Improving Neural Saliency Prediction with a Cognitive Model of Human Visual Attention

Ekta Sood[1], **Lei Shi**[1], **Matteo Bortoletto**[1], **Yao Wang**[1], **Philip Müller**[2], **Andreas Bulling**[1]
[1]University of Stuttgart, Institute for Visualization and Interactive Systems (VIS), Germany
[2]German Research Center for Artificial Intelligence (DFKI)
{ekta.sood,lei.shi,matteo.bortoletto,yao.wang,andreas.bulling}@vis.uni-stuttgart.de
{philipp.mueller}@dfki.de

Abstract

We present a novel method for deep image saliency prediction that leverages a cognitive model of visual attention as an inductive bias. This is in stark contrast to recent purely data-driven models that have achieved performance improvements mainly by increased model capacity, resulting in high computational costs and the need for large scale, domain specific training data. We demonstrate that by leveraging a cognitive model of visual attention, our method achieves competitive performance to the state-of-the-art across several natural image datasets while only requiring a fraction of the parameters. Furthermore, we set the new state of the art for saliency prediction on information visualizations, demonstrating the effectiveness of our approach for cross-domain generalization.We further provide large-scale cognitively plausible synthetic gaze data on corresponding images in the full MSCOCO and FigureQA datasets, which we used for pre-training. These results are highly promising and underline the significant potential of bridging between first-principle cognitive and data-driven models for computer vision tasks, potentially also beyond saliency prediction, and even visual attention.

Keywords: neural networks, computational models of cognition, computer vision, saliency prediction, human visual attention

Introduction

Predicting human visual attention on natural images has been widely studied in computer vision (e.g., saliency prediction) (Borji, 2019; Borji & Itti, 2012). While early works have proposed models that were closely inspired by the human visual system (Frintrop, Werner, & Martin Garcia, 2015; Itti, Koch, & Niebur, 1998), latest models rely on deep neural networks trained on large-scale natural image datasets (Jia & Bruce, 2020; Linardos, Kümmerer, Press, & Bethge, 2021; Lou, Lin, Marshall, Saupe, & Liu, 2022). These models consist of up to 84 million parameters (Jia & Bruce, 2020) or make use of vision transformers (Lou et al., 2022) or several separate backbone networks to improve performance (Linardos et al., 2021). In addition to the computational burden introduced by these models due to their high complexity, they also require ground truth gaze information for training that has to be collected using eye trackers in a costly and cumbersome process (Judd, Durand, & Torralba, 2012a). The requirement for real gaze data can be partly alleviated by simulating gaze using mouse clicks as a proxy (Jiang, Huang, Duan, & Zhao, 2015; Kim et al., 2017). However, mouse click data still has to be collected from a large number of users, not necessarily reflects gaze well, and is currently only available at scale for natural images. For example, in the growing area of saliency prediction on information visualizations (Matzen, Haass, Divis, Wang, & Wilson, 2017), only few datasets were published so far (Borkin et al., 2015; Shin, Chung, Hong, & Elmqvist, 2022; Wang, Koch, Bâce, Weiskopf, & Bulling, 2022).

It is widely agreed that eye movements are contingent on unobservable (i.e. covert) attention shifts, which are in turn controlled by underlying cognitive processes (Salvucci, 2001). In parallel line of work, researchers have therefore developed cognitive models of human attention allocation on images (Kieras & Meyer, 1994; Nyamsuren & Taatgen, 2013a; Salvucci, 2000). Instead of learning patterns within training data by maximizing prediction performance, these first principle-models are rule-based and aim to reproduce basic, domain-independent human attentive processes as faithfully as possible. For example, the Eye Movements and Movement of Attention (EMMA) model (Salvucci, 2000) relies on the cognitive architecture ACT-R (J. R. Anderson & Lebiere, 2014) to predict shifts of overt visual attention and synthesize spatio-temporal eye movements from these shifts. As such, cognitive models of human visual attention hold two main promises for the human saliency prediction field. First, they have the potential to alleviate the need for ever-increasing training data and model sizes in image saliency prediction. Second, due to their general nature, they may be beneficial for target domains where only little human gaze data is available. Despite this potential, no attempt to integrate cognitive models of visual attention into deep image saliency prediction models has been made so far.

We propose the first method that integrates a cognitive attention model into the training process of a neural saliency model. Our method consists of three steps: (1) generating synthetic gaze data for both natural images and information visualizations datasets using the EMMA model, (2) pre-training computationally light-weight saliency models on the synthetic saliency maps, and (3) finetuning the pre-trained models on the target saliency benchmark datasets. The results show that our approach achieves competitive performance to several state-of-the-art saliency models (Fosco et al., 2020; Jia & Bruce, 2020) on three benchmark natural image datasets, while only requiring a fraction of the model parameter. We additionally show consistent improvements in cross-domain evaluations on information visualizations – a target domain for which only little training data are available, outperforming the previous

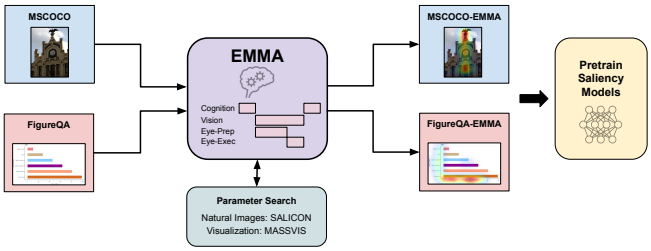

Figure 1. Our pipeline for pre-training saliency prediction models with synthetic gaze data generated by the EMMA cognitive model. EMMA's control flow consists of four processes that run in parallel: cognition that leads to attention shifts, vision that encodes objects, eye movement preparation and eye movement execution. We find the best parameters for EMMA by using ground truth gaze data from SALICON and MASSVIS datasets. Then we use EMMA to generate synthetic gaze data on datasets which do not provide it, i.e. MSCOCO and FigureQA. Neural saliency models are then trained on the resulting MSCOCO-EMMA and FigureQA-EMMA datasets.

state of the art on MASSVIS (Borkin et al., 2015).

Our contributions are threefold. First, for the first time we integrate a cognitive model of visual attention into the training process of a neural saliency model. Second, we conduct extensive evaluations showing that our method achieves competitive results to the SOTA for saliency prediction on natural images with a significantly smaller model size (only around one-third of the parameters). Furthermore, our method sets the new state of the art on saliency prediction on information visualizations, indicating its effectiveness in cross-domain scenarios. Third, we release a large-scale dataset consisting of 130k high-quality simulated gaze data obtained using the EMMA model on MSCOCO.

Related work

Cognitive models of visual attention

Models of human attentive processes have a long history in cognitive science (Feigenbaum, 1959). These models are commonly used to model visual search on stimuli like chess boards (Simon & Feigenbaum, 1964), or colour features and shapes (Kieras & Meyer, 1994) but are rarely evaluated on natural images (Cutsuridis, 2009). One of the most popular architectures for modelling human cognition is ACT-R (B. Anderson, 2014; J. R. Anderson & Lebiere, 2014). Motivated by the fact that attention shifts only indirectly correspond to observable eye movements, Salvucci (2001) proposed the EMMA model (Salvucci, 2000) that linked eye movement generation with the cognition of unobservable visual attention shifts within ACT-R: While ACT-R served as a cognitive processor producing attention shifts, EMMA generated corresponding spatio-temporal eye movements. EMMA is a well-established model and has been extensively studied for a range of visual search tasks on both images and text (Kotseruba & Tsotsos, 2020; Nyamsuren & Taatgen, 2013b; Salvucci, 2001). EMMA has also triggered several further works on human reasoning modelling (Bubb, 2021; Nyamsuren & Taatgen, 2014), theory of mind (J. R. Anderson et al., 2004), and text saliency modelling (Sood, Tannert, Müller, & Bulling, 2020).

We leverage EMMA in the training procedure of neural saliency models. EMMA allows us to generate saliency maps on large-scale image datasets that do not offer ground truth human gaze information.

Saliency prediction

While early work on predicting 2D human saliency maps on natural images has focused on bottom-up combination of basic image features (Itti & Koch, 2000; Itti et al., 1998), latest methods use neural networks trained on large datasets with human-generated saliency ground truth. These models consist of a large number of parameters. For example, EML-NET (Jia & Bruce, 2020) has more than 84 million parameters while SimpleNet (Reddy, Jain, Yarlagadda, & Gandhi, 2020) relies on a PNASNet-5 backbone with 86 million parameters (Liu et al., 2018). To achieve further performance improvements, recent work has resorted to building increasingly complex models that, e.g., combine multiple backbone networks (Linardos et al., 2021) or introduce vision transformers into the model architecture (Lou et al., 2022). Two notable exceptions with a modest model size are MSI-NET (24.9M parameters) (Kroner, Senden, Driessens, & Goebel, 2020) and the Multi-Duration Saliency Excited Model (MD-SEM) (30.9M parameters) (Fosco et al., 2020).

While most prior work has focused on natural images (Droste, Jiao, & Noble, 2020; Fosco et al., 2020; Kroner et al., 2020), saliency prediction is also important for other types of visual stimuli, such as graphical user interfaces (Xu, Sugano, & Bulling, 2016) or information visualizations (Matzen et al., 2017). Information visualizations often contain empty areas and a mix of text and graphical elements, causing models trained on natural images to perform poorly (Polatsek, Waldner, Viola, Kapec, & Benesova, 2018). The field of saliency prediction on information visualizations is new, with only few datasets and limited ground truth gaze information available (Matzen et al., 2017; Wang, Bâce, & Bulling, 2023). The largest dataset, MASSVIS, only has 393 information visualizations (Borkin et al., 2015).

Rather than increasing model complexity even further or requiring even larger human saliency datasets, we propose a fundamentally different approach: By leveraging a cognitive model of attention, we can synthesize any number of training samples for arbitrary visual stimuli, including information visualizations. In contrast to manipulating the stimuli themselves with human gaze ground truth to augment the training set (Che et al., 2019), our approach allows to synthesize gaze on arbitrary images for which no human gaze is available at all. Using our approach, we achieve competitive performance to the state of the art for saliency prediction on natural images while requiring significantly less parameters. In addition, our approach allows for effective adaptation to other image domains, such as information visualizations shown here.

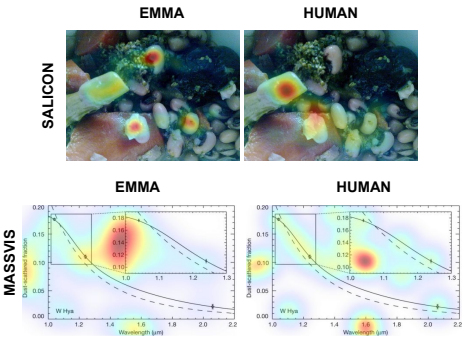

Figure 2. Synthetic gaze maps generated by EMMA and corresponding human ground truth gaze maps on sample images from the SALICON and MASSVIS datasets.

Method

At the core of our approach is the idea to use a cognitive model to generate synthetic gaze data on datasets for which human ground truth gaze data is not available. Subsequently, any neural saliency prediction model can be trained on this augmented datasets. To use EMMA for synthetic gaze data generation on images, a modification was made to take objects and their locations as input. The Faster R-CNN object detector was used to obtain object bounding boxes and class labels. The first bounding box was selected randomly, then EMMA chose the next fixation location based on its visual attention and eye movement modules, which model fixations, fixation durations, and re-fixations. Since EMMA has parameters that need to be optimized to the application domain, we made use of auxiliary datasets for which saliency ground truth maps exists (in our case SALICON [21] and MASSVIS [8]). Once optimized, EMMA was used to generate synthetic gaze data for MSCOCO and FigureQA. The synthetic gaze datasets were then used to pre-train neural saliency prediction models[1]. An overview of this approach is shown in Figure 1.

EMMA with ACT-R

EMMA assumes that eye movements are triggered by attention shifts and consists of two ACT-R components: The visual attention module models how humans process, encode, and shift attention between visual targets while the eye movement module describes how humans move their eyes to new targets. When an attention shift is triggered, EMMA produces synthetic gaze data based on the time T_{enc} to encode an object, the time to prepare and execute the eye movement (T_{prep} and T_{exec}) and the locations of the objects. The time to encode object i is given by $T_{enc} = -K\log(f_i)e^{kd_i}$, where f_i is the frequency of the object being encoded, normalized between 0 and 1, K and k are scaling constants, and d_i is the eccentricity of the object, measured in units of visual angle between the current eye position and the object that the focus should be shifted on. It is calculated by $d_i = \arctan(\|F_c - F_n\|_2/d_v)$, where d_v is the viewing distance to screen, F_c and F_n the current and next object locations. When the visual attention is shifted from one object to another, a saccade is produced to switch gaze to the new object. The landing point of the saccade is sampled from a Gaussian distribution around the new object location F_n. The eye movement consists in two stages: preparation and execution. Preparation takes place together with object encoding and the saccade is executed after preparation is finished. The preparation time T_{prep} is a parameter set to 135 ms, and the execution time T_{exec} is set to $70 + 2d_i$ ms. If a new visual attention shift arrives during saccade preparation, the eye movement is cancelled. This no longer applies during execution.

The attention shifts that EMMA uses are generated by the ACT-R model by means of a free-viewing visual search module. Starting from a random object in the visual scene, the module shifts attention to the object with the least Euclidean distance to the current attention point. After looking at all the objects in the visual scene, the process is terminated.

Generating saliency maps with EMMA

Being a top-down attention cognitive model, EMMA does not have a pixel-level understanding of the image and relies solely on object bounding boxes and corresponding label annotations. To produce such annotations, we make use of the Faster-RCNN (P. Anderson et al., 2018) object detector. EMMA outputs a sequence of fixation locations and corresponding durations. To construct saliency maps from this sequence, we place a Gaussian kernel at each fixation location and weight it with the corresponding duration.

Estimating EMMA parameters

The optimal parameters for EMMA are found by using an auxiliary dataset with ground truth saliency data in the target domain (natural images or information visualizations). The synthetic saliency maps are generated for different sets of EMMA parameters and compared with the ground truth saliency maps using Earth Mover's Distance (EMD). The parameter set with the smallest EMD is selected as the optimal set for the target domain. For natural images, the optimal parameters are estimated using the SALICON dataset and for information visualizations the MASSVIS.Figure 2 shows a comparison between EMMA-generated and human gaze data.

Training saliency models with EMMA

We leveraged EMMA with optimized parameters to synthesize cognitively plausible saliency maps for large datasets, across two different domains, which do not include human ground truth. As a result, we created two new large-scale datasets, MSCOCO-EMMA and FigureQA-EMMA, containing synthetic saliency maps for all images in MSCOCO and FigureQA, respectively. We then used these augmented datasets to train and evaluate performance gains in two lightweight saliency prediction models, MSI-NET and MD-SEM. Although both models have approximately one third of the parameters compared to the SOTA model, our approach allows them to achieve competitive results on the SALICON validation set.

[1]Code, datasets, our pretrained models and additional supporting material can be found at perceptualui.org/publications/sood23_cogsci

Table 1
*Saliency prediction performance for MSI-NET and MD-SEM on the SALICON validation set, with and without pre-training on synthetic data from EMMA and finetuning (FT) on target dataset. Best results are in **bold**, second best underlined.*

Method	AUC ↑	CC ↑	KL ↓	NSS ↑	SIM ↑
MSI-NET	0.855	0.840	0.265	1.784	0.740
MSI-NET$_{EMMA}$	**0.866**	0.886	0.198	**1.891**	0.776
MD-SEM	0.858	0.843	0.268	1.818	0.732
Ours	0.865	**0.894**	**0.193**	**1.891**	**0.780**

Subsequently, pre-trained MD-SEM on MSCOCO-EMMA synthetic data and evaluated its performance compared to SOTA models on three well established saliency benchmark datasets. For cross-domain performance on information visualizations, we pre-trained MD-SEM both on MSCOCO-EMMA and FigureQA-EMMA. Further training details are provided in supporting material[1].

Datasets and experiments

We evaluated our approach both on saliency prediction benchmark datasets containing natural images (Borji & Itti, 2015; Jiang et al., 2015; Judd, 2009) and information visualizations (Borkin et al., 2015). In the following, we provide a brief description of the datasets, implementation, and training settings. We measure the quality of EMMA-generated data and saliency prediction performance using five metrics commonly used in the literature (Kummerer, Wallis, & Bethge, 2018): Kullback-Leibler divergence (KL), Pearson's Correlation Coefficient (CC), Similarity Metric (SIM), Normalized Scanpath Saliency (NSS), and Area under ROC Curve (AUC)[2].

MSCOCO-EMMA and FigureQA-EMMA

As mentioned in Sec.Method, we leveraged EMMA to produce synthetic gaze data for MSCOCO and FigureQA. The results are MSCOCO-EMMA and FigureQA-EMMA, consisting of 130k and 100k images, scanpaths and saliency maps, respectively. Details about the optimal parameters used for EMMA are shown in supporting details[1]. To validate the quality of the synthetic data, we compared EMMA-generated gaze data and randomly sampled locations on SALICON and MASSVIS datasets respectively (see supporting material for further details[1]). EMMA outperforms by a large margin in reproducing human attention allocation on images both on SALICON (CC: 0.432 vs. 0.167; KL: 1.624 vs. 2.068; SIM: 0.457 vs. 0.332) and MASSVIS (CC: 0.400 vs. 0.096; KL: 0.641 vs. 1.703; SIM: 0.568 vs. 0.307). [3]

Saliency benchmark datasets

SALICON (Jiang et al., 2015) contains 10k training, 5k validation, and 5k testing images. In contrast to other widely used attention datasets, it includes mouse clicks as a proxy to human visual attention. In line with previous work, we used SALICON both for pre-training and evaluation (Jia & Bruce, 2020; Kroner et al., 2020). MIT300 (Judd, Durand, & Torralba, 2012b), MIT1003 (Judd, 2009) and CAT2000 (Borji & Itti, 2015) are popular datasets for evaluating saliency prediction models. MIT300 contains 300 images with eye-tracking data from 39 observers. MIT1003 contains 1,003 natural scene images with real eye-tracking data of 15 observers. CAT2000 consists of 2,000 images of 20 categories – such as indoor, outdoor, and cartoons – including 100 images each. Each image comes with eye tracking data of 12 observers. To evaluate our approach on information visualizations, we made use of MASSVIS (Borkin et al., 2015), the currently largest visualization dataset. It provides eye-tracking data recorded during a memorability task on 393 visualizations.

Natural images. We pre-trained our models on the MSCOCO-EMMA data generated with the best set of EMMA parameters found via parameter search on SALICON (see supporting materials for further details[1]). Subsequently, we finetuned the pre-trained models using the same procedure employed in previous work (Jia & Bruce, 2020; Kroner et al., 2020) (see supporting material for further details[1]). After evaluating the models on the SALICON validation set, we choose the best model to continue finetuning on the CAT2000 dataset and further on the MIT1003 dataset, following the splits and training procedure employed in previous work (Cornia, Baraldi, Serra, & Cucchiara, 2018; Linardos et al., 2021)(see supporting materials for further details[1]).

Information visualizations. When training our method for information visualizations, we started from the best model trained on natural images (with SALICON finetuning) and further finetuned it on information visualizations (see supporting details[1]). Since there are currently no large-scale human attention datasets in information visualizations, we used EMMA to synthesize gaze data for approximately 100k visualizations from FigureQA. We identified the best parameters for EMMA by using a parameter search on MASSVIS. After pre-training on FigureQA-EMMA we finetuned our model on ground truth MASSVIS data, as described in Section Method. Further training details are discussed in supporting details[1].

Results and Discussion

We evaluated our hybrid saliency prediction method on established saliency benchmark datasets containing natural images and on information visualizations to evaluate cross-domain generalizability. When selecting approaches for comparison amongst the large number of existing saliency prediction methods, we focused on those approaches that showed competitive performance across the three common benchmarks (test sets) SALICON, MIT300, and CAT2000.

[2]Additionally, for SALICON, we report Information Gain (IG) and shuffled AUC (sAUC), as they are returned by the SALICON evaluation server

[3]MSCOCO-EMMA and FigureQA-EMMA are publicly available. Access, full metrics and additional details about these datasets are provided in supporting material[1]

Table 2
*Prediction performance on the SALICON test set. Best results are highlighted in **bold**, second best underlined, and third best in italic. Params indicates the number of parameters.*

Method	sAUC ↑	AUC ↑	CC ↑	IG ↑	KL ↓	NSS ↑	SIM ↑	Params
DeepGaze IIE (Linardos et al., 2021)	**0.767**	**0.869**	0.872	0.766	0.285	*1.996*	0.733	104.5M
TranSalNet (Lou et al., 2022)	0.747	0.868	**0.907**	*0.788*	0.373	2.014	**0.803**	72.5M
EML-NET (Jia & Bruce, 2020)	*0.746*	*0.866*	0.886	0.736	0.520	2.050	0.780	84.7M
MSI-NET (Kroner et al., 2020)	0.736	0.865	*0.889*	0.793	*0.307*	1.931	0.784	24.9M
SAM-RESNET (Cornia et al., 2018)	0.741	0.865	0.899	0.538	0.610	1.990	0.793	70.1M
MD-SEM (Fosco et al., 2020)	*0.746*	0.864	0.868	0.660	0.568	**2.058**	0.774	30.9M
Ours	0.736	*0.866*	0.899	**0.812**	**0.272**	1.931	*0.791*	30.9M

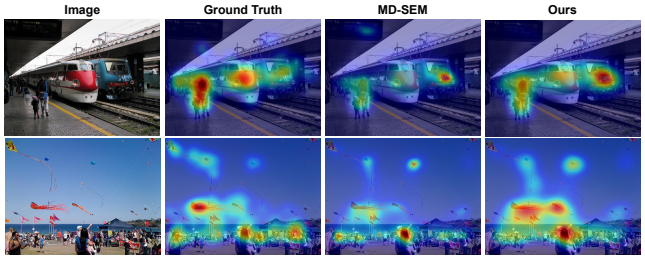

Figure 3. A visualization of example images from the SALICON val set with the corresponding human ground truth maps, and predictions from MD-SEM and Our approach.

Performance on saliency benchmark datasets

We evaluated our approach on improvement for light-weight models with cognitive pretraining on the SALICON validation set (Table 1). For both MD-SEM (30.9M parameters) and MSI-NET (24.9M parameters), our method consistently improves across all metrics. These experiments showcase for the first time that models with lower capacity benefit greatly when initially pretrained with synthetic data generated from the EMMA cognitive model. In general, MD-SEM performed better than MSI-NET. This holds true without EMMA pretraining (3 out of 5 metrics), and with EMMA pretraining (3 out of 5 metrics, and one tie). As a result, we based all further ablation studies and state-of-the-art comparisons on our combination of MD-SEM and EMMA. In addition, we investigated the effect of EMMA pre-training on larger models, namely EML-NET with different backbones. We did not find consistent improvements, suggesting that pre-training with a cognitive model specifically helps lower capacity models.[4].

Comparison against SOTA models. Table 2 shows results on the SALICON test set. Our method achieves best results on IG (0.812) and KL (0.272). It also achieves performance comparable to SOTA on AUC (0.866 vs. 0.869 for DeepGazeIIE), CC (0.899 vs. 0.907 for TransSalNet), and SIM (0.791 vs. 0.803 for TransSalNet). Importantly, we improve over vanilla MD-SEM (without pre-training on MSCOCO-EMMA) in 5 of 7 metrics, with especially large margins for IG (0.812 to 0.660) and KL (0.272 to 0.568). Thus, pre-training with MSCOCO-EMMA helps MD-SEM (30.9M parameters) to close the gap to- and even outperform approaches with more than twice the number of parameters, e.g. DeepGaze IIE (104.5M) or EML-NET (84.7M). We show qualitative results in Fig. 3.

Table 3 shows results on the MIT300 test set[5]. Our approach outperforms vanilla MD-SEM on all metrics and achieves the third highest score for all metrics except NSS. DeepGaze IIE remains the dominant model in terms of performance. On the other hand, it is the largest model, with over 100M parameters. Our approach still achieved comparable results on AUC, NSS, CC and SIM with 30.9M parameters. Compared to EML-NET (72.5M parameters), we obtain higher scores in sAUC, CC, KL and SIM and competitive in AUC.

Table 4 compares the performance of different models on CAT2000 test. We outperform the previous state of the art (DeepGaze IIE) on three metrics (AUC, NSS, and SIM) and is competitive on sAUC and CC[7]. These results demonstrate the effectiveness of leveraging EMMA to pre-train saliency prediction models.

Cross-Domain Performance

As a domain-independent model of gaze behavior, EMMA holds promise to alleviate the need for costly large-scale data collection in the respective target domain. We evaluate the benefits of our approach in the domain of information visualizations for which only little human gaze data is available. In Table 5 we show results of our method compared to the previous SOTA (Matzen et al., 2017) [8] and to different ablations on MASSVIS [9]. Our approach clearly outperforms the previous state of the art on 4 out of 5 metrics. Our approach consistently improves over ablated versions. Omitting the pre-training with EMMA-generated data on FigureQA decreases performance

[4]We provide detailed results on these additional experiments in supporting material[1]

[5]As Fosco et al. (2020) did not report MD-SEM results on the MIT300 test set, we trained the model ourselves and sent predictions from the best model on MIT1003 validation set to the test server[6]

[7]If we consider results reported in previous works but not in the leaderboard, we obtain third best results on sAUC, KL, and NSS and second best result on AUC.

[8]Code available at `http://www.cs.sandia.gov/~atwilso/get_dvs.html`

[9]MASSVIS is currently the largest available dataset of human saliency on information visualizations

Table 3
*Performance on the MIT300 test set. Best results are highlighted in **bold**, second best underlined, and third best in italic. Params indicates the number of parameters.*

Method	sAUC ↑	AUC ↑	CC ↑	KL ↓	NSS ↑	SIM ↑	Params
DeepGaze IIE (Linardos et al., 2021)	**0.7942**	**0.8829**	**0.8242**	**0.3474**	**2.5265**	**0.6993**	104.5M
TranSalNet (Lou et al., 2022)	0.7467	0.8734	0.8070	1.0141	*2.4134*	0.6895	72.5M
EML-NET (Jia & Bruce, 2020)	0.7469	0.8762	0.7893	0.8439	2.4876	0.6756	84.7M
MSI-NET (Kroner et al., 2020)	0.7787	0.8738	0.7790	0.4232	2.3053	0.6704	24.9M
SAM-RESNET (Cornia et al., 2018)	0.7396	0.8526	0.6897	1.1710	2.0628	0.6122	70.1M
MD-SEM (Fosco et al., 2020)	0.7483	0.8646	0.738	0.6962	2.1339	0.6445	30.9M
Ours	*0.7490*	*0.8748*	*0.7997*	*0.6741*	2.3518	*0.6879*	30.9M

Table 4
*Prediction performance on the CAT2000 test set. Best results are highlighted in **bold**, second best underlined, and third best in italic. Params indicates the number of parameters.*

Method	sAUC ↑	AUC ↑	CC ↑	KL ↓	NSS ↑	SIM ↑	Params
DeepGaze IIE (Linardos et al., 2021)	**0.6498**	*0.8640*	*0.7564*	0.5137	1.9619	0.6392	104.5M
EML-NET (Jia & Bruce, 2020)	0.58	0.78	0.87	0.95	**2.38**	*0.74*	84.7M
MSI-NET (Kroner et al., 2020)	0.59	0.82	0.87	**0.36**	2.30	0.75	24.9M
SAM-RESNET (Cornia et al., 2018)	0.58	**0.88**	**0.89**	—-	**2.38**	**0.77**	70.1M
MD-SEM (Fosco et al., 2020)	0.6141	0.8535	0.6388	0.7924	1.7907	0.5895	30.9M
Ours	*0.6046*	0.8687	0.7351	*0.6384*	*2.0657*	0.6411	30.9M

Table 5
*Prediction performance on the MASSVIS test set for the previous SOTA (DVS), MD-SEM, and different ablations of our approach. no FigureQA-EMMA indicates no pre-training with EMMA-generated data on Figure QA; no FT indicates no finetuning on MASSVIS. Best results are highlighted in **bold**. Results for our method (Ours) are significant with $p < .001$ from all other methods, and $p < .05$ on CC and SIM from Ours (no FigureQA-EMMA).*

Methods	AUC ↑	CC ↑	KL ↓	NSS ↑	SIM ↑
DVS (Matzen et al., 2017)	0.738	0.563	0.480	**0.996**	0.603
MD-SEM (Fosco et al., 2020)	0.728	0.727	0.097	0.913	0.786
Ours (no FigureQA-EMMA, no FT)	0.679	0.580	0.300	0.717	0.677
Ours (no FT)	0.710	0.588	0.237	0.872	0.698
Ours (no FigureQA-EMMA)	0.734	0.774	0.072	0.940	0.805
Ours	**0.743**	**0.790**	**0.061**	0.986	**0.811**

significantly in all metrics. However, a certain amount of ground truth target domain data is crucial, as indicated by the even worse performance when not finetuning on MASSVIS. Finally, neither finetuning on MASSVIS nor pre-training with FigureQA-EMMA reaches lower performance. These ablation results document the importance of domain-specific training data generated with EMMA cognitive model. Qualitative results are shown in supporting material[1].

Conclusion

We presented a novel hybrid saliency prediction method that leverages a cognitive model of visual attention. In stark contrast to prior methods that obtained improved results by increasing the model's complexity and capacity, our approach shows superior or competitive performance across several datasets using the lightweight MD-SEM saliency prediction architecture. Evaluations on both natural images and information visualizations datasets demonstrated the potential for our method in both single- and cross-domain settings. Furthermore, we provide augmented versions of the full MSCOCO and FigureQA datasets with cognitively plausible synthetic saliency data. These results underline the significant potential for bridging between cognitive and data-driven models, potentially also beyond simulated visual human attention.

Acknowledgements

E. Sood was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. L. Shi was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. M. Bortoletto was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme

under grant agreement No 801708. Y. Wang was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 251654672 - TRR 161. P. Müller was funded by the German Ministry for Education and Research (BMBF; grant number 01IS20075). A. Bulling was funded by the European Research Council (ERC; grant agreement 801708).

We would like to especially thank Simon Tannert, Pavel Denisov, Manuel Mager and Mihai Bace for their valuable insights and helpful discussions. Lastly, we would like to thank the anonymous reviewers for their helpful feedback.

References

Anderson, B. (2014). Act-r: A cognitive architecture..

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036.

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought*. Psychology Press.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Cvpr.*

Borji, A. (2019). Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, *43*(2), 679–700.

Borji, A., & Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, *35*(1), 185–207.

Borji, A., & Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.

Borkin, M. A., Bylinskii, Z., Kim, N. W., Bainbridge, C. M., Yeh, C. S., Borkin, D., ... Oliva, A. (2015). Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics*, *22*(1), 519–528.

Bubb, H. (2021). Human models. In H. Bubb, K. Bengler, R. E. Grünen, & M. Vollrath (Eds.), *Automotive ergonomics* (pp. 219–256). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from https://doi.org/10.1007/978-3-658-33941-8_5 doi: 10.1007/978-3-658-33941-8_5

Che, Z., Borji, A., Zhai, G., Min, X., Guo, G., & Le Callet, P. (2019). How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, *29*, 2287–2300.

Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, *27*(10), 5142–5154.

Cutsuridis, V. (2009). A cognitive model of saliency, attention, and picture scanning. *Cognitive Computation*, *1*, 292-299.

Droste, R., Jiao, J., & Noble, J. A. (2020). Unified image and video saliency modeling. In *European conference on computer vision* (pp. 419–435).

Feigenbaum, E. A. (1959). An information processing theory of verbal learning..

Fosco, C., Newman, A., Sukhum, P., Zhang, Y. B., Zhao, N., Oliva, A., & Bylinskii, Z. (2020). How much time do you have? modeling multi-duration saliency. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4473–4482).

Frintrop, S., Werner, T., & Martin Garcia, G. (2015). Traditional saliency reloaded: A good old model in new shape. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 82–90).

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489-1506.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Jia, S., & Bruce, N. D. (2020). Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, *95*, 103887.

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015, June). Salicon: Saliency in context. In *The ieee conference on computer vision and pattern recognition (cvpr).*

Judd, T. (2009). Learning to predict where humans look. *Proc. ICCV, 2009*.

Judd, T., Durand, F., & Torralba, A. (2012a). A benchmark of computational models of saliency to predict human fixations. In *Mit technical report.*

Judd, T., Durand, F., & Torralba, A. (2012b). A benchmark of computational models of saliency to predict human fixations. In *Mit technical report.*

Kieras, D. E., & Meyer, D. E. (1994). The epic architecture for modeling human information-processing and performance: A brief introduction..

Kim, N. W., Bylinskii, Z., Borkin, M. A., Gajos, K. Z., Oliva, A., Durand, F., & Pfister, H. (2017). Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *24*(5), 36. doi: 10.1145/3131275

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17–94.

Kroner, A., Senden, M., Driessens, K., & Goebel, R. (2020). Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, *129*, 261–270.

Kummerer, M., Wallis, T. S. A., & Bethge, M. (2018, September). Saliency benchmarking made easy: Separating models, maps and metrics. In *Proceedings of the european conference on computer vision (eccv).*

Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021). Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 12919–12928).

Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., ... Murphy, K. (2018). Progressive neural architecture search. In *Proceedings of the european conference on computer vision (eccv)* (pp. 19–34).

Lou, J., Lin, H., Marshall, D., Saupe, D., & Liu, H. (2022). Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*.

Matzen, L. E., Haass, M. J., Divis, K. M., Wang, Z., & Wilson, A. T. (2017). Data visualization saliency model: A tool for evaluating abstract data visualizations. *IEEE transactions on visualization and computer graphics*, *24*(1), 563–573.

Nyamsuren, E., & Taatgen, N. A. (2013a). Pre-attentive and attentive vision module. *Cognitive systems research*, *24*, 62–71.

Nyamsuren, E., & Taatgen, N. A. (2013b). Pre-attentive and attentive vision module action editor: Nele rußwinkel..

Nyamsuren, E., & Taatgen, N. A. (2014). Human reasoning module. *Biologically Inspired Cognitive Architectures*, *8*, 1–18.

Polatsek, P., Waldner, M., Viola, I., Kapec, P., & Benesova, W. (2018). Exploring visual attention and saliency modeling for task-based visual analysis. *Computers & Graphics*, *72*, 26–38.

Reddy, N., Jain, S., Yarlagadda, P., & Gandhi, V. (2020). Tidying deep saliency prediction architectures. In *2020 ieee/rsj international conference on intelligent robots and systems (iros)* (p. 10241–10247). IEEE Press. Retrieved from https://doi.org/10.1109/IROS45743.2020.9341574 doi: 10.1109/IROS45743.2020.9341574

Salvucci, D. D. (2000). A model of eye movements and visual attention. In *Proceedings of the international conference on cognitive modeling* (p. 259).

Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, *1*(4), 201–220.

Shin, S., Chung, S., Hong, S., & Elmqvist, N. (2022). A scanner deeply: Predicting gaze heatmaps on visualizations using crowdsourced eye movement data. *IEEE Transactions on Visualization and Computer Graphics*, 1-11. doi: 10.1109/

TVCG.2022.3209472
Simon, H. A., & Feigenbaum, E. A. (1964). An information-processing theory of some effects of similarity, familiarization, and meaningfulness in verbal learning. *Journal of Verbal Learning and Verbal Behavior*, *3*, 385-396.
Sood, E., Tannert, S., Müller, P., & Bulling, A. (2020). Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, *33*, 6327–6341.
Wang, Y., Bâce, M., & Bulling, A. (2023). Scanpath prediction on information visualisations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 1–15. doi: 10.1109/TVCG.2023.3242293
Wang, Y., Koch, M., Bâce, M., Weiskopf, D., & Bulling, A. (2022). Impact of gaze uncertainty on aois in information visualisations. In *2022 symposium on eye tracking research and applications* (pp. 1–6).
Xu, P., Sugano, Y., & Bulling, A. (2016). Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. In *Proc. acm sigchi conference on human factors in computing systems (chi)* (p. 3299-3310). doi: 10.1145/2858036.2858479