Learning from Unique Perspectives: User-aware Saliency Modeling

Shi Chen*
University of Minnesota
chen4595@umn.edu

Nachiappan Valliappan
Google Research
nac@google.com

Shaolei Shen
Google
shaoleis@google.com

Xinyu Ye
Google
yexinyu@google.com

Kai Kohlhoff
Google Research
kohlhoff@google.com

Junfeng He†
Google Research
junfenghe@google.com

Abstract

Everyone is unique. Given the same visual stimuli, people's attention is driven by both salient visual cues and their own inherent preferences. Knowledge of visual preferences not only facilitates understanding of fine-grained attention patterns of diverse users, but also has the potential of benefiting the development of customized applications. Nevertheless, existing saliency models typically limit their scope to attention as it applies to the general population and ignore the variability between users' behaviors. In this paper, we identify the critical roles of visual preferences in attention modeling, and for the first time study the problem of user-aware saliency modeling. Our work aims to advance attention research from three distinct perspectives: (1) We present a new model with the flexibility to capture attention patterns of various combinations of users, so that we can adaptively predict personalized attention, user group attention, and general saliency at the same time with one single model; (2) To augment models with knowledge about the composition of attention from different users, we further propose a principled learning method to understand visual attention in a progressive manner; and (3) We carry out extensive analyses on publicly available saliency datasets to shed light on the roles of visual preferences. Experimental results on diverse stimuli, including naturalistic images and web pages, demonstrate the advantages of our method in capturing the distinct visual behaviors of different users and the general saliency of visual stimuli.

1. Introduction

With the pervasiveness of a visual attention network in the brain, attention has become an important interface for understanding people's behavioral patterns. A collection of studies focus on leveraging human attention to optimize graphical designs [7, 16, 36], web page layouts [8, 37, 45], and user experience in immersive environments [22, 38, 44]. They demonstrate its usefulness for a broad range of applications, and more importantly, highlight the intertwined nature between attention and users' preferences [11]. As shown in Figure 1, attention modeling can be formulated as a hierarchy of tasks, *i.e.,* from a sophisticated understanding of individuals' behaviors (personalized attention), to modeling the visual behaviors of larger groups (user-group attention), and the saliency of visual stimuli (general saliency). With the great diversity in attentional behaviors among different groups (*e.g.,* attention of users with diverse characteristics, children vs elderly, male vs female, etc.), knowledge of visual preferences can play an essential role in enabling a more fine-grained understanding of attention.

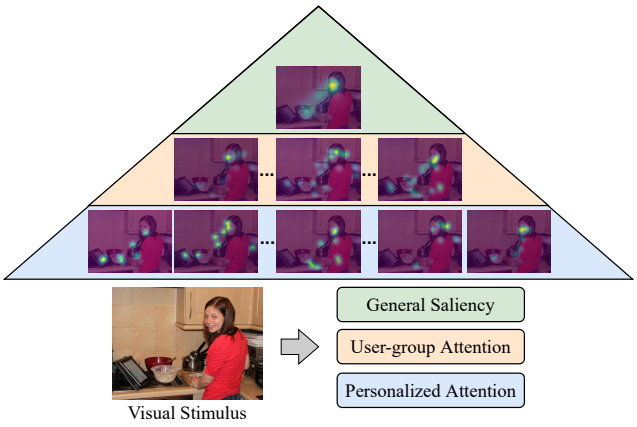

Figure 1. A hierarchy for user-aware saliency modeling, where each level focuses on a different perspective of attention.

To accurately capture human attention on visual stimuli, considerable efforts have been placed on building saliency prediction models [10, 19, 21, 29, 30]. While achieving optimistic results for modeling attention of the general popu-

*Work done during an internship at Google Research.
†Corresponding author

lation, there are two key challenges remaining largely unresolved: (1) Existing models ignore the variability of users' visual behaviors, and hence do not have the ability to identify fine-grained attention patterns of distinct users; and (2) Apart from the shortage of models for user-aware saliency modeling, there has also been no attempt to formulate a training paradigm to understand the composition of attention, which hampers the integration of attention from diverse users. To fill the gap, we concentrate on a new research problem for modeling attention of adaptively selected users, and tackle the challenge with a new computational model together with a progressive learning method.

At the heart of our saliency model is the incorporation of visual preferences with personalized filters and adaptive user masks. Unlike conventional methods designed for predicting a single saliency map representing attention of all users, it takes advantage of personalized filters to encode individuals' attention patterns. The attention patterns are adaptively integrated based on a user mask indicating the presence of users in the current sample, which enables attention prediction for various combinations of users. The aforementioned paradigm serves as the foundation for bridging individuals' preferences with visual saliency, and augments models with more abundant information about fine-grained visual behaviors. It not only shows promise in modeling attention of specific users, but also benefits the inference of the general saliency.

A key challenge in user-aware saliency modeling is the lack of understanding when aggregating attention from diverse users. The issue becomes more critical when further considering the joint effects of stimuli and user preferences on visual attention [11], where the former factor may overshadow the impacts of the latter one, leading to difficulties in capturing the variability of users' attention. Inspired by human learning that acquires knowledge through a set of carefully designed curricula [2], we propose to tackle the aforementioned issues with a progressive learning approach. The essence of our method is to encourage a model to learn the composition of attention from a dynamic set of users, from individuals to user groups representing the general population. Through optimizing on dynamically evolving annotations, it provides opportunities for models to learn both the unique attention patterns of different users and the saliency of visual stimuli.

To summarize, our major contributions are as follows:

- We identify the significance of characterizing visual preferences for attention modeling, and develop a novel model that can predict attention of various users.
- We present a progressive learning method to understand the composition of attention and capture its variability between different users.
- We perform extensive experiments and analyses to investigate the roles of visual preferences on tackling the challenges of user-aware and general saliency, and addressing the issues of incomplete users. Results demonstrate that user-aware saliency modeling is advantageous in all the above three aspects.

2. Related Works

Our work is most related to previous efforts on visual saliency prediction, which make contributions in both data collection and computational modeling.

Saliency prediction datasets. With the overarching goal of facilitating the development of attention modeling methods, many studies have contributed saliency prediction datasets with diverse visual stimuli. The pioneering work [25] presents a gaze estimation dataset for naturalistic images together with an online benchmark [24]. Several subsequent studies propose to characterize images into finer categories based on visual scenes [3], visual semantics [41], or sentiments [14]. To overcome the difficulties of large-scale data collection, Jiang *et al.* [23] use mouse-tracking as a substitute for gaze estimation and construct currently the largest saliency prediction dataset. In addition to the aforementioned works that study attention on naturalistic images, several studies focus on attention for broader types of visual stimuli, including graphical designs [7, 16, 36], web pages [8, 37, 45], and immersive environments [22, 38, 44]. There are also attempts that go beyond attention collected during free-viewing, and study goal-directed attention when performing various tasks such as driving [1] and visual reasoning [9]. These data efforts enable the developments of a series of computational methods for attention modeling.

Saliency prediction models. To estimate attention distribution on different visual stimuli, a large body of research concentrates on building saliency prediction models. Early studies leverage handcrafted features that encode high- and low-level visual cues [5, 17, 20, 40]. More recent approaches opt to automatically learn discriminative features based on deep neural networks. In particular, Huang *et al.* [19] utilize a convolution neural network (CNN) with multi-scale inputs to model the coarse-to-fine semantics. Kümmerer *et al.* [28] demonstrate the usefulness of deep features learned from image recognition for saliency prediction. Cornia *et al.* [10] develop a recurrent neural network to iteratively refine the visual features for saliency prediction. Jia *et al.* [21] augment model with rich visual semantics extracted from multiple levels of a CNN. Lou *et al.* [30] investigate the effectiveness of vision Transformers [13] for saliency prediction. Apart from the studies on general saliency prediction, a few works [32, 42, 43] attempt to tackle the problem of personalized saliency, which involves predicting attention of different users on naturalistic images. They adopt a simplified setting which only takes into account individuals' attention and assumes that the general saliency maps are given

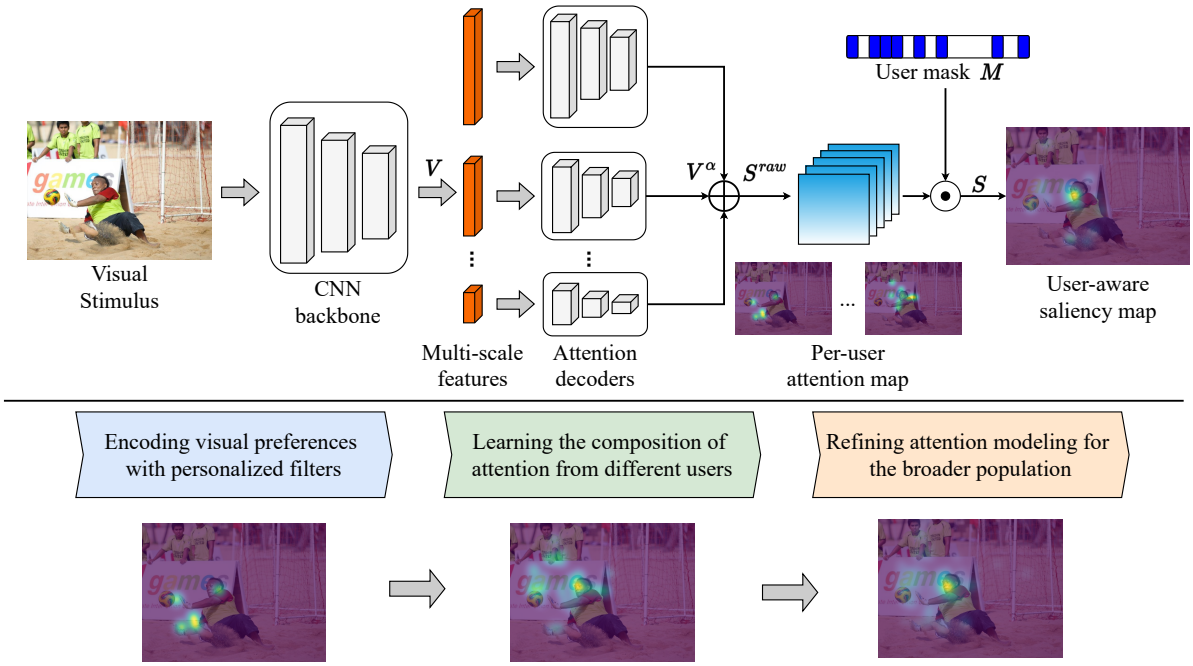

Figure 2. An overview of the proposed framework. The top figure illustrates the architectural design of our saliency model for predicting attention of diverse users, while the bottom one outlines our progressive learning method for understanding the composition of attention.

as inputs. While showing usefulness in capturing visual attention, the aforementioned methods either consider only attention aggregated from all users (*i.e.,* general saliency) or rely on assumptions that are inapplicable to less-constrained real-world scenarios. They also utilize diverse paradigms for modeling different types of attention, which prevents a joint reasoning on the variability of users' visual behaviors.

Our study differentiates from prior research in three key aspects: (1) We identify the significance of visual preferences for attention modeling, and for the first time tackle the challenge of user-aware saliency modeling. The new problem unifies previous studies on general and personalized saliency prediction, with the number of users being a dynamic factor; (2) Without imposing constraints on inputs (*e.g.,* availability of general saliency maps [32, 42, 43] and naturalistic images as visual stimuli [10, 21, 28]), our model directly infers attention from visual stimuli and is able to generalize to broader scenarios (*e.g.,* naturalistic images and web pages); and (3) In addition to a novel computational model, we also present a progressive learning approach to understand the composition of attention and capture the variability of attention.

3. Methodology

Attention modeling would benefit from the knowledge about individuals' visual preferences and the capability to adaptively integrate them. This section presents an integral framework consisting of two important components: (1) A new saliency prediction model that encodes visual preferences and dynamically predicts attention patterns of diverse combinations of users (Section 3.1), and (2) A principled approach for learning the composition of attention in a progressive manner (Section 3.2).

3.1. Encoding Visual Preferences with Personalized Filters

A primary goal of our study is to endow models with the flexibility to predict attention for various users. We tackle the challenge with a new model paradigm that leverages a collection of personalized filters to identify individuals' visual preferences, and decomposes the overall attention patterns into fine-grained ones for different users. The key differentiators of our model lie in its (1) applicability to broader scenarios (*e.g.,* from general saliency to attention of diverse users) and (2) flexibility to different settings (*e.g.,* without assuming the availability of general saliency maps).

Figure 2 (top) provides an overview of the proposed model. The principal idea behind our model is to take advantage of personalized filters to encode attention patterns of different users, and adaptively incorporate them based on the selection of users. Specifically, we build our model on top of the state-of-the-art EML-Net [21], which is a top-performing model on the MIT saliency benchmark [24]. Given an input image, our model extracts multi-scale fea-

tures from different layers of a convolutional neural network pretrained on ImageNet [12], encoding both high- and low-level visual cues. The raw features of different scales $V_t \in \mathbb{R}^{D\times H\times W}$ (D, H and W correspond to the feature channel, height, and width, t denotes index for scales) are processed independently with their corresponding attention decoders f_t^α consisting of a sequence of convolutional and upsampling layers (see Section 4.1 for details), and then the decoded attention features V_t^α are aggregated across all scales to form the unregularized attention maps S^{raw}:

$$V_t^\alpha = f_t^\alpha(V_t) \tag{1}$$

$$S^{raw} = \sum_t V_t^\alpha \tag{2}$$

Different from conventional methods [10, 19, 21, 28] that directly predict a single saliency map for the general population (*i.e.,* $S^{raw} \in \mathbb{R}^{H\times W}$), our method takes into account discriminative attention patterns of each individual user (*i.e.,* $S^{raw} \in \mathbb{R}^{N\times H\times W}$, where N is the total number of users). We achieve the goal by expanding the filters in the last convolutional layer of each attention decoder, *i.e.,* from a single filter to N personalized filters. To adaptively predict the visual attention for various combinations of users, our model takes in a user mask M in addition to the visual stimuli as input. The user mask is a binary vector $M \in [0,1]^N$ representing the presence of users, where $M_n = 1$ implies that attention of the n^{th} user is considered in the current sample. It plays the role as an indicator for integrating the unregularized attention maps from the selected users and generating the final saliency map $S \in \mathbb{R}^{H\times W}$:

$$S = softmax(\sum_n M_n \cdot S_n^{raw}) \tag{3}$$

where $\cdot$ denotes the element-wise multiplication.

To bridge individuals' visual preferences with the overall attention, we optimize our model with supervision on both the intermediate per-user predictions and the final output. We follow [10] and define the loss function L_{sal} as a combination of saliency metrics including Normalized Scanpath Saliency [33], Correlation Coefficient [31] and KL-Divergence [27] (see [10] and our supplementary materials). For per-user supervision, we apply a softmax activation function independently on unregularized attention map of each user to obtain their corresponding predictions:

$$S_n^{user} = softmax(S_n^{raw}) \tag{4}$$

Our overall objective simultaneously optimizes both the intermediate per-user predictions and the final output:

$$\begin{aligned} L = {} & L_{sal}(S, Sal, Fix) \\ & + \lambda \sum_n L_{sal}(S_n^{user}, Sal_n^{user}, Fix_n^{user}) \end{aligned} \tag{5}$$

where Sal and Fix are the ground truth saliency and fixation maps aggregated from selected users, Sal_n^{user} and Fix_n^{user} are those for individual user n. Note that only users selected in the user mask (*i.e.,* $M_n = 1$) will be considered for optimization. λ is the balancing factor.

The aforementioned model illustrates a general paradigm for connecting user preferences and visual saliency. It serves as the foundation for user-aware saliency modeling, and supports the development of our progressive learning method, as detailed in the next subsection.

3.2. Learning the Composition of Attention

Visual attention is driven by both salient regions that most people find attractive and individuals' unique interests. Understanding how to incorporate attention patterns of different users is an important challenge for user-aware saliency modeling. Without such knowledge, a model can overfit to attention driven by visual stimuli, and has difficulties learning the variability of users' attention. For instance, as there exists a considerable overlap between attention of different users, a model can exploit such a shortcut and predict an averaged attention map (of all users) regardless of the selection of users (see supplementary materials). To tackle the challenge, we propose a principled learning method that augments our saliency model with understanding of the composition of attention.

Drawing inspirations from human learning, which is highly organized and based on curricula that gradually introduce diverse concepts, our method emphasizes progressive learning with dynamically evolving objectives. As illustrated in Figure 2 (bottom), it can be viewed as a three-phase training procedure: (1) As the initial step of learning, the model is encouraged to establish the correspondence between personalized modules and individuals' visual preference; (2) Later on, in the second phase, it is driven to take advantage of the learned preferences and construct saliency maps for diverse selections of users; (3) In the final phase, we allow the model to refine its prediction on attention of broader users, unifying personalized [42] and general saliency prediction [23, 25, 41]. Instead of dividing the learning into three separate stages (*e.g.,* with a sequence of pretraining and fine-tuning), we instantiate the aforementioned paradigm by manipulating the characteristics of training samples throughout the learning process. Specifically, we first sample the number of users K from a specific range, and then randomly select of fixations K users (with a uniform distribution) to construct the ground truth fixation and saliency maps on the fly.

Through controlling the number of users sampled among different training epochs, our method allows a continuous and progressive learning of the composition of attention: (1) We start with annotations from a small number of users (*e.g.,* 1-4 users) and adaptively optimizing different compo-

nents of the networks (*e.g.,* personalized filters indicated by the input user masks), which allows the model to encode individuals' visual preferences. It also attenuates the effects of overfitting to attention driven by visual stimuli, as the model has no access to annotations from all users and thus can not exploit the correlation between individuals' attention and the averaged one; (2) With the progress of training, we gradually increase both the lower and the upper bounds for sampling range of K (*e.g.*, from 1-4 to 5-9 users). Note that the users are randomly selected on a sample basis for every training iteration (*i.e.,* one sample in a single batch). The scheme facilitates understanding of how to combine visual preferences of each individual to the attention patterns of different user groups; (3) During the last several training epochs, we fix the bounds for sampling to relatively large values (*e.g.,* $80\% - 100\%$ users), allowing the refinement of predictions for the general saliency.

Our learning paradigm enables a smooth transition from individuals' preferences to attention of a broad range of users. Together with the proposed model, our full method shows promise in capturing attention of both diverse users (Section 4.2) and the general population (Section 4.3)

4. Experiment

In this section, we present the implementation details (Section 4.1), and carry out extensive experiments to study the effectiveness of the proposed framework. Our experiments and analyses aim to shed light on the following research questions that have yet to be answered:

- Does learning the composition of attention help capture attention of diverse users? (Section 4.2)
- Are visual preferences beneficial for general saliency prediction? (Section 4.3)
- Can user-aware saliency modeling overcome the issues of incomplete users? (Section 4.4)

We also include additional analyses and ablation studies in the supplementary materials.

4.1. Implementation

Dataset. We demonstrate the effectiveness of our method with two types of visual stimuli, *i.e.,* naturalistic images and web pages. Since most saliency datasets only offer aggregated fixation annotations, we choose two publicly available saliency datasets, OSIE [39, 41] and FiWI [37], as our testbeds, which provide per-user annotations for user-aware saliency modeling. OSIE contains eye-tracking data collected on 700 naturalistic images with rich semantics. We use the version collected in [39], which has more diverse attention patterns from 32 users (about 14 users are presented with each sample) due to their less constrained experimental settings. FiWI is a popular eye-tracking dataset for web pages, which contains data from 11 users on 149 samples. We also conduct experiments on the recently introduced web page saliency dataset [8] with 450 samples, and report the results in our supplementary materials.

Evaluation. Saliency prediction is typically evaluated with multiple metrics [6]. Following [8, 19, 21, 30], we adopt seven popular saliency metrics, including Normalized Scanpath Saliency (NSS) [33], KL-Divergence (KLD) [27], Similarity (SIM) [35], Correlation Coefficient (CC) [31], Area Under the ROC Curve (AUC) [15], AUC-Judd [25], and shuffled AUC (sAUC) [4]. For the OSIE dataset, we follow [41], and randomly split the samples into 600 and 100 for training and evaluation, respectively. In terms of the FiWI dataset, following [37] and [8], we consider two evaluation settings, including 5-fold cross-validation and evaluation with fixed training/test sets. Besides estimating the capability of our method on capturing the attention patterns of diverse users (Section 4.2), we also study its usefulness on general saliency prediction (Section 4.3).

Model Configuration. We build our method on the state-of-the-art EML-Net [21] saliency model. We use ResNet-50 [18] as the backbone, which is pretrained on ImageNet classification [12] and offers profound understanding of visual semantics. To show the generalizability of our method, we also experiment with a different saliency model (*i.e.,* a variation of SimpleNet [34], see the supplementary materials). Multi-scale features are extracted from 4 layers, where each corresponds to the end of a block inside the backbone. The features are processed with a set of attention decoders, where the decoder of each scale contains two consecutive convolutional layers with 16 and N filters (N is the number of users), and an upsampling layer that rescales features to have the same spatial resolution as the image.

Training. Following [10], we train models for 50 epochs with batch size 10 using the Adam optimizer [26]. λ factor in equation (5) is empirically set to 0.2. Learning rate is initially set to 10^{-4}, and decayed by a factor of 0.8 for every 10 epochs. For our learning approach proposed in Section 3.2, we set the initial bounds to [1, 4] and [1, 6] for FiWI and OSIE, respectively, and increase both the upper and lower bounds by 2 every 10 epochs. The bounds are fixed to [9, 11] and [11, 15] for FiWI and OSIE after reaching the corresponding values. The aforementioned hyperparameters are determined based on 5-fold cross-validation. For a fair comparison with the state-of-the-art [8, 37] and a more flexible setting, we do not pretrain our models on external saliency datasets like the SALICON [23] dataset.

4.2. Does Learning the Composition of Attention Help Capture Attention of Diverse Users?

We first study the effectiveness of our method for capturing the visual preferences of diverse users. Specifically, we randomly sample K users (K=1, 3, 5) for each evaluation

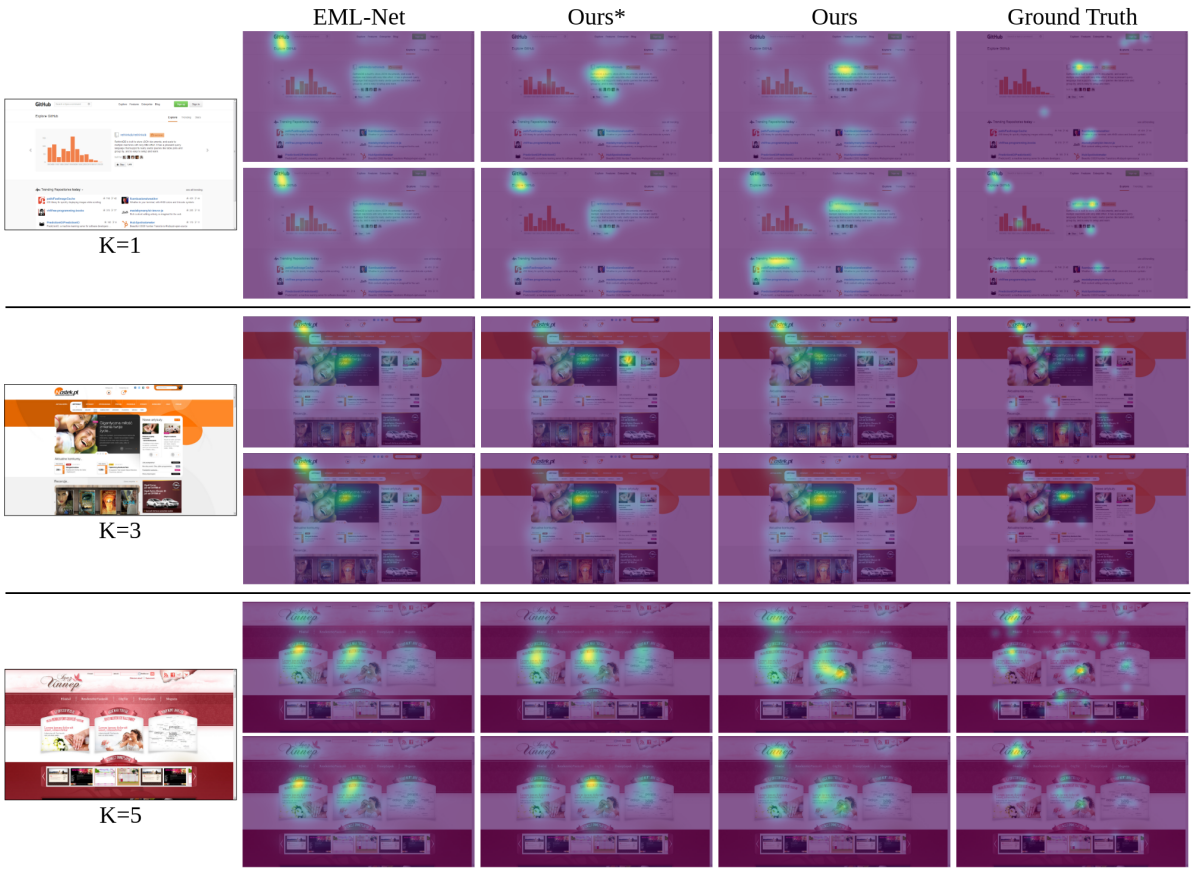

Figure 3. Qualitative results for predicting attention of diverse users. We show results for two different user groups for each sample, and also consider variable numbers of users $K = 1, 3, 5$. Note that EML-Net is user-agnostic and predicts the same result for two groups.

image, and estimate the predictions with ground truth annotations aggregated from the selected users. We repeat the procedure 5 times to reduce the variance of results, and report the average performance. Following [41] and [37], we use fixed training/test splits on the OSIE dataset and perform 5-fold cross-validation on the FiWI dataset. We compare our full method (Ours) with three approaches, including the user-agnostic EML-Net baseline [21], an ensemble of EML-Net trained on individual users (Single-user, with one model for each user), and our model without incorporating progressive learning (Ours*). Three key observations can be made on the results reported in Table 1 and Table 2:

- **There exist considerable discrepancies between general and user-aware saliency modeling.** Despite the inter-user agreement on visual attention, the user-agnostic EML-Net achieves inferior performance on both datasets, especially FiWI with web pages associated with richer semantics and more diverse viewing behaviors. The observation highlights the significance of visual preferences for attention modeling.
- **Learning how to integrate attention is important**

Table 1. User-aware saliency results on FiWI [37].

	K=1		K=3		K=5	
	NSS	CC	NSS	CC	NSS	CC
EML-Net	1.481	0.307	1.506	0.454	1.498	0.519
Single-user	1.962	0.373	1.749	0.518	1.745	0.591
Ours*	1.908	0.366	1.785	0.526	1.784	0.609
Ours	**2.059**	**0.392**	**1.829**	**0.540**	**1.815**	**0.620**

Table 2. User-aware saliency results on OSIE data [39].

	K=1		K=3		K=5	
	NSS	CC	NSS	CC	NSS	CC
EML-Net	1.755	**0.367**	1.763	0.532	1.718	0.594
Single-user	1.456	0.314	1.565	0.492	1.574	0.566
Ours*	1.640	0.294	1.768	0.537	1.812	0.625
Ours	**1.809**	0.310	**1.809**	**0.550**	**1.826**	**0.629**

Table 3. General saliency results on FiWI data [37].

	NSS	KLD	CC	AUC-Judd	sAUC
DeepGaze II	1.229	-	0.488	0.797	0.625
SAM-ResNet	1.246	-	0.595	0.791	0.673
UMSI	0.938	-	0.457	0.755	0.675
AGD-F	1.606	-	**0.735**	0.767	**0.748**
EML-Net	1.653	0.603	0.661	0.847	0.675
EML-Net+SALICON	1.722	0.567	0.689	0.848	0.697
Ours	**1.752**	**0.564**	0.699	**0.851**	0.704

Table 4. General saliency results on OSIE data [39].

	NSS	KLD	SIM	CC	AUC
SALICON	1.641	0.575	0.600	0.685	0.846
SAM-ResNet	1.811	**0.480**	0.648	0.758	0.860
UMSI	1.788	0.513	0.631	0.746	0.856
EML-Net	1.737	0.537	0.619	0.717	0.854
Ours	**1.840**	0.506	**0.652**	**0.761**	**0.860**

for capturing visual preferences. A key component of our framework is the progressive learning method for understanding the composition of attention. According to the comparative results between our methods with and without progressive learning, it helps models better identify the attentional preferences of both individuals (*i.e.,* K=1) and user groups (*i.e.,* K=3, 5). As illustrated in Figure 3, unlike EML-Net and our method without progressive learning (Ours*) that fail to capture the visual preferences and generate similar attention maps regardless of the selection of users, our full method is able to predict attention maps that accurately reflect the regions of interest of diverse users.

- **Additive ensemble falls short of modeling user-aware attention.** Training one model for each user offers a straightforward way to model user preferences, and slightly outperforms our method without progressive learning when considering individuals' attention (*i.e.,* K=1) on FiWI. However, without learning how to integrate attention, it lacks the capability to predict attention maps with multiple users and is outperformed by our full method across all settings. Additionally, such an additive ensemble also fails to benefit general saliency prediction due to its high computational overhead and the inability to jointly consider attention of different users (see our supplementary materials).

These observations highlight the significance of modeling visual preferences and the composition of attention, and demonstrate the advantages of our method for user-aware saliency modeling. While we do not consider demographic information in this study, our method is general and can be extended to broader scenarios for customized applications. For instance, instead of developing personalized filters, we can cluster users based on certain criteria (*e.g.,* gender, age, personal interest) and optimize models to predict attentional preferences of users of specific characteristics.

4.3. Are Visual Preferences Beneficial for General Saliency Prediction?

General saliency prediction focuses on attention aggregated from all users, and yet little attention has been paid to the diversity of attention patterns between users. In this subsection, we further investigate if knowledge of visual preferences is beneficial for the task. We compare our user-aware method with six state-of-the-art models, including SALICON [19], Deep Gaze II [28], SAM [10], EML-Net [21], UMSI [16], and AGD-F [8]. For a fair comparison, we adopt the training and test sets provided in [8] for evaluation on FiWI dataset. Table 3 and Table 4 report results on the FiWI [37] and OSIE [39, 41] datasets.

Compared to the EML-Net baseline, our model that takes into account visual preferences shows consistent improvements among all settings. Its performance is also competitive against existing state-of-the-art. As illustrated in Figure 4, the performance gain of our method can be attributed to the capability to capture fine-grained attention patterns (*i.e.,* focuses on the man, little girls, and woman in the three examples, respectively) in addition to the salient regions attended by the majority of users (*i.e.,* text, the food, and the boy). Moreover, the advantages of our method is also a result of increased data efficiency. Comparative results show that it outperforms EML-Net pretrained on the SALICON [23] dataset with 15000 images (EML-Net+SALICON), despite only relying on training samples in the FiWI dataset with 149 images. Such a unique feature can play an important role in generalizing towards domains with diverse characteristics (*e.g.,* web page vs naturalistic image), where suitable data for pretraining is not necessarily available.

4.4. Can User-aware Saliency Modeling Overcome the Issues of Incomplete Users?

Collecting human behavioral data demands stringent paradigms, and building large attention datasets has been challenging for decades [23]. As a result, it is common for saliency datasets to have samples associated with incomplete users. For instance, there are 32 users recruited in [39], while only about 14 users are presented for each sample. Different from prior works [19, 21, 28, 30] that overlook the issue, we examine if knowledge of visual preferences can be helpful for tackling the problem.

Specifically, we randomly drop K^* users for each train-

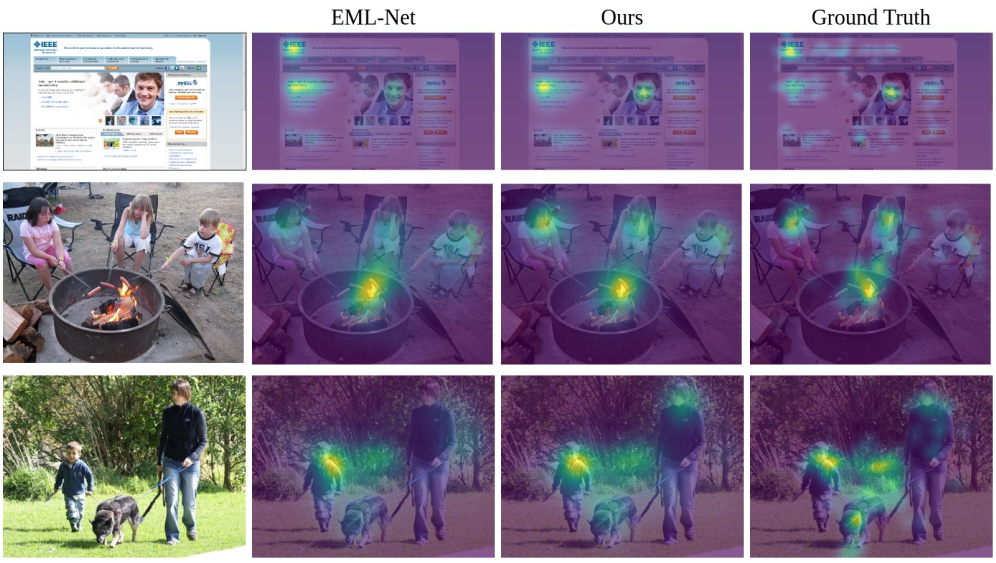

Figure 4. Qualitative results for general saliency prediction.

Table 5. Results for saliency prediction with incomplete users on seen and unseen images.

		K=1		K=3		K=5		Unseen	
		NSS	CC	NSS	CC	NSS	CC	NSS	CC
$K^* = 5$	EML-Net	1.733	**0.359**	1.719	0.518	1.712	0.589	1.616	0.601
	Ours	**1.759**	0.341	**1.781**	**0.530**	**1.797**	**0.614**	**1.698**	**0.632**
$K^* = 7$	EML-Net	1.662	**0.342**	1.659	0.501	1.658	0.573	1.584	0.551
	Ours	**1.723**	0.330	**1.734**	**0.512**	**1.739**	**0.594**	**1.708**	**0.589**

ing sample in the FiWI dataset, and consider two evaluation settings: (1) attention of missing users on seen samples (we use different numbers of missing users, *i.e.,* K=1, 3, 5), and (2) attention of all users on unseen test samples (*i.e.,* Unseen). The first setting is particularly useful for scenarios where we are interested in comparing users' preferences. On the other hand, the second setting aims to study models' robustness against the scarcity of annotations.

According to comparative results reported in Table 5, the proposed user-aware method is advantageous in both evaluation settings. Despite the slightly lower CC scores for individuals' attention (K=1) on seen images, it achieves considerable improvements among all other evaluation settings. This is likely because predictions from the user-agnostic EML-Net tend to have a broader coverage, which is favored by CC metric emphasizing the overall distribution. Differently, our method has the capability to more accurately attend to regions of interest of different users, resulting in higher NSS scores as well as CC scores for more users. It also makes better use of the available data, and is more robust against the scarcity of annotations (*i.e.,* higher generalization performance on unseen images).

5. Conclusion

This paper identifies the critical roles of user preferences in visual attention, and for the first time tackles the challenge of user-aware saliency modeling. It unifies the conventional tasks for predicting general and personalized attention, and proposes an integral framework that jointly solves attention modeling for individual users, different groups of users, and the general population. By making progress with both a novel saliency prediction model and a progressive learning method, our framework illustrates a principled paradigm for establishing the connections between user preferences and visual saliency. Experimental results in diverse settings demonstrates the advantages of the proposed method in inferring the attention patterns of a variety of users, and highlights the significance of incorporating the variability of attention. We hope that our study can open new avenues for attention research, and benefit the development of new applications based on user-aware saliency models.

References

[1] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. Dr(eye)ve: A dataset for attention-based tasks with applications to autonomous and assisted driving. In *CVPR workshop*, pages 54–60, 2016. 2

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, page 41–48, 2009. 2

[3] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR workshop*, 2015. 2

[4] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *ICCV*, pages 921–928, 2013. 5

[5] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *NeurIPS*, volume 18, 2005. 2

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *TPAMI*, 41(3):740–757, 2019. 5

[7] Zoya Bylinskii, Nam Wook Kim, Peter O'Donovan, Sami Alsheikh, Spandan Madan, Hanspeter Pfister, Frédo Durand, Bryan C. Russell, and Aaron Hertzmann. Learning visual importance for graphic designs and data visualizations. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 2017. 1, 2

[8] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, Aruna Balasubramanian, Gregory J. Zelinsky, and Dimitris Samaras. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 1, 2, 5, 7

[9] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *ECCV*, pages 91–107, 2020. 2

[10] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018. 1, 2, 3, 4, 5, 7

[11] Benjamin de Haas, Alexios L. Iakovidis, D. Samuel Schwarzkopf, and Karl R. Gegenfurtner. Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, 116(24):11687–11692, 2019. 1, 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4, 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[14] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *CVPR*, pages 7521–7531, 2018. 2

[15] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. 5

[16] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. Predicting visual importance across graphic design types. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, page 249–260, 2020. 1, 2, 7

[17] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[19] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, pages 262–270, December 2015. 1, 2, 4, 5, 7

[20] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489–1506, 2000. 2

[21] Sen Jia and Neil D.B. Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020. 1, 2, 3, 4, 5, 6, 7

[22] Ming Jiang, Shi Chen, Jinhui Yang, and Qi Zhao. Fantastic answers and where to find them: Immersive question-directed visual attention. In *CVPR*, pages 2977–2986, 2020. 1, 2

[23] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015. 2, 4, 5, 7

[24] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012. 2, 3

[25] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009. 2, 4, 5

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[27] Solomon Kullback. *Information Theory and Statistics*. Wiley, 1959. 4, 5

[28] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Deepgaze II: reading fixations from deep features trained on object recognition. *Arxiv*, 2016. 2, 3, 4, 7

[29] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In *ICLR workshop*, 2015. 1

[30] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494:455–467, 2022. 1, 2, 5, 7

[31] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007. 4, 5

[32] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention. *IEEE International Conference on Consumer Electronics*, 20(8), 2020. 2, 3

[33] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005. 4, 5

[34] Navyasri Reddy, Samyak Jain, Pradeep Yarlagadda, and Vineet Gandhi. Tidying deep saliency prediction architectures. In *IROS*, 2020. 5

[35] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000. 5

[36] Eldon Schoop, Xin Zhou, Gang Li, Zhourong Chen, Bjoern Hartmann, and Yang Li. Predicting and explaining mobile ui tappability with vision modeling and saliency analysis. In *Conference on Human Factors in Computing Systems (CHI)*, 2022. 1, 2

[37] Chengyao Shen and Qi Zhao. Webpage saliency. In *ECCV*, pages 33–46, 2014. 1, 2, 5, 6, 7

[38] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018. 1, 2

[39] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, and Vidhya Navalpakkam. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, 11, 2020. 5, 6, 7

[40] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, pages 2798–2805, 2014. 2

[41] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):1–20, 2014. 2, 4, 5, 6, 7

[42] Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu. Personalized saliency and its prediction. *TPAMI*, 41(12):2975–2989, 2019. 2, 3, 4

[43] Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao. Beyond universal saliency: Personalized saliency prediction with multi-task cnn. In *IJCAI*, pages 3887–3893, 2017. 2, 3

[44] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360° videos. In *ECCV*, pages 504–520, 2018. 1, 2

[45] Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson W. H. Lau. Task-driven webpage saliency. In *ECCV*, pages 300–316, 2018. 1, 2