Saliency in Augmented Reality

Huiyu Duan
huiyuduan@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Wei Shen*
wei.shen@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Xiongkuo Min
minxiongkuo@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Danyang Tu
danyangtu@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Jing Li
lj225205@alibaba-inc.com
Alibaba Group
Hangzhou, China

Guangtao Zhai*
zhaiguangtao@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

ABSTRACT

With the rapid development of multimedia technology, Augmented Reality (AR) has become a promising next-generation mobile platform. The primary theory underlying AR is human visual confusion, which allows users to perceive the real-world scenes and augmented contents (virtual-world scenes) simultaneously by superimposing them together. To achieve good Quality of Experience (QoE), it is important to understand the interaction between two scenarios, and harmoniously display AR contents. However, studies on how this superimposition will influence the human visual attention are lacking. Therefore, in this paper, we mainly analyze the interaction effect between background (BG) scenes and AR contents, and study the saliency prediction problem in AR. Specifically, we first construct a Saliency in AR Dataset (SARD), which contains 450 BG images, 450 AR images, as well as 1350 superimposed images generated by superimposing BG and AR images in pair with three mixing levels. A large-scale eye-tracking experiment among 60 subjects is conducted to collect eye movement data. To better predict the saliency in AR, we propose a vector quantized saliency prediction method and generalize it for AR saliency prediction. For comparison, three benchmark methods are proposed and evaluated together with our proposed method on our SARD. Experimental results demonstrate the superiority of our proposed method on both of the common saliency prediction problem and the AR saliency prediction problem over benchmark methods. Our dataset and code are available at: https://github.com/DuanHuiyu/ARSaliency.

CCS CONCEPTS

• **Computing methodologies → Mixed / augmented reality**; **Perception**; **Interest point and salient region detections**.

KEYWORDS

Augmented Reality, dataset, saliency prediction, visual confusion

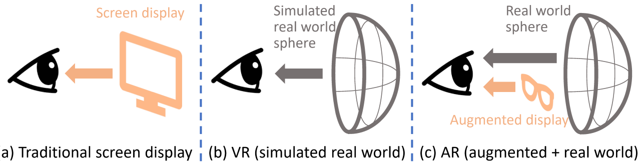

Screen display
Simulated real world sphere
Real world sphere
Augmented display
(a) Traditional screen display
(b) VR (simulated real world)
(c) AR (augmented + real world)

Figure 1: Illustration of the perception differences between traditional screen display, VR display, and AR display.

ACM Reference Format:
Huiyu Duan, Wei Shen, Xiongkuo Min, Danyang Tu, Jing Li, and Guangtao Zhai. 2022. Saliency in Augmented Reality. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547955

1 INTRODUCTION

With the evolution of multimedia technology, head-mounted display (HMD) technologies (*e.g.*, Virtual Reality (VR), Augmented Reality (AR), Mixed Reality (MR), *etc.*) have attracted more attention recently and have become promising next-generation display solutions [6, 61]. Among these technologies, AR aims at enriching the real-world information by superimposing virtual contents on it, which promises to be the next-generation mobile platform. Since AR technology can display augmented contents while keeping real-world information, it has great potentials in many applications, such as communication, entertainment, health care, education, engineering design, *etc.*

The superimposition of background (BG) scenes (*i.e.*, real-world scenes) and AR contents (*i.e.*, virtual-world contents) in AR display will cause visual confusion [47, 59], which alters the way that humans perceive both of the real world and virtual world [14–16]. The primary value of AR is to enrich the real-world information, however, inappropriate augmented design may affect the Quality of Experience (QoE) of users [15], arouse inattentional blindness [41], and may even cause danger [57]. Therefore, it is important to analyze and model the visual attention behaviour of humans in AR environment.

Visual attention analysis and prediction have long been important tasks in both multimedia and computer vision research [1, 32, 49], since they can give new insights about human attention mechanisms and contribute to many multimedia applications, such as image compression, object and motion detection, *etc.* Many previous works on saliency datasets [4, 33] and saliency prediction models [9, 35] have been conducted on screen display as shown

*Corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547955

in Figure 1 (a). Recently, with the popularity of head-mounted displays, some visual attention datasets [12, 49] and saliency prediction models [65] towards VR technology have been proposed as demonstrated in Figure 1 (b). Most of these works focused on the saliency tasks on omnidirectional images/videos [49], or egocentric videos [63]. Although AR technology has great potentials, most of existing AR studies mainly focus on egocentric video understanding [25], while they may ignore that the augmented contents are also important for AR. Zhu *et al.* [68] have constructed a saliency dataset for omnidirectional videos with augmented bounding box contents on them. However, as shown in Figure 2 (a), it is hard to acquire real-world omnidirectional view in AR applications, and augmented contents are usually complex and have various superimposition degrees rather than overlaying simple bounding boxes. Duan *et al.* [15] have studied the visual confusion effect in AR technology. However, they only explored the influence of visual confusion on quality of experience (QoE), while the visual attention research is still lacking. As shown in Figure 1 (c), the visual confusion caused by the superimposition of AR contents and BG scenes is significant in AR, however, the understanding of saliency in Augmented Reality environment is still limited.

Modeling the visual attention in AR can help better design, display, and adaptively adjust the virtual contents to accord the expectation of human vision [36], as well as contribute to AR QoE assessment [15] and augmented image compression method design [68] *etc.* Therefore, in this work, we aim at analyzing the human visual attention behavior in AR thoroughly and building an accurate saliency prediction paradigm for AR. To achieve this objective, we are facing the following research challenges:

(i) *Building a dataset for AR saliency.* Although there are many famous AR products, *e.g.*, HoloLens [43], Magic Leap [38], Epson AR [20], *etc.*, it is hard to conduct controllable eye-tracking experiments with these devices in real scenes. Moreover, the experimental scenarios may also be limited by the experimental environment.

(ii) *Understanding the effect of visual confusion on AR saliency.* As a common observation, a higher opacity value for augmented contents may make them clearer and attract more attention, while a lower opacity value for augmented contents may make background scenes clearer and attract more attention. Despite this general consensus, we observe that our understanding of the influence of the mixing level is still limited, and this general consensus cannot be generalized to various complex applications.

(iii) *Modeling AR saliency.* Since humans can perceive two layers in Augmented Reality, *i.e.*, an AR layer and a BG layer, it is necessary and significant to study how to jointly exploit these two parts of information to build an accurate saliency model for AR.

In this work, to solve the dataset challenge, we first propose to conduct the AR eye-tracking experiment under the VR environment. As shown in Figure 1 (b) and (c), since VR is used to simulate real world scenes, it is also competent to be used as BG scenes in the AR experiments. Specifically, we collect 450 omnidirectional images as BG images, and 450 common images including 150 graphic images, 150 natural images, 150 webpage images as AR images. To better understand the influence of mixing levels on the saliency in AR, three mixing levels are used during the eye tracking experiment, and the experiment of each mixing level is conducted among different subjects. To better predict the visual attention in AR, we propose a vector quantization-based saliency model and utilize multi-decoders to integrate obtained information for joint prediction. For comparison, three benchmark paradigms for AR saliency prediction are proposed and evaluated based on our dataset. Experimental results demonstrate that our proposed model achieves state-of-the-art performance compared to other baseline methods. The contributions are summarized as follows.

- We build the first AR saliency dataset that considers the visual confusion effect in AR.
- We analyze the influence of visual confusion on visual attention for various stimuli.
- A saliency prediction model for predicting AR saliency is proposed.
- Three benchmark methods are proposed and evaluated based on our dataset, and our proposed model achieves state-of-the-art performance.

2 RELATED WORK

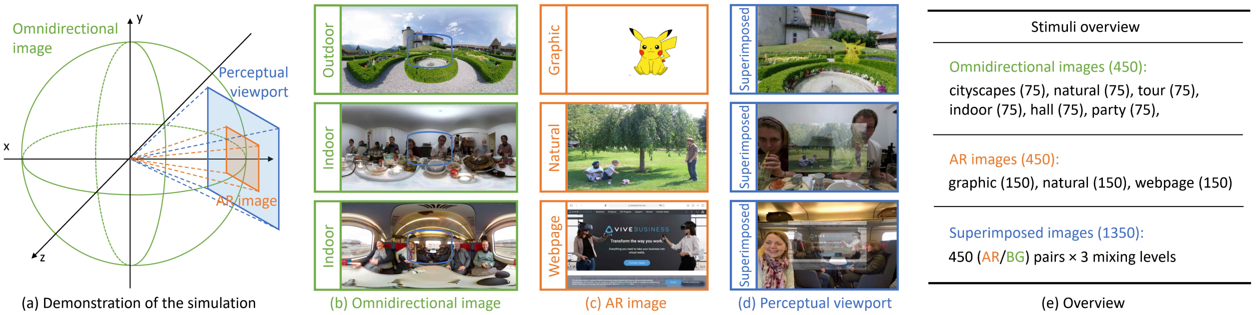
2.1 Eye-tracking Datasets

Traditional Saliency Datasets. Humans have remarkable abilities to search and focus on salient regions in a scene [3, 42], which allows us to efficiently process a large amount of information. This neural mechanism is known as visual attention. To understand and model visual attention behavior, many eye-tracking datasets have been constructed. MIT1003 [35] is a large-scale saliency dataset, which contains 1003 images. MIT300 and CAT2000 [4] are two widely used benchmark datasets, which contains 300 and 2000 test images respectively. SALICON [33] is currently the largest crowd-sourced saliency dataset, which contains 10000 training images, 5000 validation images and 5000 test images collected through mouse tracking using Amazon Mechanical Turk (AMT). This dataset is widely used to pretrain saliency models.

VR/AR Saliency Datasets. Recently, with the popularity of HMDs and the concept of XR, many eye-tracking datasets have been constructed towards the applications in these new display technologies [67]. Salient360 [49] is one of the earliest omnidirectional datasets for saliency prediction, which contains 98 stimuli including indoor, outdoor and people scenes. For each omnidirectional image, at least 40 subjects were recruited to view the stimuli for 25 seconds. AOI [60] is a large-scale omnidirectional saliency dataset, which contains 600 images and the corresponding head/eye fixations obtained from 30 subjects. All of these datasets are constructed towards the VR saliency detection task. Zhu *et al.* [68] have established a saliency dataset for 50 omnidirectional videos with bounding boxes. However, this is still an omnidirectional saliency task. Moreover, the bounding box contents can only cover part of AR applications, and they only overlay the AR contents on BG scenes, which ignores the mixing value between AR and BG. In this work, we argue that the interaction between AR contents and BG scenes is important in AR display, and study this important task among various scenarios within perceptual viewport images.

2.2 Saliency Prediction Models

Classical Saliency Models. Most traditional methods have modeled visual saliency based on the bottom-up mechanism. The early models mainly relied on extracting simple low-level feature maps such as intensity, color, and orientation *etc* [32]. Some subsequent



Stimuli overview

Omnidirectional images (450):
cityscapes (75), natural (75), tour (75),
indoor (75), hall (75), party (75),

AR images (450):
graphic (150), natural (150), webpage (150)

Superimposed images (1350):
450 (AR/BG) pairs × 3 mixing levels

(a) Demonstration of the simulation (b) Omnidirectional image (c) AR image (d) Perceptual viewport (e) Overview

Figure 2: The illustration of the AR simulation in VR environment. (a) The demonstration of the relationship between the omnidirectional image, the AR image, and the perceptual viewport image. (b) Examples of the omnidirectional images. (c) Examples of the AR images. (d) Examples of the perceptual viewport images. Note that the perceptual viewports of the subjects are changed dynamically with the head movement, however, the relative positional relationship between the omnidirectional image and the AR image is fixed. (e) An overview of the stimuli in our dataset.

models incorporated middle-level and high-level features to better predict visual attention [35, 40]. These classical methods including Attention for Information Maximization (AIM) [2], Graph-based Visual Saliency (GBVS) [27], Judd model [35], *etc.*, are still highly influential in current visual attention research.

Deep Saliency Models. With the development of deep neural network (DNN), the saliency prediction task has achieved significant improvement recently [37, 50, 53, 55, 56]. Huang *et al.* [31] proposed a two stream convolutional neural network (CNN) to extract coarse and fine features to compute saliency map. Cornia *et al.* [11] used long short-term memory (LSTM) to enhance the extracted feature maps from a dilated CNN to predict saliency. Pan *et al.* [46] proposed to use the generative adversarial network (GAN) to calculate the saliency map. Che *et al.* [9] studied the influence of transformation on visual attention and proposed a GazeGAN model based on U-Net for saliency prediction. These models based on the top-down mechanism have been widely used in various research fields recently [1].

2.3 Augmented Reality

This work mainly concerns head-mounted AR application rather than mobile phone based AR application. Early head-mounted AR devices, *e.g.*, Google Glass [24], generally display augmented contents for only one eye and used another eye for perceiving real-world scenes based on binocular visual confusion [7, 15]. However, binocular rivalry caused by binocular visual confusion may strongly affect the QoE [15, 36]. Recently, most AR technologies are built based on monocular visual confusion to avoid occluding when displaying augmented contents, such as Microsoft HoloLens [43], Magic Leap [38], Epson AR [20], *etc.*, since monocular rivalry is much weaker than binocular rivalry [45]. In this paper, we mainly consider these monocular visual confusion based technologies. This type of device generally has external cameras and internal gyroscope sensors to register the location of augmented contents in real-world environment, thus it is easy to get the position relationship between the AR contents and BG scenes [15, 61].

3 SARD: SALIENCY IN AR DATASET

3.1 Experimental Methodology

In real applications, since AR usually needs external cameras on HMDs to register and locate real-world scenes for augmented rendering, thus it is possible to capture background scenes, obtain AR contents, and acquire the relationship between AR contents and BG scenes. An intuitive way to conduct AR eye tracking experiment is wearing AR devices in various environments and then collecting eye movement data. However, this way suffers from uncontrollable experimental environments and limited experimental scenarios [15]. Another way to simulate AR scenarios is using a big screen with displaying superimposed AR/BG images on it. However, big screens cannot create immersive experience, which may not be appropriate in this study. As discussed in Sec. 1 and Figure 1, VR can simulate real world scenes [17, 18], which is capable to simulate the BG scenes in AR applications. Therefore, as demonstrated in Figure 2 (a), we adopt the method of conducting AR eye-tracking experiments in VR environment for controllable experimental environments and diverse experimental scenarios.

3.2 Data Collection

Stimuli. We first collect 450 omnidirectional images from [60] as BG scenes, which contains six categories of scenes, *i.e.*, cityscapes, natural landscapes, human tour scenes, indoor scenes, indoor hall scenes, and human party scenes. Example images are shown in Figure 2 (b). We then collect 450 common images online as AR contents, which consist of three types of images including graphic images, natural images, and webpage images as demonstrated in Figure 2 (c). For graphic images, only graphic areas are non-transparent, and for natural and webpage images, the whole images are non-transparent. The 450 omnidirectional images and 450 AR images are randomly paired to generate 1350 various perceptual scenarios with three mixing values, and the perceptual viewport images are formulated as:

$$I_S = \alpha I_\alpha I_{\text{AR}} + (1 - \alpha I_\alpha) I_{\text{BG}}, \qquad (1)$$

where I_{AR} and I_{BG} are AR and BG images, respectively, I_α is the intrinsic transparency matrix for AR contents, $\alpha \in \{0.25, 0.5, 0.75\}$ is the mixing value to generate superimposed images. I_α is usually a 0-1 matrix, where 0 means transparent and 1 means non-transparent. Since perceptual viewports are usually larger than the field-of-view (FOV) of AR contents, we pad each AR image to the perceptual viewport size with 0 values for both of the color space and the transparency space, and keep the raw color and transparency values for original contents. The examples of generated superimposed images are shown in Figure 2 (d). Note that this generation method for superimposed images is only used for the construction of the dataset.

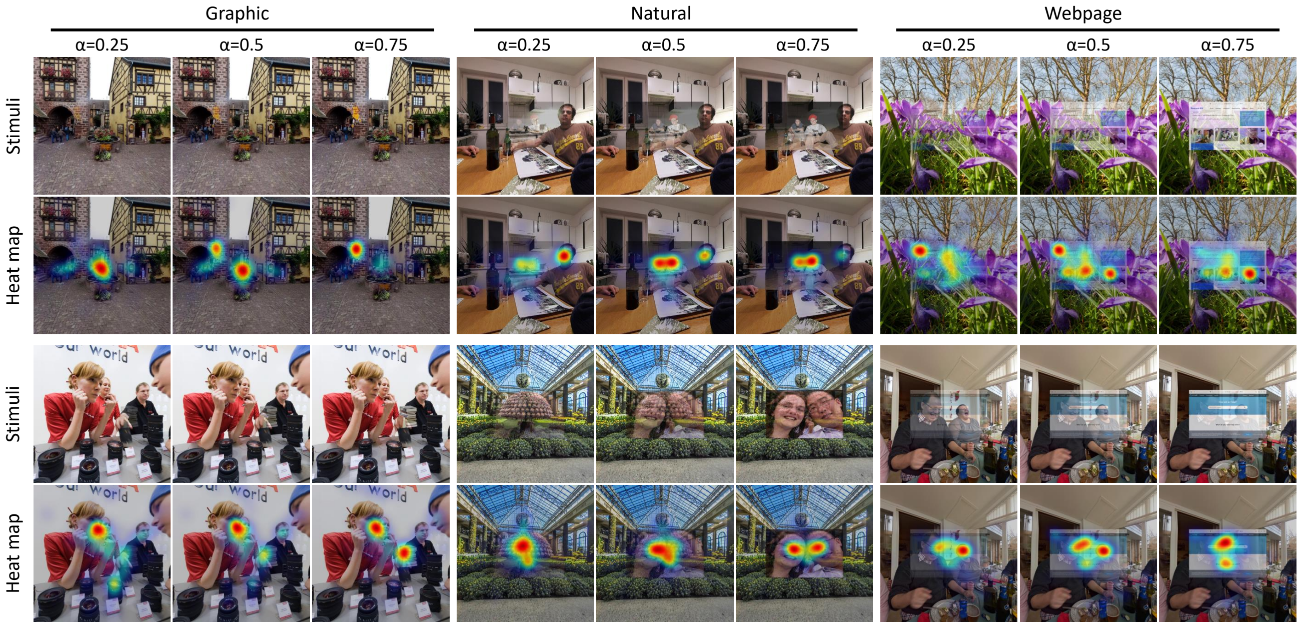


Figure 3: Qualitative comparisons of saliency maps for stimuli with various mixing values. The augmented contents can be derived from the comparison between stimuli with different mixing values.

During the eye-tracking experiment, the superimposition process is conducted in Unity3D [51] as illustrated below in "*Procedure*".

Apparatus. We use a HTC VIVE Pro Eye [30] as the hardware apparatus to display omnidirectional scenes and AR stimuli, as well as to collect eye movement data. The resolution of the displays inside HTC VIVE Pro Eye is 1440 × 1600 pixels per eye which covers 110° FOV. The refresh rate of the displays is 90 Hz. Moreover, this HMD has Tobii eye-tracker inside it with the sampling frequency of 90 Hz.

Subjects. Since 3 mixing levels are imposed for each AR/BG pair, subjects may remember the scenario if they have accessed it before and it may influence the reliability of collected data. Thus, we recruit a large number of subjects, *i.e.*, 60 subjects (20 females and 40 males), and each subject randomly watches only one mixing level of AR/BG pairs. As a result, each subject views totally 450 scenarios without scene repeat, and each superimposed viewport is viewed by 20 subjects. Before participating in the test, all subjects have read and signed a consent form which explained the human study. All participants have normal or correct-to-normal visual acuity during the experiment.

Procedure. The software system is designed using Unity3D [51] to control the experimental procedure and record all data. For each AR/BG pair and one mixing value α, we set the omnidirectional (BG) image to cover the whole space, and set the AR image at the center viewport of the BG with adjusting the transparency value of it (the AR image) to α in Unity. Before the formal experiment, a simple training session is conducted for subjects to make them familiarize with the HMD and scenarios. During the formal experiment, for each subject, 450 scenarios are randomly divided into 3 sessions with 150 scenarios per session. Unlike previous omnidirectional saliency prediction task in VR, our work focuses more on the saliency within the perceptual viewport of AR. Therefore, in our study, to make the perceptual viewport relatively fixed, subjects are seated in a fixed chair facing the center viewport of the BG rather than a swivel chair which is usually used for VR experiments [49, 60], and they are encouraged to rotate their head freely but cannot turn their body. The duration for viewing each superimposed scenario is set to 5 seconds. After viewing each superimposed image, we insert a gray omnidirectional image with a red dot located at longitude = 0° and latitude = 0°, and no AR content is displayed. The subjects are encouraged to fixate on the red dot before the next image. At the beginning of each session, we re-calibrate the eye-tracker to ensure the reliability of the acquired data.

3.3 Data Processing and Analysis

Data Processing. The raw eye movement data are recorded in the format of [pitch, yaw, roll], thus we first convert the raw data to latitude and longitude coordinates. Then we process the gaze data to extract the fixation points. Fixation occurs when user's eyes fixate at a specific region for a short period of time. We derive fixation by removing saccade (fast eye movement change) from the data. Specifically, we first calculate the distance and velocity between consecutive gaze points. Then the mean absolute deviation (MAD) [54] in gaze position is calculated within a seven-sample sliding window (80 ms) and potential fixations are defined as windows with a MAD less than $50°/s$ [28, 48]. Fixations with durations shorter than 100 ms are excluded [48, 58]. Finally, a 2D Gaussian kernel with 3.34° of visual angle [49, 66] is imposed to all fixations to generate the saliency map for an image.

Qualitative Data Analysis. Figure 3 demonstrates some sampled saliency maps of the corresponding perceptual viewport images. Similar to the general consensus as mentioned in the introduction, we first find that a higher opacity value (*i.e.*, higher α, lower transparency) generally leads more attention on augmented contents, while a lower opacity value makes the background scenes more salient. Furthermore and more importantly, we also find that

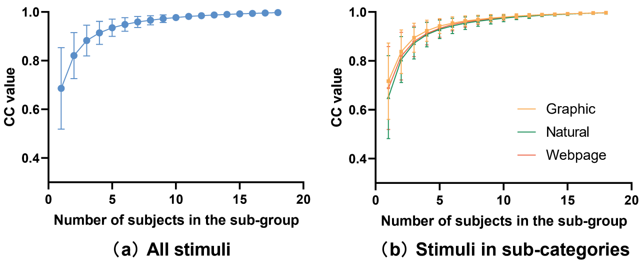


Figure 4: CC values alongside increased numbers of subjects per scenario over our SAR dataset. (a) CC values between sub-group saliency maps and overall saliency maps for all stimuli. (b) CC values between sub-group saliency maps and overall saliency maps for stimuli in sub-categories.

visual attention in AR environment (*i.e.*, superimposed stimuli) is jointly and significantly influenced by the AR image, BG image and mixing value. As shown in the examples of the first two rows in Figure 3, for $\alpha = 0.25$, subjects tend to fixate more on salient regions of the background images, for $\alpha = 0.75$, subjects tend to focus more on salient regions of the augmented images, for $\alpha = 0.5$, subjects tend to fixate on the salient regions of both AR and BG images. Thus we can clearly observe the saliency transform procedure from these three examples. For the three examples of the last two rows in Figure 3, saliency maps for $\alpha = 0.25$ and $\alpha = 0.5$ are more similar compared to $\alpha = 0.75$, which means that the transition is slight when α is less than 0.5. Therefore, Figure 3 qualitative illustrates that visual attention in AR environment is jointly influenced by the AR image, BG image and mixing value rather than only influenced by the mixing value.

Quantitative Data Analysis. As shown in Figure 4, we first analyze the consistency of eye fixation distributions across subjects when the number of subjects increases. The consistency of eye fixations between two groups is measured by calculating the linear correlation coefficient (CC). Figure 4 (a) shows the CC values between sub-group (with fewer subjects) saliency maps and overall (all subjects) saliency maps, which are calculated and averaged among all stimuli. It can be observed that the CC value, *i.e.*, consistency, increases and converges along with the increased subject number. Moreover, according to the consistency value and standard deviation value, we recommend that at least 10 subjects are required to obtain good consistency for conducting AR saliency experiment. We also analyze the consistency of eye fixation distributions across subjects for stimuli in sub-categories and show the results in Figure 4 (b). It can be observed that there is no obvious difference between different types of stimuli.

We further analyze the correlation between the saliency maps of the stimuli with different mixing values. Two correlation metrics including the linear correlation coefficient (CC) and the similarity measurement (SIM) are used for measuring the correlation. Figure 5 (a) shows the CC comparisons between different mixing levels for all stimuli. The three CC values are calculated between mixing value 1 (m1: $\alpha = 0.25$) and mixing value 3 (m3: $\alpha = 0.75$), mixing value 1 (m1: $\alpha = 0.25$) and mixing value 2 (m2: $\alpha = 0.5$), mixing value 2 (m2: $\alpha = 0.5$) and mixing value 3 (m3: $\alpha = 0.75$), respectively, and then averaged among all stimuli. It can be observed that the averaged CC of "m1 & m3" is significantly less than that of "m1 & m2" and "m2 & m3", which quantitatively illustrates that the mixing value significantly influences the visual attention in AR, and saliency maps of superimposed images with middle mixing values (α near 0.5) tend to fuse the saliency maps of superimposed images with side mixing values (α away from 0.5). Figure 5 (b) shows the CC comparisons between different mixing levels for the stimuli in sub-categories. It can be observed that the averaged CC of graphic images is significantly larger than that of natural and webpage stimuli. The reason may be that the stimuli with graphic augmented contents in our dataset have less superimposed areas compared to natural and webpage contents thus have less influence on visual attention for various mixing values.

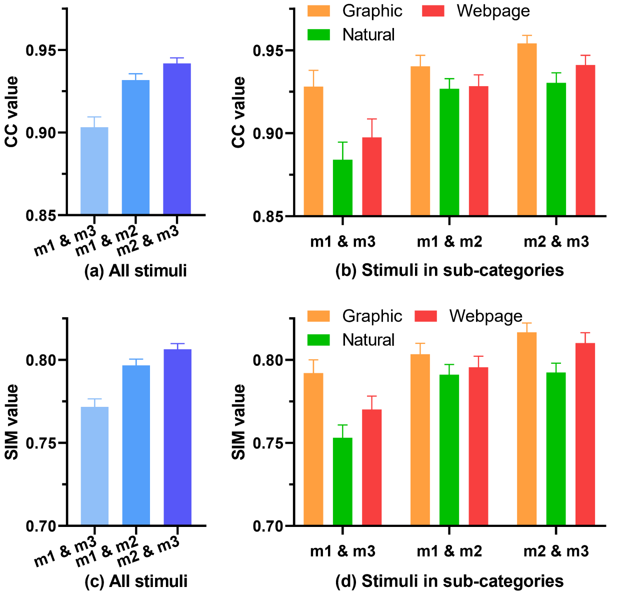


Figure 5: CC and SIM comparisons between different mixing levels (m1: $\alpha = 0.25$; m2: $\alpha = 0.5$; m3: $\alpha = 0.75$). (a) CC comparisons for all stimuli. (b) CC comparisons for stimuli in sub-categories of augmented contents. (c) SIM comparisons for all stimuli. (d) SIM comparisons for stimuli in sub-categories of augmented contents.

4 VECTOR QUANTIZED SALIENCY (VQSAL) PREDICTION IN AR

As discussed above in Sec. 3.3, the AR image, BG image and mixing value in a perceptual scenario jointly influence the visual attention in AR. Thus, it is important to consider how to integrate these three types of information for saliency prediction in AR. In this section, we first propose a vector quantization (VQ) based method for visual saliency prediction as described in Sec. 4.1 and Sec. 4.2. Most recent top-down saliency models, *e.g.*, SALICON [31], ML-Net [10], SAM-VGG/ResNet [11], SalGAN [46], rely on well-pretrained encoders to work well. These encoders are usually pretrained on the ImageNet classification task [13]. Different from these methods, in this work, we find that using unsupervised discrete representation learning as the pretraining method can lead even better saliency prediction performance. Then, a specifically designed multi-decoder fusion network for AR saliency prediction is proposed in Sec. 4.3.

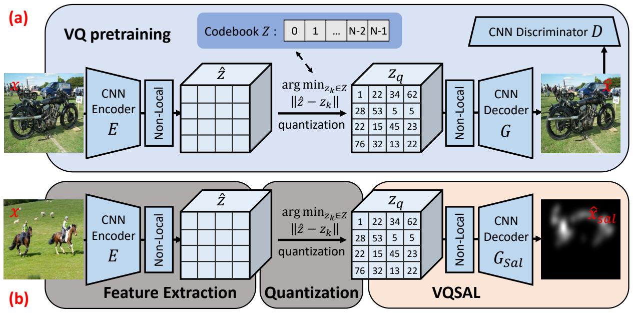


Figure 6: Overview of the proposed VQSal model. Our VQSal model is first pretrained on a large-scale unlabeled image dataset to learn a context-rich codebook for images. Then we freeze the feature extraction and quantization parts, and only finetune the decoder part to perform saliency prediction.

4.1 Learning a Discrete Representation Model with Perceptually Rich Information

The procedure of the vector quantized pretraining is demonstrated in Figure 6 (a). Instead of building on individuals pixels, neural discrete representation learning [52] aims to represent any image $x \in \mathbb{R}^{H\times W\times 3}$ by a spatial collection of codebook entries $z_{\mathbf{q}} \in \mathbb{R}^{h\times w\times n_z}$ from the codebook $\mathcal{Z}$, where n_z is the dimensionality of codes and $\mathcal{Z} = \{z_k\}_{k=1}^{K} \subset \mathbb{R}_z^{n}$ is the learned perceptually rich code book. Specifically, a given image x is first encoded by the encoder E to get the feature vector $\hat{z} = E(x) \in \mathbb{R}^{h\times w\times n_z}$. Then each spatial code $\hat{z}_{ij} \in \mathbb{R}_z^{n}$ in $\hat{z}$ is quantized by $\mathbf{q}(\cdot)$ to its closest codebook entry z_k in the codebook $\mathcal{Z}$ via $z_{\mathbf{q}} = \mathbf{q}(\hat{z}) := \left(\arg\min_{z_k\in\mathcal{Z}} \|\hat{z}_{ij} - z_k\|\right) \in \mathbb{R}^{h\times w\times n_z}$. Finally, the image can be reconstructed from these codebook entries by $\hat{x} = G(z_{\mathbf{q}})$, where $\hat{x}$ is the output of the whole model. The overall discrete representation learning pipeline is:

$$\hat{x} = G(z_{\mathbf{q}}) = G\left(\mathbf{q}(E(x))\right). \tag{2}$$

Since both of the codebook ($\mathcal{Z}$) and the model (*i.e.*, E and G) are required to be learned. The vector quantized loss function can be represented as:

$$\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|^2 + \mathcal{L}_{\text{perceptual}}, \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) = \mathcal{L}_{\text{rec}} + \ \ &\|\text{sg}[E(x)] - z_{\mathbf{q}}\|_2^2 \\ &+ \beta\|\text{sg}[z_{\mathbf{q}}] - E(x)\|_2^2,\end{aligned} \tag{4}$$

where $\mathcal{L}_{\text{perceptual}}$ is the well-known perceptual loss [34, 64], sg[·] indicates the stop-gradient, $\|\text{sg}[z_{\mathbf{q}}] - E(x)\|_2^2$ is the "commitment loss" with weighting factor β [52].

To get good reconstruction quality for this discrete representation learning, we follow the VQGAN [22] to learn a perceptually rich codebook via GAN as follows:

$$\mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) = \left[\log D(x) + \log(1 - D(\hat{x}))\right]. \tag{5}$$

4.2 Transfer Learning for Saliency Prediction

Through the discrete representation model learned above, we can represent any image x using a spatial collection of codebook entries $z_{\mathbf{q}}$, and directly reconstruct the image using these codes (visual tokens) via the decoder G. The representation code $z_{\mathbf{q}}$ includes the extremely compressed but perceptually rich information of an image, which can be directly used to decode and predict visual saliency information. Moreover, since the decoder learned during the unsupervised discrete reconstruction process can well recover most of image information, its knowledge can be easily transferred to predict visual saliency maps and learn the saliency relationship from these visual tokens. Specfically, in transfer learning, we freeze the feature extraction and quantization parts, and only finetune the decoder part to perform saliency prediction as shown in Figure 6 (b). Our proposed VQSal model can be represented as:

$$\hat{x}_{\text{sal}} = G_{\text{Sal}}(z_{\mathbf{q}}) = G_{\text{Sal}}\left(\mathbf{q}(E(x))\right), \tag{6}$$

where G_{Sal} is the decoder for saliency density prediction, and E, $\mathbf{q}$ are frozen encoder and quantization networks, respectively. The loss function of the saliency prediction in our paper is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda\mathcal{L}_{\text{sal}}, \tag{7}$$

where $\mathcal{L}_{\text{sal}} = \mathcal{L}_{\text{CC}} + \mathcal{L}_{\text{KL}}$, CC and KL are two widely used metrics for measuring the accuracy of the predicted saliency maps [9]. The weighting factor λ is empirically set as 0.2 in this paper.

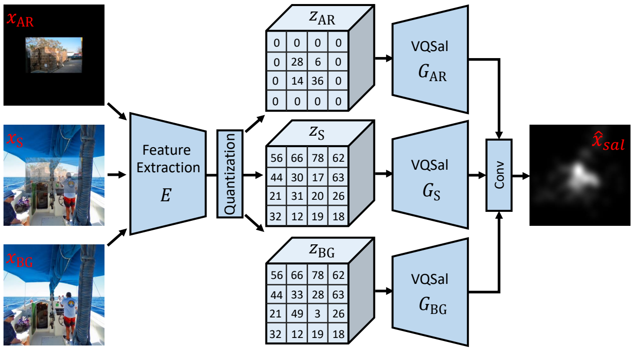


Figure 7: Overview of the proposed VQSal-AR model. Similar to the VQSal model, we first extract the features for three images (AR image, BG image, and superimposed image), and then quantize them to get the corresponding representation codes. These codes are decoded and integrated to predict AR saliency.

Table 1: Quantitative comparison results of different models on the Salicon [33] dataset. We bold the best result for each metric.

Model \ Metric	AUC ↑	CC ↑	IG ↑	KL ↓	NSS ↑	SIM ↑
SALICON [31]	0.824	0.716	34.77	5.723	1.404	0.637
ML-Net [10]	0.809	0.675	34.79	5.707	1.509	0.591
SAM-VGG [11]	0.856	0.836	34.98	5.571	1.802	0.738
SAM-ResNet [11]	0.855	0.844	35.04	5.535	1.809	0.738
SalGAN [46]	0.858	0.848	35.16	5.449	1.763	0.743
GazeGAN [9]	0.853	0.826	34.71	5.760	1.736	0.732
VQSal (Ours)	**0.863**	**0.869**	**35.18**	**5.434**	**1.863**	**0.766**

4.3 VQSal for AR Saliency Prediction

In AR saliency prediction task, three types of images including the AR image, BG image, and superimposed image can be obtained or calculated from a scenario (see **Sec. 3.1 & 3.2**), and all these three components significantly influence visual attention in AR (see **Sec. 3.3**, the superimposed image contains mixing value information).

Table 2: Quantitative comparisons between our proposed VQSal-AR model and three types of benchmark methods. We bold the best result and underline the second-best result for each metric in each type. The best result for each metric throughout the table is colored in red and the second-best result for each metric throughout the table is colored in blue.

Type	Type I					Type II					Type III				
Model \ Metric	AUC ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑	AUC ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑	AUC ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
IT [32]	0.541	0.509	0.075	0.272	0.105	0.628	0.528	0.262	0.859	0.238	0.622	0.533	0.308	1.004	0.369
AIM [2]	0.701	0.557	0.237	0.689	0.371	0.826	0.559	0.534	1.441	0.468	0.838	0.551	0.572	1.531	0.467
GBVS [27]	0.766	0.510	0.332	0.848	0.424	0.850	0.547	0.601	1.623	0.515	0.853	0.550	0.595	1.575	0.510
SR [29]	0.637	0.513	0.120	0.359	0.329	0.791	0.550	0.458	1.279	0.456	0.811	0.552	0.522	1.443	0.470
SUN [62]	0.679	0.534	0.188	0.640	0.352	0.806	**0.568**	0.491	1.462	0.445	0.827	**0.569**	0.561	1.635	0.465
PFT [26]	0.634	0.512	0.107	0.324	0.327	0.796	0.551	0.478	1.342	0.457	0.820	0.551	0.569	1.587	0.484
SMVJ [8]	0.803	0.529	0.439	1.123	0.457	0.853	0.549	0.611	1.626	0.520	0.853	0.549	0.584	1.523	0.503
Judd [35]	0.863	0.531	0.541	1.415	0.416	0.878	0.536	0.637	1.680	0.447	**0.881**	0.529	0.601	1.563	0.430
SWD [19]	0.853	0.540	0.566	1.483	0.466	**0.880**	0.554	0.722	1.946	0.528	0.880	0.545	0.693	1.830	0.507
Murray [44]	0.650	**0.559**	0.181	0.520	0.356	0.829	0.549	0.521	1.364	0.443	0.847	0.544	0.573	1.485	0.454
CA [23]	0.707	0.533	0.231	0.674	0.378	0.822	0.564	0.528	1.488	0.486	0.832	0.563	0.558	1.558	0.488
CovSal [21]	**0.864**	0.520	**0.695**	**1.859**	**0.608**	0.872	0.538	**0.768**	**2.143**	**0.649**	0.872	0.532	**0.751**	**2.035**	**0.641**
HFT [39]	0.765	0.522	0.347	0.937	0.429	0.857	0.554	0.646	1.788	0.551	0.860	0.559	0.654	1.795	0.548
SALICON [31]	0.828	0.567	0.520	1.364	0.499	0.867	0.563	0.655	1.725	0.557	0.871	0.567	0.651	1.706	0.549
ML-Net [10]	0.853	**0.579**	0.655	2.141	0.569	0.856	0.586	0.642	1.943	0.506	0.877	0.590	0.728	2.259	0.561
SAM-VGG [11]	0.892	0.555	0.846	2.412	0.716	0.881	0.580	0.747	2.102	0.635	0.897	0.577	0.840	2.362	0.690
SalGAN [46]	0.887	0.573	0.765	2.111	0.662	0.881	0.580	**0.751**	2.068	0.636	0.891	0.585	0.782	2.150	0.658
GazeGAN [9]	0.887	0.548	0.809	2.219	0.661	0.882	0.576	0.741	2.058	0.631	0.894	0.571	0.804	2.214	0.669
VQSal (Ours)	**0.899**	0.576	**0.873**	**2.606**	**0.741**	**0.885**	**0.587**	0.747	**2.120**	**0.644**	0.900	**0.597**	0.840	2.419	0.697
VQSal-AR (Ours)	-	-	-	-	-	-	-	-	-	-	**0.903**	0.591	**0.893**	**2.687**	**0.758**

Therefore, a multi-decoder fusion network based on the VQSal model is further proposed to integrate AR image information, BG image information, and superimposed image information for AR saliency prediction. Figure 7 shows the overview of this VQSal-AR model. Specifically, three images are first fed into the feature extraction and quantization modules to get visual tokens for them, respectively:

$$\{z_{\mathrm{AR}}, z_{\mathrm{BG}}, z_{\mathrm{S}}\} = \{\mathbf{q}(E(x_{\mathrm{AR}})), \mathbf{q}(E(x_{\mathrm{BG}})), \mathbf{q}(E(x_{\mathrm{S}}))\}, \qquad (8)$$

where $x_{\mathrm{AR}}, x_{\mathrm{BG}}, x_{\mathrm{S}}$ are input images, and $z_{\mathrm{AR}}, z_{\mathrm{BG}}, z_{\mathrm{S}}$ are obtained visual tokens. These visual tokens are then fed into multi-decoders and finally integrated to produce the salieny map as follows:

$$\hat{x}_{\mathrm{sal}} = \mathcal{F}(G_{\mathrm{AR}}(z_{\mathrm{AR}}), G_{\mathrm{BG}}(z_{\mathrm{BG}}), G_{\mathrm{S}}(z_{\mathrm{S}})), \qquad (9)$$

where $G_{\mathrm{AR}}, G_{\mathrm{BG}}, G_{\mathrm{S}}$ are three decoders, $\mathcal{F}$ is the final convolution integration, $\hat{x}_{\mathrm{sal}}$ is the predicted AR saliency map.

5 EXPERIMENTS

5.1 Benchmark Methodology

Although AR images, BG images, and mixing values jointly and significantly influence the visual attention in AR (see Sec. 3.3), whether saliency models should consider all these parameters as input and how to calculate AR saliency accordingly still need to be discussed. Given an AR image I_{AR}, a BG image I_{BG}, and a mixing value α, the superimposed perceptual viewport image I_{S} can be calculated via Eq. (1). To get the AR saliency map $\hat{s}$, three benchmark methods for a given saliency model $\mathcal{S}$ are defined as:

(i) Type I: only using I_{S}, which is formulated as:

$$\hat{s} = \mathcal{S}(I_{\mathrm{S}}). \qquad (10)$$

(ii) Type II: using I_{AR}, I_{BG}, and α, which is formulated as:

$$\hat{s} = \alpha\mathcal{S}(I_{\mathrm{AR}}) + (1-\alpha)\mathcal{S}(I_{\mathrm{BG}}). \qquad (11)$$

(iii) Type III: using I_{AR}, I_{BG}, and I_{S}, which is formulated as:

$$\hat{s} = \mathrm{SVR}(\mathcal{S}(I_{\mathrm{AR}}), \mathcal{S}(I_{\mathrm{BG}}), \mathcal{S}(I_{\mathrm{S}})). \qquad (12)$$

For classical saliency models, they are directly calculated on the corresponding images. For DNN models, they are retrained on SALICON [33] first. Then for I_{AR} and I_{BG}, these DNN models are directly calculated on these images using pretrained weights to get $\mathcal{S}(I_{\mathrm{AR}})$ and $\mathcal{S}(I_{\mathrm{BG}})$, and for I_{S}, they are finetuned and calculated on our dataset using the protocol in Sec. 5.2 to get $\mathcal{S}(I_{\mathrm{S}})$.

5.2 Experimental Results & Analysis

Experiments on SALICON [33]. We first conduct experiments on SALICON to validate the effectiveness of our VQSal model. SALICON [33] is currently the largest saliency dataset with 10000, 5000, 5000 images for training, validation, and test, respectively. For fair comparison, six state-of-the-art saliency models including SALICON [31], ML-Net [10], SAM-VGG [11], SAM-ResNet [11], SalGAN [46], GazeGAN [9] are retrained on SALICON training set, and tested on SALICON validation set. Six widely used metrics including AUC, CC, IG, KL, NSS, SIM [5] are used to compare the performance of these six models with our proposed VQSal model. Table 1 demonstrates that our VQSal model achieves state-of-the-art performance compared to other models among all six metrics.

Experiments on our SARD. We further conduct experiments on our SARD to validate the effectiveness and superiority of our VQSal and VQSal-AR models on the AR saliency prediction task. The benchmark study is first conducted among 13 classical saliency models including IT [32], AIM [2], GBVS [27], SR [29], SUN [62], PFT [26], SMVJ [8], Judd [35], SWD [19], Murray [44], CA [23], CovSal [21], HFT [39], and 6 DNN saliency models including SALICON [31], ML-Net [10], SAM-VGG [11], SalGAN [46], GazeGAN [9], as well as our VQSal model. Five widely used metrics including

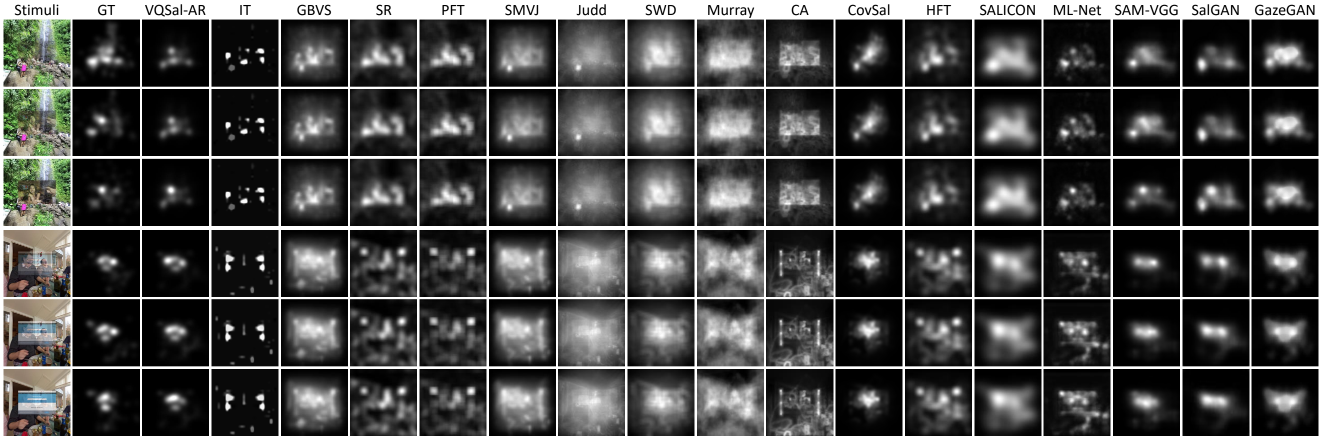


Figure 8: Qualitative comparisons for different models on our SARD.

AUC, sAUC, CC, NSS, SIM [5] are used to compare the performance of these baseline models with our proposed VQSal-AR model. For learning-based methods (*i.e.*, SVR and DNN models), we divide the SARD into 5 splits with an equal number of three stimulus categories in each split and without scenario repeat. Then we run a 5-folds cross validation experiment with 4 splits for training and 1 split for testing in each validation fold. This 5-folds experiment can cover the whole dataset and get the prediction results for all images. Moreover, all DNN models are pretrained on SALICON [33] first and then finetuned on our SARD.

Table 2 demonstrates the quantitative results of three types of benchmark methods and our VQSal-AR model. For classical models, Type II and Type III models generally perform better than Type I models. Comparing the Type II and Type III for classical models, for certain models, Type III method performs better, while for other models, Type II method is more efficient. For DNN models, Type I and Type III models generally perform better than Type II models. The reason may be that the Type I method is finetuned on our dataset and the Type III method includes the Type I saliency map as one input feature. Comparing the Type I and Type III for deep models, we find that for most saliency models, Type III performs better in terms of AUC and sAUC, while for other three evaluation metrics, both methods have their advantages. Moreover, our VQSal models achieves state-of-the-art performance for all three types of benchmark methods, and our VQSal-AR model performs much better compared to all other methods in terms of almost all metrics. Figure 8 also demonstrates the superiority of our VQSal-AR model.

5.3 Ablation Analysis

Ablation for VQSal. We first conduct an ablation study for the VQSal model and demonstrate the results as follows.

Model \ Metric	AUC ↑	CC ↑	IG ↑	KL ↓	NSS ↑	SIM ↑
w/o pretraining	0.792	0.606	34.16	6.147	1.155	0.594
w/o $\mathcal{L}_{\text{rec}}$	0.837	0.744	34.46	5.935	1.432	0.699
w/o $\mathcal{L}_{\text{sal}}$	0.854	0.854	34.28	6.067	1.839	0.744
VQSal (all combined)	**0.863**	**0.869**	**35.18**	**5.434**	**1.863**	**0.766**

We first observe the significant performance drop without the vector quantized pretraining on ImageNet, which manifests that learning a perceptually rich codebook is important for transfer learning on saliency prediction to work. Moreover, we also see that both of the $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{sal}}$ losses have significant improvement for our VQSal.

Ablation for VQSal-AR. First of all, comparing the results of the Type I method for VQSal and the VQSal-AR as shown in Table 2, we observe that our VQSal-AR strategy can significantly improve the performance on the AR saliency prediction task. We further show the effect of the SALICON dataset pretraining on our VQSal-AR model as follows (note that the pretrained weights of all three decoders in VQSal-AR are obtained from the pretrained decoder in VQSal, which is pretrained on SALICON).

Model \ Metric	AUC ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
w/o pretraining on SALICON	0.895	0.583	0.853	2.587	0.719
VQSal-AR	**0.903**	**0.591**	**0.893**	**2.687**	**0.758**

We see a significant performance drop for our VQSal-AR model if without pretraining on SALICON, thus as mentioned in Sec. 5.2, all DNN models are pretrained on SALICON and then finetuned on our dataset to get the results in Table 2 for fair comparison.

6 CONCLUSION

Visual attention analysis and prediction are important tasks for multimedia systems. In this paper, we mainly study human visual attention behavior in AR and its related saliency prediction task. We first construct a saliency in AR dataset (SARD), which contains 1350 superimposed images covering 450 AR/BG scenario pairs, and a large-scale eye-tracking experiment among 60 subjects is also conducted. Through qualitative and quantitative analysis, we conclude that visual attention in AR environment is jointly and significantly influenced by the AR contents, BG scenes and mixing values. For better predicting saliency in AR, we propose a general saliency prediction model *VQSal* and generalize it to the model *VQSal-AR* for AR application. Three benchmark methods are proposed and evaluated on our SARD, and our proposed VQSal-AR achieves state-of-the-art performance compared to these methods.

Our work considers the basic image saliency prediction task while real application scenarios are closer to the video saliency detection task due to head movements. Our future works will extend this study and focus on predicting augmented video saliency.

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China 2021YFE0206700, NSFC 61831015, 61901260, 62176159, Natural Science Foundation of Shanghai 21ZR1432200, Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102, and Alibaba Group through Alibaba Innovative Research Program.

REFERENCES

[1] Ali Borji. 2019. Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 43, 2 (2019), 679–700.

[2] Neil Bruce and John Tsotsos. 2005. Saliency based on information maximization. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 18 (2005).

[3] Claus Bundesen, Signe Vangkilde, and Anders Petersen. 2015. Recent developments in a computational theory of visual attention (TVA). *Vision research* 116 (2015), 210–218.

[4] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. 2015. Mit saliency benchmark. (2015).

[5] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41, 3 (2018), 740–757.

[6] Ozan Cakmakci and Jannick Rolland. 2006. Head-worn displays: a review. *Journal of display technology* 2, 3 (2006), 199–216.

[7] Julie Carmigniani, Borko Furht, Marco Anisetti, Paolo Ceravolo, Ernesto Damiani, and Misa Ivkovic. 2011. Augmented reality technologies, systems and applications. *Multimedia tools and applications* 51, 1 (2011), 341–377.

[8] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. 2007. Predicting human gaze using low-level saliency combined with face detection. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 20 (2007).

[9] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. 2020. How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing (TIP)* 29 (2020), 2287–2300.

[10] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2016. A deep multi-level network for saliency prediction. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*. 3488–3493.

[11] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing (TIP)* 27, 10 (2018), 5142–5154.

[12] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. A dataset of head and eye movements for 360 videos. In *Proceedings of the ACM Multimedia Systems Conference*. 432–437.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[14] Huiyu Duan, Xiongkuo Min, Wei Shen, and Guangtao Zhai. 2022. A Unified Two-Stage Model for Separating Superimposed Images. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2065–2069.

[15] Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, and Patrick Le Callet. 2022. Confusing Image Quality Assessment: Towards Better Augmented Reality Experience. arXiv:2204.04900

[16] Huiyu Duan, Wei Shen, Xiongkuo Min, Yuan Tian, Jae-Hyun Jung, Xiaokang Yang, and Guangtao Zhai. 2022. Develop then Rival: A Human Vision-Inspired Framework for Superimposed Image Decomposition. *IEEE Transactions on Multimedia (TMM)* (2022).

[17] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yucheng Zhu, Yi Fang, and Xiaokang Yang. 2018. Perceptual Quality Assessment of Omnidirectional Images. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*. 1–5.

[18] Huiyu Duan, Guangtao Zhai, Xiaokang Yang, Duo Li, and Wenhan Zhu. 2017. IVQAD 2017: An immersive video quality assessment database. In *Proceedings of the IEEE International Conference on Systems, Signals and Image Processing (IWSSIP)*. 1–5.

[19] Lijuan Duan, Chunpeng Wu, Jun Miao, Laiyun Qing, and Yu Fu. 2011. Visual saliency detection by spatially weighted dissimilarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 473–480.

[20] EPSON. 2022. *EPSON AR*. https://epson.com/moverio-augmented-reality.

[21] Erkut Erdem and Aykut Erdem. 2013. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision* 13, 4 (2013), 11–11.

[22] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 12873–12883.

[23] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. 2011. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 34, 10 (2011), 1915–1926.

[24] Google. 2022. *Google Glass*. https://www.google.com/glass.

[25] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2021. Ego4d: Around the world in 3,000 hours of egocentric video. arXiv:2110.07058

[26] Chenlei Guo, Qi Ma, and Liming Zhang. 2008. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–8.

[27] Jonathan Harel, Christof Koch, and Pietro Perona. 2006. Graph-based visual saliency. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 19 (2006).

[28] Amanda J Haskins, Jeff Mentch, Thomas L Botch, and Caroline E Robertson. 2020. Active vision in immersive, 360 real-world environments. *Nature Scientific Reports* 10, 1 (2020), 1–11.

[29] Xiaodi Hou and Liqing Zhang. 2007. Saliency detection: A spectral residual approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–8.

[30] HTC. 2022. *HTC VIVE Pro Eye*. https://www.vive.com/us/product/vive-pro-eye/overview/.

[31] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 262–270.

[32] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 20, 11 (1998), 1254–1259.

[33] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1072–1080.

[34] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 694–711.

[35] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2106–2113.

[36] Ernst Kruijff, J Edward Swan, and Steven Feiner. 2010. Perceptual issues in augmented reality revisited. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*. 3–12.

[37] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. 2017. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing (TIP)* 26, 9 (2017), 4446–4456.

[38] Magic Leap. 2022. *Magic Leap*. https://www.magicleap.com.

[39] Jian Li, Martin D Levine, Xiangjing An, Xin Xu, and Hangen He. 2012. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35, 4 (2012), 996–1010.

[40] Ming Liang and Xiaolin Hu. 2015. Predicting eye fixations with higher-level visual features. *IEEE Transactions on Image Processing (TIP)* 24, 3 (2015), 1178–1189.

[41] Arien Mack. 2003. Inattentional blindness: Looking without seeing. *Current directions in psychological science* 12, 5 (2003), 180–184.

[42] René Marois and Jason Ivanoff. 2005. Capacity limits of information processing in the brain. *Trends in cognitive sciences* 9, 6 (2005), 296–305.

[43] Microsof. 2022. *HoloLens*. https://www.microsoft.com/hololens.

[44] Naila Murray, Maria Vanrell, Xavier Otazu, and C Alejandro Parraga. 2011. Saliency estimation using a non-parametric low-level vision model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 433–440.

[45] Robert P O'Shea, Amanda Parker, David La Rooy, and David Alais. 2009. Monocular rivalry exhibits three hallmarks of binocular rivalry: Evidence for common processes. *Vision research* 49, 7 (2009), 671–681.

[46] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. 2017. Salgan: Visual saliency prediction with generative adversarial networks. arXiv:1701.01081

[47] Eli Peli and Jae-Hyun Jung. 2017. Multiplexing prisms for field expansion. *Optometry and vision science: official publication of the American Academy of Optometry* 94, 8 (2017), 817.

[48] Matthew F Peterson, Jing Lin, Ian Zaun, and Nancy Kanwisher. 2016. Individual differences in face-looking behavior generalize from the lab to the world. *Journal of Vision (JoV)* 16, 7 (2016), 12–12.

[49] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. A dataset of head and eye movements for 360 degree images. In *Proceedings of the ACM on Multimedia Systems Conference*. 205–210.

[50] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. 2022. End-to-End Human-Gaze-Target Detection With Transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2202–2210.

[51] Unity. 2022. *Unity*. https://unity.com/.

[52] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Proceedings of the NeurIPS*.

[53] Eleonora Vig, Michael Dorr, and David Cox. 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2798–2805.

[54] Benjamin Voloh, Marcus Watson, Seth König, and Thilo Womelsdorf. 2019. MAD saccade: statistically robust saccade threshold estimation. (2019).

[55] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2015. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3183–3192.

[56] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. 2016. Saliency detection with recurrent fully convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 825–841.

[57] Yuwei Wang, Yimin Wu, Cheng Chen, Bohan Wu, Shu Ma, Duming Wang, Hongting Li, and Zhen Yang. 2021. Inattentional Blindness in Augmented Reality Head-Up Display-Assisted Driving. *International Journal of Human–Computer Interaction* (2021), 1–14.

[58] Sam V Wass, Tim J Smith, and Mark H Johnson. 2013. Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods* 45, 1 (2013), 229–250.

[59] Russell L Woods, Robert G Giorgi, Eliot L Berson, and Eli Peli. 2010. Extended wearing trial of Trifield lens device for 'tunnel vision'. *Ophthalmic and physiological optics* 30, 3 (2010), 240–252.

[60] Mai Xu, Li Yang, Xiaoming Tao, Yiping Duan, and Zulin Wang. 2021. Saliency prediction on omnidirectional image with generative adversarial imitation learning. *IEEE Transactions on Image Processing (TIP)* 30 (2021), 2087–2102.

[61] Tao Zhan, Kun Yin, Jianghao Xiong, Ziqian He, and Shin-Tson Wu. 2020. Augmented reality and virtual reality displays: Perspectives and challenges. *Iscience* (2020), 101397.

[62] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. 2008. SUN: A Bayesian framework for saliency using natural statistics. *Journal of vision* 8, 7 (2008), 32–32.

[63] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. 2017. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4372–4381.

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.

[65] Xue Zhang, Gene Cheung, Yao Zhao, Patrick Le Callet, Chunyu Lin, and Jack ZG Tan. 2021. Graph learning based head movement prediction for interactive 360 video streaming. *IEEE Transactions on Image Processing (TIP)* 30 (2021), 4622–4636.

[66] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. 2018. Saliency detection in 360 videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 488–503.

[67] Yucheng Zhu, Guangtao Zhai, Yiwei Yang, Huiyu Duan, Xiongkuo Min, and Xiaokang Yang. 2021. Viewing behavior supported visual saliency predictor for 360 degree videos. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* (2021).

[68] Yucheng Zhu, Dandan Zhu, Yiwei Yang, Huiyu Duan, Qiangqiang Zhou, Xiongkuo Min, Jiantao Zhou, Guangtao Zhai, and Xiaokang Yang. 2019. A saliency dataset of head and eye movements for augmented reality. *arXiv* (2019).