Gaze Target Detection Based on Head-Local-Global Coordination

Yaokun Yang and Feng Lu*

State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University, Beijing, China
{yangyaokun,lufeng}@buaa.edu.cn

Abstract. This paper introduces a novel approach to gaze target detection leveraging a head-local-global coordination framework. Unlike traditional methods that rely heavily on estimating gaze direction and identifying salient objects in global view images, our method incorporates a FOV-based local view to more accurately predict gaze targets. We also propose a unique global-local position and representation consistency mechanism to integrate the features from head view, local view, and global view, significantly improving prediction accuracy. Through extensive experiments, our approach demonstrates state-of-the-art performance on multiple significant gaze target detection benchmarks, showcasing its scalability and the effectiveness of the local view and view-coordination mechanisms. The method's scalability is further evidenced by enhancing the performance of existing gaze target detection methods within our proposed head-local-global coordination framework.

Keywords: Gaze Target Detection · Head-Local-Global Coordination · Contrastive Learning

1 Introduction

Traditional research in human intention detection has predominantly focused on estimating eye gaze direction [5, 21, 22, 33, 34]. Determining the precise location a person is looking at, referred to as the gaze target, offers a more intuitive approach to delve into deeper human intentions. Previous methodologies often require specialized equipment (*e.g.*, eye trackers, VR/AR devices, or expensive RGB-D cameras) or controlled environments (*e.g.*, restricted subject placements). Recent advancements, however, have showcased the capability to estimate gaze targets from more accessible sources in daily life, namely single images in unconstrained settings, promising greater scalability for the task.

Traditional approaches [2, 6, 7, 10, 16, 19, 23, 25, 27] typically identify salient objects in input images conditioned on estimated gaze directions. Despite significant progress, several critical challenges remain. Most existing methods utilize

* Corresponding Author.
This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62372019.

Fig. 1: Failure cases of existing methods [2, 7, 10, 16, 23]. The red dots denote the predicted gaze target, yellow dots denote the ground truth. While effective in simple scenes, variations in human head position and gaze direction in unconstrained environments limit the efficacy of these approaches. In complex scenes, salient objects may mislead the network, resulting in erroneous predictions.

the estimated gaze direction to compute a person's gaze cone, *i.e.*, a cone-shaped field of view (FOV), emanating from their head position. Each point within this cone is assigned a weight based on gaze direction, and this weighted gaze cone map guides image feature extraction. While effective in simple scenes, variations in human head position and gaze direction in unconstrained environments limit the efficacy of these approaches. In complex scenes, salient objects may mislead the network, resulting in erroneous predictions. Fig. 1 shows typical cases.

Building upon this analysis, traditional approaches employ two perspectives in gaze target detection: extracting facial features from cropped head images (head view) and estimating gaze direction; extracting scene features from original images (global view) and inferring gaze targets. However, we posit the existence of a third valuable perspective. Assuming accurate gaze direction estimation, the image region within the calculated gaze cone theoretically encompasses all task-relevant features, including facial features, head position, gaze target features, and gaze target position. To effectively utilize this information, we propose defining the smallest rectangular area encompassing the person's head position and entire gaze cone as the local view.

Compared to traditional methods, our novel approach integrates a local view to aid neural networks in reasoning about gaze targets. Fig. 2 visually delineates the head view, local view, and global view in gaze target detection. The local view retains all task-relevant features while minimizing task-irrelevant elements (*e.g.*, salient objects outside the gaze cone, complex backgrounds) present in the global view. In scenarios with diverse human head positions and gaze directions, the local view effectively guides neural network attention towards the person's gaze cone, enhancing prediction accuracy and reliability.

This paper presents a novel gaze target detection method based on head-local-global coordination. Our approach learns to infer gaze targets from head view, local view, and global view, coordinating their spatial positions and image representations. Initially, our network extracts facial features from the head view and estimates 3D gaze direction. Subsequently, we derive the local view from the original global view based on head position and calculated gaze cones. Finally, our method reasons about gaze targets from both the extracted local view and original global view through a shared network. Additionally, to align spatial relationships between head view, local view, and global view features in high-

Fig. 2: Illustration of the head view, local view, global view in gaze target detection.

dimensional spaces, we introduce a global-local position consistency mechanism. Furthermore, to encourage learning of high-quality consistent features, we introduce a global-local representation consistency mechanism between local view and global view. Extensive experiments demonstrate that our method achieves SOTA performance on multiple important gaze target detection benchmarks.

It is noteworthy that our method exhibits strong scalability. By integrating the additional local view and view-coordination mechanism, existing gaze target detection methods can be optimized within our head-local-global coordination framework, enhancing model performance.

In summary, our main contributions are as follows:

- We introduce a FOV-based local view explicitly into gaze target detection.
- We propose a novel gaze target detection method based on head-local-global coordination, achieving state-of-the-art performance on multiple significant gaze target detection benchmarks.
- Our head-local-global coordination framework demonstrates strong scalability, capable of enhancing performance in existing methods.

2 Related Work

Gaze Target Detection. Gaze target detection provides an intuitive means to explore human intentions. Recasens *et al*. [25] pioneered this field by introducing the GazeFollow dataset, a substantial collection of images annotated with head positions and corresponding gaze targets. Lian *et al*. [19] augmented viewpoint supervision using a multi-scale field of view attention mechanism. Chong *et al*. [7] extended the task by introducing a video dataset and addressing out-of-frame scenarios. Fang *et al*. [10] incorporated monocular depth estimation as supplementary prior information, while Bao *et al*. [2] leveraged intricate analytical calculations of 3D geometry. Most existing methods primarily rely on searching for potential gaze targets in the global view.

Gaze Estimation. Appearance-based gaze estimation has long been a focal point in computer vision [5, 11, 21, 22, 33, 34]. However, the majority of available gaze estimation datasets [18, 26, 32] are obtained within controlled laboratory environments, involving meticulous configurations of multi-view cameras, 3D positions of human subjects, and designated gaze targets. Consequently, these datasets consist solely of single face images from a limited range of scenes. Zhang *et al*. [31] introduced a high quality multi-face parsing dataset MPSGaze, which contains full images of multiple people with 3D gaze ground truth.

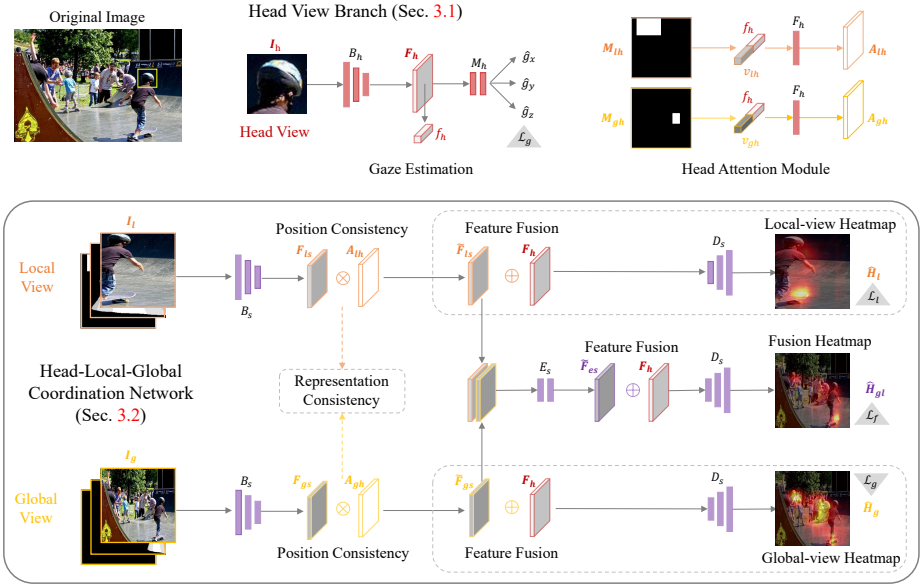

Fig. 3: Overall architecture of our proposed gaze target detection method based on head-local-global coordination framework. The complete method consists of a head view branch (top) and a head-local-global coordination network (bottom). Our head view branch extracts facial features, encodes the 3D gaze direction, extracts the FOV-based local view, and encodes the local-view/global-view head attention maps. Our head-local-global coordination network is designed to detect potential gaze targets from both the extracted local view and original global view. Please note that, the parts within the dashed box are only executed during the training phase.

Contrastive Learning. Contrastive learning has demonstrated remarkable efficacy in self-supervised and semi-supervised learning scenarios. It involves acquiring representations by contrasting multi-views of samples as positive pairs against different negative samples [3, 13]. Another interpretation involves maximizing mutual information between latent representations [1, 15]. He *et al.* [14] extended the utilization of negative samples from mini-batches to a memory bank with substantial momentum updates. Chen *et al.* [4] highlighted the significance of a non-linear projection head and the benefits of a large batch size.

3 Method

The overall architecture of our proposed gaze target detection method is depicted in Fig. 3. It comprises a head view branch and a head-local-global coordination network. The head view branch (Sec. 3.1) extracts facial features from the head view image and encodes the 3D gaze direction. Subsequently, we calculate the 3D field of view (FOV) based on the gaze direction and scene depth information. The local view is then extracted from the global view based on the head position and 3D FOV. We introduce a head attention module to fuse facial features with the

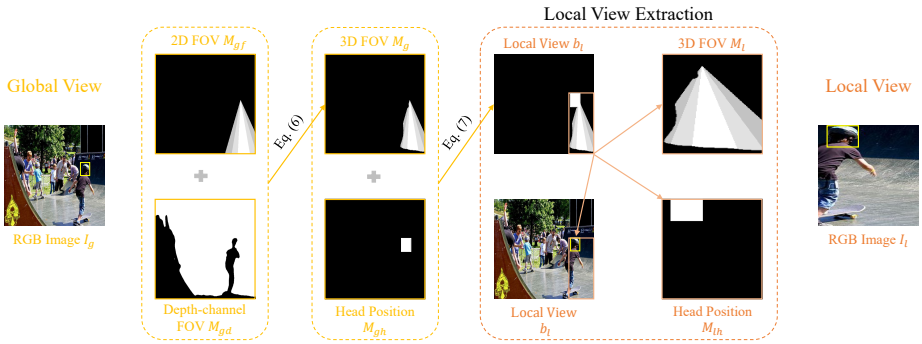

Fig. 4: Local view extracted from the global view.

local-view and global-view head positions, respectively, and encode the resulting head attention maps.

Next, we utilize the head-local-global coordination network (Sec. 3.2) to detect the gaze target from both the extracted local view and original global view. The RGB image, head position map, and 3D FOV map from these two views are concatenated as inputs for the corresponding branches, respectively. Scene features are extracted from both two views through a shared backbone, and then fused with their corresponding head attention maps to maintain position consistency. Furthermore, we introduce a global-local contrastive learning method to enhance representation consistency between these two views. Afterwards, we sequentially integrate the facial features, local-view scene features and global-view scene features together. The fused features are subsequently decoded to predict potential gaze target heatmaps. To enhance supervision for the learning process of local view and global view, we also decode the potential gaze targets from these two views separately in a supervised manner during the training phase.

3.1 Head View Branch

Gaze estimation. The head view branch first extracts a $1024\times7\times7$ dimensional facial feature map $\boldsymbol{F}_h$ from the head view image $\boldsymbol{I}_h$ through a separate feature extractor $\mathcal{B}_h(\cdot)$. It then encodes the 3D gaze direction $[\hat{g}_x, \hat{g}_y, \hat{g}_z]$ using a multi-layer perceptron (MLP) network $\mathcal{M}_h(\cdot)$.

$$\boldsymbol{F}_h = \mathcal{B}_h(\boldsymbol{I}_h), \quad [\hat{g}_x, \hat{g}_y, \hat{g}_z] = \mathcal{M}_h(\boldsymbol{F}_h). \tag{1}$$

Local View Extraction. The gaze cone is calculated based on the global-view head position and estimated 2D gaze direction, generating a weighted 2D FOV map. Specifically, we compute the angle $\varepsilon^{(x,y)}$ between the 2D gaze direction $[\hat{g}_x, \hat{g}_y]$ and the vector $[v_x, v_y]$ from the person's head center (x_c, y_c) to any point (x, y) in the global-view image,

$$[v_x, v_y] = \frac{[x - x_c, y - y_c]}{\sqrt{(x - x_c)^2 + (y - y_c)^2}}, \tag{2}$$

$$\varepsilon^{(x,y)} = arccos([v_x, v_y] \cdot [\hat{g}_x, \hat{g}_y]), \tag{3}$$

Since the gaze target is more credible to be in a neighbour region of the gaze, we assign more weights to points with smaller ε. This will produce the weighted 2D FOV map $\boldsymbol{M}_{gf}^{(x,y)}$ as follows,

$$\boldsymbol{M}_{gf}^{(x,y)} = \max(1 - \alpha \frac{\varepsilon^{(x,y)}}{\pi}, b). \tag{4}$$

Here, $\varepsilon^{(x,y)} \in [0, 180°]$. All points with $\boldsymbol{M}_{gf}^{(x,y)} > b$ form a conical region, which is the gaze cone mentioned above. α decides the half-angle of gaze cone, b is a positive weight offset. We empirically set $\alpha = 3.0$ and achieve a 30° half-angle of gaze cone. We also set $b = 0.5$ as the weight of all points outside the gaze cone, to improve the robustness of the 2D FOV map.

Meanwhile, a depth-channel FOV is obtained. Specifically, we use a well-generalized monocular depth estimation model MiDaS [24] to recover the scene depth information $\boldsymbol{D}$ from the global-view image $\boldsymbol{I}_g$. Subsequently, we produce the depth-channel FOV $\boldsymbol{M}_{gd}^{(x,y)}$ based on the relative depth relationship between the person's head center (x_c, y_c) and any point (x, y) in the global-view image, and the depth-channel gaze direction $\hat{g}_z$,

$$\boldsymbol{M}_{gd}^{(x,y)} = \begin{cases} 1, & if\ (\boldsymbol{D}^{(x,y)} - \boldsymbol{D}^{(x_c,y_c)}) \cdot \hat{g}_z >= 0; \\ 0, & otherwise. \end{cases} \tag{5}$$

Afterwards, we calculate the global-view 3D FOV map $\boldsymbol{M}_g$ based on the weighted 2D FOV map $\boldsymbol{M}_{gf}$ and depth-channel FOV map $\boldsymbol{M}_{gd}$,

$$\boldsymbol{M}_{g}^{(x,y)} = \begin{cases} 1, & if\ \boldsymbol{M}_{gf}^{(x,y)} > b \bigcap \boldsymbol{M}_{gd}^{(x,y)} = 1; \\ 0, & otherwise. \end{cases} \tag{6}$$

We set the smallest rectangular area b_l in the global view based on the global-view head position $\boldsymbol{M}_{gh}$ and 3D FOV map $\boldsymbol{M}_g$, as the local view,

$$b_l = [\min(x), \min(y), \max(x), \max(y)], \quad \boldsymbol{M}_{g}^{(x,y)} = 1 \bigcup \boldsymbol{M}_{gh}^{(x,y)} = 1. \tag{7}$$

Finally, this rectangular area b_l is used to crop the global-view RGB image $\boldsymbol{I}_g$, head position map $\boldsymbol{M}_{gh}$, 3D FOV map $\boldsymbol{M}_g$ to obtain the local-view RGB image $\boldsymbol{I}_l$, head position map $\boldsymbol{M}_{lh}$, 3D FOV map $\boldsymbol{M}_l$, respectively. All these global-view and local-view images are resized to 224×224 dimensions, serving as inputs to our network. The process of local view extraction is illustrated in Fig. 4.

Head Attention Module. We introduce a head attention module to fuse the extracted facial features with the local-view head position map and global-view head position map separately, then encode the local-view head attention map and global-view head attention map, respectively. Specifically, we first perform average pooling on $\boldsymbol{F}_h$ across the channel dimension to obtain a 1024 dimensional embedding vector f_h. Then, the head position maps $\boldsymbol{M}_{lh}$ and $\boldsymbol{M}_{gh}$ are resized to 28×28 dimensions separately and flattened into 784 dimensional head position vectors v_{lh} and v_{gh}, respectively. Subsequently, v_{lh} and v_{gh} are

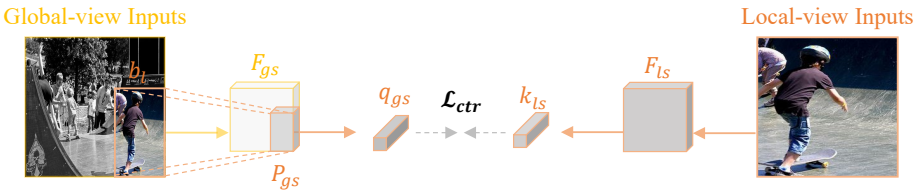

Fig. 5: Global-local contrastive learning method between the local view and global view to enhance their representation consistency. We crop out the feature patches $\boldsymbol{P}_{gs}$ corresponding to the local-view rectangular area b_l from the global-view scene feature maps $\boldsymbol{F}_{gs}$. Subsequently, we set a contrastive learning objective $\mathcal{L}_{ctr}$ between $\boldsymbol{P}_{gs}$ and the local-view scene feature maps $\boldsymbol{F}_{ls}$ to maximize their mutual information.

concatenated with f_h across the channel dimension separately and fed into an FC layer $\mathcal{F}_h(\cdot)$ to encode the $1 \times 7 \times 7$ dimensional local-view head attention map $\boldsymbol{A}_{lh}$ and global-view head attention map $\boldsymbol{A}_{gh}$, respectively.

3.2 Head-Local-Global Coordination Network

Scene Feature Extraction from Local View/Global View. We propose a novel head-local-global coordination network to detect the gaze target from both the local view and global view. For the local view branch, we concatenate the local-view RGB image $\boldsymbol{I}_l$, head position map $\boldsymbol{M}_{lh}$ and 3D FOV map $\boldsymbol{M}_l$ along the channel dimension and feed them into a shared backbone $\mathcal{B}_s(\cdot)$ to extract the $1024 \times 7 \times 7$ dimensional local-view scene features $\boldsymbol{F}_{ls}$. We perform the same operation on the global view branch as well,

$$\boldsymbol{F}_{ls} = \mathcal{B}_s(\boldsymbol{I}_l \oplus \boldsymbol{M}_{lh} \oplus \boldsymbol{M}_l), \quad \boldsymbol{F}_{gs} = \mathcal{B}_s(\boldsymbol{I}_g \oplus \boldsymbol{M}_{gh} \oplus \boldsymbol{M}_g). \tag{8}$$

Global-Local Position Consistency. To maintain position consistency between the local view and global view, we multiply the extracted scene feature maps $\boldsymbol{F}_{ls}$ and $\boldsymbol{F}_{gs}$ with the corresponding head attention maps $\boldsymbol{A}_{lh}$ and $\boldsymbol{A}_{gh}$ along the channel dimension, separately, to obtain the $1024 \times 7 \times 7$ dimensional weighted feature maps $\tilde{\boldsymbol{F}}_{ls}$ and $\tilde{\boldsymbol{F}}_{gs}$, respectively.

Global-Local Representation Consistency. Subsequently, to encourage the shared backbone $\mathcal{B}_s(\cdot)$ to better learn task-consistent image features from both the local view and global view, we introduce a self-supervised contrastive learning method between these two views to enhance their representation consistency. Considering the diversity of human head position and gaze direction in the input image from the original global view, the gaze cone may only occupy a small part of the geometric space of the global-view image. Therefore, the global-view image may contain a large number of task-irrelevant features (*i.e.*, noises), making it difficult for the network to learn high-quality representations in the high-dimensional feature space. Based on the local view extraction method above, we crop out the feature patches $\boldsymbol{P}_{gs}$ corresponding to the local-view rectangular area b_l from the global-view scene feature maps $\boldsymbol{F}_{gs}$. $\boldsymbol{P}_{gs}$ contains all task-consistent features in the high-dimensional feature space extracted from the

global view inputs. Subsequently, we set a contrastive learning objective $\mathcal{L}_{ctr}$ between $\boldsymbol{P}_{gs}$ and the local-view scene feature maps $\boldsymbol{F}_{ls}$ to maximize their mutual information. $\boldsymbol{F}_{ls}$ and $\boldsymbol{P}_{gs}$ are projected into the same high-dimensional embedding space through an averaging pooling operation on the channel dimension, resulting in the query q_{ls} and key k_{gs}, respectively. This process is shown in Fig. 5. The global-local contrastive loss function can be calculated as follows,

$$\mathcal{L}_{ctr} = -log\frac{exp(q_{ls}(k_{gs})^T/\tau)}{\Sigma_{i=1}^{N} exp(q_{ls}(k_{gs}^i)^T/\tau)}, \tag{9}$$

where $i \in \{1, ..., N\}$, denoting all images in the sample batch. τ is a temperature hyper-parameter (set as 0.1 in our experiments). By training with $\mathcal{L}_{ctr}$, the shared backbone $\mathcal{B}_s(\cdot)$ is encouraged to learn high-quality task-consistent representations from both the local view and global view.

Head-Local-Global Feature Fusion. We sequentially fuse the extracted facial features $\boldsymbol{F}_h$, the weighted local-view scene features $\tilde{\boldsymbol{F}}_{ls}$, and the weighted global-view scene features $\tilde{\boldsymbol{F}}_{gs}$ together to decode a potential gaze target heatmap $\hat{\boldsymbol{H}}_{gl}$. First, the $1024 \times 7 \times 7$ dimensional $\tilde{\boldsymbol{F}}_{ls}$ and $1024 \times 7 \times 7$ dimensional $\tilde{\boldsymbol{F}}_{gs}$ are concatenated on the channel dimension. These are then fed into a 1×1 CNN layer $\mathcal{E}_s(\cdot)$ to encode a $1024 \times 7 \times 7$ dimensional feature map $\tilde{\boldsymbol{F}}_{es}$. Next, the encoded scene feature maps $\tilde{\boldsymbol{F}}_{es}$ are concatenated with the facial feature maps $\boldsymbol{F}_h$, and fed into a CNN-based decoder $\mathcal{D}_s(\cdot)$ comprising a 1×1 CNN layer and three D-CNN layers. This decodes a $1 \times 64 \times 64$ dimensional gaze target heatmap $\hat{\boldsymbol{H}}_{gl}$. Additionally, the concatenated feature maps are fed into an MLP network and a softmax classifier to predict a confidence score $\hat{y}_{gl}$ used to identify the out-of-frame target. We set the heatmap regression loss for $\hat{\boldsymbol{H}}_{gl}$ and the classification loss for $\hat{y}_{gl}$ to supervise the learning process of the head-local-global coordination network.

Independent Supervision of Local View/Global View. To further strengthen the supervision of predicting the gaze target from both the local view and global view, we decode potential gaze targets from these views separately and set corresponding heatmap regression losses during training. Specifically, for the local view branch, we concatenate the weighted local-view scene feature maps $\tilde{\boldsymbol{F}}_{ls}$ with the facial feature maps $\boldsymbol{F}_h$. These are fed into the shared decoder $\mathcal{D}_s(\cdot)$ to decode a $1 \times 64 \times 64$ dimensional local-view gaze target heatmap $\hat{\boldsymbol{H}}_l$. Similarly, the same operation is performed on the global view branch to obtain a $1 \times 64 \times 64$ dimensional global-view gaze target heatmap $\hat{\boldsymbol{H}}_g$. We set heatmap regression losses for $\hat{\boldsymbol{H}}_l$ and $\hat{\boldsymbol{H}}_g$, respectively. For the classification task, we no longer introduce additional supervision due to its relative simplicity compared to heatmap regression.

3.3 Overall Loss Function

Our overall learning objective can be formulated as follows:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{ang}([\hat{g}_x, \hat{g}_y, \hat{g}_z], [g_x, g_y, g_z]) + \mathcal{L}_{cls}(\hat{y}_{gl}, y_g) + \mathcal{L}_{reg}(\hat{\boldsymbol{H}}_{gl}, \boldsymbol{H}_g) \\ + \mu\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_l, \boldsymbol{H}_l) + \nu\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_g, \boldsymbol{H}_g) + \lambda\mathcal{L}_{ctr},\end{aligned} \tag{10}$$

where $[g_x, g_y, g_z]$ denotes the ground truth 3D gaze direction. y_g represents the ground truth out-of-frame category. $\boldsymbol{H}_g$ and $\boldsymbol{H}_l$ denote the heatmaps generated from the ground truth gaze target positions from the global view and local view, respectively. $\mathcal{L}_{ang}$ stands for the angle error loss. $\mathcal{L}_{cls}$ denotes the out-of-frame-target classification loss using the softmax loss function. $\mathcal{L}_{reg}$ represents the heatmap regression loss using the mean square error loss function. $\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_{gl}, \boldsymbol{H}_g)$ denotes the heatmap regression loss of head-local-global coordination prediction. $\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_l, \boldsymbol{H}_l)$ denotes the heatmap regression loss of local-view prediction. $\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_g, \boldsymbol{H}_g)$ denotes the heatmap regression loss of global-view prediction. $\mathcal{L}_{ctr}$ represents the global-local contrastive loss function, as expressed in Eq. (9). The hyper-parameters μ, ν, and λ are set as $0.5, 0.5$, and 1.0 respectively in our experiments to balance the different losses.

3.4 Inference Method

In the inference phase, we commence by extracting facial features and estimating 3D gaze direction from the head view image using the head view branch. Subsequently, leveraging the head position and gaze direction, we extract the local view from the original global view. Ultimately, employing the head-local-global coordination network, we detect the gaze target from both the extracted local view and the original global view.

4 Experiments

Datasets. In this study, we utilized the well-established gaze target detection datasets GazeFollow [7] and VideoAttentionTarget [7]. GazeFollow is a comprehensive gaze-tracking dataset, comprising 130,339 individuals across 122,143 images sourced from diverse existing datasets such as ImageNet [8], COCO [20], PASCAL [9], SUN [30], *etc.*. Following dataset partitioning, 4,782 annotated individuals from GazeFollow were allocated for testing purposes, with the remainder serving for training. Notably, each person in the test images underwent 10 human annotations to assess human performance. VideoAttentionTarget extends the task to out-of-frame scenarios. This dataset encompasses 1,331 video clips procured from various sources on YouTube, accompanied by 164,541 frame-level head bounding box annotations. Additionally, we utilized several gaze estimation datasets to pretrain our head view branch, including Gaze360 [18], ETH-XGaze [32], and MPSGaze [31]. MPSGaze is a high-quality multi-face parsing dataset containing full images of multiple people with 3D gaze ground truth.

Evaluation metrics. The evaluation of our method was conducted using the following metrics: L_2 **Distance:** This metric quantifies performance by evaluating the L_2 distance between the predicted gaze target point and the corresponding ground truth annotation. **Angle Error:** We computed the angle error between the predicted gaze direction and the ground truth gaze vector from the face location to the gaze point. **Out-of-frame AP:** The accuracy of identifying out-of-frame instances was assessed through the utilization of average precision (AP). **AUC:** We also employed the area under curve (AUC) criteria proposed by Judd *et al.* [17] to assess the confidence of the predicted heatmap.

Methods	Views	GazeFollow				VideoAttentionTarget		
		AUC↑	Min. L_2↓	Avg. L_2↓	Angle↓	AUC↑	Dist.↓	AP↑
Chong *et al.* [6]	G	0.896	0.112	0.187	-	0.830	0.193	0.705
Lian *et al.* [19]	G	-	0.906	0.145	17.6°	-	-	-
VideoAtt [7]	G	0.921	0.077	0.137	-	0.860	0.134	0.853
Fang *et al.* [10]	G	0.922	0.067	0.124	14.9°	0.905	0.108	0.896
ESCNet [2]	G	0.928	-	0.122	14.6°	0.885	0.120	0.869
Tu *et al.* [29]	G	0.917	0.069	0.133	-	0.904	0.126	0.854
Jin *et al.* [16]	G	0.923	0.064	0.120	14.8°	0.900	0.104	0.895
Miao *et al.* [23]	G	0.934	0.065	0.123	-	0.917	0.109	0.908
Tafasca *et al.* [27]	G	0.936	0.064	0.125	-	0.914	0.109	0.834
Ours-g	G	0.927	0.071	0.129	14.5°	0.908	0.113	0.894
Ours-l	L	0.933	0.065	0.122	13.1°	0.914	0.107	0.905
Ours	**G+L**	**0.939**	**0.059**	**0.114**	**12.4°**	**0.920**	**0.101**	**0.916**
Human		0.924	0.040	0.096	11.0°	0.921	0.051	0.925

Table 1: Quantitative comparisons on GazeFollow [25] and VideoAttentionTarget [7] benchmarks. G: global view. L: local view. Ours-g: our method w/o local view. Ours-l: our method w/o global view. **Ours**: our complete method. The best results are given in red and the second best results are given in blue.

Implementation Details. Our method was implemented using PyTorch. All inputs were resized to 224×224, and all backbones utilized the ResNet-50 network. Due to the lack of annotations for 3D gaze direction in gaze target detection datasets, we pretrained our head view branch on multiple gaze estimation datasets, including Gaze360, ETH-XGaze, and MPSGaze, to improve its generalization ability on gaze target detection datasets. Our monocular depth estimation module utilized the pretrained well-generalized model Midas [12]. Our network was trained on two NVIDIA Titan Xp GPUs, using a mini-batch size of 32 and an initial learning rate of 0.0001. We employed SGD as our optimizer, with a weight decay of 0.0001 and a momentum factor of 0.9. The entire training phase spanned 70 epochs on the GazeFollow dataset and was completed within approximately 18 hours, with the learning rate being scaled down by a factor of 0.1 at both the 50th and 60th epochs. Our complete method achieved an image inference time of approximately 22ms on a single NVIDIA Titan Xp GPU.

4.1 Comparison with Existing Methods

We compared our method with all recent gaze target detection methods, including Chong *et al.* [6], Lian *et al.* [19], VideoAtt [7], Fang *et al.* [10], ESCNet [2], Tu *et al.* [29], Jin *et al.* [16], Miao *et al.* [23], Tafasca *et al.* [27], Tonini *et al.* [28]. All existing methods predict gaze targets from the original global view only. It is worth noting that the method [28] relied on an additional object detector to obtain the object position and category priors. Furthermore, the experimental settings in [28, 29], referred to as 'Real', differed significantly from the standard settings in all other methods. Therefore, it was unfair to directly compare the experimental results in [28] with our method and other existing methods.

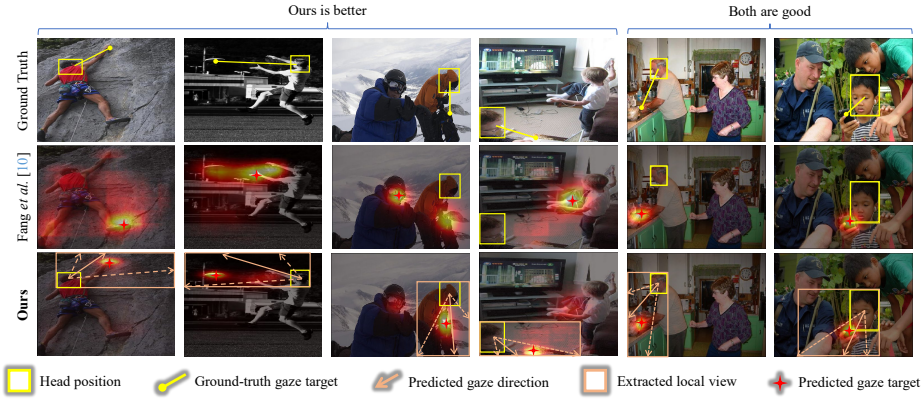

Fig. 6: Qualitative comparisons between our proposed method (3rd row) and the existing SOTA method [10] (2nd row). Faced with the diversity of human head position and gaze direction, our method more accurately predicted the true gaze target through the coordination of both the extracted local view and original global view, as demonstrated in the qualitative results.

Methods	FPS	GFLOPs	Avg. L_2↓
VideoAtt [7]	67	9.4	0.137
Tu *et al.* [29]	16	79.7	0.133
Ours	45	13.3	0.114

Table 2: FLOPs and FPS comparison on a NVIDIA Titan Xp GPU.

Methods	L_2 Distance ↓	
	GF	VAT
Global-local fusion	0.138	0.121
Head-global fusion	0.129	0.113
Head-local fusion	0.122	0.107
Ours	**0.114**	**0.101**

Table 3: Ablation of head-local-global feature fusion.

Quantitative Comparisons. Tab. 1 shows the quantitative results on the GazeFollow benchmark [25] and VideoAttentionTarget benchmark [7]. Our method achieved the state-of-the-art (SOTA) performance on the GazeFollow benchmark in terms of all evaluation metrics. Compared to existing SOTA methods [2,16,27] which also rely on scene depth prior and 3D gaze direction, our method exhibited a relative improvement of 5.0% on average L_2 distance metric, 7.8% on minimum L_2 distance metric, and 15.1% on angle error metric. It is worth noting that the variant of our method ('Ours-l'), which detects gaze targets solely from the local view, has achieved almost the same good performance as existing SOTA methods. Additionally, our method also achieved SOTA performance on the VideoAttentionTarget benchmark. Compared to existing SOTA methods [16,23], our method achieves a relative improvement of 2.9% on L_2 distance metric. These results quantitatively demonstrate the effectiveness of the local view we introduced and head-local-global coordination we proposed. We also provide the FLOPs and inference speed comparison in Tab. 2.

Qualitative Comparisons. Figure 6 visualizes the qualitative comparisons between our proposed method (3rd row) and the existing SOTA method [10] (2nd row). The 1st row shows the original global-view image and the ground-truth gaze target. Faced with the diversity of human head position and gaze

Methods	GazeFollow		VideoAttentionTarget	
	Angle Error ↓	Average L_2 ↓	Angle Error ↓	L_2 Distance ↓
Gaze360	16.7°	0.132	14.4°	0.115
ETH-XGaze	15.9°	0.126	13.8°	0.110
MPSGaze	14.8°	0.118	12.9°	0.104
Ours	**14.2°**	**0.114**	**12.4°**	**0.101**

Table 4: Ablation of gaze estimation module ('Angle Error') and its impact on gaze target detection ('L_2' Distance).

Methods	L_2 Distance (Average) ↓	
	GazeFollow	VideoAttentionTarget
W/o position consistency	0.192	0.183
W/o representation consistency	0.123	0.108
Ours	**0.114**	**0.101**

Table 5: Ablation of global-local position and representation consistency.

direction, our method more accurately predicted the true gaze target through the coordination of both the extracted local view and original global view, as demonstrated in the qualitative results. These qualitative demonstrate the effectiveness of the local view and head-local-global coordination we introduced.

4.2 Ablation Study

Ablation of Head-Local-Global Feature Fusion. We evaluated the effectiveness of the head-local-global feature fusion in our network by implementing several variants, as shown in Tab. 3. Among them, 'head-local fusion' means fusing facial features and local-view features only. 'Head-global fusion' means fusing facial features and global-view features only. 'Global-local fusion' means fusing global-view features and local-view features only. 'Ours' represents our complete head-local-global feature fusion. These results quantitatively demonstrated that the proposed feature fusion strategy effectively improved the network's performance in predicting the gaze target.

Ablation of Gaze Estimation. We evaluated the effectiveness of gaze estimation in our head view branch by implementing several variants, as shown in Tab. 4. Among them, 'Gaze360', 'ETH-XGaze' and 'MPSGaze' represent pretraining our head view branch on the Gaze360 dataset [18], ETH-XGaze dataset [32] and MPSGaze dataset [31] respectively. 'Ours' represents our complete cross-dataset pretraining method, *i.e.* pretraining the model together on these three datasets. Our cross-dataset pretrained model exhibited better generalization performance of 3D gaze estimation on gaze target detection datasets.

We also validated the impact of gaze estimation accuracy on the performance of our method in gaze target detection, as shown in Tab. 4. These results demonstrated that the more accurate the gaze direction, the better the performance of our method in gaze target detection. This phenomenon may be due to the fact that the accuracy of gaze direction affects both the accuracy of FOV calculation and the accuracy of local view extraction.

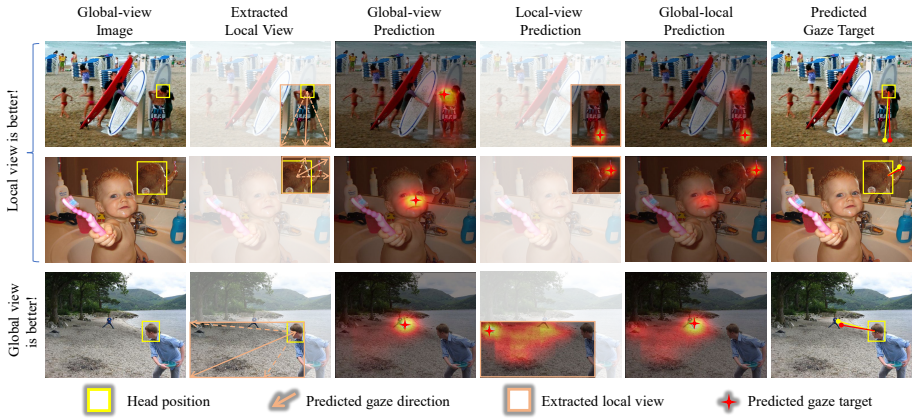

Fig. 7: Visualizing the sub-modules of our head-local-global coordination network.

Methods	L_2 Distance (Average) ↓	
	GazeFollow	VideoAttentionTarget
W/o $\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_l, \boldsymbol{H}_l)$ & $\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_g, \boldsymbol{H}_g)$	0.126	0.110
Ours	**0.114**	**0.101**

Table 6: Ablation of independent supervision of local view/global view.

Ablation of Position/Representation Consistency. We evaluated the effectiveness of the proposed global-local position consistency and representation consistency by implementing several variants, as shown in Tab. 5. Among them, 'w/o position consistency' means abandoning the head attention mechanisms in both the local view branch and global view branch. 'W/o representation consistency' means abandoning the global-local contrastive learning mechanism between the local-view scene features and global-view scene features. These results demonstrated that our proposed position consistency and representation consistency mechanisms significantly contributed to the network's ability to align spatial relationships and learn high-quality image representations between the local-view features and global-view features in high-dimensional feature spaces.

Ablation of Independent Supervision of Local View/Global View. We evaluated the effectiveness of the heatmap regression loss $\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_l, \boldsymbol{H}_l)$ for local-view prediction and $\mathcal{L}_{reg}(\hat{\boldsymbol{H}}_g, \boldsymbol{H}_g)$ for global-view prediction by implementing the variant abandoning both of them, as shown in Tab. 6. The introduction of these losses effectively strengthened the supervision of the learning process for predicting the gaze target in our head-local-global coordination network.

4.3 Module Visualization

Fig. 7 visualizes the sub-modules of our head-local-global coordination network, including the original global-view image, the extracted local-view image (pink bounding box), the global-view prediction, the local-view prediction (pink bounding box), the head-local-global coordination prediction, the predicted gaze target (red point) and ground-truth gaze target (yellow point). These outcomes

Methods	Origin			Head-Local-Global Coordination		
	AUC↑	Min. L_2↓	Avg. L_2↓	AUC↑	Min. L_2↓	Avg. L_2↓
VideoAtt [7]	0.921	0.077	0.137	**0.932**	**0.065**	**0.122**
Fang *et al.* [10]	0.922	0.067	0.124	**0.936**	**0.062**	**0.117**

Table 7: Evaluation of the scalability of our head-local-global coordination framework on the GazeFollow dataset [25].

qualitatively demonstrate the effectiveness of our proposed gaze target detection method based on head-local-global coordination.

4.4 Evaluation of Framework Scalability

To evaluate the scalability of the head-local-global coordination framework we proposed, we optimized the existing methods [7,10] by introducing the local view and view-coordination mechanism, as shown in Tab. 7. Among them, 'Origin' represents the original method. 'Head-Local-Global Coordination' represents the optimized method. These results quantitatively demonstrated that by integrating the additional local view and view-coordination mechanism, existing gaze target detection methods can be optimized within our head-local-global coordination framework, enhancing model performance.

5 Limitation and Feature Work

Due to the fact that the extraction of local view in our method relies on accurate gaze direction, our method may produce erroneous predictions in the cases of low facial visibility leading to significant errors in gaze estimation.

Besides, most existing datasets in this task have limited ecological validity, since they have no ground truth gaze data but are photo collections that humans have annotated the gaze target with their best guesses. As a result, the manual annotations are not entirely accurate. Researchers in this new research field are conducting continuous works for better datasets and it takes time.

6 Conclusion

Our research presents a significant advancement in the field of gaze target detection. By introducing a FOV-based local view and employing a head-local-global coordination approach, we address the limitations of traditional gaze target detection methods. Our novel framework not only improves accuracy in identifying gaze targets but also demonstrates strong scalability by enhancing existing methods. The extensive experimental results confirm our method's superior performance on key benchmarks. Despite its reliance on accurate gaze direction estimation, which could be a limitation in cases of low facial visibility, our method marks a substantial step forward in understanding and interpreting human gaze in unconstrained environments. Future work will focus on overcoming these limitations and exploring the integration of more robust gaze estimation techniques to further enhance the method's applicability and performance.

References

1. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. Advances in neural information processing systems **32** (2019) 4
2. Bao, J., Liu, B., Yu, J.: Escnet: Gaze target detection with the understanding of 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14126–14135 (2022) 1, 2, 3, 10, 11
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 4
4. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020) 4
5. Cheng, Y., Zhang, X., Lu, F., Sato, Y.: Gaze estimation by exploring two-eye asymmetry. IEEE Transactions on Image Processing **29**, 5259–5272 (2020) 1, 3
6. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: Proceedings of the European conference on computer vision (ECCV). pp. 383–398 (2018) 1, 10
7. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5396–5406 (2020) 1, 2, 3, 9, 10, 11, 14
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 9
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010) 9
10. Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., Zhai, G.: Dual attention guided gaze target detection in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11390–11399 (2021) 1, 2, 3, 10, 11, 14
11. Fischer, T., Chang, H.J., Demiris, Y.: Rt-gene: Real-time eye gaze estimation in natural environments. In: Proceedings of the European conference on computer vision (ECCV). pp. 334–352 (2018) 3
12. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019) 10
13. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006) 4
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 4
15. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018) 4
16. Jin, T., Yu, Q., Zhu, S., Lin, Z., Ren, J., Zhou, Y., Song, W.: Depth-aware gaze-following via auxiliary networks for robotics. Engineering Applications of Artificial Intelligence **113**, 104924 (2022) 1, 2, 10, 11

17. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th international conference on computer vision. pp. 2106–2113. IEEE (2009) 9
18. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6912–6921 (2019) 3, 9, 12
19. Lian, D., Yu, Z., Gao, S.: Believe it or not, we know what you are looking at! In: Asian Conference on Computer Vision. pp. 35–50. Springer (2018) 1, 3, 10
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 9
21. Lu, F., Okabe, T., Sugano, Y., Sato, Y.: Learning gaze biases with head motion for head pose-free gaze estimation. Image and Vision Computing **32**(3), 169–179 (2014) 1, 3
22. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE transactions on pattern analysis and machine intelligence **36**(10), 2033–2046 (2014) 1, 3
23. Miao, Q., Hoai, M., Samaras, D.: Patch-level gaze distribution prediction for gaze following. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 880–889 (2023) 1, 2, 10, 11
24. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020) 6
25. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? Advances in neural information processing systems **28** (2015) 1, 3, 10, 11, 14
26. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1821–1828 (2014) 3
27. Tafasca, S., Gupta, A., Odobez, J.M.: Childplay: A new benchmark for understanding children's gaze behaviour. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20935–20946 (2023) 1, 10, 11
28. Tonini, F., Dall'Asen, N., Beyan, C., Ricci, E.: Object-aware gaze target detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21860–21869 (2023) 10
29. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2192–2200. IEEE (2022) 10, 11
30. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010) 9
31. Zhang, M., Liu, Y., Lu, F.: Gazeonce: Real-time multi-person gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4197–4206 (2022) 3, 9, 12
32. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: European Conference on Computer Vision. pp. 365–381. Springer (2020) 3, 9, 12
33. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4511–4520 (2015) 1, 3

34. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 51–60 (2017) 1, 3