Gaze Target Detection by Merging Human Attention and Activity Cues

Yaokun Yang, Yihan Yin, Feng Lu*

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

Abstract

Despite achieving impressive performance, current methods for detecting gaze targets, which depend on visual saliency and spatial scene geometry, continue to face challenges when it comes to detecting gaze targets within intricate image backgrounds. One of the primary reasons for this lies in the oversight of the intricate connection between human attention and activity cues. In this study, we introduce an innovative approach that amalgamates the visual saliency detection with the body-part & object interaction both guided by the soft gaze attention. This fusion enables precise and dependable detection of gaze targets amidst intricate image backgrounds. Our approach attains state-of-the-art performance on both the Gazefollow benchmark and the GazeVideoAttn benchmark. In comparison to recent methods that rely on intricate 3D reconstruction of a single input image, our approach, which solely leverages 2D image information, still exhibits a substantial lead across all evaluation metrics, positioning it closer to human-level performance. These outcomes underscore the potent effectiveness of our proposed method in the gaze target detection task.

Introduction

Eye gaze assumes a pivotal role in elucidating human activities. Although traditional studies (Lu et al. 2014a,b; Cheng et al. 2020; Zhang et al. 2015, 2017) have predominantly centered around estimating the gaze direction, discerning the precise location that a person fixates upon—termed as the gaze target—offers a more intuitive avenue for delving into profound human attention. Consequently, the detection of human gaze targets in real-world contexts has emerged as a formidable endeavor within the realm of computer vision. Furthermore, this approach has discovered extensive applications across diverse domains such as human-computer interaction (Fathi, Li, and Rehg 2012; Schauerte and Stiefelhagen 2014), analysis of social awareness (Marin-Jimenez et al. 2019, 2014; Fan et al. 2018), and medical research.

Traditionally, the task of gaze target detection has predominantly revolved around visual saliency detection along the gaze direction (Recasens et al. 2015; Lian, Yu, and Gao 2018; Chong et al. 2020). Furthermore, recent advancements (Fang et al. 2021; Bao, Liu, and Yu 2022) have integrated monocular depth estimation as an auxiliary information source to enhance the computation of the scene's three-dimensional geometry. Despite achieving noteworthy performance gains, prevailing methods continue to grapple with the precise and dependable detection of gaze targets amidst intricate image backgrounds. This challenge can be attributed to the lack of consideration given to the intricate connection between human attention and activity cues.

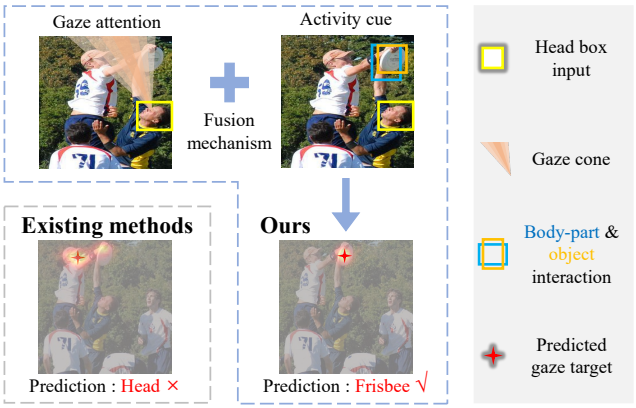


Figure 1: Comparison between existing methods and ours.

The gaze target detection task serves as a means to elucidate the connection between human attention and activity cues. Specifically, by observing an individual's gaze attention, we can glean insights into their activities. Moreover, comprehending an individual's activity cues helps us to anticipate their gaze target. Based on above analysis, as Illustrated in Fig. 1, we consider merging human attention and activity cues in the gaze target detection task. In this study, we introduce an innovative approach that amalgamates the visual saliency detection with the body-part & object interaction both guided by the soft gaze attention. This fusion enables precise and dependable detection of gaze targets amidst intricate image backgrounds.

Based on our observations, when individuals are engrossed in specific activities, their gaze attention tends to be fixated on objects they are actively interacting with (see Fig. 2 (a, b, c)). However, scenarios exist where the gaze target might involve non-interactive objects (see Fig. 2 (d)) or be directed towards the conduct of another individual. Thus, it

*Corresponding Author.

Figure 2: Visualizing the intricate connection between human attention and activity cues. The gaze target of an individual can either be directed at an object that interacts with specific body parts (see images (a, b, c)), or it might involve a non-interactive object (see image (d)). Moreover, objects that interact with the entire body might not necessarily align with their gaze attention (see image (e)).

becomes imperative to establish a mechanism that unearths the intricate connection between human gaze attention and activity cues. Recognizing that a significant portion of interactive objects might not align with the individual's gaze attention (*e.g.*, Fig. 2 (e)), we introduce a pioneering body-part & object interaction attention mechanism specially designed for gaze target detection. Our approach centers on identifying interactions between five specific body parts—namely, the head, hands and feet—and each object within the scene. This process is guided by the individual's gaze attention and aims to effectively discern the potential gaze target among all interactive objects.

Moreover, a significant portion of samples present challenges of low facial visibility in the wild due to factors like blurriness, orientation, or obstructions, among others. Gaze estimation methods (Cheng et al. 2020; Zhang et al. 2015) that solely rely on facial characteristics are susceptible to failure under such circumstances. To address this limitation, we introduce a resilient soft gaze attention mechanism. This technique extracts gaze-consistent features from both the human face and five specific head keypoints—namely, the nose, eyes, and ears. The resultant gaze attention is harnessed to assess the probability of a salient region or an interaction hotspot housing potential gaze targets.

We stand as pioneers in merging human attention and activity cues into the gaze target detection task. In this study, we introduce an innovative approach that amalgamates the visual saliency detection with the body-part & object interaction both guided by the soft gaze attention. This fusion enables precise and dependable detection of gaze targets amidst intricate image backgrounds. Notably, our approach attains state-of-the-art performance on both the Gazefollow benchmark (Recasens et al. 2015) and the GazeVideoAttn benchmark (Chong et al. 2020). In comparison to recent methods that rely on intricate 3D reconstruction of a single input image, our approach, which solely leverages 2D image information, still exhibits a substantial lead across all evaluation metrics, positioning it closer to human-level performance. These outcomes underscore the potent effectiveness of our proposed method in gaze target detection.

This paper makes the following primary contributions:

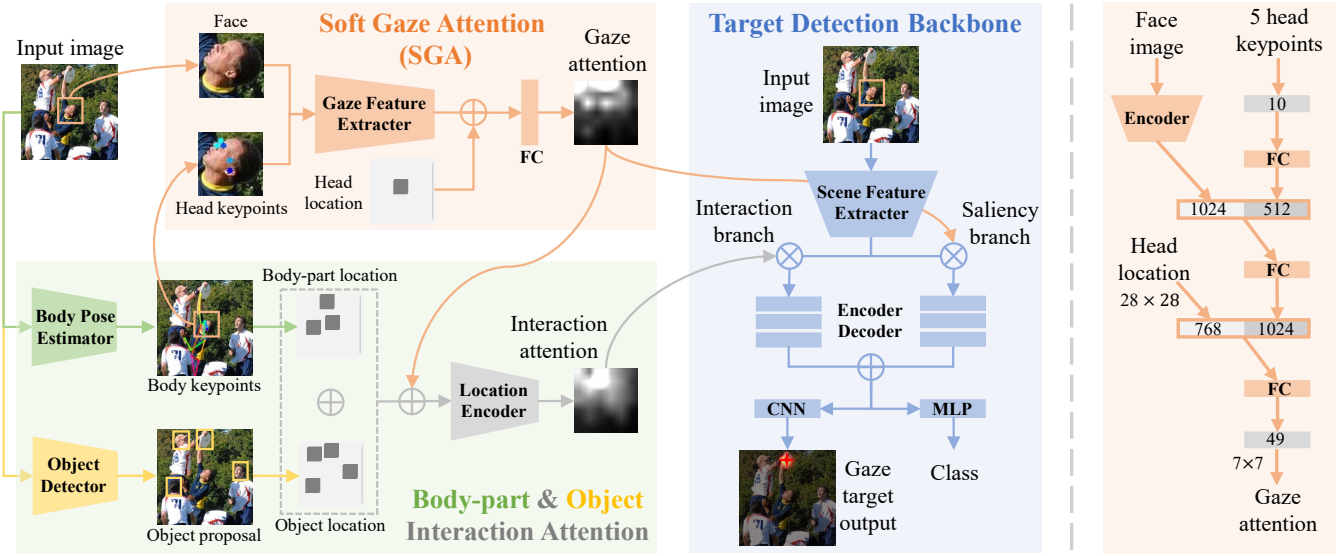
- We propose a novel approach which utilizes gaze and activity cues to solve the gaze target detection task. Our strategy to integrate gaze direction and human-object interaction reflects the natural idea of combining human attention and activity.
- We design a robust gaze attention mechanism which extracts the gaze features from both the human face and specific head keypoints.
- We introduce a specialized body-part & object interaction module which is able to uncover the connection between human attention and activity cues.

Related Work

Gaze Target Detection The gaze target detection task offers a more intuitive approach to delve into profound human attention. Recasens (Recasens et al. 2015) pioneered the exploration of this general problem and presented the expansive GazeFollow image dataset, featuring annotations of head positions and corresponding gaze targets. Lian (Lian, Yu, and Gao 2018) harnessed multi-scale FOV attention to enhance view supervision. Chong (Chong et al. 2020) extended the task to out-of-frame scenarios through a video dataset. Fang (Fang et al. 2021) introduced monocular depth estimation as additional prior information. Bao (Bao, Liu, and Yu 2022) utilized intricate analytical calculations for 3D geometry. Despite these achievements in performance, prevailing methods still encounter challenges in accurately detecting gaze targets amid complex image backgrounds.

Gaze Estimation The problem of appearance-based gaze estimation has long been a focal point in computer vision (Lu et al. 2014a,b; Cheng et al. 2020; Fischer, Chang, and Demiris 2018; Zhang et al. 2015, 2017). Nevertheless, the majority of available gaze estimation datasets (Kellnhofer et al. 2019; Sugano, Matsushita, and Sato 2014; Zhang et al. 2020) are obtained within controlled laboratory environments, encompassing 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.

Human-Object Interaction The task of recognizing human-object interactions (Yao and Fei-Fei 2010, 2012; Gupta and Malik 2015; Gkioxari et al. 2018; Gao, Zou, and Huang 2018; Chao et al. 2018; Qi et al. 2018) can be represented as detecting hhuman, verb, objecti triplets. Gupta and Malik (Gupta and Malik 2015) first tackle the HOI detection problem — detecting people doing actions and the object instances they are interacting with. Gkioxari (Gkioxari et al. 2018) introduces an action-specific density map over target object locations based on the appearance of a detected person. In addition to using object instance appearances, Chao (Chao et al. 2018) also encode the relative spatial relationship between a person and the object with a CNN.



(a) Gaze Target Detection Approach **(b) SGA-Module**

Figure 3: Overview. Our gaze target detection approach consists of three main modules: Soft Gaze Attention, Body-part & Object Interaction Attention, and Target Detection Backbone. The target detection backbone comprises two branches: the saliency branch and the interaction branch. Finally, we combine the target heatmaps generated by these two branches, utilize a CNN network to predict the ultimate gaze target heatmap, and employ an MLP to determine if the gaze target falls out of the frame. SGA-Module is the architecture of our soft gaze attention module. This module is designed to generate a soft gaze attention map by leveraging information from both the human face and specific head keypoints.

Approach

Illustrated in Figure 3, our gaze target detection approach consists of three main modules: Soft Gaze Attention, Body-part & Object Interaction Attention, and Target Detection Backbone. The target detection backbone encompasses two distinctive branches: the saliency branch and the interaction branch. Our soft gaze attention module is designed to predict gaze attention by leveraging information from both the human face and five specific head keypoints (the nose, eyes and ears). The resulting gaze attention map $\boldsymbol{A_g}$ plays a pivotal role in guiding the body-part & object interaction module and the target detection backbone.

Our body-part & object interaction attention module initiates by employing a pre-trained body pose estimator to calculate the body keypoints of the individual, denoted as $\boldsymbol{v_{bk}}$, and a pre-trained object detector to derive object proposals within the scene. Guided by the soft gaze attention $\boldsymbol{A_g}$, this module discerns interactions between five distinct body parts (*i.e.*, the head, hands and feet) and all objects present within the scene. Subsequently, the body-part & object interaction attention $\boldsymbol{A_{hoi}}$ is generated and employed to guide the interaction branch within our target detection backbone.

Our target detection backbone initiates by extracting scene features from the entire scene input. Guided by soft gaze attention $\boldsymbol{A_g}$, our saliency branch determines whether the extracted saliency regions encompass potential gaze targets. In parallel, guided by body-part & object interaction attention $\boldsymbol{A_{hoi}}$, our interaction branch gauges the likelihood that the detected interaction hotspots constitute potential gaze targets. Finally, we combine the target heatmaps generated by these two branches, utilize a CNN network to predict the ultimate gaze target heatmap, and employ an MLP to determine if the gaze target falls out of the frame.

Soft Gaze Attention

The architecture of our soft gaze attention module is illustrated in Figure 3. We employ the lightweight MobileNet (Howard et al. 2019) to extract features from the provided face image $\boldsymbol{I_{face}}$, which has been pre-resized to 64×64 pixels. Then, the extracted feature maps undergo an average pooling operation, resulting in a 1024-dimensional feature vector $\boldsymbol{v_f}$. Simultaneously, as depicted in Figure 3, we derive five specific head keypoints (*i.e.*, the nose, eyes and ears) from the computed body keypoints of the individual, which is accomplished by a pre-trained body pose estimator. Subsequently, these five head keypoints are encoded and transformed into a 512-dimensional feature vector $\boldsymbol{v_{hk}}$ through a fully connected (FC) layer. The vectors $\boldsymbol{v_f}$ and $\boldsymbol{v_{hk}}$ are then concatenated and further projected into a 1024-dimensional feature vector $\boldsymbol{v_g}$ via an additional FC layer.

Following this, the head location map $\boldsymbol{M_h}$ is resized to dimensions of 28×28 pixels and encoded into a 768-dimensional vector $\boldsymbol{v_h}$. These vectors, $\boldsymbol{v_g}$ and $\boldsymbol{v_h}$, are concatenated and projected into a 49-dimensional vector $\boldsymbol{v_{atn}}$ through a subsequent FC layer. Finally, the vector $\boldsymbol{v_{atn}}$ is resized to yield the 7×7 pixel gaze attention map $\boldsymbol{A_g}$.

In situations where faces have limited visibility, our soft gaze attention module showcases heightened resilience. This enhanced resilience stems from the module's ability to leverage the spatial correlation between head keypoints and facial orientation, setting it apart from traditional gaze estimation methods (Cheng et al. 2020) that exclusively emphasize the

Figure 4: Comparison between the Baseline and our method. First row: showcases a selection of samples extracted from the Gazefollow test set along with their corresponding true annotations. Second row: presents the gaze target heatmap forecasted by the Baseline. Third row: displays the prediction generated by our method. Our method distinctly outperforms the Baseline when the actual gaze target is a diminutive interactive object concealed within intricate image backgrounds.

extraction of gaze-consistent features from the human face.

Body-part & Object Interaction Attention

As depicted in Fig. 3, our approach initially utilizes a pre-trained body pose estimator to calculate the body keypoints of the individual. Subsequently, we determine the location map $\boldsymbol{M_{bp}}$ for five specific body parts (*i.e.*, the head, hands and feet) using the keypoint coordinates. Simultaneously, employing a pre-trained object detector, we acquire object proposals from the scene and generate an object location map $\boldsymbol{M_o}$ for all detected objects. Subsequently, we concatenate the object location map $\boldsymbol{M_o}$ with the body-part location map $\boldsymbol{M_{bp}}$ in channel dimension, resulting in the formation of the body-part & object location pair.

Through Eq. 1, these paired location maps, along with the gaze attention map $\boldsymbol{A_g}$, are concatenated and passed through a location encoder denoted as $\mathcal{F}_{loc}(\cdot)$, leading to the generation of the interaction attention map $\boldsymbol{A_{hoi}}$ that pertains to the body-parts of the individual and the detected objects.

$$\boldsymbol{A_{hoi}} = \mathcal{F}_{loc}((\boldsymbol{M_{bp}} \oplus \boldsymbol{M_o}) \oplus \boldsymbol{A_g}). \tag{1}$$

Our body-part & object interaction attention mechanism enhances the precision of identifying potential gaze targets among a range of interactive objects.

Target Detection Backbone

Illustrated in Fig.3, we commence by concatenating the head location map $\boldsymbol{M_h}$ of the given individual with the complete scene image $\boldsymbol{I_{rgb}}$. Subsequently, we utilize the feature extractor $\mathcal{F}_{scn}(\cdot)$ to extract the convolutional scene feature maps denoted as $\boldsymbol{m_{scn}}$,

$$\boldsymbol{m_{scn}} = \mathcal{F}_{scn}(\boldsymbol{I_{rgb}} \oplus \boldsymbol{M_h}). \tag{2}$$

The saliency branch $\mathcal{F}_{sal}(\cdot)$ is composed of two 1×1 CNN layers and three transposed CNN layers. Guided by the soft gaze attention $\boldsymbol{A_g}$, this branch encodes and decodes a target heatmap $\boldsymbol{H_{sal}}$ through Eq. 3, to ascertain if the extracted saliency regions contain potential gaze targets.

$$\boldsymbol{H_{sal}} = \mathcal{F}_{sal}(\boldsymbol{m_{scn}} \otimes \boldsymbol{A_g}). \tag{3}$$

The interaction branch $\mathcal{F}_{hoi}(\cdot)$ shares the same architecture as the saliency branch. Guided by the body-part & object interaction attention $\boldsymbol{A_{hoi}}$, this branch encodes and decodes another target heatmap $\boldsymbol{H_{hoi}}$ through Eq. 4, to determine the probability that the identified interaction hotspots represent potential gaze targets.

$$\boldsymbol{H_{hoi}} = \mathcal{F}_{hoi}(\boldsymbol{m_{scn}} \otimes \boldsymbol{A_{hoi}}). \tag{4}$$

Finally, through Eq. 5, we combine these two predicted heatmaps and input them into a fusion network $\mathcal{F}_{fus}(\cdot)$ comprising two 1×1 CNNs, to generate the ultimate prediction $\boldsymbol{H_{fus}}$ for the gaze target.

$$\boldsymbol{H_{fus}} = \mathcal{F}_{fus}(\boldsymbol{H_{sal}} \oplus \boldsymbol{H_{hoi}}). \tag{5}$$

Meanwhile, we also input $\boldsymbol{H_{sal}} \oplus \boldsymbol{H_{hoi}}$ into a MLP classifier to determine if the gaze target falls out of the frame.

Overall Loss Function

To provide supervision for the saliency branch, we employ a regression loss function $\mathcal{L}_{sal}$ that computes the mean square error between the gaze-guided scene saliency map $\boldsymbol{H_{sal}}$ and the ground truth gaze target heatmap $\boldsymbol{H^*}$,

$$\mathcal{L}_{sal} = MSE(\boldsymbol{H_{sal}}, \boldsymbol{H^*}). \tag{6}$$

Since there are no annotations pertaining to human activities in gaze target detection datasets, we do not provide distinct supervision for our interaction branch. We achieve supervision over the fusion of the interaction branch and the saliency branch through the inclusion of an additional loss function, denoted as $\mathcal{L}_{fus}$, in our fusion prediction. The loss function $\mathcal{L}_{fus}$ computes the mean square error between the fusion target heatmap $\boldsymbol{H_{fus}}$ and the ground truth $\boldsymbol{H^*}$,

$$\mathcal{L}_{fus} = MSE(\boldsymbol{H_{fus}}, \boldsymbol{H^*}). \tag{7}$$

We define the classification loss function of the gaze target as $\mathcal{L}_{cls}$. The overall loss function is formulated as follows,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{sal} + \lambda_3 \mathcal{L}_{fus}, \tag{8}$$

where λ_1, λ_2 and λ_3 are hyper-parameters. We empirically set $\lambda_1 = \lambda_2 = \lambda_3 = 1.0$.

Methods	Supervision				GazeFollow		VideoAttentionTarget	
	Activity	Depth	3D	Eye	Min. Dist. ↓	Avg. Dist. ↓	Dist. ↓	AP ↑
Random					0.391	0.484	0.458	0.621
Fixed bias					0.219	0.306	0.326	0.624
Baseline					0.077	0.137	0.147	0.848
Chen	√				0.074	0.136	-	-
Fang		√		√	0.067	0.124	0.108	0.896
Tu					0.069	0.133	0.126	0.854
Bao		√	√		-	0.122	0.120	0.869
Miao		√			0.065	0.123	0.109	0.908
Ours*					0.068	0.126	**0.106** (20.9% ↓)	**0.910** (7.3% ↑)
Ours	√				**0.061** (20.8% ↓)	**0.118** (13.9% ↓)	-	-

Table 1: Evaluation on the GazeFollow dataset and the VideoAttentionTarget dataset. Ours*: our method without the body-part & object interaction attention and the interaction branch. Ours: our complete method. The data in parentheses represents the proportion of improvement in the performance of our method compared to the Baseline. Activity: the individual's activity cues. Depth: depth prior information of the scene. 3D: 3D reconstruction of the scene. Eye: additional eye annotations.

Experimental Results

Preparation

Datasets This paper employs two well-established datasets for gaze target detection, namely GazeFollow (Recasens et al. 2015) and VideoAttentionTarget (Chong et al. 2020). GazeFollow constitutes a large-scale gaze-tracking dataset that comprises 130,339 individuals within 122,143 images. These images are sourced from a diverse range of existing datasets, *e.g.*, ImageNet (Deng et al. 2009), COCO (Lin et al. 2014), PASCAL (Everingham et al. 2010), SUN (Xiao et al. 2010), *etc.*. After partitioning, 4,782 annotated individuals are designated for testing, with the remainder allocated for training. Furthermore, ten human annotations are solicited per individual in the test images to facilitate an evaluation of 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.

Evaluation Metrics The evaluation of our proposed model's performance is conducted using the following metrics. **Dist.**: This metric quantifies the performance by evaluating the L_2 distance between the predicted gaze target point and the corresponding ground truth annotation. **Out of frame AP**: The accuracy of identifying out-of-frame instances is assessed through the utilization of average precision (AP). These metrics provide a comprehensive assessment of our model's performance across various aspects.

Implementation Details Our implementation is carried out using the PyTorch framework. We utilize ResNet-50 (He et al. 2016) as our scene feature extractor. All input scene images are resized to dimensions of 224×224, while our input face image is resized to 64×64. During training, we employ a mini-batch size of 32 on a single NVIDIA Titan Xp GPU, initializing with a learning rate of 0.0001. Our training regimen spans 90 epochs on the GazeFollow dataset, with learning rate adjustments at the 80th and 90th epochs, involving a multiplication by 0.1. Our entire training process takes approximately 18 hours. As our optimizer, we rely on the Adam algorithm (Kingma and Ba 2014), with an Adam weight decay set at 0.0001 and an Adam momentum of 0.9. During inference, our complete model achieved an image processing time of less than 75ms on a single NVIDIA GPU.

Comparison Methods

Baseline We adopt the method introduced in Video (Chong et al. 2020) as our Baseline. The Baseline approach generates gaze attention solely from the human face and predicts the gaze target exclusively by extracting the gaze-guided salience feature of the scene. It is evident that the disparity in performance between our comprehensive model and the Baseline stems from the integration of the interaction branch guided by our proposed body-part & object interaction attention, along with the incorporation of five specific head keypoints into our soft gaze attention module.

Gaze Target Detection Methods Furthermore, we conduct comparisons with five recent methods: Chen (Chen et al. 2021), Fang (Fang et al. 2021), Tu (Tu et al. 2022), Bao (Bao, Liu, and Yu 2022), and Miao (Miao, Hoai, and Samaras 2023). These methods have all demonstrated notable performance within the confines of within-dataset evaluations.

Performance Comparison with SOTA Methods

Evaluation on GazeFollow Dataset As demonstrated in Table 1, our method exhibits a substantial lead over the second-best competitor across all evaluation metrics, positioning it closer to human-level performance. Compared to the Baseline approach in Video (Chong et al. 2020), our method achieves a relative enhancement of 20.8% for the minimum L_2 distance and 13.9% for the average L_2 distance. Even compared with the state-of-the-art method Bao (Bao, Liu, and Yu 2022), which relies on intricate 3D reconstruction of a single input image, our approach,which solely leverages 2D image information, still attains a relative advancement of 3.3% for the average L_2 distance.

Evaluation on VideoAttentionTarget Dataset The VideoAttentionTarget dataset (Chong et al. 2020) exhibits a deficiency in terms of diverse human activities, thereby placing limitations on the efficacy of our proposed body-part & object interaction module. As depicted in Table 1,

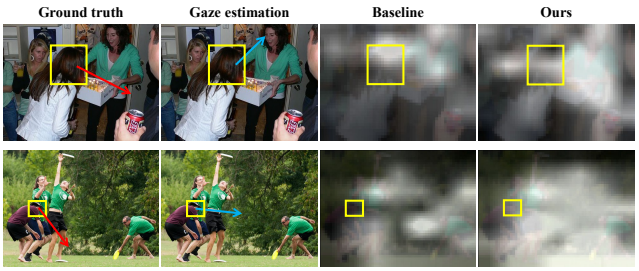


Figure 5: Visualizing the comparison between our soft gaze attention method (fourth column), the conventional gaze estimation approach (second column) and the Baseline (third column), in scenarios where faces have limited visibility.

Methods	Min. Dist. ↓	Avg. Dist. ↓
Baseline	0.077	0.137
Ours	**0.068**	**0.126**

Table 2: Ablation study of soft gaze attention module on the GazeFollow dataset. Baseline: soft gaze attention in the Baseline method. Ours: our proposed soft gaze attention.

the performance of our model without the body-part & object interaction module is represented by "Ours*". In comparison to the Baseline approach in Video (Chong et al. 2020), "Ours*" still attains a relative enhancement of 20.9% for the L_2 distance and 7.3% for the average precision concerning out-of-frame identification.

Qualitative Experimental Results A qualitative comparison between the Baseline and our method is presented in Figure 4. The initial row showcases a selection of samples extracted from the Gazefollow test set, along with their corresponding true annotations. The second and third rows respectively depict the gaze target heatmaps forecasted by the Baseline and our model. Our method notably outperforms the Baseline when the actual gaze target is a diminutive interactive object enshrouded in intricate image backgrounds. This substantial improvement is attributed to the fusion of human attention and activity cues within our approach.

Ablation Study

Soft Gaze Attention As shown in Fig.5, to evaluate the precision and resilience of our novel soft gaze attention approach (fourth column), which synergizes facial features and head keypoints, we conduct a comparative analysis with the conventional gaze estimation method (Zhang et al. 2020) (second column), as well as the soft gaze attention module within the Baseline method (third column). Both of these alternatives focus solely on extracting gaze-consistent features from the human face. In scenarios involving faces with reduced visibility, our proposed method demonstrates enhanced resilience attributed to its utilization of the spatial correlation between head keypoints and facial orientation. Besides, the quantitative comparison is shown in Tab.2. To ensure fairness, we exclude the body-part & object interaction attention and the interaction branch from our approach. This adjustment aligns our resulting model's framework with that of the Baseline method, namely scene saliency detection guided by soft gaze attention. These outcomes underscore the exceptional accuracy and robustness of our soft gaze attention method, even when confronting challenging instances of limited facial visibility in real-world conditions.

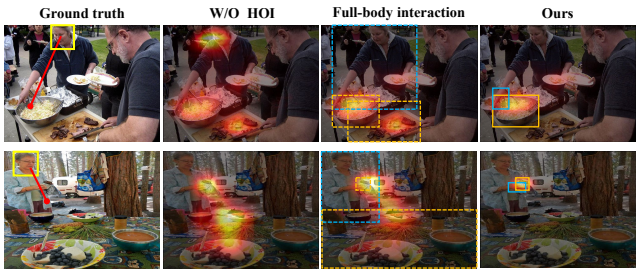


Figure 6: Visualizing the comparison between the interaction branch guided by our proposed body-part & object interaction attention (fourth column), the variant employing full-body object interaction (third column) and another variant without the entire interaction module (second column).

Methods	Min. Dist. ↓	Avg. Dist. ↓
W/O HOI	0.068	0.126
Full-body HOI*	0.066	0.124
Full-body HOI	0.063	0.121
Ours*	0.064	0.122
Ours	**0.061**	**0.118**

Table 3: Ablation study of our interaction branch on the GazeFollow dataset. Ours: our complete model with the interaction branch guided by our proposed body-part & object interaction attention. Full-body HOI: the variant of our interaction branch employing the full-body object interaction attention. *: the variant of our interaction attention lacking the guidance of gaze attention. W/O HOI: the variant of our method without the entire interaction module.

Interaction Branch As depicted in Fig. 6, in order to validate the effectiveness of the interaction branch which is guided by the body-part & object interaction attention, we juxtapose our approach (fourth column) with two variants: one lacking the entire interaction module (second column), and the other utilizing the full-body object interaction attention (third column). The focal point of our body-part & object interaction attention lies in the discernment of interactions between five specific body components (the head, hands and feet) and all detected objects. This attention mechanism enables a heightened precision in identifying potential gaze targets within all interactive objects. Furthermore, we scrutinize the performance of our proposed body-part & object interaction module in comparison to a variant operating without the guidance of gaze attention. The quantitative results are shown in Tab.3. This analysis underscores the effectiveness of infusing gaze attention into the body-part & object interaction module.

Module Visualization

The visualization of various stages within our network is presented in Figure 7, encompassing elements *e.g.*, the soft gaze attention map, gaze target heatmaps derived from both

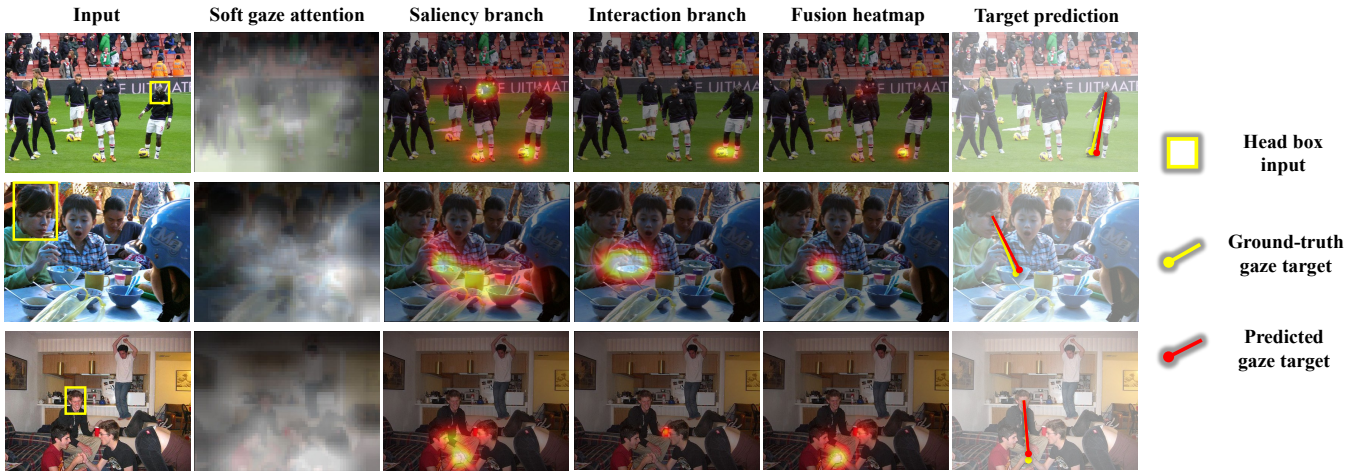


Figure 7: Visualization of our soft gaze attention, saliency branch prediction, interaction branch prediction, fusion heatmap, and the predicted gaze target. The third row presents a scenario wherein the gaze target is a non-interacting object.

the saliency branch and the interaction branch, the fusion heatmap, and the predicted gaze target. The initial two rows offer a demonstration of the adeptness of our proposed interaction branch in accurately discerning gaze targets amidst intricate image backgrounds. Conversely, the third row presents a scenario wherein the gaze target is a non-interacting object. Recognizing that instances where the true gaze target lacks interaction with the given individual are not uncommon in natural settings, the fusion of predictions from both the saliency branch and the interaction branch emerges as a strategy to yield enhanced robustness.

Computational Complexity

In order to analyse the computation complexity, we examine the inference speed of each module seperately. For our method, we use the pre-trained lightweight body pose estimator RTMPose (Jiang et al. 2023) and object detector YOLOv3 (Redmon et al. 2016). On the other hand, competing methods introduced some other modules, *e.g.*, face detection and depth estimation from the scene (Fang et al. 2021), body pose estimation and 3D reconstruction from the scene (Bao, Liu, and Yu 2022), ViT backbone (Tu et al. 2022). In order to measure their computation complexity, we also select recent high-speed implementations for them, and compared their inference speed on a single NVIDIA Titan XP GPU. The results are shown in Table 4, where our method shows its advantage in terms of inference speed.

Discussion

Incorporating human gaze target annotations into tasks that encompass human activities (*e.g.*, human-object interaction, action recognition/prediction, scene understanding, *etc.*) proves more advantageous for investigating the connection between human attention and activity cues, compared to datasets containing solely gaze target annotations. This augmentation is anticipated to evolve into a promising and innovative research domain within the realms of computer vision and human-computer interaction.

Method	Input Size	Time/image
Tu	224×224	ViT(63ms)
Fang	224×224	F($\sim 10ms$) + D($\sim 13ms$) + G($\sim 8ms$)
Bao	224×224	P($\sim 10ms$) + 3D($\sim 30ms$) + G($\sim 8ms$)
Ours	224×224	**P**($\sim 10ms$) **+ O**($\sim 11ms$) **+ G**($\sim 8ms$)

Table 4: Evaluation of inference speed w.r.t different modules. ViT: ViT backbone (Tu et al. 2022). F: face detection module (Deng et al. 2020). D: depth estimation module (Godard et al. 2019). G: gaze target detection backbone (our implementation). 3D: 3D reconstruction module (Sun et al. 2021). P: human pose estimation module (Jiang et al. 2023). O: object detection module (Redmon et al. 2016).

Conclusion

In this study, we propose a novel approach which utilizes gaze and activity cues to solve the gaze target detection task. Our strategy to integrate gaze direction and human-object interaction reflects the natural idea of combining human attention and activity. Our method achieves state-of-the-art performance on both the GazeFollow benchmark and the GazeVideoAttn benchmark. In comparison to recent methods which rely on intricate 3D reconstruction of a single input image, our approach which only leverages 2D image information still exhibits a substantial lead across all evaluation metrics, positioning it closer to human-level performance. These outcomes prove the effectiveness of our method in the gaze target detection task.

Acknowledgments

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

References

Bao, J.; Liu, B.; and Yu, J. 2022. ESCNet: Gaze Target Detection With the Understanding of 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14126–14135.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, 381–389. IEEE.

Chen, W.; Xu, H.; Zhu, C.; Liu, X.; Lu, Y.; Zheng, C.; and Kong, J. 2021. Gaze estimation via the joint modeling of multiple cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1390–1402.

Cheng, Y.; Zhang, X.; Lu, F.; and Sato, Y. 2020. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29: 5259–5272.

Chong, E.; Wang, Y.; Ruiz, N.; and Rehg, J. M. 2020. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5396–5406.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Fan, L.; Chen, Y.; Wei, P.; Wang, W.; and Zhu, S.-C. 2018. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6460–6468.

Fang, Y.; Tang, J.; Shen, W.; Shen, W.; Gu, X.; Song, L.; and Zhai, G. 2021. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11390–11399.

Fathi, A.; Li, Y.; and Rehg, J. M. 2012. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, 314–327. Springer.

Fischer, T.; Chang, H. J.; and Demiris, Y. 2018. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, 334–352.

Gao, C.; Zou, Y.; and Huang, J.-B. 2018. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*.

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8359–8367.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3828–3838.

Gupta, S.; and Malik, J. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324.

Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; and Chen, K. 2023. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv preprint arXiv:2303.07399*.

Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6912–6921.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lian, D.; Yu, Z.; and Gao, S. 2018. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, 35–50. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lu, F.; Okabe, T.; Sugano, Y.; and Sato, Y. 2014a. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing*, 32(3): 169–179.

Lu, F.; Sugano, Y.; Okabe, T.; and Sato, Y. 2014b. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10): 2033–2046.

Marin-Jimenez, M. J.; Kalogeiton, V.; Medina-Suarez, P.; and Zisserman, A. 2019. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3477–3485.

Marin-Jimenez, M. J.; Zisserman, A.; Eichner, M.; and Ferrari, V. 2014. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3): 282–296.

Miao, Q.; Hoai, M.; and Samaras, D. 2023. Patch-level Gaze Distribution Prediction for Gaze Following. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 880–889.

Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 401–417.

Recasens, A.; Khosla, A.; Vondrick, C.; and Torralba, A. 2015. Where are they looking? *Advances in neural information processing systems*, 28.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Schauerte, B.; and Stiefelhagen, R. 2014. "Look at this!" learning to guide visual saliency in human-robot interaction. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 995–1002. IEEE.

Sugano, Y.; Matsushita, Y.; and Sato, Y. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1821–1828.

Sun, J.; Xie, Y.; Chen, L.; Zhou, X.; and Bao, H. 2021. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15598–15607.

Tu, D.; Min, X.; Duan, H.; Guo, G.; Zhai, G.; and Shen, W. 2022. End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2192–2200. IEEE.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Yao, B.; and Fei-Fei, L. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 17–24. IEEE.

Yao, B.; and Fei-Fei, L. 2012. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE transactions on pattern analysis and machine intelligence*, 34(9): 1691–1703.

Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; and Hilliges, O. 2020. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, 365–381. Springer.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4511–4520.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. 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*, 51–60.