This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Dual Attention Guided Gaze Target Detection in the Wild

Yi Fang[1*], Jiapeng Tang[1*], Wang Shen[1], Wei Shen[2†], Xiao Gu[1], Li Song[1†], Guangtao Zhai[1†]
[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{yifang, tangjp, shenwang, wei.shen, gugu97, song_li, zhaiguangtao}@sjtu.edu.cn

Abstract

Gaze target detection aims to infer where each person in a scene is looking. Existing works focus on 2D gaze and 2D saliency, but fail to exploit 3D contexts. In this work, we propose a three-stage method to simulate the human gaze inference behavior in 3D space. In the first stage, we introduce a coarse-to-fine strategy to robustly estimate a 3D gaze orientation from the head. The predicted gaze is decomposed into a planar gaze on the image plane and a depth-channel gaze. In the second stage, we develop a Dual Attention Module (DAM), which takes the planar gaze to produce the filed of view and masks interfering objects regulated by depth information according to the depth-channel gaze. In the third stage, we use the generated dual attention as guidance to perform two sub-tasks: (1) identifying whether the gaze target is inside or out of the image; (2) locating the target if inside. Extensive experiments demonstrate that our approach performs favorably against state-of-the-art methods on GazeFollow and VideoAttentionTarget datasets.

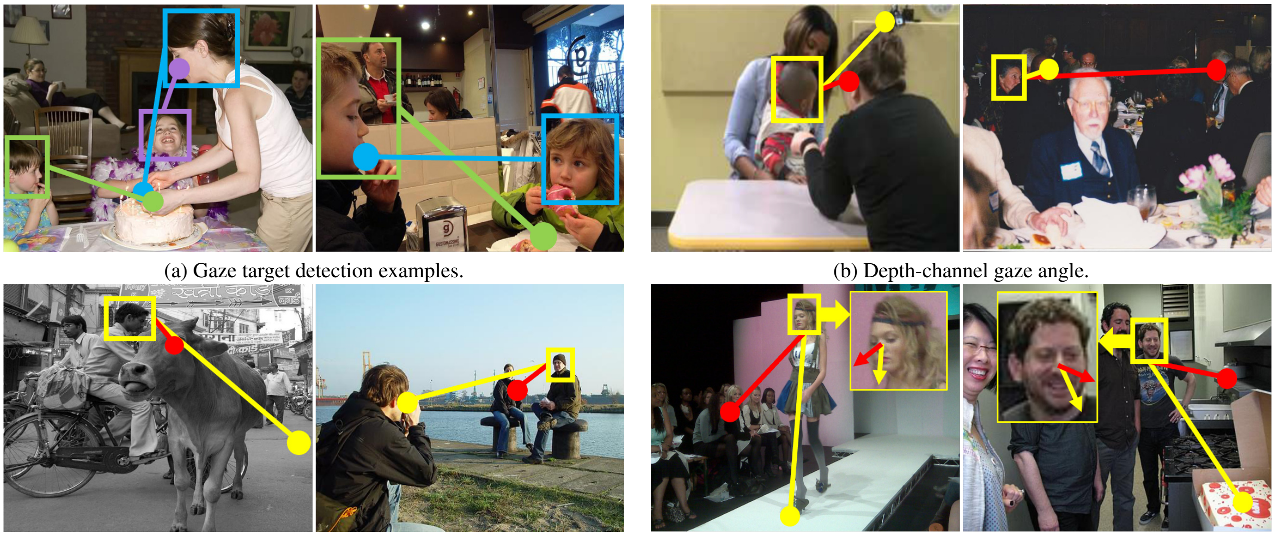
1. Introduction

Gaze cues indicate what a person is interested in, and thus function as an important means to evaluate intentions and predict human behaviors in various social contexts [12]. For these reasons, gaze analysis has widely been used in neurophysiology studies [36, 10], relevant saliency prediction [8, 34], and social awareness tracking [7, 29, 30]. However, most existing works need particular equipment (e.g., eye tracker [13], VR/AR device, or costly RGB-D cameras [42]) or specialized settings (e.g., human-robot interaction [31, 40], or constrained subject locations [32, 1]). In contrast, we are concerned with gaze target detection from a more readily available source in daily life, i.e., a single image in the wild. As depicted in Figure 1 (a), given a scene and the head location for each person (bounding box), we aim to predict where they are looking, including identifying out-of-frame targets and locating inside-frame targets (dot).

Existing methods [37, 28, 5, 6, 46] typically reason about salient objects in the scene conditioned on an estimated gaze orientation. While significant advances have been made, there are still three critical problems to be considered. (1) Most prior works explore the gaze direction in 2D representations and barely encode the depth-channel gaze. They fail to capture whether the marked person is looking forward, backward, or sideward (see Figure 1 (b)). An intuitive solution proposed by Chong *et al.* [5] simply incorporates the 3D gaze as an additional feature channel but does not reasonably combine with scene contexts. Thus, we need an explicit 3D gaze representation coupling with a more effective way to exploit it. (2) Previous methods search for salient objects mainly from 2D visual cues. It is hard for them to capture exact spatial information for lack of scene depth understanding. For instance, two or more candidate objects at different depths may exist along the subject's gaze direction (see Figure 1 (c)). Thus, we need to model the person-relative depth of surroundings for 3D scene understanding. (3) Existing approaches directly learn a mapping function from head features to gaze direction. They are hard to cope with the fixation inconsistency between the eyes and the head (see Figure 1 (d), e.g., facing forward but looking downward). Thus, we need to learn the dependency between eyes and the head for a more accurate prediction.

Based on the above observations, we propose a three-stage scheme to simulate the human gaze inference behavior in 3D space. When one infers the gaze target of another person, he/she first predicts a gaze orientation and then estimates the target by analyzing the 3D geometry of the scene along the gaze direction. Similarly, in the first stage, we learn to estimate a 3D gaze direction from the head image. The predicted gaze is decomposed into a planar gaze on the 2D image plane and a depth-channel gaze. Then we propose the *Dual Attention Module (DAM)* to model the person's depth-aware perspective in the scene as the second stage. Specifically, we aggregate two parallel attention components. One is a *field of view (FOV) generator*

[*] Equal contribution. [†] Corresponding anthor.



(a) Gaze target detection examples.

(b) Depth-channel gaze angle.

(c) Scene depth understanding.

(d) Inconsistent head pose & eyeball orientation.

Figure 1. **Examples of gaze target detection ((a)) and failure cases of existing methods** [37, 28, 5, 6] **((b)-(d)).** Given an image and the ground truth location of a head (bounding box), our approach learns to predict where the person is looking, including identifying out-of-frame targets and locating inside-frame targets (dot). Yellow lines and red lines in (b)-(d) represent **ground truth gaze** and the **predicted gaze**, respectively. (b) shows misjudgement (e.g., located at the front woman in the first example) by existing methods [5, 6] resulting from the lack of depth-channel gaze (e.g., the marked baby is looking backward not forward). (c) shows defect of these methods [5, 28] in scene depth understanding. The first example is a cycling man looking at the side ground, while these methods wrongly predict at the front cow. Existing methodology [5, 37] is incapable to precisely estimate gaze when inconsistency occurs between head pose (e.g., the girl is facing forward in the first example) and eyeball orientation (e.g., the girl is actually looking downward). Our novel network architecture, guided by a Dual Attention Module, well solves above problems and achieves accurate detection results.

for FOV attention, and the other is a *Depth Rebasing* component for depth attention. The *FOV Generator* takes the planar gaze to produce a perspective scope on the image plane. The *Depth Rebasing* segments the scene into front-, middle-, back-grounds based on the pre-estimated depth priori, and subsequently deduces the focused ground from the dept-channel gaze. The depth attention effectively assists in masking 2D-salient interfering objects at unmatched depths in the FOV attention, bridging the 2D cues with 3D relations. In the last stage, we take the dual attention as guidance for two sub-tasks. We formulate the first sub-task, i.e., identifying out-of-frame targets, as a binary classification problem, and the second sub-task, i.e., locating inside-frame targets, as a heatmap regression problem.

Specifically for the first stage, we introduce a coarse-to-fine strategy to estimate 3D gaze in the wild. It is a challenging task for large camera-to-subject distances, diversities in illumination, free subject motions and the resulting variations of appearance in unconstrained images. We present a coarse-grained component to estimate a coarse gaze from the head image. This component does not rely on facial key points but on visible head features. Even with completely occluded eyes and faces, the model outputs a relatively accurate prediction. To cope with the possible inconsistency between the eyes and the head, we take an eye detector and refine the coarse gaze with fine-grained eye features using a transform layer. The model learns the intrinsic correlations between eyes and the face explicitly when eyes are visible. In this way, the proposed 3D gaze estimator can manage in-the-wild images and boost the gaze estimation accuracy.

Our contributions can be summarized as follows:

- We design a novel Dual Attention Module (DAM) that explicitly embodies the person's field of view regulated by depth information in 3D space. To the best of our knowledge, we are among the first to incorporate scene depth understanding in 2D gaze target detection.
- We introduce a coarse-to-fine strategy to estimate 3D gaze orientation. The robust gaze estimator shows competitive generalization properties on images in the wild, especially for eye-included or occluded cases.
- We demonstrate that the proposed method performs favorably against state-of-the-art methods on the GazeFollow [37] benchmark and the VideoAttentionTarget [6] benchmark.

2. Related Work

The key component of our network is the *Dual Attention Module (DAM)*, which filters candidate targets over depth and field of view simultaneously. Naturally, we will introduce related works on gaze target detection, monocular depth estimation, and 3D gaze estimation in this section.

Gaze Target Detection. Some researches explore gaze target detection for specific applications, for example, detecting people looking at each other [30, 29], identifying the common gaze point of multiple human observers [49, 7], estimating the gaze target in several given positions through human-robot interaction tasks [32, 31]. Recent works [37, 28, 5, 6, 46] generalize gaze target detection to images in the wild. These works typically develop a two-stage scheme, in which the gaze direction is estimated first and then combined with a saliency model. Specifically, Recasens *et al.* [37] pioneer tackling the general problem. They publish a large-scale image dataset with annotations of head position and corresponding gaze targets. Lian *et al.* [28] use planar multi-scale gaze direction fields to strengthen gaze supervision on the saliency model. Chong *et al.* [5, 6] extend to cases where the person may look somewhere out of the image. Although efficient to some extent, these works based on 2D visual cues lack scene depth understanding and depth-channel gaze supervision, resulting in ambiguity in fore/background points.

In contrast, the proposed *DAM* explicitly utilizes the depth information and 3D gaze, and produces a target-focused spatial attention map. Our model reliably excludes distractions at improper depth and locate the gaze-at region.

Monocular Depth Estimation. As stated above, recovering scene depth information is greatly needed to clarify the spatial relations in objects. Concretely, the *Depth Rebasing* component takes a priori scene depth map as input. We need to estimate depth from the input RGB image, similar to monocular depth estimation. Monocular depth estimation is an ill-posed problem for a single RGB image can be generated from an infinite number of realistic scenarios [17]. Early works [19, 21, 23] make an effort to exploit some statistically meaningful monocular cues (e.g., perspective, object sizes, and object localization). Recently, CNN based approaches [11, 25, 15, 26] show significant improvements in this field. However, most of these methods are limited in lab-made scenarios or illumination conditions, but can not generalize well to images in the wild. It is theoretically reasonable that a CNN model that has learned plentiful enough modalities can estimate scene depth from a single RGB image in arbitrary scenarios.

In this work, we employ the well-generalized model of Ranftl *et al.* [35], which is trained across diverse datasets and 3D movies, to predict a priori depth map for the scene image. In the proposed *Depth Rebasing* component, we remodel the relative position relationships between the person and external surroundings. Then, we produce a depth-channel perspective scope (depth attention) by referring to the depth-channel gaze. We illustrate the conclusion that the depth information significantly boosts model performance in Section 4.3. Moreover, to study the impact of depth estimation performance on our method, we adopt four state-of-the-art monocular depth estimation methods to predict the priori depth map and use them in our model in Section 4.5.

3D Gaze Estimation. Unlike most existing gaze target detection methods, we learn a 3D gaze direction to represent the gaze behavior. 3D gaze estimation methods can be divided into geometric methods and appearance-based gaze methods. Geometric methods [18, 43, 48], which rely on key points detection, can mostly achieve relatively high accuracy with little data but are restricted to lab settings. Appearance-based methods often learn a more robust and direct mapping function from eyes or face images to gaze directions. Some practices [45, 38, 22, 33] approximate gaze by head pose, allowing for coverage over a wide range of head poses. These methods, however, are hard to predict accurately, since eyeball orientation can differ from head orientation by 35° [41]. Eye-involved methods [47, 14, 4] demonstrate that eye information can boost gaze prediction from the head only. However, eyes will become increasingly occluded at extreme head poses. Eye-involved approaches are restricted to primarily front view rather than free-head conditions.

We propose a well-generalized gaze estimator with a high capacity to cope with natural scenes, including eye occlusion and the possible large gap between eye orientation and head orientation. We employ a coarse-to-fine strategy that approximates a basic gaze by the head pose and refines it with fine-grained eye shifts when eyes are visible.

3. Method

This section presents the architecture of our biologically-inspired model, which consists of three stages, as shown in Figure 2. In the first stage, given the head image of the marked person, we train a 3D gaze estimator to estimate the sight. In the second stage, we propose a *Dual Attention Module (DAM)*, which is the key component to model searching at the gaze direction in the scene, by two parallel attention components. In the third stage, we feed the generated dual attention map stacked with the scene image into a shared backbone for feature extraction. Two heads take the features respectively for two sub-tasks: (1) classify whether the gaze target is within the image or outside; (2) if inside, regress a pixel-level gaze target location.

3.1. 3D Gaze Estimation

We propose a robust coarse-to-fine strategy to estimate the 3D gaze for unconstrained head images in the wild (**yellow panel** in Figure 2). If eyes are invisible, the gaze will be coarsely approximated by the head pose. By contrast, fine-grained eye features are additionally considered for a more accurate gaze direction.

The *Head Pose Extractor* takes the head image as input and estimates a coarse-grained head orientation relative

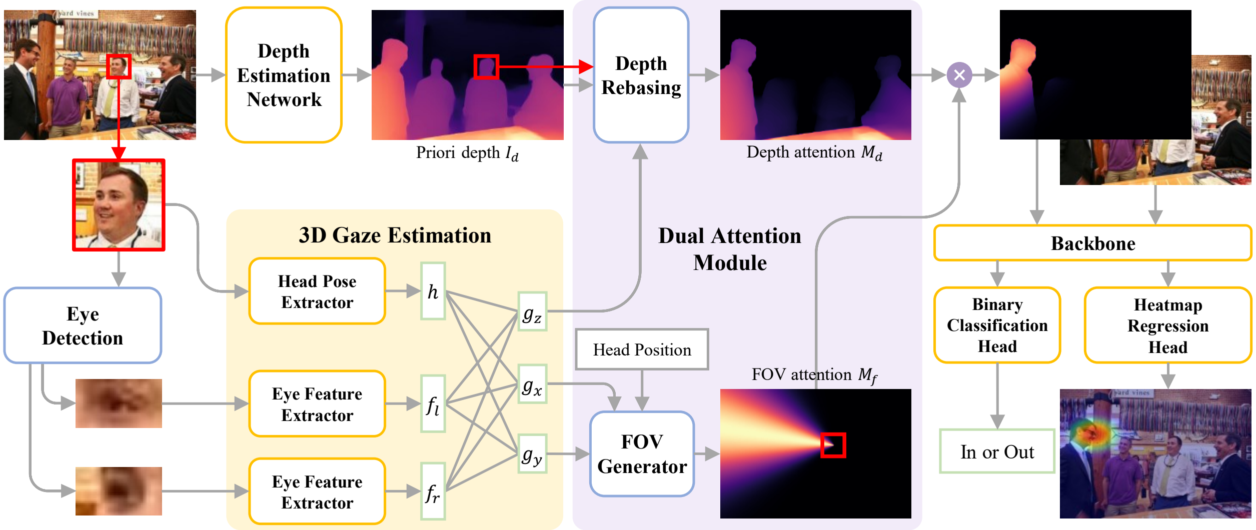


Figure 2. **Architecture for gaze target detection in the wild.** In the first stage (**yellow panel**), given the head image of the marked person, we employ a coarse-to-fine strategy to estimate the 3D gaze orientation. In the second stage (**violet panel**), the proposed *Dual Attention Module (DAM)* produces a field of view by FOV attention and masks interfering objects at unmatched depths by depth attention. In the third stage, we feed the dual attention map stacked with the scene image into a shared backbone for two sub-tasks: (1) classify whether the gaze target is within the image or outside; (2) if inside, regress a pixel-level location.

to the camera coordinate system. We denote the estimated head pose vector as $\boldsymbol{h} = (yaw,\ pitch)$, which separately indicates the horizontal and vertical rotation angle. Then, we algorithmically detect the positions of left and right eyes from the head image (see next Paragraph *Eye Detection*). The cropped eye patches are fed separately into two parallel *Eye Feature Extractors* to generate left-eye feature vector $\boldsymbol{f_l}$ and right-eye feature vector $\boldsymbol{f_r}$.

Inspired by a research work [44] which proves that fully connected layers perform geometric gaze transformation better than hand-crafted gaze transform operation [47], we employ a multi-layer perceptron with one hidden layer (denoted as $\mathcal{F}(\cdot)$) to refine the coarse-grained head orientation with fine-grained eye features:

$$\boldsymbol{g} = \mathcal{F}(\boldsymbol{h},\ \mathbb{1} \cdot (\boldsymbol{f_l} \oplus \boldsymbol{f_r})), \tag{1}$$

where $\mathbb{1} \in \{0, 1\}$ indicates whether eyes are detected, and $\oplus$ is the concatenation operation. Finally, the 3D gaze estimator outputs a normalized gaze vector $\boldsymbol{g}$ in the form of (g_x, g_y, g_z), where xy denote the image plane, and z is for depth direction. A positive g_z indicates that the marked person is looking backward, while negative, forward.

Eye Detection. Since eye annotations are not provided in existing relevant datasets, we need to detect and annotate the locations of the left and right eyes if the subject is facing the camera. In order to make less false positives, we adopt a double-check strategy by referring to 3D head pose and facial landmark detection jointly.

Firstly, given a head image, we use the head pose detector of Kellnhofer *et al.* [22] to extract the subject's facial orientation. If this direction is beyond an appropriate range (60°, determined by experiments) relative to the camera, eyes are assumed self-occluded. We use the facial landmark detector of Bulat *et al.* [2] for a second check. We decide whether the detected eye landmarks are reasonable from the distance between the center of left and right eye landmarks.

To increase the detection accuracy, we rotate and rescale the head image 10 times, and average normalized locations of detected eye landmarks. Rectangles tight around aligned eye landmarks will be extended to fixed-size eye patches and subsequently cropped from the normalized head image.

3.2. Dual Attention Module

The *Dual Attention Module (DAM)* (**violet panel** in Figure 2) learns a target-focused attention map that models a third-view person's gaze tracking behavior in 3D space. DAM can be decomposed into a FOV generator and a depth rebasing component in a mutually binding way. FOV generator generates a planar polarized region to simulate the marked person's field of view. Depth rebasing aims to mask interfering objects out of gaze-depth range.

FOV Generator. The field of view (FOV) can be regarded as an infinitely extended solid cone starting from the head position. Its conic sections at different depths are elliptic slices of different sizes. The projection of these slices on the camera plane will form a sector region. Based on the above analysis, we present the FOV generator.

Denote the generated FOV attention map as $\boldsymbol{M_f}$. Given the head position (h_x, h_y) and the estimated planar gaze direction (g_x, g_y), we first compute the angular difference θ

between the gaze direction and the vector from one point to head position:

$$\theta^{(i,j)} = \arccos(\frac{(i-h_x, j-h_y)\cdot(g_x, g_y)}{\|(i-h_x, j-h_y)\|_2 \cdot \|(g_x, g_y)\|_2}), \quad (2)$$

where (i, j) is the coordinate of each point in $\boldsymbol{M_f}$. Since a smaller θ indicates that the point is more likely to be a fixation point, we assign more weights to points closer to the estimated sight line and less to the farther. The FOV attention map can be generated as:

$$\boldsymbol{M_f}^{(i,j)} = \max(1 - \frac{\alpha\theta^{(i,j)}}{\pi},\ 0), \quad (3)$$

where α decides the angle of view. We empirically set α to 6 and achieve a viewing angle of $60°$.

Depth Rebasing. Beyond planar attention guidance, we design a depth rebasing component to introduce scene depth understanding, helping further select candidate objects.

Firstly, we use a state-of-the-art monocular depth estimator of [35] to extract a priori normalized depth map (denoted as $\boldsymbol{I_d} \in [0, 1]$) from the scene image. Note that lower depth values mean farther from the camera. Based on the depth of the target person, we rebase the depth map by calculating the depth difference map:

$$\boldsymbol{F_d} = \boldsymbol{I_d} - \frac{1}{N_\Omega}\sum_{(i,j)\in\Omega} \boldsymbol{I_d}^{(i,j)}, \quad (4)$$

where (i, j) is a pixel index in the head bounding box Ω that contains N_Ω elements. The mean depth of Ω serves as a threshold for depth rebasing. Naturally, pixels of $\boldsymbol{F_d}$ greater than zero values are considered as the foreground points, and conversely, the background. The middle ground includes those pixels close to zero values. Thus, we obtain three different scene segmentation maps:

$$\boldsymbol{M_{front}} = \max(\boldsymbol{F_d},\ 0), \quad (5)$$

$$\boldsymbol{M_{mid}} = \max(1 - \tau\boldsymbol{F_d}^2,\ 0), \quad (6)$$

$$\boldsymbol{M_{back}} = \max(-\boldsymbol{F_d},\ 0), \quad (7)$$

where τ decides the chosen depth range around head depth and is assigned to 16.

Finally, according to the depth-channel gaze value g_z (provided by 3D gaze estimator), we can select corresponding front/mid/back scene as our depth attention map:

$$\boldsymbol{M_d} = \begin{cases} \boldsymbol{M_{front}}, & g_z \in (-1,\ -\delta) \\ \boldsymbol{M_{mid}}, & g_z \in (-\delta,\ +\delta) \\ \boldsymbol{M_{back}}, & g_z \in (+\delta,\ +1) \end{cases}, \quad (8)$$

where δ is a empirical threshold to determine which scene to choose. We set δ to 0.3 in our experiment.

Dual Attention Attachment. In order to search for salient objects within the field of view at a proper depth, we aggregate the FOV attention map $\boldsymbol{M_f}$ and the depth attention map $\boldsymbol{M_d}$ to generate a dual attention map:

$$\boldsymbol{M_{dual}} = \boldsymbol{M_f} \otimes \boldsymbol{M_d}, \quad (9)$$

where $\otimes$ denotes the element-wise product. In this way, only those points with large activations in both FOV and depth attention maps will be activated in the dual attention map. In other words, interfering objects within the FOV attention but at unmatched depth will be masked with the help of depth attention. Subsequently, the output dual attention map concatenated with the scene image will be fed into a backbone for regression.

3.3. Gaze Target Detection

In this stage, the generated dual attention map and the scene image are concatenated and passed through a backbone to perform feature extraction. The extracted attentive features are shared across the *Binary Classification Head* and the *Heatmap Regression Head*.

In detail, the *Binary Classification Head* consists of two convolutional layers followed by a fully connected layer to classify whether the gaze target is within the image or outside. For the *Heatmap Regression Head*, we apply another two convolutional layers followed by three deconvolutional layers to predict where the target person is looking and output a logits map. The point of the maximum value in this heatmap is our predicted gaze point.

We employ the binary cross entropy loss for the *Binary Classification Head*, denoted as $\mathcal{L}_{Cls}$. The *Heatmap Regression Head* loss function $\mathcal{L}_{Reg}$ is computed with the mean squared error loss. Besides, we introduce the planar angular loss $\mathcal{L}_{Ang}$ at the output of the 3D gaze estimator for a more precise FOV attention map:

$$\mathcal{L}_{Ang} = 1 - \frac{(d_x, d_y)\cdot(g_x, g_y)}{\|(d_x, d_y)\|_2 \cdot \|(g_x, g_y)\|_2}, \quad (10)$$

where (d_x, d_y) is the ground truth planar gaze direction, i.e., the offset between the ground truth head and gaze point positions, and (g_x, g_y) are the 2D projection of the estimated 3D gaze direction. The overall loss function is defined as:

$$\mathcal{L} = \lambda_1\mathcal{L}_{Cls} + \lambda_2\mathcal{L}_{Reg} + \lambda_3\mathcal{L}_{Ang}, \quad (11)$$

where $\lambda_1, \lambda_2, \lambda_3$ are their weight parameters, respectively.

3.4. Implementation details

We implement our model[1] on PyTorch. For 3D gaze estimation in Section 3.1, the two detected eye patches will be cropped from the head image. If not detected, we will fill

[1] https://github.com/Crystal2333/DAM

the patches with black pixels instead. Then the eye patches are resized to 36 × 60 and fed into two parallel ResNet-18 separately, i.e., the *Eye Feature Extractors*. The *Head Pose Extractor* is a ResNet-34 followed by three fully connected layers and takes the cropped head image (resized to 224 × 224) as input. For gaze target detection in Section 3.3, the generated dual attention map and the scene image are both resized to 224 × 224, and passed through a ResNet-50 backbone pretrained on ImageNet [9]. The *Heatmap Regression Head* outputs a heatmap with size 64 × 64.

In the training stage, we first pretrain the 3D gaze estimator on the Gaze360 dataset [22]. Second, the proposed model except the *Binary Classification Head*, is trained on the GazeFollow dataset [37] until convergence. Finally, we finetune the full model on the VideoAttentionTarget dataset [6]. The whole network is optimized by Adam [24], with learning rate of 0.0001 and batch size of 128.

4. Experiments

4.1. Datasets and Evaluation Metrics

We quantitatively evaluate our full model on the GazeFollow dataset [37] and the VideoAttentionTarget dataset [6]. Besides, we evaluate our 3D gaze estimator on the Gaze360 dataset [22]. To estimate more fine-grained gaze direction, we additionally extend the annotation labels of existing datasets with each annotated person's left and right eye bounding boxes. We follow the standard training/testing split of each dataset for fair comparisons.

Datasets. The GazeFollow dataset [37] includes 122,143 images, with 130,339 annotations of head locations and corresponding gaze points. Note that since GazeFollow focuses on gaze targets inside the image, we only use it for the *Heatmap Regression Head*. The VideoAttentionTarget dataset [6] consists of 1,331 video clips collected from various sources on YouTube. The annotations of VideoAttentionTarget include 164,541 frame-level head bounding boxes, 109,574 in-frame gaze targets, and 54,967 out-of-frame gaze indicators. The Gaze360 dataset [22] is a large-scale gaze-tracking dataset in the wild. It contains 172K full-face images captured in various indoor and outdoor environments with 3D gaze annotations across the full gamut of gaze orientations (360°) relative to the camera.

Evaluation Metrics. We adopt the following metrics to evaluate the performance of the proposed model. **AUC:** We use the area under curve (AUC) criteria proposed by Judd *et al.* [20] to assess the confidence of the predicted heatmap. **Dist.:** We evaluate the L_2 distance between predicted gaze point and ground truth gaze annotation. **Ang.:** We compute the angular error between predicted gaze direction and ground truth gaze vector from head position to gaze point. **Out of frame AP:** We utilize the average precision (AP) to assess the accuracy of out-of-frame identifying.

Table 1. **Evaluation on the GazeFollow dataset [37] for single-image gaze target detection.** The numbers in **red** and blue represent the best and second-best results.

Method	AUC ↑	Dist. ↓	Min Dist. ↓	Ang. ↓
Random [37]	0.504	0.484	0.391	69.0°
Center [37]	0.633	0.313	0.230	49.0°
Fixed bias [37]	0.674	0.306	0.219	48.0°
Recasens *et al.* [37]	0.878	0.190	0.113	24.0°
Chong *et al.* [5]	0.896	0.187	0.112	-
Lian *et al.* [28]	0.906	0.145	0.081	17.6°
VideoAtt* [6]	0.921	0.137	0.077	-
Ours	**0.922**	**0.124**	**0.067**	**14.9°**
Human	0.924	0.096	0.040	11.0°

Table 2. **Evaluation on the VideoAttentionTarget dataset [6] for single-image gaze target detection.**

Method	*in frame*		*out of frame*
	AUC ↑	Dist. ↓	AP ↑
Random [6]	0.505	0.458	0.621
Fixed bias [6]	0.728	0.326	0.624
Chong *et al.* [5]	0.830	0.193	0.705
VideoAtt* [6]	0.854	0.147	0.848
VideoAtt [6]	0.860	0.134	0.853
Ours	**0.905**	**0.108**	**0.896**
Human	0.921	0.051	0.925

4.2. Dual Attention Guided Model Evaluation

To evaluate how the proposed model performs the task of single-image gaze target detection, we compare against several state-of-the-art baselines [37, 28, 5, 6] on GazeFollow [37] and VideoAttentionTarget [6]. It is worth noting that since the method of [6] (denoted as VideoAtt) uses a spatio-temporal architecture for dynamic prediction in video, we retain its spatial part (denoted as VideoAtt*) for fair comparisons. Moreover, since GazeFollow aims at in-frame gaze target detection, we remove the *Binary Classification Head* when using this dataset.

Experimental results are summarized in Table 1 and Table 2. We can observe that: (1) In terms of all evaluation metrics, our method surpasses the second-best competitor on both datasets by a large margin, closer towards human performance. We achieve a relative improvement of 9.49% for L_2 distance on GazeFollow, and 19.40% on VideoAttentionTarget. (2) We compare our model's performance across the two datasets. Our approach produces a mean Euclidean error of 0.108 on VideoAttentionTarget, better than that of 0.124 on GazeFollow. A potential reason lies in the fact that VideoAttentionTarget contains higher-resolution and lower-noise images, leading to a more accurate 3D gaze for attention map generation. (3) As to out-of-frame identifying, the results in Table 2 demonstrate that the proposed method outperforms all baselines and gives the best AP of 0.896. This validates our model's superiority in 3D spatial understanding of a monocular image.

We present some examples of our predictions in Fig-

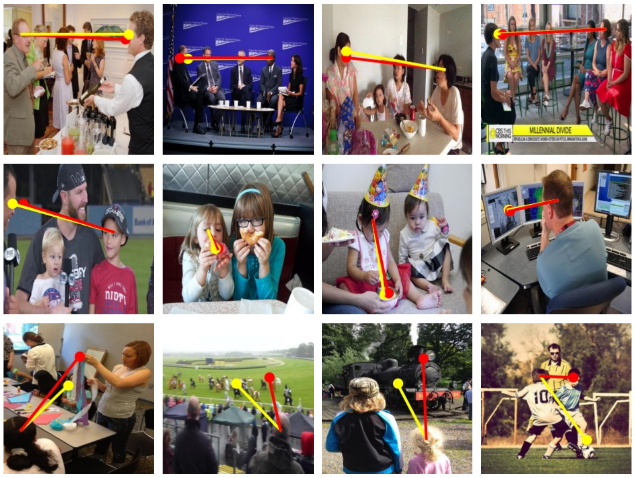


Figure 3. **Qualitative results.** The yellow lines indicate **ground truth gaze**, and the red lines are our **predicted gaze**. Images in the first row illustrate depth-aware scene understanding of our method. The second row show that our model can effectively estimate gaze in the wild. We present some failure cases in the third row.

ure 3. The proposed approach can reliably identify out-of-frame cases and detect gaze targets inside the image. Images in the first row illustrate depth-aware scene understanding of our method. For the first picture, the proposed model accurately locates the gaze-at waiter, regardless of interfering objects (people talking in the background) on the 2D gaze path at improper depth. Pictures in the second row show that our model can effectively estimate accurate gaze in the wild. In spite of head-eye inconsistency (e.g., facing forward but looking sideward) and self-occlusion (e.g., facing backward or downward), our approach still accurately predicts the gaze orientation. We present some failure cases in the third row, where it is difficult to distinguish two or more meaningful objects close together when faces are invisible, even for a human observer.

4.3. Detailed Analysis

Ablation Study. To better understand the effectiveness of different components in the proposed model, we train the following variations on VideoAttentionTarget [6]: 1) DAM-*None*: We remove the Dual Attention Module (DAM), and extract features directly from the scene image through the backbone. 2) Depth-*None*: We remove the depth attention map generated by the depth rebasing component in DAM, and use a uniformly-weighted map instead. 3) FOV-*None*: We replace the FOV attention map in DAM with a uniformly-weighted map. 4) Eye-*None*: We remove the two eyesight feature extraction networks, but approximate gaze by coarse-grained head pose only. 5) Scene-*None*: We remove the concatenation of dual attention map and scene image, and only take the dual attention map as the backbone's input. 6) $\mathcal{L}_{Ang}$-*None*: We initialize the proposed model as before, and train from scratch without angular supervision.

Table 3. **Quantitative results of ablation study on the VideoAttentionTarget dataset** [6].

Method	*in frame*		*out of frame*
	AUC ↑	Dist. ↓	AP ↑
DAM-*None*	0.775	0.237	0.690
Depth-*None*	0.853	0.136	0.841
FOV-*None*	0.837	0.143	0.846
Eye-*None*	0.878	0.124	0.872
Scene-*None*	0.892	0.119	0.875
$\mathcal{L}_{Ang}$-*None*	0.864	0.131	0.859
Ours full	**0.905**	**0.108**	**0.896**

Table 4. **Comparisons between our 3D gaze estimator and other state-of-the-art methods on the Gaze360 dataset** [22].

Method	All 360° ↓	Front 180° ↓	Front Facing ↓
Mean [22]	59.0°	40.5°	19.0°
Ruiz *et al.* [39]	49.3°	30.7°	22.7°
Gaze360 [22]	13.5°	11.4°	11.1°
Ours	**11.3°**	**9.6°**	**9.2°**

The quantitative results are provided in Table 3. It can be clearly seen that all components of our network architecture are necessary for an outstanding result. Not surprisingly, the most contributing performance improvement comes with the Dual Attention Module (DAM). This is determined by the nature of DAM: imitating human gaze behavior by searching salient objects in 3D space. In addition, scene-*None* produces a comparable result to the full model, further proving the importance of dual attention map. Despite this, the stacked scene image still serves as a useful supplement to promoting the final prediction.

Visualization. We provide a visualization of different stages of our network in Figure 4, including the depth attention map, the FOV attention map, the dual attention map, the output heatmap, and the predicted gaze targets. The first three rows indicate our predictions of gaze targets inside the image. For example, in the first row, the annotated man is looking forward at the computer. Our depth attention map emphasizes the front scene relative to the man (table, computer, leg). Our FOV attention map generates the field of view along his gaze direction (computer, woman, painting). After the element-wise product, the generated dual attention map activates those areas with high activations in above two maps, and effectively narrows down the salient region to the computer. Finally, the output heatmap accurately focuses on the fixation point. Besides, we show an out-of-frame example in the last row. Our prediction correctly identifies that the marked man is looking outside the image.

4.4. Gaze Direction Evaluation

We compare the proposed 3D gaze estimation module against other state-of-the-art methods [39, 22] on the Gaze360 dataset [22] to evaluate the performance of gaze estimation in the wild. Prediction errors in degrees are re-

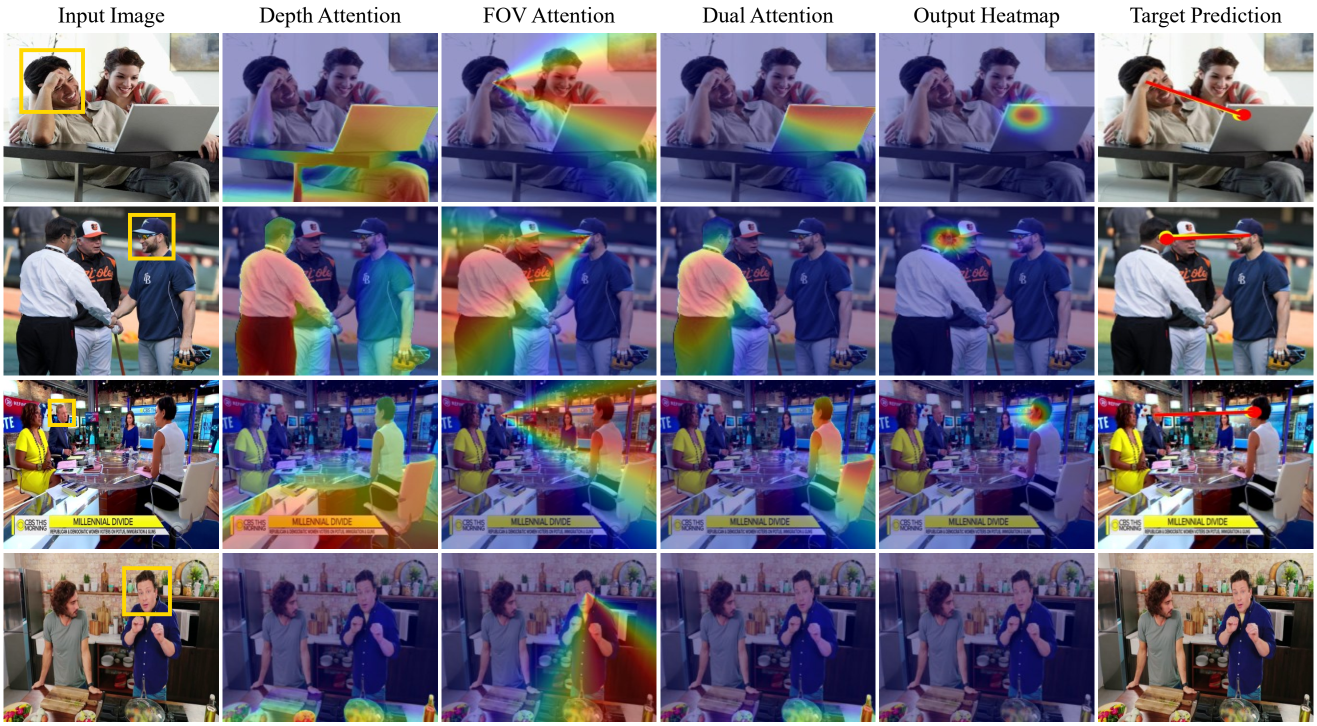


Figure 4. **Visualization results of attention modules.** The first two rows are on GazeFollow [37] and the last two rows are on VideoAttentionTarget [6]. For each row, we show an input image with an annotated head bounding box, the depth attention map, the FOV attention map, and the integrated dual attention map. Besides, the output heatmap, the **prediction result** and the **ground truth** are also provided.

ported in Table 4. Evaluation scopes of all 360°, front 180°, and front facing represent cases where the person is viewing freely, within 90° and 20°, respectively. The first noticeable fact is that our method performs favorably against other competitors, producing an angular error of 11.3°. This demonstrates that it is possible to use this module solely as a 3D gaze estimator. Moreover, we observe that a significant performance boost is achieved in front facing setting. We attain a reasonably low error of 9.2° by explicitly incorporating eyeball orientation with the head pose.

4.5. Priori Depth Evaluation

In order to investigate whether our model is robust to different priori depth maps, we present a comparative study using four recent monocular depth estimation methods [16, 3, 27, 35]. As mentioned in previous works on depth estimation, the three supervised methods (i.e., PSMNet, MC, MiDaS) remarkably exceed the unsupervised method (i.e., MonoDepth) in terms of generalization performance. The quantitative results on VideoAttentionTarget [6] are reported in Table 5. According to the presented performance boost, we reach the conclusion that our model benefits from a better depth map. Even with the depth map obtained by the unsupervised method [16], our model still outperforms the state-of-the-arts at the time of submission. This demonstrates that the proposed model is highly robust to depth maps of different quality.

Table 5. **Depth Analysis.** Comparisons of depth maps of different quality for gaze target detection on VideoAttentionTarget [6].

Depth	*in frame*		*out of frame*
	AUC ↑	Dist. ↓	AP ↑
MonoDepth [16]	0.872	0.127	0.861
PSMNet [3]	0.893	0.120	0.875
MC [27]	0.897	0.116	0.884
MiDaS [35]	**0.905**	**0.108**	**0.896**

5. Conclusion

In this paper, we propose a three-stage method for gaze target detection in the wild. In the first stage, we introduce a coarse-to-fine strategy to provide robust 3D gaze estimation for unconstrained images. In the second stage, we design the Dual Attention Module (DAM) which models the person's depth-aware field of view. In the third stage, we identify out-of-frame gaze targets and locate within-image targets. Extensive quantitative and qualitative evaluations demonstrate that the proposed method performs favorably against the existing approaches.

Acknowledgments. We are grateful for the assistance from Wenbo Bao. This work was supported in part by the National Natural Science Foundation of China under Grant 61831015, Grant 61771305 and Grant U1908210.

References

[1] Ernesto Brau, Jinyan Guan, Tanya Jeffries, and Kobus Barnard. Multiple-gaze geometry: Inferring novel 3d locations from gazes observed in monocular video. In *ECCV*, 2018. 1

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 4

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018. 8

[4] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *AAAI*, 2020. 3

[5] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *ECCV*, 2018. 1, 2, 3, 6

[6] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *CVPR*, 2020. 1, 2, 3, 6, 7, 8

[7] Meir Cohen, Ilan Shimshoni, Ehud Rivlin, and Amit Adam. Detecting mutual awareness events. *IEEE TPAMI*, 2012. 1, 3

[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE TIP*, 2018. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[10] Huiyu Duan, Xiongkuo Min, Yi Fang, Lei Fan, Xiaokang Yang, and Guangtao Zhai. Visual attention analysis and prediction on human faces for children with autism spectrum disorder. *TOMM*, 2019. 1

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 3

[12] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neurosci. Biobehav. Rev.*, 2000. 1

[13] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 1

[14] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*, 2018. 3

[15] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 3

[16] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 8

[17] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *ICLR*, 2020. 3

[18] Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. In *ETRA*, 2006. 3

[19] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH*. 2005. 3

[20] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009. 6

[21] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE TPAMI*, 2014. 3

[22] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019. 3, 4, 6, 7

[23] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 2012. 3

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 6

[25] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 3

[26] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust depth prediction. In *ACCV*, 2018. 3

[27] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 8

[28] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *ACCV*, 2018. 1, 2, 3, 6

[29] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suárez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *CVPR*, 2019. 1, 3

[30] Manuel Jesús Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. Detecting people looking at each other in videos. *IJCV*, 2014. 1, 3

[31] Benoît Massé, Silèye Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE TPAMI*, 2017. 1, 3

[32] Benoit Massé, Stéphane Lathuilière, Pablo Mesejo, and Radu Horaud. Extended gaze following: Detecting objects in videos beyond the camera field of view. In *FG*, 2019. 1, 3

[33] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. *BMVC*, 2018. 3

[34] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019. 1

[35] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. 3, 5, 8

[36] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.*, 1998. 1

[37] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *NIPS*, 2015. 1, 2, 3, 6, 8

[38] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *CVPRW*, 2018. 3

[39] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *CVPRW*, 2018. 7

[40] Boris Schauerte and Rainer Stiefelhagen. "look at this!" learning to guide visual saliency in human-robot interaction. In *IROS*, 2014. 1

[41] John S Stahl. Amplitude of human head movements associated with horizontal saccades. *Exp. Brain Res.*, 1999. 3

[42] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *CVPR*, 2018. 1

[43] Dong Hyun Yoo and Myung Jin Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *CVIU*, 2005. 3

[44] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *CVPR*, 2020. 4

[45] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *CVPRW*, 2017. 3

[46] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *IJCV*, 2019. 1, 3

[47] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *ICCV*, 2017. 3, 4

[48] Zhiwei Zhu and Qiang Ji. Eye gaze tracking under natural head movements. In *CVPR*, 2005. 3

[49] Ning Zhuang, Bingbing Ni, Yi Xu, Xiaokang Yang, Wenjun Zhang, Zefan Li, and Wen Gao. Muggle: Multi-stream group gaze learning and estimation. *IEEE TCSVT*, 2019. 3