FixationNet: Forecasting Eye Fixations in Task-Oriented Virtual Environments

Zhiming Hu, Andreas Bulling, Sheng Li*, Guoping Wang

Fig. 1: Our model's eye fixation prediction performances in different scenes. The green dot represents the ground truth of eye fixation, the red dot denotes the result of our novel model, FixationNet, and the blue dot refers to the state-of-the-art method [21]. In practice, our model exhibits higher accuracy than the state-of-the-art method.

Abstract—Human visual attention in immersive virtual reality (VR) is key for many important applications, such as content design, gaze-contingent rendering, or gaze-based interaction. However, prior works typically focused on free-viewing conditions that have limited relevance for practical applications. We first collect eye tracking data of 27 participants performing a visual search task in four immersive VR environments. Based on this dataset, we provide a comprehensive analysis of the collected data and reveal correlations between users' eye fixations and other factors, i.e. users' historical gaze positions, task-related objects, saliency information of the VR content, and users' head rotation velocities. Based on this analysis, we propose *FixationNet* – a novel learning-based model to forecast users' eye fixations in the near future in VR. We evaluate the performance of our model for free-viewing and task-oriented settings and show that it outperforms the state of the art by a large margin of 19.8% (from a mean error of 2.93° to 2.35°) in free-viewing and of 15.1% (from 2.05° to 1.74°) in task-oriented situations. As such, our work provides new insights into task-oriented attention in virtual environments and guides future work on this important topic in VR research.

Index Terms—Fixation forecasting, task-oriented attention, visual search, convolutional neural network, deep learning, virtual reality

1 Introduction

Immersive virtual reality (VR) can provide users with higher sense of presence than traditional 2D displays. It gives users a chance to explore a virtual 3D world and has become an important 3D user interface in recent years. Human visual attention in immersive VR is crucial for many important applications, including level-of-detail management [29], VR content design [43], gaze guidance [16], gaze-contingent rendering [34, 46], redirected walking [45], and gaze-based interaction [12, 24, 31, 36]. Consequently, visual attention analysis and prediction has become a popular research topic in VR [19–22, 43, 52]. However, previous works typically focused on free-viewing conditions and few works have studied the more challenging but also more practically relevant task-oriented situations in which users' visual attention is influenced by a specific task [17, 21, 27].

Currently, the most commonly used solution for eye tracking in immersive virtual reality is to employ an eye tracker. However, eye trackers themselves can only provide users' current and historical gaze positions and cannot directly forecast users' gaze positions in the future. Information on users' future eye fixations is valuable for intelligent user interfaces [37] and has significant relevance for a number of areas, including visual attention enhancement [14], pre-computation of gaze-contingent rendering [21, 34], dynamic event triggering [17], as well as human-human and human-computer interaction [33, 44]. An intuitive method of forecasting eye fixations is to only employ users' current gaze. However, current gaze has been proven to be only effective at short time intervals and cannot efficiently encode long-term gaze behavior [20]. Recently, Hu et al. proposed a learning-based model to forecast users' *future* gaze positions [21] but their method is also geared to free-viewing conditions.

To address the limitations of existing methods, in this work we propose the first learning-based model to forecast users' eye fixations in task-oriented virtual environments. We specifically focus on visual search, which is a frequent and important routine behavior in people's everyday life, e.g. when looking for your smartphone, trying to find a friend in a crowd, or searching for food in the fridge. While visual search is an active area of vision research [50], most findings about visual search are derived from 2D viewing conditions. Visual search in immersive virtual reality has not been fully explored.

We first collect eye tracking data of users in a task-oriented virtual environment. Specifically, 27 participants were asked to perform a visual search task in four immersive virtual environments, containing two static scenes and two dynamic scenes. Using this dataset, we analyse users' eye fixations and show that fixations are closely correlated with other factors, i.e. previous gaze positions, task-related objects, saliency information of the VR content, and users' head rotation velocities. Based on our analysis, we then propose *FixationNet* – a novel learning-based method for fixation forecasting in VR that consists of a feature extraction network and a fixation prediction network. We further conduct extensive experiments to evaluate the performance of our model. Our results show that our model significantly outperforms the state-of-the-art method, achieving an improvement of 19.8% (from a mean error of 2.93° to 2.35°) in free-viewing conditions and an improvement of

- *Zhiming Hu, Sheng Li, Guoping Wang are with Peking University, China. E-mail: {jimmyhu | lisheng | wgp}@pku.edu.cn.*
- *Andreas Bulling is with the University of Stuttgart, Germany. E-mail: andreas.bulling@vis.uni-stuttgart.de.*
- *Sheng Li is the corresponding author.*

- *Project URL: https://cranehzm.github.io/FixationNet*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

15.1% (from 2.05° to 1.74°) in task-oriented situations.

The specific contributions of our work are three-fold:

- We provide a novel dataset that contains users' eye tracking data in task-oriented virtual environments, containing both static and dynamic scenes.
- We analyse human visual attention during visual search on this dataset and reveal interesting correlations between users' eye fixations and other factors, such as historical gaze positions and task-related objects.
- We present *FixationNet*, a novel learning-based model for forecasting human eye fixations in task-oriented virtual environments that outperforms the state of the art by a large margin.

2 Related Work

2.1 Computational Modeling of Visual Attention

In the area of vision research, computational modeling of visual attention has been well-studied in the past few decades. Generally, models of visual attention can be classified into bottom-up and top-down models. Bottom-up models predict human visual attention by employing low-level image features such as intensity, contrast, color, and orientation [6, 23]. For example, Itti et al. proposed one of the earliest bottom-up saliency prediction models [23]. This model combines multiscale image features including color, intensity, and orientation to predict saliency maps. Cheng et al. presented a model to detect salient regions based on global contrast [6]. In contrast, top-down models focus on high-level image features such as specific tasks and scene context [13, 35]. Peters et al. incorporated task-dependent influences into a computational model to predict visual attention [35]. Ehinger et al. took scene context into consideration to predict visual attention for a person detection task [13]. Recent advances in machine learning have spurred interest in predicting fixation sequences instead of static saliency maps [8, 28] as well as in forecasting attention e.g. during mobile device interactions [44] or multi-person conversations [33].

In the field of VR, gaze prediction has also been explored in some aspects. Some researchers focused on non-interactive situations such as 360° images and 360° videos. Sitzmann et al. adapted existing saliency predictors to predict saliency maps of 360° images [43]. Xu et al. presented a model that takes 360° video frames as input to predict gaze displacement [52]. This model is specialized for 360° videos and it is not suitable for real-time calculation due to its time cost. In the aspect of interactive virtual environments, Hu et al. proposed an eye-head coordination model to predict gaze positions in static virtual scenes without using any eye trackers [22]. Recently, Hu et al. presented a gaze prediction model called DGaze for dynamic virtual scenes [21]. DGaze can take advantage of users' historical gaze positions provided by an eye tracker to predict users' gaze positions in the future. However, DGaze is derived from free-viewing conditions (no specific task) and its performance will deteriorate when applied to task-oriented situations [21]. In contrast with DGaze, our model concentrates on task-oriented situations and takes task-related data into consideration to forecast users' eye fixations in the future.

2.2 Gaze Prediction in Task-Oriented Situations

Gaze prediction in task-oriented situations has also been explored by many researchers. Borji et al. [4] and Koulieris et al. [27] both focused on gaze prediction in video games. Borji et al. employed players' input such as 2D mouse position and joystick buttons to predict visual attention while Koulieris et al. utilized game state variables to predict users' gaze in video games. Deng et al. proposed a random forest-based model to predict drivers' fixation positions in a driving environment [10]. Zheng et al. presented an end-to-end learning framework to predict visual saliency on webpages under different task conditions such as information browsing and form filling [54]. Xu et al. explored spatio-temporal modeling and prediction of visual attention for a text editing task in 2D graphical user interfaces [51]. Recently, Bâce et al. studied users' visual attention during everyday mobile device interactions [2]. In contrast with previous works, we focus on a visual search task in immersive virtual environments.

2.3 Visual Search

Visual search requires a subject to detect a target among many distractors and it is a typical perceptual task in people's daily life. In vision research, it has been one of the most popular research topics in the past few decades. In their seminal work, Treisman et al. analysed human attention in visual search task and proposed a feature-integration theory of attention [47]. Wolfe proposed a guided search model for visual search task and presented a computer simulation of substantial parts of the model [49]. Vickery et al. focused on the target template set-up process in visual search and concluded that detailed visual information is utilized to find the target [48]. Hollingworth revealed that visual memory guides attention during the process of visual search [18]. Recently, Wolfe et al. discussed some factors that guide attention in visual search, which include bottom-up features, top-down guidance, and the previous history of search [50]. Hadnett-Hunter et al. explored the effect of search task on visual attention in desktop monitor-based virtual environments [17]. A recent line of works has explored predicting and visually decoding the target of visual search from eye fixations [39–41]. However, most of the findings about visual search are derived from 2D viewing conditions and only a few studies focus on 3D environments. Kit et al. evaluated the role of scene memory in guiding eye movements in an immersive virtual environment [26]. Li et al. reported that spatial memory of the scene influences visual search strategies in large-scale environments [30]. In this work, we analyse the characteristics of human visual attention during visual search in immersive virtual reality and utilize that characteristics to forecast human eye fixations.

3 Data Collection

3.1 Stimuli

To collect the gaze data, we used four immersive virtual environments as our stimuli (Fig. 2). The environments contain two outdoor scenes, i.e. a tropical island and a desert, and two indoor scenes, i.e. a warehouse and a gym, that are commonly used in VR research and applications [17, 21, 22, 32]. Considering that human gaze behavior was shown to be different in dynamic and static scenes [1, 15, 21], we ensured to have both a dynamic and static indoor as well as outdoor scenes, respectively. In each static scene, we randomly placed three types of static objects, i.e. chests and footballs, for the visual search task. In each dynamic scene, three types of dynamic objects, i.e. deer and cats, were employed. We controlled the animals' movements using their own animations and navigated their paths using our own Unity script to make the animals wander in the environments in a random manner [21]. In each scene, the three types of objects share some common features (e.g., color, size, and shape) with each other and this obliges a participant to identify targets by performing a serial search using eye fixations [3].

Fig. 2: Four immersive virtual scenes used for data collection, containing two dynamic scenes (top) and two static scenes (bottom).

3.2 Apparatus and Participants

Our data collection experiments were conducted on a platform with an Intel(R) Core(TM) i7-10875H @ 2.30GHz CPU and an NVIDIA GeForce RTX 2060 GPU. HTC Vive was employed to display the scenes and Vive controller was utilized for user interaction. Users' gaze data was collected using a 7invensun VR eye tracker running at 100 Hz and providing an accuracy of 0.5°. Users' head motion was recorded using HTC Vive's Lighthouse tracking system at a sampling rate of 100 Hz. We employed the Unity3D game engine to render the test scenes in real-time and utilized our own Unity scripts to record the information on the task-related objects at a frequency of 100 Hz. The scene content, i.e., the image sequences viewed by the participants were recorded by a screen-recorder at 60 fps. The snapshot of our experimental setup is demonstrated in Fig. 3.

Fig. 3: Experimental setup used in our study.

We recruited 27 participants (15 males and 12 females, aged between 17 and 32 years) to take part in our experiments. All of the users reported normal or corrected-to-normal vision. We calibrated the eye tracker for each user before he or she started the experiment.

3.3 Procedure

Each participant was asked to explore two scenes containing one dynamic scene and one static scene that were randomly chosen from the four scenes. Before experiments in one scene, participants were given at least five minutes to get familiar with the virtual environment and the three types of objects in this scene. In each scene, a user was required to complete three trials that corresponded to three types of targets. Specifically, in one trial, one type of object was utilized as the search target while the other two types of objects were treated as distractors. The total number of targets in one trial is equal to the total number of distractors. In the virtual environment, we placed a target object in front of the user's start location in order to inform the user of the search target in this trial. Users could teleport themselves to any location in their field of view by pointing at the destination using a Vive controller. They could also switch between four preset locations (including the start location) to fully explore the virtual environment. During their exploration in the environment, the participants were required to search for the target objects and, once a target was found, they could cast a ray using the Vive controller to hit the object. If a target was hit, it will disappear. If a distractor was hit, nothing will happen. This feedback mechanism helps the participants remember the search target in the trial. Each trial lasted for about two minutes and the number of objects (targets and distractors) was sufficient for the visual search task. The time that the users had utilized and the number of targets that they had found were displayed near the Vive controller (Fig. 3) to help them become aware of the progress of the search task. The test scenes were silent and the users were provided with a pair of earplugs to avoid auditory disturbance.

During the experiments, we recorded the scene content, i.e. the image sequences viewed by the users, users' head rotation velocities, information on task-related objects, and users' gaze positions (measured in visual angles). Given that this paper focuses on visual search, the targets and distractors in the search task were treated as task-related objects. Information on these objects that we use includes the object's position (its center) in horizontal and vertical direction, its distance from the observer, and a tag indicating whether it is a target or a distractor. For simplicity, only the information on the nearest five task-related objects was recorded.

In total, our dataset contains 27 participants' exploration data in 162 ($27 \times 2 \times 3$) trials. Each trial data contains about 12,000 gaze positions (100 Hz sampling rate), 12,000 task-related object information (100 Hz), 12,000 head velocities (100 Hz), and 7,200 frames of scene screenshots (60 fps). Our dataset is named **FixationNet-dataset** and is available online at https://cranehzm.github.io/FixationNet.

4 Analysis of Eye Fixation

4.1 Fixation Distribution and Fixation-Gaze Correlation

Human eye movements can be classified into two types: fixations (pauses over regions of interest) and saccades (rapid eye movements between fixations). Compared with a fixation, little or no visual processing can be achieved during a saccade [38]. Therefore, in order to analyse users' visual attention, we first extracted users' eye fixations from the raw gaze data. Specifically, we employed a thresholding method based on velocity and duration to detect fixations [38]. We set the threshold velocity for gaze speed to $75^\circ/s$ [21] and required a minimum fixation duration of 200 ms [38]. Consequently, we obtained 1,661,223 fixation positions from the raw gaze data. For clarity, we utilize the term "gaze position" to denote a raw gaze data and employ the term "fixation position" to refer to a gaze data that lies in a fixation period.

To gain a sound understanding of users' fixations, we first analysed the distribution of the fixation positions. The left of Fig. 4 illustrates fixation positions' distribution on the head mounted device's (HMD's) screen, which is smoothed using a Gaussian filter with sigma equal to one degree of visual angle [5]. We can see that most of users' fixation positions lie in the central region of the HMD's screen and this suggests that the information in the central region is more likely to attract users' visual attention than the information in the peripheral region. Moreover, we find that the fixation positions, whose center is $(0.14^\circ, 10.13^\circ)$, exhibit a slight bias towards the upper visual field, as revealed in prior work [17]. This reflects that, during the visual search task, the participants are more likely to pay attention to the VR content in front of them. We further extracted some cluster centers from the fixation positions using a k-means clustering algorithm with k set to 128. We can see from the right of Fig. 4 that, similar to the fixation positions, the cluster centers are mostly located on the screen center and their distribution exhibits an upward bias.

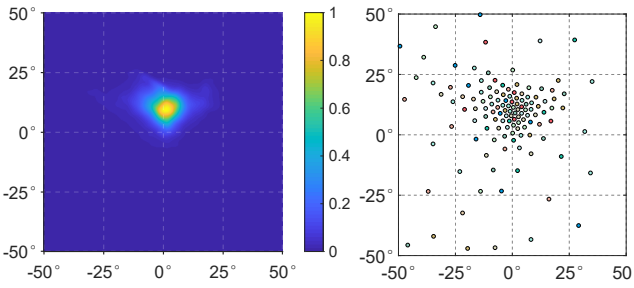

Fig. 4: Left: The distribution of users' fixation positions on the HMD's screen. Right: The cluster centers extracted from the fixation positions. The fixation positions and the cluster centers mostly lie in the central region and their distributions exhibit an upward bias.

We further analysed the temporal characteristics of users' visual attention. Specifically, we calculated the correlation between users' fixation positions and their historical gaze positions using Spearman's rank correlation coefficient. Spearman's correlation measures the monotonic relationship between two variables and outputs a value between -1 (perfect monotone decreasing relationship) and $+1$ (perfect monotone increasing relationship). The results are illustrated in Fig. 5. We can see that the horizontal and vertical correlations are very strong (> 0.9) when the time interval between fixation position and historical gaze position is short (≤ 150 ms). However, with the increase of time interval,

the correlations deteriorate significantly. This is caused by the fact that the duration of a fixation is usually in the range of $200-400$ *ms* [38]. If the time interval is very large (>400 *ms*), users may have changed their fixations and their fixation positions are therefore less correlated with the historical gaze positions. These results indicate that historical gaze positions are more effective in forecasting users' fixation positions in the near future than predicting users' long-term fixations.

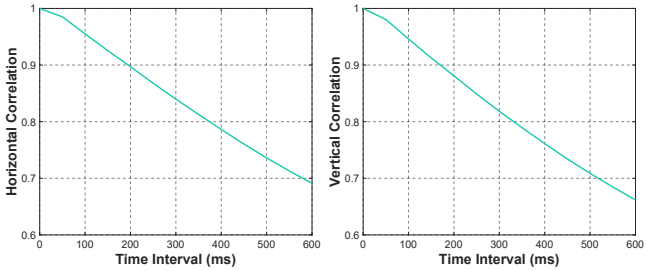

Fig. 5: The correlations between users' fixation positions and historical gaze positions in the horizontal (left) and vertical (right) directions. The fixation positions are highly correlated with historical gaze positions at short time intervals.

4.2 Fixation-Task Correlation

To analyse the correlation between users' fixation positions and the task-related objects, we calculated Spearman's correlations between fixation positions and historical task-related object positions. As illustrated in Fig. 6, in both the horizontal and vertical directions, users' fixation positions are correlated with task-related objects, indicating that task-related objects attract users' visual attention. Fig. 6 also reveals that users' fixation positions have higher correlations with nearer task-related objects. This suggests that users are more likely to pay attention to the task-related objects that are closer to them.

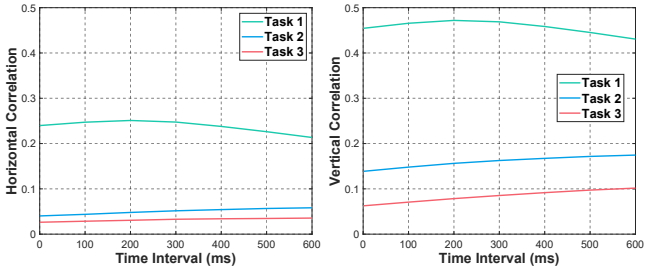

Fig. 6: The horizontal (left) and vertical (right) correlations between fixation positions and historical task-related objects. Tasks 1-3 are the nearest three objects, ranked from nearest to farthest. Users' fixation positions have correlations with task-related objects.

To analyse the differences between targets and distractors' influences on users' visual attention, we calculated the correlations between fixation positions and the nearest task-related object in situations when the nearest object is a target or a distractor, respectively. Fig. 7 illustrates that users' visual attention is more likely to be attracted by the targets than the distractors. Moreover, we find that fixation-target and fixation-distractor correlation curves arrive at their peaks at the time interval of 200 *ms*, which means users' fixation positions lag behind the task-related objects. This indicates users' fixation positions follow the task-related objects, i.e. users' visual attention is directed by the task-related objects [11].

4.3 Fixation-Saliency and Fixation-Head Correlation

The bottom-up saliency information of the scene usually attracts users' visual attention [21, 23, 52]. To analyse the correlation between users' fixation positions and the saliency information of the VR content, we

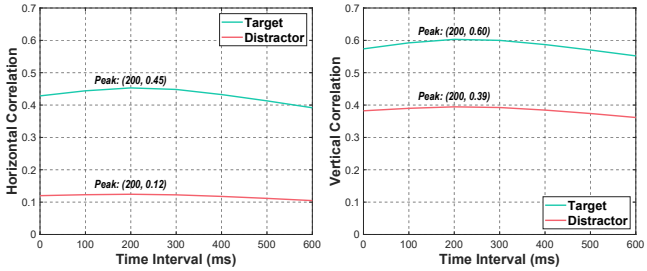

Fig. 7: The horizontal (left) and vertical (right) correlations between fixation positions and the nearest targets and distractors. Fixation positions have higher correlations with the targets than the distractors.

utilized SAM-ResNet [8], which is one of the state-of-the-art saliency predictors, to extract saliency maps from VR images. Specifically, since users' fixation positions are mostly located on the screen center (Fig. 4), we calculated the saliency maps of the central region with a radius of 35°. Based on saliency values, we evenly divided the pixels of a saliency map into 5 salient regions, i.e. salient regions 1-5, ranked from high to low saliency values. We further calculated the distribution of users' fixation positions on the salient regions. As illustrated in Fig. 8, most of the fixation positions lie in the regions with high saliency values. The results indicate that the salient regions of the VR content attract users' visual attention.

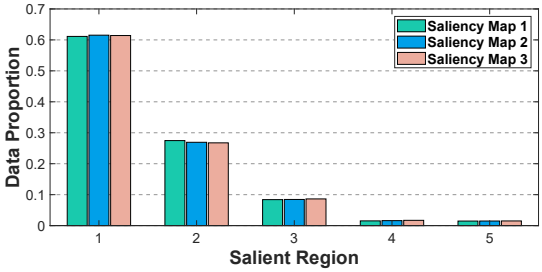

Fig. 8: The distribution of users' fixation positions on the salient regions of the historical VR content. Saliency maps 1-3 correspond to the real-time saliency map, saliency map in the past 200 *ms*, and saliency map in the past 400 *ms*, respectively. The fixation positions are mostly located in the regions with high saliency values.

We also analysed the correlation between users' fixation positions and their head rotation velocities. Spearman's rank correlation coefficient was employed to measure the correlation and the results are illustrated in Fig. 9. We can see that, in both the horizontal and vertical directions, users' historical head velocities are correlated with their fixation positions. This result demonstrates that users' historical head rotation velocities can be applied to forecast their eye fixations.

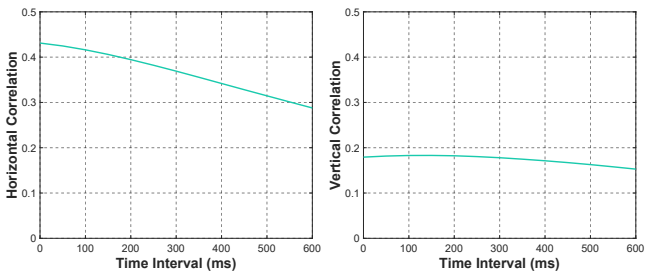

Fig. 9: The correlations between users' fixation positions and their historical head rotation velocities in the horizontal (left) and vertical (right) directions. The fixation positions are correlated with historical head velocities.

5 FixationNet Model

5.1 Problem Formulation

We formulate the problem of forecasting human eye fixations in immersive virtual environments as a regression problem, whose goal is to predict a user's fixation position in the near future given his or her historical gaze positions. A fixation position (f_x, f_y) is denoted by user's visual angles in the horizontal and vertical directions. Human eye movements can be classified into fixations and saccades. We only forecast fixation positions, which refer to gaze data that lies in fixation periods, because little or no visual processing can be achieved during saccade periods [38]. In addition to historical gaze positions, other related factors such as user's head rotation velocities and the VR content can also be employed to facilitate fixation prediction. Since this paper focuses on task-oriented situations, we also take task-related data into consideration when forecasting fixations. In our experiments, we evaluated our model's ability of forecasting users' eye fixations in the future 150 *ms*, 300 *ms*, 450 *ms*, and 600 *ms*.

5.2 Feature Extraction Network

The first component of our model is a feature extraction network, which extracts features from VR images, historical task-related data, historical gaze data, and historical head data for further fixation prediction (Fig. 10).

To take advantage of the VR content viewed by the observers, we employed SAM-ResNet [8], which is one of the state-of-the-art saliency predictors, to calculate the saliency maps of VR images and then utilized a CNN layer to extract saliency features. Since users' fixation positions mostly lie in the central region of the screen (Fig. 4), we only took images of the central region as input. The calculation of saliency maps is very time-consuming. Therefore, to improve computational efficiency, we sampled the VR images every 200 *ms* and only computed saliency maps of the sampled images. For each prediction, we employed saliency maps in the past 400 *ms*, i.e. two saliency maps, down-sampled them to the size of 24×24, and then fed them to a CNN layer. The CNN layer has a kernel size of 1×1, and eight output channels. A batch normalization layer was added after the CNN layer and then ReLU was applied to activate the neurons. After activation, a max-pooling layer with kernel size two was employed, and then a dropout layer with dropout rate 0.5 was applied to improve the network's generalization ability.

As revealed in Sect. 4.2, users' fixation positions are correlated with task-related objects, i.e. targets and distractors in the visual search task. Therefore, in light of the good performance of 1D CNN for processing time series data [7, 21], we utilized two 1D CNN layers, each with kernel size of one, to extract features from historical task-related objects. Specifically, the information on task-related objects in the past 400 *ms* ($\Delta t_1 = 400$ *ms*) were taken as the CNN layers' input and each data point contains the information on the nearest three task-related objects ($T_i \in R^{12}$). The two CNN layers have 64 and 32 output channels, respectively. Each CNN layer was followed by a batch normalization layer, a ReLU activation function, and a max-pooling layer with kernel size two.

To make use of historical gaze data, a 1D CNN layer with kernel size of one and 32 output channels was applied to extract features from users' gaze positions in the past 400 *ms* ($\Delta t_2 = 400$ *ms*, $G_i \in R^2$). A batch normalization layer was added after the CNN layer and ReLU was utilized as the activation function. A max-pooling layer with kernel size two was applied after activation.

Sect. 4.3 reveals that users' head rotation velocities have correlations with their fixation positions. Therefore, we employed a 1D CNN layer, which has a kernel size of one and 64 output channels, to extract features from head velocities in the past 400 *ms* ($\Delta t_3 = 400$ *ms*, $H_i \in R^2$). This CNN layer was followed by a batch normalization layer, a ReLU activation function, and a max-pooling layer with kernel size two.

5.3 Fixation Prediction Network

After the feature extraction network, we employed a fixation prediction network to forecast users' eye fixations based on the extracted features and prior knowledge of the fixation data (Fig. 10).

We first utilized a fully connected (FC) layer with 128 neurons to integrate the extracted features. A batch normalization layer was added after the FC layer, ReLU was applied as the activation function, and a dropout layer with dropout rate 0.5 was employed to improve generalization ability.

Previous work on taxi destination prediction reveals that prior knowledge of the destinations can be used to improve the accuracy of destination prediction models [9]. Therefore, inspired by this finding, we integrated prior knowledge of the fixation data in the architecture of our model. Specifically, to take advantage of prior information on the distribution of users' fixation positions, we extracted cluster centers ($(c_i)_{i=1}^{C}$, $c_i \in R^2$, C is the number of cluster centers) from the fixation data and then predicted fixation position using the weighted centroid of the cluster centers and user's current gaze position:

$$\hat{f} = g_0 + \sum_{i=1}^{C} p_i c_i, \tag{1}$$

where $\hat{f}$ is the predicted fixation position; g_0 is user's current gaze position; c_i is the position of a cluster center and p_i is its corresponding weight with $\sum_{i=1}^{C} p_i = 1$ and $p_i \geq 0$. The weights of the cluster centers ($(p_i)_{i=1}^{C}$) were calculated by a classification layer. The classification layer is a fully connected layer with C number of neurons. It integrates features from the previous layer and utilizes Softmax activation function to generate the weight of each cluster center:

$$p_i = \frac{exp(e_i)}{\sum_{j=1}^{C} exp(e_j)}, \tag{2}$$

where $(e_j)_{j=1}^{C}$ are the output of the classification layer before activation.

In our experiments, we employed a k-means clustering algorithm to extract cluster centers from the corresponding training fixation data. We set $k = 128$ for k-means algorithm in the experiments and obtained 128 cluster centers ($C = 128$) for fixation prediction.

5.4 Loss Function and Training Algorithm

To train our model, we employed the angular error between the ground truth line of sight and the predicted line of sight, i.e. the displacement in visual angles, as our loss function (Angular Loss):

$$L(f, \hat{f}) = d_{angular}(f, \hat{f}), \tag{3}$$

where $L(f, \hat{f})$ is the angular loss; f is the ground truth fixation position; $\hat{f}$ is the predicted fixation position; $d_{angular}(f, \hat{f})$ is the angular distance between f and $\hat{f}$.

We employed Adam with weight decay $5.0e^{-5}$ as our optimizer to minimize the angular loss. We set the initial learning rate to 0.01 and decayed the learning rate by γ every epoch: $lr = lr_0 * \gamma^{epoch-1}$, where lr is the current learning rate; lr_0 is the initial learning rate; γ is the multiplicative factor of learning rate decay; $epoch$ is the current epoch. We set γ to 0.80 and trained the model for 30 epochs in total using a batch size of 512. Our model was implemented using the PyTorch framework. The source code of our model and the pre-trained models are available online at https://cranehzm.github.io/FixationNet.

6 Experiments and Results

We conducted extensive experiments to evaluate the performance of our model. Specifically, we first compared our model with the state-of-the-art method DGaze [21] and two baselines on our dataset using a cross-user evaluation and a cross-scene evaluation. We also evaluated our model's performance in free-viewing conditions. An ablation study was performed to validate the effectiveness of each component in our model.

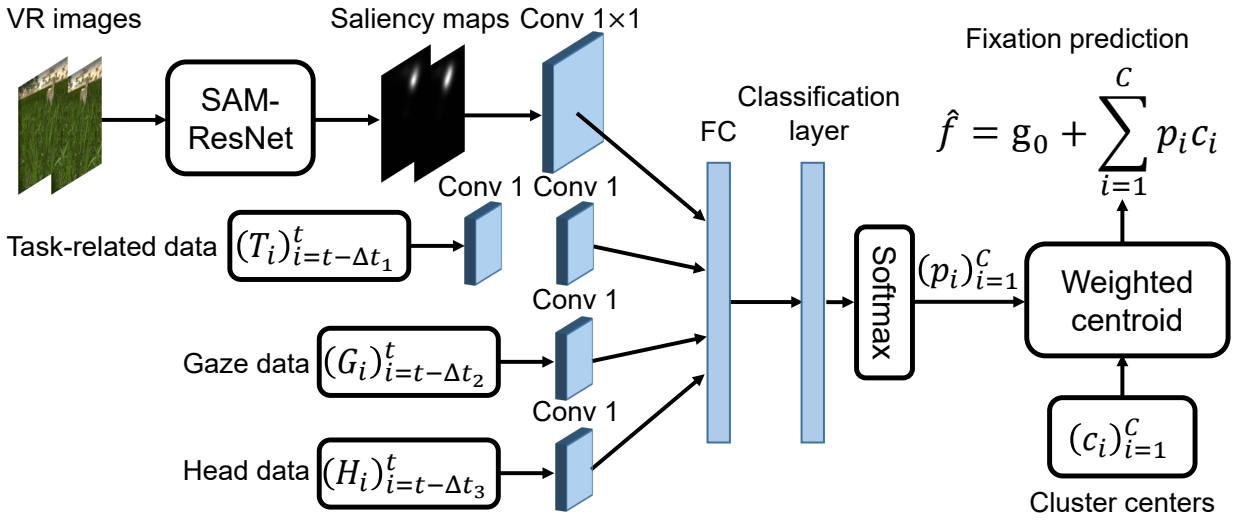

Fig. 10: Architecture of the proposed model **FixationNet**. FixationNet consists of a feature extraction network and a fixation prediction network. The feature extraction network extracts features from VR images, historical task-related data, historical gaze data, and historical head data for further fixation prediction. The fixation prediction network combines the extracted features and the pre-computed cluster centers of the fixations to forecast users' eye fixations.

6.1 Comparison, Baselines, and Evaluation Metric

We compared the performance of our proposed model **FixationNet** with the state-of-the-art approach **DGaze** [21]. We also tested two baselines: one was to employ user's current gaze position (**Current Gaze**) and the other was to utilize the mean of historical gaze positions (**Mean Gaze**). In practice, the mean of gaze positions in the past 50 *ms* was utilized as the mean gaze baseline. To evaluate the performance of fixation prediction, we employed the angular error between the ground truth and the predicted fixation position as our evaluation metric [21,22]. The smaller the angular error, the better the prediction performance.

6.2 Cross-User Evaluation

To evaluate our model's generalization capability for different users, we set users' fixation positions in the future 150 *ms* as our model's prediction target and performed a three-fold cross-user evaluation. Specifically, we evenly divided all the data into three folds according to different users, trained our model on two folds, and tested on the remaining one fold. Our model was trained and tested for three times in total in which each fold was tested once. We collected our model's prediction results in the three tests and calculated the mean and standard deviation (SD) of the prediction errors. To compare our model with the state-of-the-art method DGaze [21], we retrained DGaze on our dataset and evaluated its performance using the three-fold cross-user evaluation. Table 1 presents the cross-user prediction performances of our model and other methods. We can see that our model FixationNet outperforms the state-of-the-art method DGaze by 15.1%, i.e. from a mean error of 2.05° to 1.74°. We performed a paired Wilcoxon signed-rank test and validated that the difference between our model and DGaze is statistically significant ($p < 0.01$). Fig. 1 highlights some of our prediction results. We also calculated the cumulative distribution function (CDF) of the prediction errors for performance comparison. The higher the CDF curve, the better the prediction performance. As illustrated in the left of Fig. 11, our model achieves better performance than DGaze in terms of CDF curve. The above results demonstrate that our model has a good prediction accuracy and a strong generalization capability for different users.

We further tested our model's prediction performances at longer time intervals. Specifically, we respectively set users' fixation positions in the future 300 *ms*, 450 *ms*, and 600 *ms* as our model's prediction target and employed a three-fold cross-user evaluation to calculate our model's prediction results. Instead of using Equation 1 in the fixation

	Ours	DGaze	Current Gaze	Mean Gaze
Mean	**1.74°**	2.05°	2.07°	2.26°
SD	3.61°	3.45°	4.82°	4.69°

Table 1: Our model and other methods' cross-user prediction performances in the future 150 *ms*. Our model achieves an improvement of 15.1% over DGaze [21] in terms of mean prediction error.

prediction network, we utilized

$$\hat{f} = \sum_{i=1}^{C} p_i c_i, \tag{4}$$

because we found that user's current gaze was less effective for forecasting fixations over 300 *ms* (See discussion on the effectiveness of current gaze in Sect. 6.5). DGaze was retrained and tested in the same manner for comparison. The mean prediction errors of our model and other methods are illustrated in the left of Fig. 12. We can see that our model outperforms other methods at different time intervals and the results are statistically significant ($p < 0.01$, paired Wilcoxon signed-rank test). We also find that the performances of all the methods deteriorate significantly with the increase of prediction time. From the prediction time of 150 *ms* to 600 *ms*, the accuracy of our model deteriorates from 1.74° to 4.94°. This is because the durations of users' fixations are often in the range of $200 - 400$ *ms* [38], which means users usually change their fixations after such a time interval. As a consequence, it will be difficult to accurately forecast users' fixation positions in the long-term future (> 400 *ms*) based on only historical features (See Sect. 7 for further discussion).

6.3 Cross-Scene Evaluation

Since our dataset was collected from four different scenes, we also evaluated our model's generalization capability for different scenes. We set users' fixation positions in the future 150 *ms* as our model's prediction target and utilized a four-fold cross-scene evaluation to train and test our model. DGaze was also retrained and tested for comparison. The results are indicated in Table 2. In terms of cross-scene prediction performance, our model achieves an improvement of 11.7% over the state-of-the-art method DGaze (from a mean error of 2.05° to 1.81°) and the result is statistically significant ($p < 0.01$, paired Wilcoxon signed-rank test). We also calculated the CDF curves of the prediction

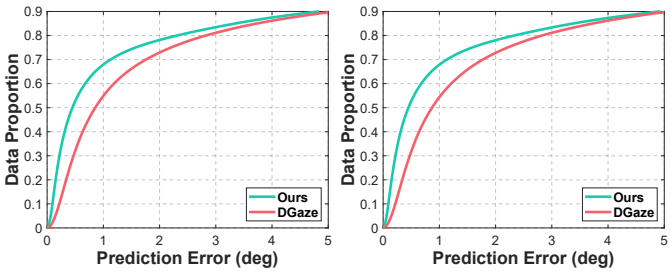

Fig. 11: The cumulative distribution functions of our model and DGaze's cross-user (left) and cross-scene (right) prediction errors in the future 150 *ms*. Our model outperforms DGaze in both cross-user performance and cross-scene performance.

errors. The right of Fig. 11 illustrates that our model outperforms DGaze in the aspect of CDF curve. The above results validate that our model has a good prediction accuracy and a strong generalization capability for different scenes.

	Ours	DGaze	Current Gaze	Mean Gaze
Mean	**1.81**°	2.05°	2.07°	2.26°
SD	3.88°	3.44°	4.82°	4.69°

Table 2: Our model and other methods' cross-scene prediction performances in the future 150 *ms*. Our model outperforms DGaze by 11.7% in the aspect of mean prediction error.

6.4 Performance in Free-Viewing Conditions

We further evaluated our model's performance on a newly released free-viewing VR eye tracking dataset called DGaze-dataset [21]. DGaze-dataset records 43 users' eye tracking data in five dynamic virtual environments under free-viewing conditions. It contains the VR content viewed by the observers, the information on the dynamic objects, users' gaze positions, and their head rotation velocities. We first extracted fixation positions from the raw gaze data using the same method as Sect. 4.1 and obtained 1,224,352 fixation positions. Then we retrained our model and DGaze on this dataset to evaluate their prediction performances. To train our model, we treated the information on the dynamic objects as a kind of task-related data and extracted features from them to forecast fixations. Equation 4 was employed in the fixation prediction network. We employed a three-fold cross-user evaluation and a five-fold cross-scene evaluation (this dataset contains five scenes) to evaluate our model and other methods' prediction performances. As illustrated in Table 3, our model performs significantly better than the state-of-the-art method DGaze ($p < 0.01$, paired Wilcoxon signed-rank test) when forecasting fixations in the future 150 *ms*, achieving an improvement of 19.8% (from a mean error of 2.93° to 2.35°) for cross-user evaluation and an improvement of 19.5% (from 2.93° to 2.36°) for cross-scene evaluation. The CDF curves of the prediction errors are illustrated in Fig. 13. In terms of CDF curves, our model performs better than DGaze for both cross-user evaluation and cross-scene evaluation. The above results validate that our model outperforms the state-of-the-art method in free-viewing conditions.

We further evaluated our model's cross-user prediction performances at longer time intervals, i.e. 300 *ms*, 450 *ms*, and 600 *ms*. The right of Fig. 12 illustrates the results. Our model exhibits higher accuracy than other methods in different prediction times and the results are statistically significant ($p < 0.01$, paired Wilcoxon signed-rank test). We also find that, similar to task-oriented situations, all the methods' performances in free-viewing conditions deteriorate dramatically when the prediction time increases.

		Ours	DGaze	Current Gaze	Mean Gaze
Cross-User	Mean	**2.35**°	2.93°	2.99°	3.20°
	SD	3.70°	3.59°	6.65°	6.08°
Cross-Scene	Mean	**2.36**°	2.93°	2.99°	3.20°
	SD	3.72°	3.63°	6.65°	6.08°

Table 3: Our model and other methods' cross-user and cross-scene prediction performances in the future 150 *ms* in free-viewing conditions. Our model outperforms other methods in both cross-user performance and cross-scene performance.

6.5 Ablation Study

We further performed an ablation study to evaluate the effectiveness of each component in our model. Specifically, we respectively removed the saliency maps of the VR images, task-related data, gaze data, head data, and cluster centers and retrained the ablated models. Users' fixation positions in the future 150 *ms* were set as the ablated models' prediction targets and a three-fold cross-user evaluation was employed to calculate the results. Table 4 presents the cross-user performances of the ablated models. We can see that our model outperforms all the ablated models and the results are statistically significant ($p < 0.01$, paired Wilcoxon signed-rank test). This validates that each component in our model helps improve our model's accuracy. In addition, we find that users' historical gaze data plays the most important role in our model. If more features related to users' gaze, such as users' eye images, are provided, our model can be further improved by considering these features (See our discussion in Sect. 7).

	Ours	Saliency	Task	Head	Cluster	Gaze
Mean	**1.74**°	1.77°	1.77°	1.83°	1.83°	2.01°
SD	3.61°	3.74°	3.72°	4.14°	3.49°	4.80°

Table 4: The cross-user prediction performances of our model and the ablated models. Our model exhibits higher accuracy than all the ablated models, meaning that each component contributes to our model's performance.

We also evaluated the effectiveness of user's current gaze in the fixation prediction network (Equation 1). Specifically, we tested our model's three-fold cross-user performances with and without current gaze at different time intervals (Table 5). We find that current gaze is effective at the prediction time of 150 *ms* and it becomes less useful at large time intervals ($\geq$ 300 *ms*). These results are statistically significant ($p < 0.01$, paired Wilcoxon signed-rank test). This is caused by the fact that the correlations between users' current gaze positions and their future fixations deteriorate dramatically with the increase of time interval (Sect. 4.1). In light of this, we recommend to employ current gaze in the fixation prediction network only when current gaze positions are highly correlated (e.g. ≥ 0.9) with the fixations to be predicted. When testing our model in free-viewing conditions, we found that current gaze is not highly correlated (< 0.9) with fixations even at the interval of 150 *ms*. Therefore, we did not utilize current gaze when evaluating our model in free-viewing conditions (Sect. 6.4).

Prediction Time	150 *ms*	300 *ms*	450 *ms*	600 *ms*
Current Gaze	**1.74**°	3.32°	4.26°	5.06°
w/o Current Gaze	1.91°	**3.27**°	**4.20**°	**4.94**°

Table 5: Our model's performances with and without current gaze in different prediction times. Current gaze is effective at short time interval and its effectiveness deteriorates when the prediction time increases.

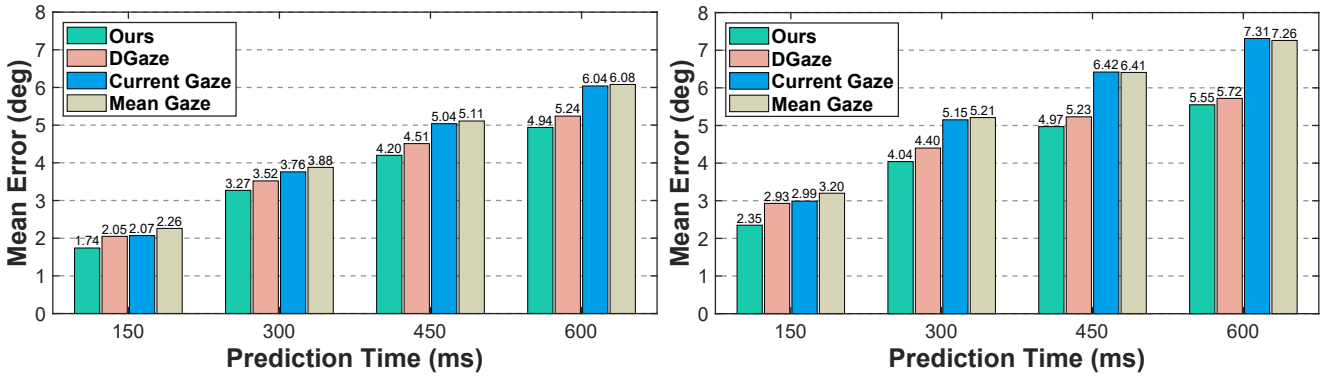

Fig. 12: Our model and other methods' cross-user prediction performances in task-oriented situations (left) and free-viewing conditions (right) at different time intervals. Our model performs better than other methods in different prediction times.

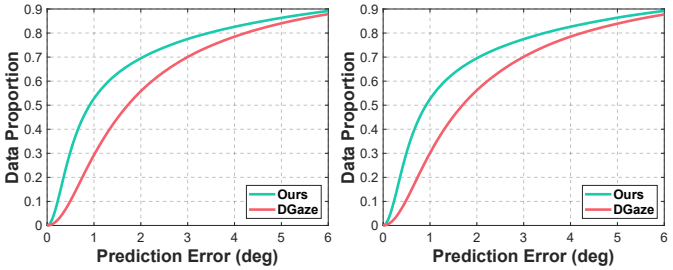

Fig. 13: The CDF curves of our model and DGaze's cross-user (left) and cross-scene (right) prediction errors in the future 150 *ms* in free-viewing conditions. Our model performs better than DGaze in terms of both cross-user performance and cross-scene performance.

6.6 Runtime Performance

We implemented our model on an NVIDIA TITAN Xp GPU platform with an Inter(R) Xeon(R) E5-2620 v4 2.10 GHz CPU and calculated its average prediction time for each fixation position. The average running time of our model is 0.02 *ms* on GPU. If our model runs on CPU, the average time cost is 0.16 *ms*. These results indicate that our model is fast enough for practical usage.

7 Discussion

This work made an important step towards forecasting task-oriented attention in VR, i.e. forecasting human eye fixations in task-oriented virtual environments. Our experimental results reveal some important aspects related to users' task-oriented visual attention.

Long-Term vs. Short-Term Fixation Prediction: From the results in Fig. 12, we find that, in both task-oriented and free-viewing situations, the performances of all the methods deteriorate significantly with the increase of prediction time. This means forecasting users' long-term eye fixations is much more difficult than predicting short-term fixations, as revealed in prior works [20,21]. This is because users usually change their fixations after a short time interval (200 - 400 *ms* [38]) to find a new region of interest. Our model only employs historical features and thus it is less effective for long-term fixation prediction. If some future features related to users' visual attention, such as the future information on task-related objects, are available, our model's long-term forecasting performance can be further boosted by considering the future features.

Task-Oriented vs. Free-Viewing Fixation Prediction: By comparing our model's task-oriented and free-viewing prediction performances (Table 1, Table 2, and Table 3), we find that our model achieves a larger improvement over the state-of-the-art method in free-viewing situations than task-oriented conditions. This is because users' eye movements in task-oriented situations are much more complicated than that in free-viewing conditions [4]. It is therefore more challenging to forecast users' task-oriented attention than to predict free-viewing attention.

Cross-Scene vs. Cross-User Fixation Prediction: From the results in Table 1 and Table 2, we find that our model's cross-user performance is better than its cross-scene performance. This is because users' eye movements in dynamic scenes behave differently from that in static scenes [1, 15, 21]. Our dataset contains both static scenes and dynamic scenes and it is therefore more challenging to achieve good cross-scene performance than to obtain good cross-user performance. When tested in free-viewing conditions (Table 3), our model achieves a good cross-scene performance, which is close to its cross-user performance, because the free-viewing dataset only contains dynamic scenes (Sect. 6.4).

Limitations: We identified several limitations of our model. First, our model was trained on the dataset that was collected during a visual search task. Our model cannot be directly applied to other kinds of tasks, such as text editing or assembly task because different tasks require different task-specific gaze behaviors. To forecast eye fixations in other tasks, our model needs to be retrained using the corresponding task-related data. In addition, in the fixation prediction network, we used pre-computed cluster centers of users' fixations (Sect. 5.3). However, it may be better to set the cluster centers as our model's parameters and learn these parameters in the training process than to utilize the pre-computed cluster centers. Furthermore, the task-related objects (targets and distractors) in our scenes were sparse and the dynamic objects (animals) were set to move at a normal speed (random walk). Our model therefore may not directly handle more cluttered scenes or faster-moving objects. We plan to explore such settings in future work. Finally, in the data collection process, we set the test scenes to silent in order to avoid auditory disturbance. Taking the influence of sound on eye fixations into consideration may further boost the performance of our model.

Future Work: Besides overcoming the above limitations, many potential avenues of future work exist. First, it will be interesting to explore the problem of forecasting users' long-term eye fixations. Existing methods are unable to accurately forecast users' long-term visual attention (Fig. 12) and it is therefore meaningful to derive more accurate long-term fixation prediction models. In addition, there is still some room to improve our model's performance by considering other factors related to users' visual attention. For example, if the information on users' body movements is available, we can extract features from this information and employ the extracted features to facilitate fixation prediction because there exists a coordination between users' eye movements and their body movements [42]. Moreover, in order to provide a general solution for existing eye tracking systems, we only acquired users' historical gaze positions from the eye tracker to forecast fixations and this ensures our model is applicable to any type of eye tracker. However, there exist many kinds of eye tracking technologies such as eye image-based eye trackers [25,53] and corneal reflection-based gaze estimators [55, 56]. For a specific eye tracker, we can also combine

other features provided by the eye tracker, e.g. users' historical eye images, to facilitate eye fixation prediction. Furthermore, our dataset may still be insufficient for training robust fixation forecasting models that are flexible to work for any type of virtual environment. Therefore, to increase the richness of our dataset, we plan to explore other different scenes as well as actual VR applications in the future. Finally, our analysis and our model are restricted to immersive virtual environments. It will be interesting to analyse and forecast users' eye fixations in other systems like augmented reality (AR) and mixed reality (MR) systems. In AR and MR interfaces, the task-related objects may be situated in the real rather than the virtual world, e.g. in a task that requires the user to search for a person in the real world. In such cases, we could employ object detection methods to obtain the location and type of the task-related object, e.g. use face detection algorithm to extract the information on the potential targets from the background in a person search task, and utilize that information to forecast human eye fixations. Our model has the potential to be converted to such systems.

8 Conclusion

In this work, we focused on the problem of forecasting human eye fixations in task-oriented virtual environments. We first presented a gaze dataset of users performing a visual search task in VR and showed that eye fixations are strongly correlated temporally as well as with task-related objects, saliency information of the VR content, and head rotation velocities. Based on these insights, we proposed a novel method to forecast users' fixations in the near future that outperformed the state-of-the-art method in both task-oriented and free-viewing conditions by a large margin. As such, our work represents an important advance and guides future research on task-oriented attention analysis and prediction in immersive virtual environments.

Acknowledgments

We would like to thank our anonymous reviewers for their valuable comments. We would also thank 7invensun for their eye tracking resource. This project was supported by the National Key R&D Program of China (No.2017YFB0203002 and No.2017YFB1002700) and the National Natural Science Foundation of China (No.61632003). A. Bulling was funded by the European Research Council (ERC; grant agreement 801708). Zhiming Hu would thank his grandmother Ms. Feng Wang for her love and care during the COVID-19 epidemic.

References

[1] R. A. Abrams and S. E. Christ. Motion onset captures attention. *Psychological Science*, 14(5):427–432, 2003.

[2] M. Bâce, S. Staal, and A. Bulling. Quantification of users' visual attention during everyday mobile device interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.

[3] M. Bernhard, E. Stavrakis, M. Hecher, and M. Wimmer. Gaze-to-object mapping during visual search in 3d virtual environments. *ACM Transactions on Applied Perception (TAP)*, 11(3):1–17, 2014.

[4] A. Borji, D. N. Sihite, and L. Itti. Probabilistic learning of task-specific visual attention. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 470–477. IEEE, 2012.

[5] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016.

[6] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[7] H. Cho and S. Yoon. Divide and conquer-based 1d cnn human activity recognition using test data sharpening. *Sensors*, 18(4):1055, 2018.

[8] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 2018.

[9] A. De Brébisson, É. Simon, A. Auvolat, P. Vincent, and Y. Bengio. Artificial neural networks applied to taxi destination prediction. *arXiv preprint arXiv:1508.00021*, 2015.

[10] T. Deng, H. Yan, and Y.-J. Li. Learning to boost bottom-up fixation prediction in driving environments via random forest. *IEEE Transactions on Intelligent Transportation Systems*, 19(9):3059–3067, 2017.

[11] T. R. Derrick and J. M. Thomas. Time series analysis: the cross-correlation function. *Innovative Analyses of Human Movement*, 2004.

[12] A. T. Duchowski. Gaze-based interaction: A 30 year retrospective. *Computers & Graphics*, 73:59–69, 2018.

[13] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009.

[14] M. S. El-Nasr, A. Vasilakos, C. Rao, and J. Zupko. Dynamic intelligent lighting for directing visual attention in interactive 3-d scenes. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(2):145–153, 2009.

[15] S. L. Franconeri and D. J. Simons. Moving and looming stimuli capture attention. *Perception & psychophysics*, 65(7):999–1010, 2003.

[16] S. Grogorick, M. Stengel, E. Eisemann, and M. Magnor. Subtle gaze guidance for immersive environments. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 1–7, 2017.

[17] J. Hadnett-Hunter, G. Nicolaou, E. O'Neill, and M. Proulx. The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception (TAP)*, 16(3):1–17, 2019.

[18] A. Hollingworth. Task specificity and the influence of memory on visual search: Comment on võ and wolfe (2012). *J Exp Psychol Hum Percept Perform*, 2012.

[19] Z. Hu. Gaze analysis and prediction in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020.

[20] Z. Hu, S. Li, and M. Gai. Temporal continuity of visual attention for future gaze prediction in immersive virtual reality. *Virtual Reality & Intelligent Hardware*, 2020.

[21] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE transactions on visualization and computer graphics*, 2020.

[22] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha. Sgaze: A data-driven eye-head coordination model for realtime gaze prediction. *IEEE transactions on visualization and computer graphics*, 25(5):2002–2010, 2019.

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

[24] M. Kassner, W. Patera, and A. Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pp. 1151–1160, 2014.

[25] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[26] D. Kit, L. Katz, B. Sullivan, K. Snyder, D. Ballard, and M. Hayhoe. Eye movements, visual search and scene memory, in an immersive virtual environment. *PLoS One*, 9(4), 2014.

[27] G. A. Koulieris, G. Drettakis, D. Cunningham, and K. Mania. Gaze prediction using machine learning for dynamic stereo manipulation in games. In *2016 IEEE Virtual Reality (VR)*, pp. 113–120. IEEE, 2016.

[28] M. Kümmerer, T. S. Wallis, L. A. Gatys, and M. Bethge. Understanding low-and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision*, pp. 4799–4808, 2017.

[29] S. Lee, G. J. Kim, and S. Choi. Real-time tracking of visually attended objects in virtual environments and its application to lod. *IEEE Transactions on Visualization and Computer Graphics*, 15(1):6–19, 2008.

[30] C.-L. Li, M. P. Aivar, M. H. Tong, and M. M. Hayhoe. Memory shapes visual search strategies in large-scale environments. *Scientific reports*, 8(1):1–11, 2018.

[31] D. Mardanbegi, K. Pfeuffer, A. Perzl, B. Mayer, S. Jalaliniya, and H. Gellersen. Eyeseethrough: Unifying tool selection and application in virtual environments. In *The 26th IEEE Conference on Virtual Reality and 3D User Interfaces*, 2019.

[32] C. Mousas, A. Koilias, D. Anastasiou, B. Hekabdar, and C.-N. Anagnostopoulos. Effects of self-avatar and gaze on avoidance movement behavior. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 726–734. IEEE, 2019.

[33] P. Müller, E. Sood, and A. Bulling. Anticipating averted gaze in dyadic interactions. In *Proc. ACM International Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 1–10, 2020. doi: 10.1145/3379155.3391332

[34] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph.*, 35(6):179:1–179:12, Nov. 2016.

[35] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2007.

[36] T. Piumsomboon, G. Lee, R. W. Lindeman, and M. Billinghurst. Exploring natural eye-gaze-based interaction for immersive virtual reality. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 36–39. IEEE, 2017.

[37] D. D. Salvucci and J. R. Anderson. Intelligent gaze-added interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 273–280, 2000.

[38] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pp. 71–78, 2000.

[39] H. Sattar, A. Bulling, and M. Fritz. Predicting the category and attributes of visual search targets using deep gaze pooling. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2740–2748, 2017. doi: 10.1109/ICCVW.2017.322

[40] H. Sattar, M. Fritz, and A. Bulling. Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. *Neurocomputing*, 387:369–382, 2020. doi: 10.1016/j.neucom.2020.01.028

[41] H. Sattar, S. Müller, M. Fritz, and A. Bulling. Prediction of search targets from fixations in open-world settings. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 981–990, 2015. doi: 10.1109/CVPR.2015.7298700

[42] L. Sidenmark and H. Gellersen. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(1):1–40, 2019.

[43] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics (IEEE VR 2018)*, 24(4):1633–1642, 4 2018.

[44] J. Steil, P. Müller, Y. Sugano, and A. Bulling. Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors. In *Proc. ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, pp. 1:1–1:13, 2018. doi: 10.1145/3229434.3229439

[45] Q. Sun, A. Patney, L.-Y. Wei, O. Shapira, J. Lu, P. Asente, S. Zhu, M. McGuire, D. Luebke, and A. Kaufman. Towards virtual reality infinite walking: dynamic saccadic redirection. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[46] N. T. Swafford, J. A. Iglesias-Guitian, C. Koniaris, B. Moon, D. Cosker, and K. Mitchell. User, metric, and computational evaluation of foveated rendering methods. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 7–14. ACM, 2016.

[47] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[48] T. J. Vickery, L.-W. King, and Y. Jiang. Setting up the target template in visual search. *Journal of vision*, 5(1):8–8, 2005.

[49] J. M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

[50] J. M. Wolfe and T. S. Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8, 2017.

[51] P. Xu, Y. Sugano, and A. Bulling. Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3299–3310, 2016.

[52] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze prediction in dynamic 360 immersive videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5333–5342, 2018.

[53] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[54] Q. Zheng, J. Jiao, Y. Cao, and R. W. Lau. Task-driven webpage saliency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 287–302, 2018.

[55] Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 918–923. IEEE, 2005.

[56] Z. Zhu, Q. Ji, and K. P. Bennett. Nonlinear eye gaze mapping function estimation via support vector regression. In *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, pp. 1132–1135. IEEE, 2006.