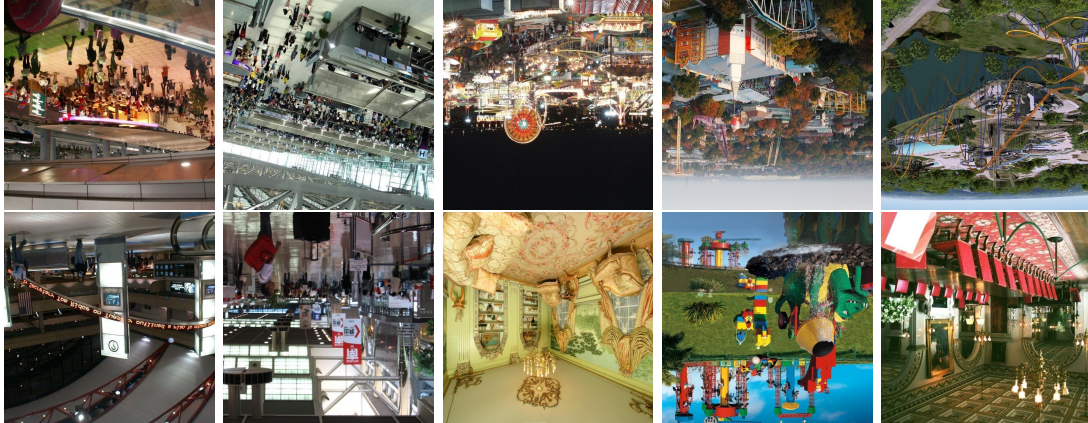


Figure C.3: Ten most complex inverted scene images, as rated by humans. Complexity maps for these images are placed in the same location as their corresponding image in Figure C.4

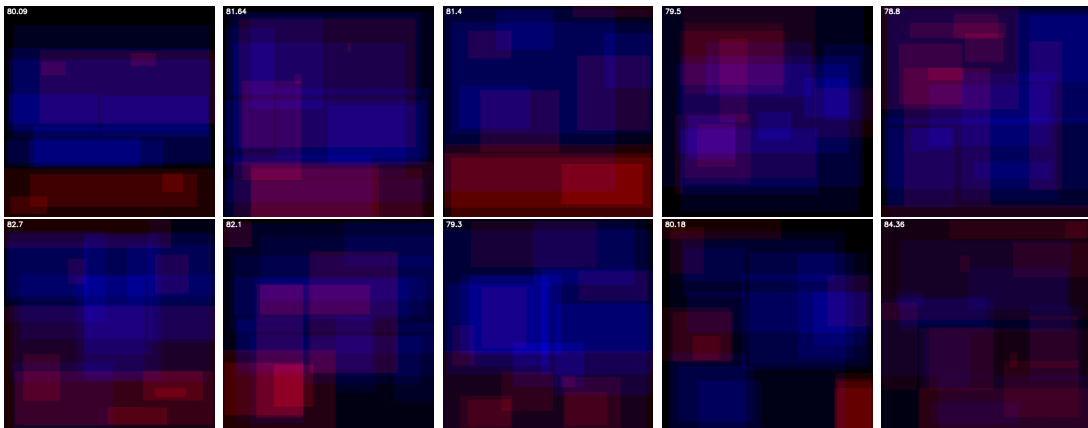


Figure C.4: Complexity maps for ten high-complexity inverted scene images. Note increased annotation coverage compared to Figure C.2

## C.2 Low Complexity Images

The following figures show the ten least complex scenes for both upright and inverted images.



## APPENDIX C. COMPLEX AND SIMPLE IMAGES

### C.2.1 Upright Scenes



Figure C.5: Ten least complex upright scene images, as rated by humans. Complexity maps for these images are placed in the same location as their corresponding image in Figure C.6

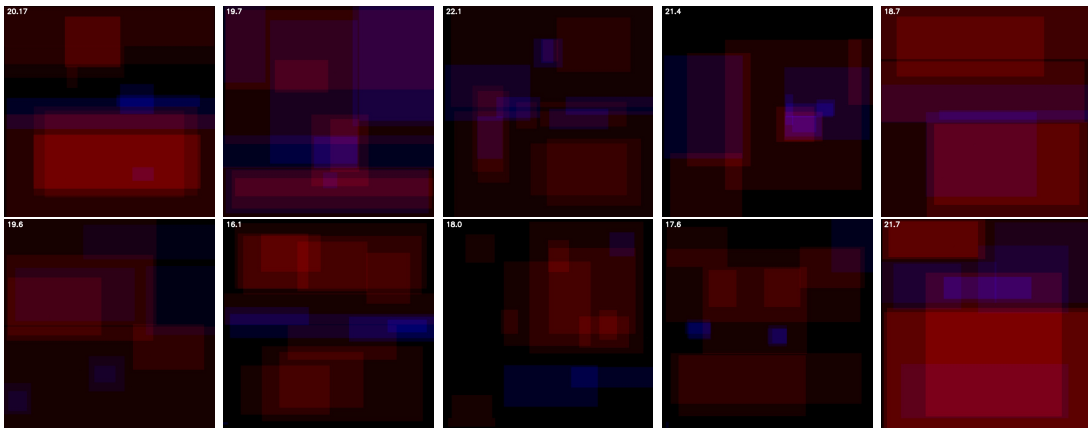


Figure C.6: Complexity maps for simplest upright scenes. Note prevalence of annotated 'simple' regions, matching low overall score.

## C.2. LOW COMPLEXITY IMAGES

### C.2.2 Inverted Scenes



Figure C.7: Ten least complex inverted scene images, as rated by humans. Complexity maps for these images are placed in the same location as their corresponding image in Figure C.8

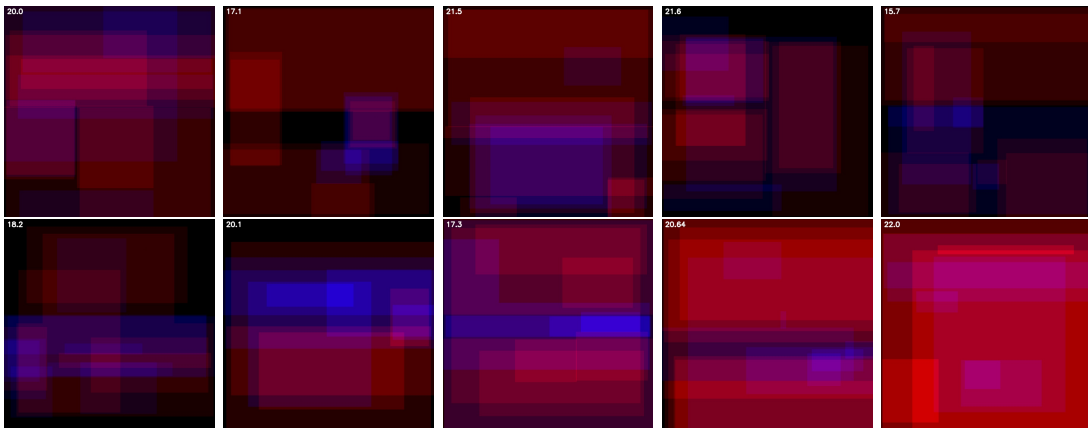


Figure C.8: Complexity maps for least complex inverted scene images. Evidence for loss of localisation ability compared to upright low-complexity scenes (Figure C.6)

### C.3 Additional Prediction Results

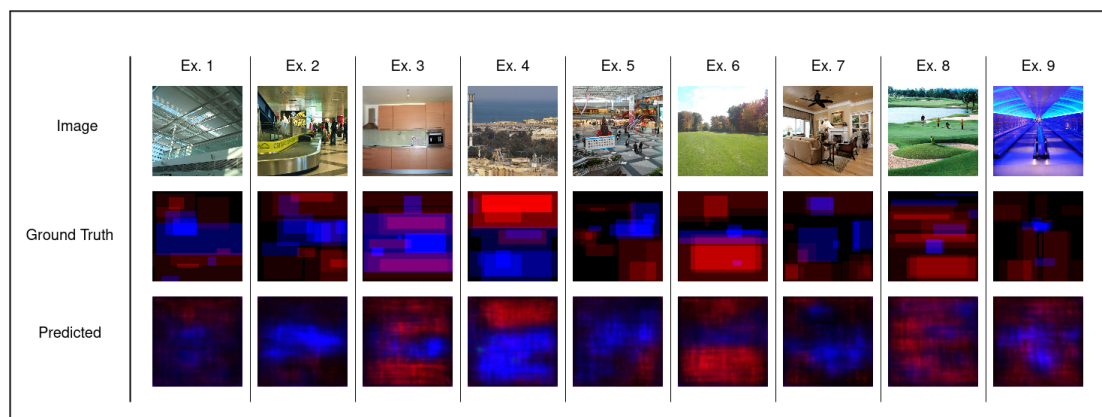


Figure C.9: Additional prediction results for complexity maps.

## MLR Table for Complexity/Memorability

This table is a supplement to Chapter 6 and shows results for a multiple linear regression conducted over all complexity metrics vs scene memorability.

Table D.1: Results of multiple linear regression. Coefficients for each variable are shown, as is the coefficient of multiple regression (R) and variance explained (R-squared). All regressions are significant. Complexity can explain a small, but significant portion of variance inherent in memorability data for DPrime, hit rate, and false alarm rate. Significant values shown in bold,  $p < 0$ : \*\*\*, 0.001: \*\*, 0.05: \*

	D-Prime	Hit Rate	False Alarm Rate
Constant	0.9394	0.2588	0.07
Complex Intensity	0.0166	-0.002	<b>-0.003**</b>
Complex Coverage	<b>-0.724*</b>	-0.041	<b>0.119***</b>
Simple Intensity	<b>-0.0284**</b>	<b>-0.005*</b>	0.002
Simple Coverage	<b>0.7745**</b>	<b>0.1818**</b>	-0.028
Complexity Scores	<b>0.0015***</b>	<b>0.004***</b>	-0.0005
R	0.2714	0.256	0.181
R-squared	0.074	0.065	0.033
Adjusted R-Squared	0.068	0.06	0.027
Observations	800		

## Bibliography

- [1] Samuel A Turner and Paul Silvia. ‘Must interesting things be pleasant? A test of competing appraisal structures’. In: *Emotion (Washington, D.C.)* 6 (Dec. 2006), pp. 670–4.
- [2] Erdem Akagündüz, A. G. Bors and K. K. Evans. ‘Defining Image Memorability using the Visual Memory Schema’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.9 (2020), pp. 2165–2178.
- [3] Krizhevsky Alex, Ilya Sutskever and Geoffrey E Hinton. ‘Imagenet classification with deep convolutional networks’. In: *volume-1; pages-1097–1105; NIPS’12 Proceedings of the 25th International Conference on Neural Information Processing Systems*.
- [4] M. Arjovsky, S. Chintala and L. Bottou. ‘Wasserstein Generative Adversarial Networks’. In: *Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 70*. 2017, pp. 214–223.
- [5] Alan D. Baddeley and Graham J. Hitch. ‘Is the Levels of Processing effect language-limited?’ In: *Journal of Memory and Language* 92 (Feb. 2017), pp. 1–13.
- [6] Wilma A Bainbridge. ‘The resiliency of image memorability: A predictor of memory separate from attention and priming’. In: *Neuropsychologia* 141 (2020), p. 107408.
- [7] Wilma A. Bainbridge, Daniel D. Dilks and Aude Oliva. ‘Memorability: A stimulus-driven perceptual neural signature distinctive from memory’. In: *NeuroImage* 149 (1st Apr. 2017), pp. 141–152.
- [8] Sathisha Basavaraju, Sibaji Gaj and Arijit Sur. ‘Object Memorability Prediction using Deep Learning: Location and Size Bias’. In: *Journal of Visual Communication and Image Representation* 59 (2019), pp. 117–127.

## BIBLIOGRAPHY

- [9] Sathisha Basavaraju, Paritosh Mittal and Arijit Sur. ‘Image memorability: The role of depth and motion’. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 699–703.
- [10] David Bau et al. ‘Understanding the role of individual units in a deep neural network’. In: *Proceedings of the National Academy of Sciences* (2020).
- [11] Y. Baveye et al. ‘Deep Learning for Image Memorability Prediction: The Emotional Bias’. In: *Proc. of the 24th ACM Int. Conf. on Multimedia*. 2016, pp. 491–495.
- [12] George David Birkhoff. *Aesthetic measure*. Harvard University Press, 2013.
- [13] Richard A. Block. ‘Intent to remember briefly presented human faces and other pictorial stimuli enhances recognition memory’. In: *Memory & Cognition* 37.5 (July 2009), pp. 667–678.
- [14] Bernhard E Boser, Isabelle M Guyon and Vladimir N Vapnik. ‘A training algorithm for optimal margin classifiers’. In: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152.
- [15] T. F. Brady et al. ‘Visual long-term memory has a massive storage capacity for object details’. In: *Proceedings of the National Academy of Sciences* 105.38 (23rd Sept. 2008), pp. 14325–14329.
- [16] Timothy F Brady, Talia Konkle and George A Alvarez. ‘A review of visual memory capacity: Beyond individual items and toward structured representations’. In: *Journal of Vision* 11.5, Article 4 (2011).
- [17] W. Brendel and M. Bethge. ‘Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet’. In: *Proc. Int. Conf. on Learning Representations (ICLR)*. 2019.
- [18] Andrew Brock, Jeff Donahue and Karen Simonyan. ‘Large scale GAN training for high fidelity natural image synthesis’. In: *arXiv preprint arXiv:1809.11096* (2018).
- [19] Z. Bylinskii et al. ‘What Do Different Evaluation Metrics Tell Us About Saliency Models?’ In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.3 (Mar. 2019), pp. 740–757.

## BIBLIOGRAPHY

- [20] Zoya Bylinskii et al. ‘Intrinsic and extrinsic effects on image memorability’. In: *Vision Research* 116 (2015), pp. 165–178.
- [21] Maurizio Cardaci et al. ‘A fuzzy approach to the evaluation of image complexity’. In: *Fuzzy Sets and Systems* 160.10 (2009). Special Issue: Fuzzy Sets in Interdisciplinary Perception and Intelligence, pp. 1474–1484.
- [22] Maurizio Cardaci et al. ‘A fuzzy approach to the evaluation of image complexity’. In: *Fuzzy Sets and Systems* 160.10 (2009), pp. 1474–1484.
- [23] Bora Celikkale, Aykut Erdem and Erkut Erdem. ‘Predicting memorability of images using attention-driven spatial pooling and image semantics’. In: *Image and vision Computing* 42 (2015), pp. 35–46.
- [24] Nadine Chang et al. ‘BOLD5000, a public fMRI dataset while viewing 5000 visual images’. In: *Scientific data* 6.1 (2019), pp. 1–18.
- [25] Xi Chen et al. ‘InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets’. In: *Advances in Neural Information Processing Systems* 29. 2016, pp. 2172–2180.
- [26] Bowen Cheng, Alex Schwing and Alexander Kirillov. ‘Per-pixel classification is not all you need for semantic segmentation’. In: *Advances in Neural Information Processing Systems* 34 (2021).
- [27] Gianluigi Ciocca et al. ‘Does color influence image complexity perception?’ In: *International Workshop on Computational Color Imaging*. Springer. 2015, pp. 139–148.
- [28] T. F. Cootes, G. J. Edwards and C. J. Taylor. ‘Active appearance models’. In: *Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 1407*. 1998, pp. 484–498.
- [29] Silvia Corchs, Gianluigi Ciocca and Francesca Gasparini. ‘Human perception of image complexity: real scenes versus texture patches’. In: *Journal of Alzheimer’s Disease*. Vol. 53. Abstracts for the Second International Meeting of the Milan Center for Neuroscience (Neuromi): Prediction and Prevention of Dementia: New Hope (Milan, July 6–8, 2016). 2016, s51.
- [30] Silvia Elena Corchs et al. ‘Predicting complexity perception of real world images’. In: *PloS one* 11.6 (2016), e0157986.

## BIBLIOGRAPHY

- [31] Corinna Cortes, Mehryar Mohri and Afshin Rostamizadeh. ‘L2 Regularization for Learning Kernels’. In: *CoRR* abs/1205.2653 (2012). arXiv: 1205.2653.
- [32] Corinna Cortes and Vladimir Vapnik. ‘Support-vector networks’. In: *Machine Learning* 20.3 (1st Sept. 1995), pp. 273–297.
- [33] Corbin A. Cunningham, Michael A. Yassa and Howard E. Egeth. ‘Massive memory revisited: Limitations on storage capacity for object details in visual long-term memory’. In: *Learning & Memory* 22.11 (Nov. 2015), pp. 563–566.
- [34] N. Dalal and B. Triggs. ‘Histograms of oriented gradients for human detection’. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05). Vol. 1. June 2005, 886–893 vol. 1.
- [35] Hy Day. ‘The importance of symmetry and complexity in the evaluation of complexity, interest and pleasingness’. In: *Psychonomic Science* 10.10 (1968), pp. 339–340.
- [36] Jia Deng et al. ‘Imagenet: A large-scale hierarchical image database’. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [37] Sagnik Dhar, Vicente Ordonez and Tamara L. Berg. ‘High level describable attributes for predicting aesthetics and interestingness’. In: *2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*. 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011. 22nd Sept. 2011, pp. 1657–1664.
- [38] I. G. Dobbins et al. ‘Predicting individual false alarm rates and signal detection theory: A role for remembering.’ In: *Memory and Cognition* 28 (2000), pp. 1347–1356.
- [39] Don Donderi. ‘Visual Complexity: A Review.’ In: *Psychological bulletin* 132 (Feb. 2006), pp. 73–97.
- [40] R. Dubey et al. ‘What Makes an Object Memorable?’ In: *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*. 2015, pp. 1089–1097.
- [41] Endel Tulving. ‘Episodic and Semantic Memory’. In: *Organization of Memory*. Academic Press, 1972.



## BIBLIOGRAPHY

- [42] Russell A Epstein et al. ‘Cortical correlates of face and scene inversion: a comparison’. In: *Neuropsychologia* 44.7 (2006), pp. 1145–1158.
- [43] Karla K. Evans and Alan Baddeley. ‘Intention, attention and long-term memory for visual scenes: It all depends on the scenes’. In: *Cognition* 180 (Nov. 2018), pp. 24–37.
- [44] J. Fajtl et al. ‘AMNet: Memorability Estimation with Attention’. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6363–6372.
- [45] Alexandra Forsythe. ‘Visual Complexity: Is That All There Is?’ In: *Engineering Psychology and Cognitive Ergonomics*. Ed. by Don Harris. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 158–166.
- [46] D. Garcia-Gasulla et al. ‘On the Behavior of Convolutional Nets for Feature Extraction’. In: *Jour. of Artificial Intelligence Research* 61 (2018), pp. 563–592.
- [47] Miguel A García-Pérez and Rocío Alcalá-Quintana. ‘Interval bias in 2AFC detection tasks: sorting out the artifacts’. In: *Attention, Perception, & Psychophysics* 73.7 (2011), pp. 2332–2352.
- [48] Lore Goetschalckx, Pieter Moors and Johan Wagemans. ‘Image memorability across longer time intervals’. In: *Memory* 26.5 (2018), pp. 581–588. eprint: <https://doi.org/10.1080/09658211.2017.1383435>.
- [49] Lore Goetschalckx et al. ‘Understanding and Predicting Image Memorability at a Large Scale’. In: *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*. 2015, pp. 5744–5753.
- [50] A. Gonzalez-Garcia, D. Modolo and V. Ferrari. ‘Do Semantic Parts Emerge in Convolutional Neural Networks?’ In: *Int. Journal of Computer Vision* 126.5 (2018), pp. 476–494.
- [51] I. Goodfellow et al. ‘Generative adversarial nets’. In: *Advances in Neural Inf. Proc. Systems (NIPS)*. 2014, pp. 2672–2680.
- [52] Daniel L. Greenberg and Mieke Verfaellie. ‘Interdependence of episodic and semantic memory: Evidence from neuropsychology’. In: *Journal of the International Neuropsychological Society : JINS* 16.5 (Sept. 2010), pp. 748–753.

## BIBLIOGRAPHY

- [53] Yağmur Güçlütürk et al. ‘Representations of naturalistic stimulus complexity in early and associative visual and auditory cortices’. In: *Scientific reports* 8.1 (2018), pp. 1–16.
- [54] I. Gulrajani et al. ‘Improved Training of Wasserstein GANs’. In: *Advances in Neural Inf. Proc. Systems (NIPS)*. 2017, pp. 5769–5779.
- [55] M. Gygli et al. ‘The Interestingness of Images’. In: *2013 IEEE International Conference on Computer Vision*. 2013 IEEE International Conference on Computer Vision. Dec. 2013, pp. 1633–1640.
- [56] Jonathan Harel, Christof Koch and Pietro Perona. ‘Graph-Based Visual Saliency’. In: *Advances in Neural Information Processing Systems* (2007), p. 8.
- [57] Daniel Cabrini Hauage and Noah Snavely. ‘Image matching using local symmetry features’. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 206–213.
- [58] Kaiming He et al. ‘Deep residual learning for image recognition’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [59] Kaiming He et al. ‘Identity mappings in deep residual networks’. In: *European conference on computer vision*. Springer. 2016, pp. 630–645.
- [60] Christopher Heaps and Stephen Handel. ‘Similarity and features of natural textures.’ In: *Journal of Experimental Psychology: Human Perception and Performance* 25.2 (1999), p. 299.
- [61] Martin Heusel et al. ‘GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium’. In: *Advances in Neural Information Processing Systems (NIPS)*. 2017, pp. 6629–6640.
- [62] Sepp Hochreiter and Jürgen Schmidhuber. ‘Long Short-Term Memory’. In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780.
- [63] Thomas Hofmann, Bernhard Schölkopf and Alexander J Smola. ‘Kernel methods in machine learning’. In: *The annals of statistics* 36.3 (2008), pp. 1171–1220.
- [64] Z. Hu and A.G. Bors. ‘Conditional Attention for Content-based Image Retrieval’. In: *Proc. British Machine Vision Conference (BMVC)*. 2020.
- [65] *Inquisit*: <https://www.millisecond.com>.

## BIBLIOGRAPHY

- [66] Helene Intraub and Christopher A. Dickinson. ‘False memory 1/20th of a second later: what the early onset of boundary extension reveals about perception’. In: *Psychological Science* 19.10 (Oct. 2008), pp. 1007–1014.
- [67] P. Isola et al. ‘What Makes a Photograph Memorable?’ In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (July 2014), pp. 1469–1482.
- [68] Phillip Isola et al. *Understanding the intrinsic memorability of images*. Tech. rep. MASSACHUSETTS INST OF TECH CAMBRIDGE, 2011.
- [69] Phillip Isola et al. ‘What makes an image memorable?’ In: *CVPR 2011*. IEEE, 2011, pp. 145–152.
- [70] Tero Karras et al. ‘Progressive Growing of GANs for Improved Quality, Stability, and Variation’. In: *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1710.10196*. 2018.
- [71] Yan Ke, Xiaoou Tang and Feng Jing. ‘The design of high-level features for photo quality assessment’. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 1. IEEE, 2006, pp. 419–426.
- [72] Todd A Kelley, Marvin M Chun and Kao-Ping Chua. ‘Effects of scene inversion on change detection of targets matched for visual salience’. In: *Journal of Vision* 3.1 (2003), pp. 1–1.
- [73] Aditya Khosla et al. ‘Memorability of image regions’. In: (2012).
- [74] Aditya Khosla et al. ‘Modifying the Memorability of Face Photographs’. In: *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*. 2013, pp. 2390–2398.
- [75] Aditya Khosla et al. ‘Understanding and Predicting Image Memorability at a Large Scale’. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, Dec. 2015, pp. 2390–2398.
- [76] D. P. Kingma and M. Welling. ‘Auto-Encoding Variational Bayes’. In: *Proc. Int. Conf. on Learning Repres. (ICLR)*. 2014.
- [77] Diederik P Kingma and Jimmy Ba. ‘Adam: A method for stochastic optimization’. In: *arXiv preprint arXiv:1412.6980* (2014).
- [78] Andrei N Kolmogorov. ‘Three approaches to the quantitative definition of information’. In: *Problems of information transmission* 1.1 (1965), pp. 1–7.

## BIBLIOGRAPHY

- [79] Talia Konkle et al. ‘Scene memory is more detailed than you think: The role of categories in visual long-term memory’. In: *Psychological science* 21.11 (2010), pp. 1551–1556.
- [80] Wilma Koutstaal et al. ‘False recognition of abstract versus common objects in older and younger adults: testing the semantic categorization account.’ In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.4 (2003), p. 499.
- [81] Wilma Koutstaal and Daniel L. Schacter. ‘Gist-Based False Recognition of Pictures in Older and Younger Adults’. In: *Journal of Memory and Language* 37.4 (1st Nov. 1997), pp. 555–583.
- [82] A. Kroner et al. ‘Contextual encoder–decoder network for visual saliency prediction’. In: *Neural Networks* 129 (2020), pp. 261–270.
- [83] C. Kyle-Davidson, A.G. Bors and K.K. Evans. ‘Predicting Visual Memory Schemas with Variational Autoencoders’. In: *Proc. British Machine Vision Conference (BMVC)*. 2019.
- [84] Cameron Kyle-Davidson, Adrian G Bors and Karla K Evans. ‘Modulating human memory for complex scenes with artificially generated images’. In: *Scientific Reports* 12.1 (2022), pp. 1–15.
- [85] Adam M Larson et al. ‘The spatiotemporal dynamics of scene gist recognition.’ In: *Journal of Experimental Psychology: Human Perception and Performance* 40.2 (2014), p. 471.
- [86] K. Lasinger et al. ‘Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer’. In: *accepted to IEEE T-PAMI, arXiv preprint arXiv:1907.01341* (2021).
- [87] Y. Lecun et al. ‘Gradient-based learning applied to document recognition’. In: *Proceedings of the IEEE* 86.11 (Nov. 1998), pp. 2278–2324.
- [88] Min Lin, Qiang Chen and Shuicheng Yan. ‘Network in network’. In: *arXiv preprint arXiv:1312.4400* (2013).
- [89] Tsung-Yi Lin et al. ‘Microsoft coco: Common objects in context’. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.
- [90] Yanxi Liu, Hagit Hel-Or and Craig S Kaplan. *Computational symmetry in computer vision and computer graphics*. Now publishers Inc, 2010.

## BIBLIOGRAPHY

- [91] Elizabeth F. Loftus. ‘Planting misinformation in the human mind: A 30-year investigation of the malleability of memory’. In: *Learning & Memory* 12.4 (1st July 2005), pp. 361–366.
- [92] David G. Lowe. ‘Distinctive Image Features from Scale-Invariant Keypoints’. In: *International Journal of Computer Vision* 60.2 (1st Nov. 2004), pp. 91–110.
- [93] J. Lu, M. Xu and Z. Wang. ‘Predicting the memorability of natural-scene images’. In: *2016 Visual Communications and Image Processing (VCIP)*. 2016 Visual Communications and Image Processing (VCIP). Nov. 2016, pp. 1–4.
- [94] Jiří Lukavský and Filip Děchtěrenko. ‘Visual properties and memorising scenes: Effects of image-space sparseness and uniformity’. In: *Attention, Perception, & Psychophysics* 79.7 (1st Oct. 2017), pp. 2044–2054.
- [95] Christopher R Madan et al. ‘Visual complexity and affect: Ratings reflect more than meets the eye’. In: *Frontiers in psychology* 8 (2018), p. 2368.
- [96] M. Mancas and O. Le Meur. ‘Memorability of natural scenes: The role of attention’. In: *2013 IEEE International Conference on Image Processing*. 2013 IEEE International Conference on Image Processing. Sept. 2013, pp. 196–200.
- [97] Yalda Mohsenzadeh et al. ‘The perceptual neural trace of memorable unseen scenes’. In: *Scientific reports* 9.1 (2019), pp. 1–10.
- [98] N. Murray, L. Marchesotti and F. Perronnin. ‘AVA: A large-scale database for aesthetic visual analysis’. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI: IEEE, June 2012, pp. 2408–2415.
- [99] Fintan Nagle and Nilli Lavie. ‘Predicting human complexity perception of real-world scenes’. In: *Royal Society Open Science* 7.5 (2020), p. 191487.
- [100] Peter Neri. ‘Semantic control of feature extraction from natural scenes’. In: *Journal of Neuroscience* 34.6 (2014), pp. 2374–2388.
- [101] Augustus Odena, Christopher Olah and Jonathon Shlens. ‘Conditional Image Synthesis With Auxiliary Classifier GANs’. In: *Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 70*. 2017, pp. 2642–2651.
- [102] Aude Oliva. ‘Gist of the scene’. In: *Neurobiology of attention*. Elsevier, 2005, pp. 251–256.

## BIBLIOGRAPHY

- [103] Aude Oliva and Antonio Torralba. ‘Modeling the shape of the scene: A holistic representation of the spatial envelope’. In: *International journal of computer vision* 42.3 (2001), pp. 145–175.
- [104] Aude Oliva and Antonio Torralba. ‘Building the gist of a scene: The role of global image features in recognition’. In: *Progress in brain research* 155 (2006), pp. 23–36.
- [105] Aude Olivia et al. ‘Identifying the perceptual dimensions of visual complexity of scenes’. In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 26. 26. 2004.
- [106] Viorica Patraucean, Rafael Grompone von Gioi and Maks Ovsjanikov. ‘Detection of mirror-symmetric image patches’. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2013, pp. 211–216.
- [107] *Prolific*: <https://www.prolific.co/>.
- [108] J. Rigau, M. Feixas and M. Sbert. ‘An Information-Theoretic Framework for Image Complexity’. In: *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*. Computational Aesthetics’05. Girona, Spain: Eurographics Association, 2005, pp. 177–184.
- [109] Jaume Rigau, Miquel Feixas and Mateu Sbert. ‘Conceptualizing Birkhoff’s Aesthetic Measure Using Shannon Entropy and Kolmogorov Complexity.’ In: Jan. 2007, pp. 105–112.
- [110] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton*. Cornell Aeronautical Laboratory, 1957.
- [111] Ruth Rosenholtz, Yuanzhen Li and Lisa Nakano. ‘Measuring visual clutter’. In: *Journal of vision* 7.2 (2007), pp. 17–17.
- [112] Michael G Ross and Aude Oliva. ‘Estimating perception of scene layout properties from global image features’. In: *Journal of vision* 10.1 (2010), pp. 2–2.
- [113] Nicole C Rust and Vahid Mehrpour. ‘Understanding image memorability’. In: *Trends in cognitive sciences* 24.7 (2020), pp. 557–568.
- [114] Elham Saracee, Mona Jalal and Margrit Betke. ‘Visual complexity analysis using deep intermediate-layer features’. In: *Computer Vision and Image Understanding* 195 (2020), p. 102949.

## BIBLIOGRAPHY

- [115] Wataru Sato and Sakiko Yoshikawa. ‘Recognition memory for faces and scenes’. In: *The Journal of general psychology* 140.1 (2013), pp. 1–15.
- [116] Daniel L Schacter and Donna Rose Addis. ‘The cognitive neuroscience of constructive memory: remembering the past and imagining the future’. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481 (29th May 2007), pp. 773–786.
- [117] Daniel L Schacter and Donna Rose Addis. ‘The ghosts of past and future’. In: *Nature* 445.7123 (2007), pp. 27–27.
- [118] Jonathan W Schooler, Delia Gerhard and Elizabeth F Loftus. ‘Qualities of the unreal.’ In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 12.2 (1986), p. 171.
- [119] Jianbo Shi and Jitendra Malik. ‘Normalized cuts and image segmentation’. In: *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000), pp. 888–905.
- [120] Aliaksandr Siarohin et al. ‘How to Make an Image More Memorable? A Deep Style Transfer Approach’. In: *Proc. of ACM on Int. Conf. on Multimedia Retrieval (ICMR)*. 2017, pp. 322–329.
- [121] Oleksii Sidorov. ‘Changing the Image Memorability: From Basic Photo Editing to GANs’. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPR-w)*. 2019, pp. 790–799.
- [122] Karen Simonyan and Andrew Zisserman. ‘Very deep convolutional networks for large-scale image recognition’. In: *arXiv preprint arXiv:1409.1556* (2014).
- [123] Joan G Snodgrass and Mary Vanderwart. ‘A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity.’ In: *Journal of experimental psychology: Human learning and memory* 6.2 (1980), p. 174.
- [124] G. Spanò, H. Intraub and J. O. Edgin. ‘Testing the “Boundaries” of boundary extension: Anticipatory scene representation across development and disorder’. In: *Hippocampus* 27.6 (2017), pp. 726–739.
- [125] Hammad Squalli-Houssaini et al. ‘Deep learning for predicting image memorability’. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 2371–2375.

## BIBLIOGRAPHY

- [126] Lionel Standing. ‘Learning 10000 pictures’. In: *Quarterly Journal of Experimental Psychology* 25.2 (May 1973), pp. 207–222.
- [127] Antonio Torralba. ‘How many pixels make an image’. In: *Visual neuroscience* 26.1 (2009), pp. 123–131.
- [128] Dorothy Tse et al. ‘Schemas and Memory Consolidation’. In: *Science* 316.5821 (2007), pp. 76–82. eprint: <https://science.sciencemag.org/content/316/5821/76.full.pdf>.
- [129] Endel Tulving. ‘Elements of episodic memory’. In: (1983).
- [130] Rolf Ulrich and Dirk Vorberg. ‘Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators’. In: *Attention, Perception, & Psychophysics* 71.6 (2009), pp. 1219–1227.
- [131] Ashish Vaswani et al. ‘Attention is all you need’. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.
- [132] Edward A. Vessel and Nava Rubin. ‘Beauty and the beholder: highly individual taste for abstract, but not real-world images’. In: *Journal of Vision* 10.2 (22nd Feb. 2010), pp. 18.1–14.
- [133] ‘VISCHEMA 1&2’. In: *URL: <https://www.cs.york.ac.uk/vischema/>*. 2019.
- [134] Melissa Le-Hoa Võ, Zoya Bylinskii and Aude Oliva. ‘Image Memorability In The Eye Of The Beholder: Tracking The Decay Of Visual Scene Representations’. In: *bioRxiv* (24th May 2017), p. 141044.
- [135] Dirk B Walther et al. ‘Natural scene categories revealed in distributed patterns of activity in the human brain’. In: *Journal of neuroscience* 29.34 (2009), pp. 10573–10581.
- [136] X. Wang et al. ‘Non-local Neural Networks’. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7794–7803.
- [137] Z. Wang et al. ‘Image Quality Assessment: From Error Visibility to Structural Similarity’. In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612.
- [138] Xiang Wei et al. ‘Improving the Improved Training of Wasserstein GANs’. In: *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1803.01541*. 2018.



## BIBLIOGRAPHY

- [139] Jeremy M Wolfe and Yoana I Kuzmova. ‘How many pixels make a memory? Picture memory for small pictures’. In: *Psychonomic bulletin & review* 18.3 (2011), pp. 469–475.
- [140] Jianxiong Xiao et al. ‘Sun database: Large-scale scene recognition from abbey to zoo’. In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE. 2010, pp. 3485–3492.
- [141] S. Yoon and J. Kim. ‘Object-Centric Scene Understanding for Image Memorability Prediction’. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Apr. 2018, pp. 305–308.
- [142] Fisher Yu et al. ‘LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop’. In: *arXiv preprint arXiv:1506.03365*. 2015.
- [143] H. Yu and S. Winkler. ‘Image complexity and spatial information’. In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. 2013, pp. 12–17.
- [144] Honghai Yu and Stefan Winkler. ‘Image complexity and spatial information’. In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE. 2013, pp. 12–17.
- [145] Matthew D Zeiler. ‘Adadelta: an adaptive learning rate method’. In: *arXiv preprint arXiv:1212.5701* (2012).
- [146] H. Zhang et al. ‘Self-Attention Generative Adversarial Networks’. In: *Proc. of Int. Conf. on Machine Learning (ICML), vol. PMLR 97*. 2019, pp. 7354–7363.
- [147] B. Zhou et al. ‘Object Detectors Emerge in Deep Scene CNNs’. In: *Proc. Int. Conf. on Learning Representations (ICLR)*. 2015.
- [148] Bolei Zhou et al. ‘Places: An image database for deep scene understanding’. In: *arXiv preprint arXiv:1610.02055* (2016).
- [149] Bolei Zhou et al. ‘Scene parsing through ade20k dataset’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 633–641.