Article:
Kyle-Davidson, Cameron, Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 and Evans, Karla orcid.org/0000-0002-8440-1711 (2025) Visual Memory Schemas for Localised Image Memorability Prediction. IEEE Transactions on Cognitive and Developmental Systems. pp. 1-13. ISSN 2379-8920

https://doi.org/10.1109/TCDS.2025.3533112

Visual Memory Schemas for Localised Image Memorability Prediction

Cameron Kyle-Davidson, Adrian G. Bors, Karla K. Evans

***Abstract*—Visual memory schemas (VMS) capture the regions of scene images that cause that scene to be remembered, providing a two-dimensional memorability map which indicates the parts of a given scene that match to mental schemas held in the mind. Despite the advantage of determining which parts of an image lead to remembering said image, VMS prediction capabilities lag behind those of single-score memorability. Compared to predicting single-score ratings for the likelihood of a person remembering an image, VMS prediction is a significantly harder task, due to increased computational complexity, minimal model development compared to single score, and lack of relevant data. In this work, we aim to improve methods for two-dimensional memorability prediction. We first significantly increase the size of a database containing VMS maps obtained from participants in a scene memorisation experiment, and then we develop an architecture which leverages existing single-score image memorability datasets to predict VMS maps. Our final model, 'DF-VMS' significantly outperforms existing VMS prediction models, with a performance increase of 11.8%. Additionally, we explore the semantic structures which are actually captured by visual memory schemas, determining the combination of scene elements which lead to remembering that scene.**

***Index Terms*—Image memorability prediction, Visual Memory Schemas, deep learning, visual memory, cognitive psychology**

I. Introduction

Images vary across a wide spectrum of memorability; some images stick in the mind and can be recalled easily at a later time, whereas others fade rapidly from memory. Identifying which visual features support or hinder the memorability of a given scene has been a recent focus of the computer vision community, over a wide variety of images, from faces to scenes [1]. By building on foundational work from decades of cognitive psychology research, large-scale computational approaches have been deployed to both understand, and predict, scene memory [2]–[4]. Improved understanding of *what* is memorable has the potential to lead to both better understanding of the operation of the visual memory circuitry of the brain, as well as several practical applications: improved advertising, creating better educational aids, even tracking cognitive decline through comparison of a patients memory performance against a standardised baseline. Recent research has lead to the development of image memorability datasets of thousands of images, each paired with a ground-truth memorability score gathered via repeat-recognition memory experiments over hundreds of observers [5].

Manuscript created May, 2024; Cameron Kyle-Davidson and Karla K. Evans are with the Department of Psychology. Adrian G. Bors is with the Department of Computer Science. All authors are with the University of York, Yorkshire, England

Computational analysis on these datasets has made progress in determining which image features lead to an image being remembered by a human. Generally, there is a high degree of consistency ($\rho = 0.75$) [2] between participants memory for images; that is, 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. Instead, high level semantic attributes such as image category, the contents of the image (i.e, the objects present), and scene dynamics (what is actually occurring in the image) appear to better correlate with image memorability [3]. Predictive models have now been developed that, given a scene image, can indicate how likely the average human observer will be to actually remember that scene.

Most image memorability models only produce a singular 'score' indicative of that image's memorability. This does not reveal which elements in the scene are actually *driving* that memorability. More recent research introduces the concept of a "Visual Memory Schema" (VMS) [6], which capture the scene regions that are responsible for a human observer's ability to recall having seen that particular scene. Generally, it is not the individual pixel intensities that lead to a scene being remembered - it is instead the collective semantic content present in the scene, and the relations between that semantic content [7]. For example, a beach scene is remembered due to the presence of sand, ocean, parasols and beach balls in various arrangements; all of which match a prototypical mental schema of what a beach *is*. Practically, VMS-based predictive models can produce more than just a single score representing the memorability of a given image; they can highlight the areas of the image which correspond to semantically relevant regions that are likely to cause a human to remember that scene. This effectively creates a two-dimensional per-pixel image memorability map for a given scene image. This allows for an improved understanding of why a certain scene was remembered, and another forgotten; refining the concept of image memorability from an abstract score into concrete image-space details.

Currently, the development of visual memory schema prediction techniques lags behind single-score methods due to increased technical difficulty (predicting 2D memorability maps is significantly harder than prior one-dimensional approaches), a current lack of knowledge of which potential computational techniques may lead to improved predictive performance, and a lack of available data. Current single-score memorability datasets have samples in the tens of thousands; for two-dimensional memorability, there are only eight hundred images

obtained from image memorisation experiments. This is due to the increased experimental complexity of gathering two-dimensional data. In this work we tackle these problems, and propose a new model for visual memory schema prediction which leverages existing single-score datasets for improved predictive performance. The main contributions of this paper are:

- An expanded visual memory schema memorability dataset that consists of 3,461 new image/memorability map pairs, based upon the Visual Memory Schema paradigm, and extracted via human observer experiments. This dataset (which we call 'VMS4k') can be seamlessly combined with the existing VMS dataset, resulting in 4,261 image/VMS pairs, an improvement in data availability of over 400%.
- A comprehensive investigation into which computational techniques offer the best performance for two-dimensional memorability prediction with a theoretical grounding in cognitive psychology.
- The development of a two-dimensional memorability prediction technique that takes advantage of both the new two-dimensional memorability map dataset, and leverages existing single-score datasets to improve 2D memorability map prediction.
- A quantification of visual memory schema characteristics, including an object-based analysis of the semantic structures contained by memorable scene regions, and which match mental schemas held in the human mind.

We first gather extensive visual memory schema data. Using this new dataset, we provide a set of comprehensive baselines for a variety of VMS map prediction methods, and finally design a new architecture which can take advantage of pre-existing single-score memorability data to boost predictive performance for Visual Memory Schemas. The rest of the paper is organised as follows: In Section II we provide an overview of the existing research studies into Visual Memory Schemas. In Section III we provide the proposed methodology. In Section IV we detail the experimental setup for collecting the dataset used in the experiments in this paper. The experimental results are provided in Section V, while the conclusions of this study are drawn in Section VI.

II. Related Work

A. Image Memorability Prediction

The field of memorability prediction burgeoned in the early 2010s, with the work of Isola *et al.* [2]. Isola gathered memorability data on over two thousand images. Each image was scored by 78 different participants in total. The mean memorability score (defined as the percentage of correct recognitions) is around 67.5%. The Spearmans rank coefficient for 25 random splits of the gathered data is 0.75, indicative of strong agreement between participants; generally, humans remember and forget the same images. Interestingly, human judgements on whether an image is likely to be memorable or not were negatively correlated with the images actual memorability data [8]. Humans are hence poor judges of whether they are likely to remember or forget an image. That is, prior to undertaking a visual memorability experiment, a person cannot accurately predict which of the images they are going to see, are going to be the most memorable. This does not mean that they cannot trust their own judgement upon recollection. Isola proposed a support vector regression approach based upon computed image features capable of predicting ground-truth memorability scores (rank correlation, $r = 0.489$), with image category being the strongest predictor of all considered.

Later, Khosla *et al.* [9] developed a probabilistic model to simulate a 'noisy memory process' , with similar performance to the above, hypothesising the likelihood of remembering an image is related to the distance between the actual perceived image and the noisy internal representation. Shortly after, Khosla *et al.* [5] introduced LaMem, a dataset of 60000 images and memorability scores derived from human observer experiments. A convolutional neural network (CNN) trained on these can reach a rank correlation with human data of 0.64. This signified the beginning of the now universal deep neural network approach to image memorability modelling. Following the approach from [5], Lukavsky *et al.* [10] developed a context-dependent method based upon late-stage CNN features, while Yoon *et al.* [11] proposed a segmentation based approach, and Squalli-Houssaini *et al.* [12] a Long Short-term Memory (LSTM)-based captioning approach. Recently, neural networks have become capable of predicting image single-score memorability to a degree that matches human-level memory performance [13], [14]. It is tempting, given the success of these predictive approaches, to believe that these methods approach some ground-truth 'objective' memorability value held within the image. However, an image cannot have a memorability value in a vacuum - a human being is required to be present in order to remember that image. Instead, these prior studies, and the study presented in this work, work under the hypothesis that human memorability, while subjective, is *consistent* enough across humans to allow for prediction.

B. Image Memorability Analysis

Computational analysis has enabled large-scale investigation into which image elements lead to a correspondingly greater or lower memorability score. Mancas *et al.* [15] finds that eye-fixation duration increases with greater memorability, and models including attention-based metrics enable reduced feature dimensionality. Both studies by Celikkale et al. [16] and Isola et al. [3] find that saliency and semantic features (scene category label) can capture scene memorability. Analyses on large scale memorability datasets have revealed both the impact of context, with contextually distinct images proving more memorable [17], as well as the contribution of perceived depth and motion [18]. Other research studies [19], [20] have found that individual objects within an image have their own degrees of memorability, which can be primarily explained with semantic features. Dubey *et al.* [19] finds that a CNN achieves reasonable predictive performance when shown segmented object images. The object with the greatest memorability is loosely predictive of the overall image memorability. However, none of these techniques so far capture the overall image region that is driving the memorability of the image.

C. Visual Memory Schemas

In cognitive science, a schema is a mental construct that facilitates the encoding of a scene [7], [21]. 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. Introduced by Akagunduz *et al.* [6] the VISCHEMA dataset consists of 800 high-resolution scene images paired with 800 memorability maps ('VMS Maps') that capture the regions of scene images that lead to an image being either remembered (a 'true schema'), or falsely remembered (a 'false schema'). False schemas lead a person to believe they have seen a scene, when in fact they have not. This VMS memorability data was gathered from 90 participants during a repeat-recognition memory experiment, for eight different real-world scene categories, with one hundred images per category. These categories range from indoor scenes such as living rooms, kitchens, and conference centres, to outdoor environments such as deserts, mountains, and golf courses. Images displaying features considered classically memorable are specifically excluded from the corpus of scene images used in the experiment; this includes recognisable landmarks, attention-drawing text, and people looking directly at the camera. This results in a more stable dataset overall, as the memorability of the scene is more likely to be affected by semantic scene content. VMS Maps are highly consistent: with a split-half reliability of 0.7, participants agree on scene regions that cause the scene to be remembered. They have only a weak correlation with eye fixations and computational saliency, and are a robust enough measure to be used to influence the overall memorability of generated scene images [22]. Predicting these maps is a significant challenge given the low data amounts and increased computational complexity compared to single-score prediction. While these maps (and captured memorable regions) are known to relate to cognitive schemas, exactly how the image contents of each region drives memorability is still an active area of research. Generally, the scene contents, arrangement, and detail present in that region is likely to have an effect [6], [23], [24]. In Akagunduz *et al.* [6], MemNet, and four VGG [25] architectures underwent transfer learning for VMS prediction. Each network could produce a 20×20 pixel VMS map that combined both true and false schemas into a singular grayscale probability map. Later work [26] employed a variational approach [27] to predict both true and false schemas separately, with an output resolution of 224×224 pixels. Despite this initial success, techniques for VMS map prediction are still relatively unexplored compared to single-score memorability prediction, due to the lack of available data and limited investigation into applicable computational tools.

III. Methods

In this section we will detail the various architectural components and designs which we test for memory map prediction. Our approach for each component is psychologically motivated, aiming to take advantage of pre-existing knowledge of how memory is processed in the brain, and our designs data-motivated; aiming to take advantage of pre-existing datasets in addition to our own. We aim to explore multi-scale information, the contribution of depth information, self-attention (to capture relational dependencies) and finally, dual memorability feedback. Our goal is to integrate the best performing components into a final end-to-end Visual Memory Schema prediction architecture.

A. Improving VMS Prediction

Prior work has shown the validity of a Variational Autoencoder (VAE) approach to VMS map prediction [26]. However, the family of models capable of specifically identifying the regions of images which are responsible for their memorisation has not been studied in depth compared to their single-score counterparts. Hence, we assess 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. Our goal is to discover both which techniques are applicable to VMS prediction, and to set a variety of comprehensive baselines for future work.

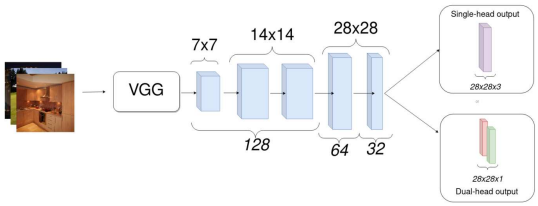

Fig. 1. End-to-end deconvolutional network showing single and dual headed outputs. The height and width of the convolution filters is given above, while the channels are given below the diagram. The dimensions of the output is given below each layer.

We choose three architectures as potential baselines against which to evaluate further developments to our proposed VMS predictor models: a deconvolutional CNN, the same CNN with incorporated multi-scale information, and a set of variational models. Our proposed deconvolutional (CNN-deconv) architecture is similar to that used in [6]. A pretrained VGG16 network feeds features into five convolutional blocks, with upscaling at specified intervals, as in the architecture shown in Figure 1. 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 3×3 aside from the final outputs, which are 1×1. To this basic structure we can easily incorporate modular architectural improvements: self-attention, and depth information.

We include multi-scale information as structures capable of influencing image memorability are likely to exist at various scales in the image (for example, a set of dining chairs and table exists at quite a different scale to the arrangement of cutlery that may lie on that table). We consider the approach of [28], enabling a deep learning architecture that can infer

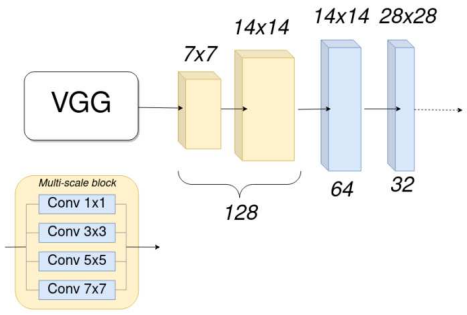

Fig. 2. Multi-scale VMS predictor with multi-scale blocks (MSB) from [28].

information from multiple scales, allowing us to assess its efficacy for visual memory schema prediction. When considering a multi-scale architecture, the three initial convolution blocks, from the architecture in Figure 2, are replaced with two multi-scale blocks (MSB). The multi-scale blocks are composed of four different convolutional layers with differing kernel sizes (1x1, 3x3, 5x5, 7x7) which each capture features present at different scales. These are then concatenated together into a singular output and the features passed forward through the network.

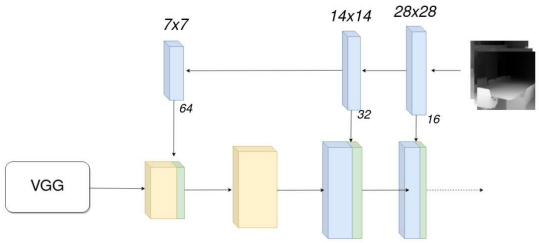

Fig. 3. Multiscale architecture modified to embed depth-map information.

Previous research has indicated that the perceived depth of the scene influences memorability score prediction performance, according to Basavaraju *et al.*, [18]. However, whether this effect holds for visual memory schemas has not been explored. Hence we propose generating depth maps for our dataset using MiDaS [29], 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. An auxiliary input is added to the model which learns increasingly deep features from depth maps. These features are then concatenated with the generative path of the model, which allows it to learn to generate memorability maps with identically-sized 'image' and 'depth' features.

Self-Attention: 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 semantic representation of the scene can indicate additional memory clues [11]. Based on these observations we introduce a self-attention component to support the detection of these structures. Non-local blocks [30] are designed to capture long-range dependencies by allowing the network to determine which features should be attended to, across the entire input. In our architectures we integrate the 'Embedded Gaussian' variant from [30] in order to determine whether long-range modelling aids VMS map prediction.

Considering a given input x and its corresponding embedding spaces $W_\phi \boldsymbol{x}_i$, the self-attention output following the study from [31], is given by:

$$\boldsymbol{y} = \lambda \;\; softmax\; (\boldsymbol{x}^T W_\theta^T W_\phi \boldsymbol{x}) g(\boldsymbol{x}) + \boldsymbol{x}, \tag{1}$$

where $g(\boldsymbol{x})$ is a linear function of the input and λ is a learnable weighting hyperparameter.

We combine the non-local blocks with our memorability predictors in two ways. Firstly, in multi-scale architectures, after the multi-scale blocks and prior to the output. Secondly, in variational architectures, we include the self-attention layer in the decoder responsible for producing the VMS map.

Loss functions: Current state-of-the-art for VMS prediction is based upon variational autoencoding (VAE) models. VAEs consist of two networks: an encoder and a decoder. The encoder estimates a latent space $\boldsymbol{z}$ corresponding to the given data $\boldsymbol{x}$ and the decoder aims to reconstruct the data from the latent space encoding. As in [26], for our VAE architectures we maximise the evidence lower bound (ELBO) on the sample log-likelihood characteristic to the classical VAE [32] :

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q_\theta(\boldsymbol{z}|\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}[q_\theta(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})], \tag{2}$$

where $p_\phi(\boldsymbol{x}|\boldsymbol{z})$ is calculated by the decoder of parameters ϕ and $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ is an inference model implemented by a neural network of parameters θ, which has Gaussian-specific prior parameters $\{\mu, \sigma\}$ for its last layer's outputs and D_{KL} is the Kullback-Leibler (KL) divergence, where

$$D_{KL}[q_\theta(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})] = \int q_\theta(\boldsymbol{z}|\boldsymbol{x}) \ln \frac{p(\boldsymbol{z})}{q_\theta(\boldsymbol{z}|\boldsymbol{x})}, \tag{3}$$

VAE models employ the standard variational loss D_{KL}, where the first term reconstructs the log-likelihood and the latter implements the Kullback-Leibler divergence between the distribution $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ and the prior $p(\boldsymbol{z})$.

Overall, in this study we consider three loss functions. For non-variational models we test binary cross-entropy and Kullback-Leibler (KL) divergence (shown to be effective for saliency map prediction [33]), whereas for VAE architectures we use the standard ELBO loss (Eq. 2). Additionally, we expand on the work of [26] by varying the size of the latent space as $|\mathbf{z}| = \{8, 32, 64, 128\}$, where $|\cdot|$ denotes the cardinality.

B. Dual-Feedback VMS Prediction

The LaMem dataset [5] 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 may be possible to use this data to support localised, *two-dimensional* memorability predictions. To that end, we propose a new architecture that can be trained both on visual memory schema and scene data, while also defining

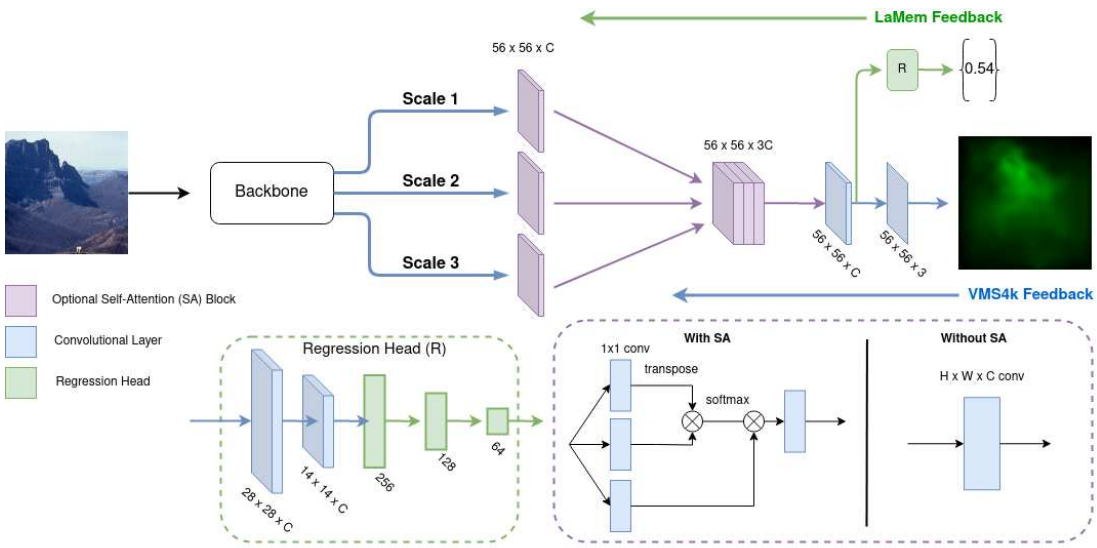

Fig. 4. Architecture of proposed Visual Memory Schema predictor with Dual Memorability Feedback.

an auxiliary loss that takes advantage of the LaMem dataset, so 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 (DF-VMS).

Our proposed architecture for VMS prediction incorporates the potential improvements discussed above: 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. 4. The network first extracts features from multiple scales, optionally computes attention maps for these features, and finally combines these multi-scale attention maps to predict the output map. In the following we describe how each separate component integrates with a dual-feedback/auxiliary loss architecture for memorability prediction. The location of various components in the architecture is provided in Figure 4.

Multi-scale Feature Extraction We consider two backbone architectures: VGG16 [25] and RESNET50 [34], and employ these to extract semantic features from the input images. We extract the semantic features at three different scales from the corresponding processing blocks in the backbone architecture. Given an input image $I_n \in \mathbb{R}^{224\times224\times3}$, for each backbone we extract feature maps at $S_1 \in \mathbb{R}^{56\times56\times256}$, $S_2 \in \mathbb{R}^{28\times28\times512}$, and $S_3 \in \mathbb{R}^{14\times14\times512}$, where S_1, S_2, and S_3 we call Scale 1, Scale 2, and Scale 3 respectively. All scaled images are passed through a 1×1 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 a 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.

Self Attention Self-attention maps are generated for each scale using the non-local block approach [30]. Given the embedding spaces $W_\phi \boldsymbol{s}_i$ for the given scaled input $\boldsymbol{s}$, and the learnable weighting hyperparameter λ_1, the self-attention output is given by:

$$\boldsymbol{y} = \lambda_1 \ \ softmax\ (\boldsymbol{s}^T W_\theta^T W_\phi \boldsymbol{s}) g(\boldsymbol{s}) + \boldsymbol{s}, \tag{4}$$

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

Each scale's self-attention embedding space is parameterised by a 1×1 convolution. If self-attention is disabled, each block is replaced by a 3×3 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 three scales. With $S_1, S_2, S_3 \in \mathbb{R}^{C\times56\times56}$, $S_m = [S_1, S_2, S_3]$, $S_m \in \mathbb{R}^{3C\times n\times n}$. This is followed by two output heads. The primary output consists of a 3×3 convolution followed by a 1×1 convolution that produces VMS map V for input image i, $V_i \in \mathbb{R}^{n\times n\times3}$. The auxiliary head consists of two stacked 3×3 convolution + max pooling blocks, followed by channel-wise global average pooling [35], 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 computational budget, and dataset size.

Loss Function We propose the following loss function for training our Dual Feedback - Visual Memory Schema (DF - VMS) architecture:

$$Loss(V, L) = \frac{1}{v}\sum_{i=1}^{v}(V_i - \hat{V}_i)^2 + \alpha\frac{1}{k}\sum_{i=1}^{k}(L_i - \hat{L}_i^2). \tag{5}$$

The first term represents the loss over the samples of ground truth and predicted memorability maps, with V representing a

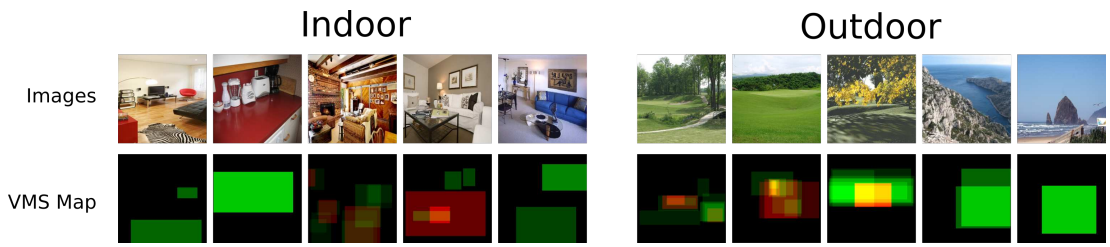

Fig. 5. Examples of images from the VMS4k Dataset. Green areas indicate that region caused the image to be remembered, red areas indicate regions that caused an image to be falsely remembered; these regions lead to a memory of having seen that scene before, despite said image never being shown to the participant.

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. α 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.

C. Quantifying Visual Memory Schemas

While visual memory schemas reveal the regions that drive scene memorability, and hence represent the schema elements used to recognise 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 dataset developed in this study, VMS4k, would be intractable both time-wise and financially. Instead, we would like to be able to computationally identify and structure the scene elements that are contained within memorable regions. This is a challenging task given that the 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 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 contained within every memorable region in the dataset; revealing the actual schemas being used to recognize our scene images. To do this, we use the MaskFormer architecture [36]. MaskFormer is a semantic segmentation network, employing a transformer-based decoder for class predictions [37]. MaskFormer is trained upon the ADE20k-Full dataset, with 847 classes, and so should be able to detect the majority of elements present in common scenes. We use MaskFormer to label all classes present in each of our scene images, over the entire dataset. This captures both object-level occurrences as well as the presence of more abstract elements such as ”skies” and ”walls”. We then calculate the overlap between a given visual memory schema annotation and the segmented object classes. This allows us to approximate the mental schema a given Visual Memory Schema annotation is capturing - and in turn, the arrangement of objects which contribute to the scene being remembered.

IV. Dataset

We consider three different datasets. VISCHEMA PLUS is used for testing the processing model architecture and components due to its reduced size. Meanwhile, the significantly larger VMS4k and LaMem [5] datasets are used together for developing a dual-feedback prediction model.

A. Experimental Setup

In the following we describe how we gather additional 2D memorability data in order to extend the initial VISCHEMA dataset [6] by taking two different approaches. The first approach is to replicate the original VISCHEMA experiment, aiming to double the data from [6]. To do this, we select a set of eight hundred images from the SUN database [38] matching the VISCHEMA categories. The initial VISCHEMA dataset [6] was divided into eight categories. Indoor scenes, both private and public: Kitchens, living room, air terminals, and conference rooms. Outdoor Scenes: public entertainment (playgrounds + amusement parks), work/home (skyscrapers + houses), populated (golf courses + pastures), and isolated (badlands and mountains). Together these cover a broad array of commonly encountered scenes. For creating the dataset, we follow the same criteria as in [6] by excluding obvious landmarks, faces looking directly at the camera, and attention-grabbing text where appropriate. Memorability data is gathered via a repeat-recognition two phase psychological experiment. Participants are shown 400 images during a ‘study’ phase. This is followed by a ‘recognition’ phase, during which 200 repeats, and 200 foils (images not shown during the study phase) are shown. Each image is shown for 3 seconds to the participant in the study. If the participant indicates they have seen an image before with sufficient confidence, they are tasked to annotate the regions of the image that they believe caused them to recall that scene. While this approach allows for sufficient data for explorative analyses, due to experiment length, category and data requirements, it is difficult to obtain significant data with this methodology. We title this 1600-image expanded dataset ‘VISCHEMA PLUS’.

For the second approach, in order to obtain a wider corpus of 2D memorability data we design a continuous image-stream, cloud based, crowd-sourced experiment. To provide sufficient initial data we relax the categories, choosing to retain only the ‘Indoor’ and ‘Outdoor’ distinction. The indoor category consists of 2000 images, mainly extracted from the SUN kitchen and living room categories, with additional scenes

Fig. 6. Repeat-recognition experiment structure. Scenes are shown to the participant in a continuous stream. After a certain interval, target scenes will repeat. If the repeat is detected, the participant is asked them to annotate the regions that lead to them recalling having seen that image.

from the conference room and airport terminal categories. 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.

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 repeat. Each image was shown to the participant for three seconds. Observers were asked to indicate whether the image they were seeing was a repeat of a scene they have seen before in the image stream, or whether that image was a new, previously unseen image. Once an image 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 that image. This procedure is illustrated in Fig. 6. In total, 93 participants undertook the experiment. Participants were paid for their participation and no personally identifiable information was gathered or stored by the authors. Participants informed consent was obtained, and they were free to withdraw from the study at any time. The study was approved by the ethics board of the University of York, UK. The experiment was distributed to users via Prolific [39]. We call this resultant larger-scale dataset 'VMS4k'.

B. The VMS4k Dataset

Of the 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. This leaves 3,461 images with corresponding maps indicating the regions that caused the participants to remember that image. Examples from both the indoor and outdoor categories with memorability maps are shown in Fig. 5. The VMS map images consist of two channels: one containing regions labelled as memorable, and another 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. In general, participants show good memory performance for the images shown during the image sequences, with the majority of participants showing a D-prime of over 2.0 (average 2.59 $\pm$ 0.063), indicating suitable performance.

V. Experimental Results

In this section we go over the results of the approaches described in Section III. We start by analysing the performance of individual components and architectural approaches using the new VISCHEMA PLUS dataset. We then evaluate the performance of an end-to-end dual-feedback architecture which includes the best performing components.

A. Improving VMS Prediction

We use the VISCHEMA 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.

Prior work evaluates the efficacy of VMS predictors with two distinct measures: the Pearson 2D correlation [6], and the mean squared error (MSE) [26]. 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) [40], metrics commonly used to evaluate saliency map models. We also employ the pixel-wise Spearman rank correlation, S^{2D}, as the measure commonly used to evaluate memorability score predictors. The 'best' 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 assessing a variety of measures for VMS prediction we consider a wider context of applications.

The deconvolutional networks are trained for 100 epochs (after which there is no improvement against the validation set), which is optimised using the Root Mean Square Propagation (RMSProp), which is an adaptive learning rate optimisation algorithm designed to address some of the issues encountered with the stochastic gradient descent (SGD) in training deep neural networks, using a learning rate of $\eta = 0.0001$. Each deconvolutional network outputs a 28×28 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 layers, which standardises the feature magnitudes for downstream processing. All networks were trained on a single NVIDIA 1080 Ti GPU.

TABLE I
PREDICTION RESULTS FOR THE VMS MEMORABILITY CHANNEL. SH: SINGLE-HEADED OUTPUT. KL: KULLBACK-LEIBLER DIVERGENCE.

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 [26])	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	**0.99**	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	**53.78**	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	**0.53**
VAE L64	83.57	0.47	0.52	-	35.06	0.51
VAE L128	96.13	0.43	0.47	-	47.22	0.49
VAE+Att L8	87.65	**0.49**	**0.53**	-	**34.17**	**0.53**
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

TABLE II
VMS FALSE MEMORABILITY CHANNEL PREDICTION RESULTS. THESE ARE PREDICTIONS FOR REGIONS WHICH CAUSE A HUMAN TO BELIEVE THEY HAVE SEEN A SCENE BEFORE, WHEN IN FACT THEY HAVE NOT - A 'FALSE MEMORY'.

Model	MSE↓	P^{2D} ↑	S^{2D} ↑	KLD↓	EMD↓	SIM↑
CNN-deconv	39.9	-0.05	-0.09	8.73	33.3	0.12
MSB	**35.96**	-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	**22.64**	0.11
VAE (from [26])	75.66	0.34	**0.37**	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	**0.37**	-	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
VAE+Att L8	74.66	**0.36**	**0.37**	-	29.73	**0.37**
VAE+Att L32	73.41	0.34	**0.37**	-	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

The summary of the memorability prediction results expressed with different measurements of performance is presented in Table I, and false memorability in Table II. In these tables, we denote by MSB when considering multiscale blocks, attention (or att) where we use non-local neural blocks, and 'Depth,' when using depth maps. VAE latent spaces are denoted by L + the latent dimension $|\mathbf{z}|$. Models trained using Kullbach-Leibler (KL) divergence are not tested with KL divergence.

For memorability, the best performing straight deconvolutional networks were trained with the KL Divergence loss, 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 [26]. The superior performance of the KL-loss may explain why VAEs remain the best overall approach. 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 provides the best results for four independent memorability metrics. The baselines (VAE aside) performed poorly at two-dimensional memorability prediction, as can be seen from Tables I and II.

The poorest performing architecture is the straight deconvolutional network. The initial introduction of multi-scale blocks improves performance slightly, while producing a single output improves performance significantly. The introduction of self-attention and depth information by the methods given in Section III improves memorability prediction, though depth alone causes significantly poorer performance when predicting false memorability. Depth and attention modules combined exceed the performance of either one alone. We achieve a Pearson 2D correlation P^{2D} of 0.49 for true memorability and 0.36 for false memorability respectively, which exceeds all previous models tested on the VISCHEMA 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.

B. Dual Feedback VMS Prediction

In the following 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 [41] with a learning rate of 5×10^{-5} with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We choose Adam over the RMSProp optimisers used in the separate architectural experiments due to its well-known history of empirical efficacy when paired with residual networks. However, we would expect the DF-VMS architecture to be robust to optimiser selection and present similar results under a wide range of optimisers. 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 [5], 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 set $\alpha = 40^{-1}$, where α is hyperparameter which controls the influence of the one-dimensional memorability scores on the training procedure. 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.

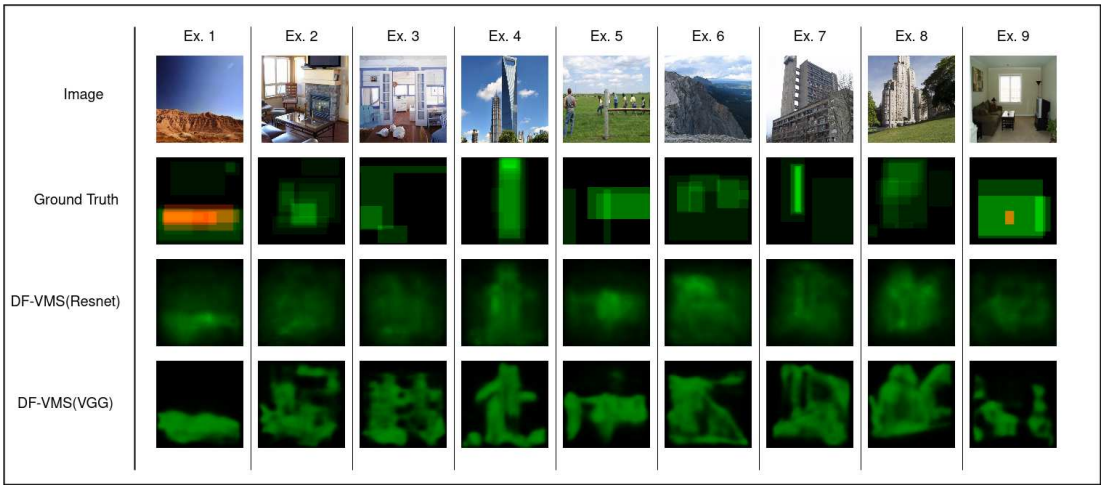

Fig. 7. Predicted VMS maps for the scene images in the top row. Ground-truth maps come from human defined VMS maps. Some maps contain false schemas (red), for visualisation purposes in this figure we only show predicted true (memorable) schemas. The best performing DF-VMS variant employs a ResNet backbone, 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.

TABLE III

VMS RECONSTRUCTION RESULTS. TRUE & FALSE REFER TO MEMORABLE AND FALSELY MEMORABLE SCHEMAS (GREEN/RED IN IMAGES). P^{2D} IS THE PEARSONS 2D CORRELATION [6], [42]. LAMEM PERFORMANCE MEASURED BY SPEARMANS CORRELATION (ρ). A DASH IN THE TABLE INDICATES THE NETWORK DOES NOT COMPUTE THAT OUTPUT. A COMPARISON WITH THE STATE-OF-THE-ART IS GIVEN AGAINST THE CURRENT BEST MODEL; VMS-VAE FROM [26]. WE INCLUDE RESULTS FOR A MORE MODERN BACKBONE, RESNET50, AND FOR A FAIR COMPARISON WITH PRIOR WORK, A VGG16 BACKBONE.

Backbone	Method	Dataset	True P2d	False P2d	LaMem (ρ)	Params.
None	Edge Detection	VMS4k	0.234	0.216	-	-
VGG16	vms-VAE	VMS4k	0.395	0.357	-	16.91M
	DF-VMS	VMS4k + LaMem	0.425	0.374	0.552	14.79M
ResNet50	DF-VMS-R	VMS4k + LaMem	**0.513**	**0.443**	0.466	8.68M

Results for reconstruction accuracy on VMS4k are shown in Table III and Table IV. While we obtain the best results with the ResNet50 backbone feature extractor, we include results using the VGG16 architecture for the purposes of comparison to prior work. In Table III, on the second row, we show the results for the previous best performing, variational autoencoder based model, 'vms-VAE' [26] on VMS4k. The results presented in Table III show that our DF-VMS model outperforms the vms-VAE model after this was trained on the VMS4k dataset, and hence benefits from additional training data not available when said model was created. Our analysis reveals that prior memorability models are not capable of taking advantage of our larger dataset, unlike the proposed DF-VMS approach. Our qualitative results indicate that DF-VMS models which use the VGG16 backbone give overconfident predictions over the object content of the image, but do not capture the memorability of broad scene regions well. 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 other 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. We also find that our ResNet DF-VMS model outperforms the VGG16 model by 8.8% and 6.9% for true and false schemas respectively, and note that our ResNet model uses fewer parameters than both prior work and the VGG16 model. Thus we show our proposed architecture outperforms prior work, and also that we can attain best-possible performance with a more suitable backbone. In Fig. 7 we show a set of predicted examples for a variety of both indoor and outdoor scene images along with their ground-truth human defined VMS maps. Many VMS maps contain annotations which indicate regions that lead to false remembering. These annotations are coloured red in the per-pixel maps. While we focus primarily on predicting true memorability (region which *cause* recognition) we can equally apply our models to predict image regions which cause a false memory. See Fig. 1 in the supplementary information for 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 (as we do not focus on memorability 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).

TABLE IV
ABLATION TESTS FOR OUR DF-VMS MODEL. IN THIS TABLE, IN THE MODEL TITLE, -XA INDICATES NO ATTENTION, -XDF NO DUAL-FEEDBACK, -XM, NO MULTI-SCALE INFORMATION, AND -XVMS, SCORE PREDICTION ONLY. P^{2D} IS THE PEARSONS 2D CORRELATION. A DASH IN THE RESULTS TABLE INDICATES NO OUTPUT.

Backbone	Method	Dataset	True P2d	False P2d	LaMem (ρ)	Params.
ResNet50	DF-VMS-R-xA	VMS4k + LaMem	0.497	0.435	0.444	8.68M
	DF-VMS-R-xDF	VMS4k	0.488	0.423	-	8.68M
	DF-VMS-R-xM	VMS4k + LaMem	0.497	0.418	0.446	8.66M
	DF-VMS-R-xVMS	LaMem	-	-	0.28	8.68M

C. Ablation Testing

To evaluate the impact of various optional model modules x = {attention, dual feedback, multi-scale information} we train the best performing model a further three times with one of the above modules **excluded** (**x**) from the model, and provide the results in Table IV. In the table, **-xA** indicates attention excluded from the model, **-xDF**, dual-feedback excluded, and **-xM**, multiscale information excluded. Additionally, we test the performance purely on the auxiliary memorability loss by disabling visual memory schema feedback (**-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 a poorer model performance, with the most drastic decrease occurring when dual feedback is disabled (0.488 vs 0.513). We also note that even with dual-feedback disabled, the ResNet50 model outperforms the VGG16 vms-VAE and DF-VMS models despite having fewer parameters, and while trained on fewer data. 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 (leading 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 in performance (5.12%, perc. increase) 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 VMS4k leads to worse single-score performance (a drop of 37%). This highlights the interaction between single-score memorability and two-dimensional memorability, and suggests that an ideal image memory prediction model should account for both *how* memorable an image is as well as *where* in that image is memorable. Spatial memorability maps gathered from humans could be applied in future work to boost single-score prediction performance for challenging datasets such as natural scene images, or where single-score ratings are naturally less consistent or not available.

D. Quantifying Visual Memory Schemas

Examples of semantic elements found and labelled inside memorable regions of the VMS4k dataset are shown in Fig. 8. For example, in the image of a field; it is obvious that not one single element 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 schemas held within the semantic knowledge store of the participants who labelled this image, and which aided in recognition of the scene. In order to extract a *generalised* schema for each category, we ask which scene components commonly occur with each other components inside memorable regions. That is, we aim to determine which local arrangement of elements most frequently causes a scene region to lead to the remembering of that 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.

In Fig. 9 we show some examples of this procedure, for the kitchen category. We limit this analysis to co-occurrences of just three objects; higher amounts of objects can also be examined (we include all data in the supplementary information). 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, by appearing together, reveal the 'schema' used to recognise scene images for a given category. With this approach, scene memorability can be tracked from a mental schema, to two-dimensional maps, and finally to human-understandable descriptions of those schemas for each VISCHEMA 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, by 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.

Being able to capture object arrangements which cause a scene to be remembered additionally allows us to empirically investigate whether there is any notable difference between these labelled regions for memorable or forgettable scenes drawn from the same category. To do this we obtain co-occurrence statistics for the top 20 and bottom 20 memorable scenes from each category. This is shown in figure 10 with

Fig. 8. 'Semantic units' contained within the regions of images that participants have labelled as causing them to successfully remember that image.

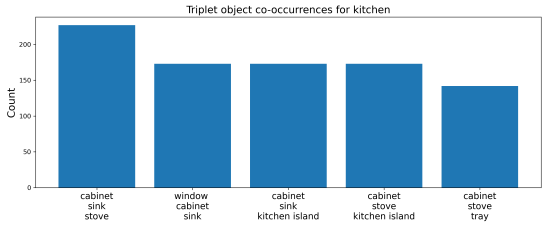

Fig. 9. These sets of objects frequently appear together inside the memorable regions of an image, for that category. For example, cabinets, sinks, and stoves most frequently occur within one labelled VMS region, considered over the entire kitchen category. This shows which combinations of objects, together, lead to an image being recognised by a human observer. By limiting to three objects, more cases of object co-occurrences can be examined.

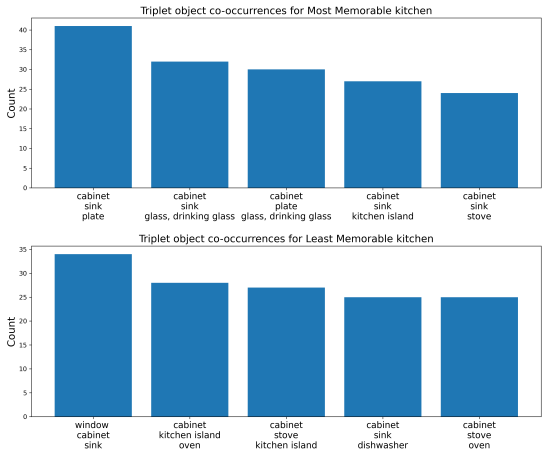

Fig. 10. Differences in remembered image regions for most memorable (top) and least memorable (bottom) kitchen scenes.

a full set of results given in the supplementary information. We first examine the impact of object counts in memorable regions to determine whether the intra-category difference in memorability may be caused by the *amount* of objects captured in a memorable region. However, we find no significant difference in object counts for either indoor (*one-way ANOVA, $p > 0.05$*) or outdoor scenes (*one-way ANOVA, $p > 0.05$*). This suggests that the regions which drive memorability for highly memorable vs poorly memorable scenes do not differ significantly in objects contained. This reinforces the hypothesis that it is the scenes *semantic* content causing the differences in memorability. Observationally, figure 10 shows that there are differences in the *objects* which are frequently annotated together in memorable vs forgettable images from the same category. For example, within the kitchen category, memorable images have regions which appear to contain more idiosyncratic object arrangements (e.g, plates, glasses, which have more variety in placement and arrangement) whereas less memorable images appear to be remembered due to arrangements of larger scale 'standard' kitchen elements (stoves, dishwashers, kitchen islands). As all the objects are part of the schema for that category, this does not imply an incongruency effect, but instead suggests this is an effect on memory of the semantic contribution of complexity [23], implying additional scene detail may be advantageous for schema-based memory.

VI. Conclusion

Visual Memory Schema maps capture the regions of scene images that cause a person to remember an image. Compared to single-score metrics of image memorability, VMS maps are significantly more informationally dense. While single-score approaches allow for an overall memorability rating, VMS maps enable the localisation of exactly which image regions contribute to scene recognition. These image regions contain memorable (or falsely memorable) local arrangements of semantic features and objects responsible for that entire scene being remembered. While more powerful than single-score metrics, they are also correspondingly more difficult to predict; as a given VMS predictor needs to be able to identify these memorable arrangements. In this work, we attempt to tackle the inherent difficulties behind VMS map prediction. We started by introducing a new dataset, which expands the existing 800 sample VISCHEMA dataset to 4000+ samples. We first provide a set of comprehensive baselines for a variety of techniques which may be used to predict VMS maps, before designing an architecture which can take advantage of pre-existing single-score memorability data to increase the predictive performance for Visual Memory Schemas. Moreover, we include an object-based analysis of the semantic content of any given visual memory schema, allowing us to understand which elements are being captured by a schema, and hence determine which arrangements of objects lead to a scene being remembered. Our DF-VMS model shows an improvement of 11.8% for memorable regions and 8.6% for falsely memorable regions, a significant increase over previous approaches to VMS prediction. Our analysis suggests that this approach is a highly effective method to produce two-dimensional memorability maps for scene images. This result supports research into understanding why a given scene is remembered and another forgotten, and enables two-dimensional analysis of scenes that lack ground-truth VMS data.

In future work it would be advantageous to gather additional data, and further assess the impact of self-attention by exploring transformer-based architectures for VMS prediction. We additionally plan to explore false schemas, and false memories in greater depth, aiming to understand which scene elements lead to the creation of a false memory in the human visual long-term memory system.

REFERENCES

[1] Z. Bylinskii, L. Goetschalckx, A. Newman, and A. Oliva, "Memorability: An image-computable measure of information utility," *Human Perception of Visual Information: Psychological and Computational Perspectives*, pp. 207–239, 2022.

[2] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2011, pp. 145–152.

[3] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1469–1482, Jul. 2014.

[4] S. Lahrache and R. El Ouazzani, "A survey on image memorability prediction: From traditional to deep learning models," in *Proc of 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2022, pp. 1–10.

[5] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proc of. IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile: IEEE, pp. 2390–2398.

[6] E. Akagündüz, A. G. Bors, and K. K. Evans, "Defining image memorability using the visual memory schema," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2165–2178, 2020.

[7] J. M. Mandler and G. H. Ritchey, "Long-term memory for pictures.," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 3, no. 4, p. 386, 1977.

[8] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," MASSACHUSETTS INST OF TECH CAMBRIDGE, Tech. Rep., 2011.

[9] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," *Advances on Neural Information Processing Systems (NIPS) pp. 296–304*, 2012.

[10] J. Lukavský and F. Děchtěrenko, "Visual properties and memorising scenes: Effects of image-space sparseness and uniformity," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2044–2054, Oct. 1, 2017. (visited on 10/03/2018).

[11] S. Yoon and J. Kim, "Object-centric scene understanding for image memorability prediction," in *Proc. of IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Apr. 2018, pp. 305–308.

[12] H. Squalli-Houssaini, N. Q. Duong, M. Gwenaëlle, and C.-H. Demarty, "Deep learning for predicting image memorability," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2371–2375.

[13] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, "AMNet: memorability estimation with attention," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6363–6372.

[14] C. D. Needell and W. A. Bainbridge, "Embracing new techniques in deep learning for estimating image memorability," *Computational Brain & Behavior*, vol. 5, no. 2, pp. 168–184, 2022.

[15] M. Mancas and O. L. Meur, "Memorability of natural scenes: The role of attention," in *Proc. of IEEE International Conference on Image Processing*, pp. 196–200.

[16] B. Celikkale, A. Erdem, and E. Erdem, "Predicting memorability of images using attention-driven spatial pooling and image semantics," *Image and vision Computing*, vol. 42, pp. 35–46, 2015.

[17] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision Research*, vol. 116, pp. 165–178, 2015.

[18] S. Basavaraju, S. Gaj, and A. Sur, "Object memorability prediction using deep learning: Location and size bias," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 117–127, 2019.

[19] R. Dubey, J. Peterson, A. Khosla, M. Yang, and B. Ghanem, "What makes an object memorable?" In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 1089–1097.

[20] M. A. Kramer, M. N. Hebart, C. I. Baker, and W. A. Bainbridge, "The features underlying the memorability of objects," *Science Advances*, vol. 9, no. 17, paper eadd2981, pp. 1–14, 2023.

[21] F. C. Bartlett, *Remembering: A study in experimental and social psychology* (Remembering: A study in experimental and social psychology). New York, NY, US: Cambridge University Press, 1932, xix, 317.

[22] C. Kyle-Davidson, A. G. Bors, and K. K. Evans, "Modulating human memory for complex scenes with artificially generated images," *Scientific Reports*, vol. 12, no. 1, Article 1583 pp. 1–15, 2022.

[23] C. Kyle-Davidson, E. Y. Zhou, D. B. Walther, A. G. Bors, and K. K. Evans, "Characterising and dissecting human perception of scene complexity," *Cognition*, vol. 231, p. 105 319, 2023.

[24] K. K. Evans and A. Baddeley, "Intention, attention and long-term memory for visual scenes: It all depends on the scenes," *Cognition*, vol. 180, pp. 24–37, Nov. 2018, ISSN: 1873-7838. DOI: 10.1016/j.cognition.2018.06.022.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[26] C. Kyle-Davidson, A. Bors, and K. Evans, "Predicting visual memory schemas with variational autoencoders," *arXiv preprint arXiv:1907.08514*, 2019.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[28] Z. Hu and A. Bors, "Conditional attention for content-based image retrieval," in *Proc. British Machine Vision Conference (BMVC)*, 2020.

[29] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1623-1637*, 2022.

[30] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[31] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. of Int. Conf. on Machine Learning (ICML), vol. PMLR 97*, 2019, pp. 7354–7363.

[32] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. Int. Conf. on Learning Repres. (ICLR)*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6114.

[33] A. Kroner, M. Senden, D. K., and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," *Neural Networks*, vol. 129, pp. 261–270, 2020.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[35] M. Lin, Q. Chen, and S. Yan, "Network in network," *International Conference on Learning Representations (ICLR)*, 2014.

[36] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1–14, 2021.

[37] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[38] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. of the IEEE Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3485–3492.

[39] *Prolific: Https://www.prolific.co/*. [Online]. Available: https://www.prolific.co/.

[40] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Reprentations (ICLR)*, 2015.

[42] C. Kyle-Davidson, A. Bors, and K. Evans, "Predicting visual memory schemas with variational autoencoders," in *Proc. British Machine Vision Conference (BMVC)*, 2019.