Multi-view Gaze Target Estimation

Qiaomu Miao[1], Vivek Raju Golani[1], Jingyi Xu[1], Progga Paromita Dutta[1], Minh Hoai[2], Dimitris Samaras[1]
[1]Stony Brook University [2] The University of Adelaide

Abstract

This paper presents a method that utilizes multiple camera views for the gaze target estimation (GTE) task. The approach integrates information from different camera views to improve accuracy and expand applicability, addressing limitations in existing single-view methods that face challenges such as face occlusion, target ambiguity, and out-of-view targets. Our method processes a pair of camera views as input, incorporating a Head Information Aggregation (HIA) module for leveraging head information from both views for more accurate gaze estimation, an Uncertainty-based Gaze Selection (UGS) for identifying the most reliable gaze output, and an Epipolar-based Scene Attention (ESA) module for cross-view background information sharing. This approach significantly outperforms single-view baselines, especially when the second camera provides a clear view of the person's face. Additionally, our method can estimate the gaze target in the first view using the image of the person in the second view only, a capability not possessed by single-view GTE methods. Furthermore, the paper introduces a multi-view dataset for developing and evaluating multi-view GTE methods. Data and code are available at https://www3.cs.stonybrook.edu/~cvl/multiview_gte.html.

1. Introduction

Gaze Target Estimation (GTE) is an important problem with applications in areas such as social behavior analysis [13, 52], human-machine interactions [1, 32], and mental disorder diagnosis [17, 63]. Earlier works studied gaze behaviors using specialized equipment like eye trackers [16, 40] or head-mounted cameras [15, 46], which are expensive and intrusive. Recent advances in deep learning [23, 26, 36, 62] have facilitated the development of GTE models to estimate gaze in the wild using ordinary scene cameras, thereby broadening their range of applications.

Several methods have been proposed for GTE using ordinary scene cameras. However, as shown in Fig.1, existing methods struggle with images in which the subject's face is not visible to the camera and multiple potential targets exist.

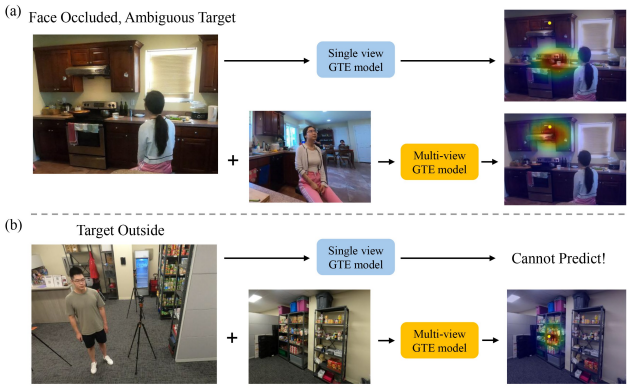


Figure 1. Benefits of multi-view GTE. Single-view GTE models struggle with input where the subject face is occluded, and cannot predict the target location outside the frame. In contrast, a multi-view GTE model can leverage another view's information to improve GTE accuracy, and predict the gaze target across views.

In addition, single view GTE methods only function when both the person and their gaze target are visible in the image. If the gaze target is outside the image, these methods cannot function at all. These limitations make current methods restrictive due to the limited field of view of a camera.

Using multiple cameras provides a solution to these limitations. Compared to single-camera systems, multi-camera setups provide broader coverage and multiple perspectives of both human faces and the scene background. This enables more accurate gaze estimation by providing a clearer view of the face, and allows predicting gaze targets that may appear in a different camera view from the subject. Furthermore, multi-camera setups are widely utilized in many environments, such as supermarkets and lecture halls, that could benefit from non-intrusive gaze target estimation (GTE).

However, there are challenges in developing a method that effectively leverage multiple cameras for GTE. Due to perspective changes between different views, directly combining input images or extracted features without accounting for the geometric relationship between them, is not beneficial. Meanwhile, in real-world applications, explicit and complete 3D reconstruction of the scene and subject is not always feasible, as the views often have limited or no over-

lap. Even if 3D reconstruction is possible, performing it for every input—particularly when the subject or other people move in the scene—would be memory and time-intensive.

In this paper, we introduce the first multi-view GTE model that effectively and efficiently leverages information from multiple camera views. Our model builds upon a transformer-based single-view GTE framework and extends it to process a pair of camera views. It incorporates a Head Information Aggregation (HIA) module that leverages the head appearance information and the geometric relationship between both views to enhance the head embedding and improve gaze direction estimation. The estimated gaze vectors are then processed by an Uncertainty-based Gaze Selection (UGS) module, which selects the more reliable gaze vector from the two views, replacing the predicted vector in the less reliable view. Additionally, the Epipolar-based Scene Attention (ESA) module integrates scene background information from different perspectives. Altogether, these modules enhance the GTE capability with both the gaze and scene information from multiple views.

Compared to single-view methods, our model leverages an additional view to achieve significant improvements, especially when the additional view captures the person's head appearance. Our model also addresses a unique scenario that existing single-view GTE models cannot handle: when the gaze target is visible in one view but the person is only visible in the second view. This scenario is challenging as triangulation cannot be used to infer the absolute depth of the person and scene due to little view overlap. To address this issue, we estimate absolute depth by comparing monocular depth maps against a pre-reconstructed 3D scene, generated from a multi-view reconstruction model [64] prior to training. This allows us to estimate the absolute depth for all new inputs in a scene by applying the 3D reconstruction model only once.

To train and evaluate our method, we introduce the first dataset for multi-view GTE: the Multi-View Gaze Target (MVGT) dataset. This dataset was collected across four real-world scenes, featuring 28 subjects with precise gaze target annotations. The images were captured simultaneously from multiple calibrated cameras positioned in the scene. We also introduce a data collection protocol that can collect data non-intrusively and obtain precise gaze targets without introducing artifacts.

In summary, our main contributions are: (1) the first exploration of the multi-view GTE task; (2) a novel GTE model that effectively leverages the human head and scene information from multiple views, surpassing current state-of-the-art single-view models; (3) the MVGT dataset, featuring images captured synchronously by multiple calibrated cameras with precise target annotations, along with a data collection protocol to collect accurate gaze target annotations without creating artifacts.

2. Related Works

Gaze target estimation was first investigated in [49], which introduced the GazeFollow dataset. Lian et al. [38] generated 2D direction fields for GTE, while Chong et al. extended the task to predicting whether the target is located in the image [8] and GTE in videos [9]. Later models adopted additional modalities such as depth [3, 14, 42, 56, 65] and human pose [3, 19, 66] for improvement. Recent methods compute 3D field-of-view (FoV) heatmap as gaze target priors by predicting a 3D gaze vector and using depth input [27, 28, 54]. Additional works leveraged transformer architectures for GTE [53, 55], and jointly predicting the head location and gaze targets [57, 59]. Some works explored unified GTE and social gaze prediction [20, 21], or GTE using fewer labels [43, 58]. All these methods focus on GTE in a single camera view. Several works have explored solutions for out-of-frame gaze targets. Recasens et al. [50] introduced a dataset and a model to predict gaze targets in future video frames based on a person seen in the current frame. Li et al. [37] investigated GTE in 360° images, and Yu et al. [71] learned a joint embedding for first- and third-person frames. However, they did not leverage the explicit geometric relationships between views from the calibrated camera parameters. Furthermore, no previous work has explored GTE with multiple third-person views, which is increasingly prevalent in real-world applications.

Multi-view settings have been extensively studied in tasks such as 3D human pose estimation [11, 24, 31, 48, 60] and 3D reconstruction [6, 18, 33, 44, 64, 68]. These methods estimate the 3D locations of human body/hand keypoints or object/scene point clouds and generally require large overlap between different views. In the gaze domain, a few recent works [4, 7, 25] have investigated improving gaze direction estimations using multi-view input on specialized datasets [47, 72]. However, these methods impose several restrictions on the input data: the subject's head cannot turn away from the camera, the face must be rectified, and the eyes should be clearly visible. These limitations prevent these methods from being directly applicable to GTE. Nonaka et al. [45] introduced a multi-view dataset for 3D gaze estimation, but the subjects wore intrusive eye-tracking glasses, and their proposed model does not consider interactions between views. To our knowledge, no previous work has investigated multi-view GTE, and no dataset is available for training and evaluating this task.

3. Multi-View Gaze Target (MVGT) Dataset

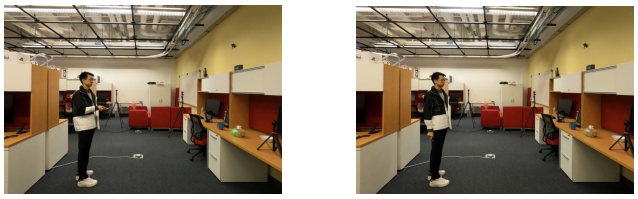
For the development and evaluation of multi-view GTE models, we collected a dataset named MVGT. The dataset contains 13,686 images of size 4000×3000 with 2,281 unique gaze targets, resulting in 68,430 pairs when pairing camera views for multi-view GTE. The images were cap-

(a) (b)

Figure 2. Dataset samples and information. (a) Example images and annotations of the subject's head location (green bounding box) and the gaze target (yellow dot) from all 6 cameras. (b) Dataset information including the density of the gaze target in the entire dataset, the camera setup in an example scene, and the number of cameras that the gaze target appears.

tured by 6 synchronized GoPro Hero 8 cameras from four scenes: a university commons room, a small convenience store, a kitchen, and a research lab space. Images are distributed approximately evenly across scenes. The cameras were controlled by a cellphone via Bluetooth for simultaneous image capturing, and were calibrated before data collection (see Supplementary). Fig. 2 shows examples of images and annotations. In total, 28 subjects participated in the data collection, with 7 subjects in each scene. We provide detailed annotations per image, including the subject's head bounding box (detected with a YOLOv5 head detector [29, 61]), the 2D gaze-target coordinate, and a visibility label. Human annotators mark the laser point as gaze target location and classify each as "target inside," "target outside," or "target occluded," representing 44.3%, 47.5%, and 8.2% of the dataset, respectively. Fig. 2(b) shows most targets are simultaneously visible in at least two camera views.

We also introduce a non-intrusive data collection protocol for obtaining precise gaze targets without artifacts. During collection, each subject was instructed to point to a random gaze target with a handheld laser pointer, then turn off the pointer while maintaining their gaze on the target. By comparing images with and without the laser point (Fig. 3), we accurately determined the gaze target location. This approach is both cost-effective and more precise than letting annotators subjectively infer the targets [9, 49]. The dataset only contains images without laser points, while images with laser points were used solely to establish the ground truth. This protocol avoids artifacts on the gaze object compared to using an image-inpainting model to remove the laser point [28]. Meanwhile, it is easily applicable to new scenes and allows for future extension of the dataset. To introduce pose variability, subjects were asked to stand for half of the samples and sit for the other half.



(a) taken with laser pointer on (b) taken with laser pointer off

Figure 3. Images of a subject looking at the same gaze target, one with the laser pointer on (a) and one with the laser pointer off (b).

4. Multi-view Gaze Target Estimation

In this section, we describe our method for multi-view GTE. To maximize applicability, our model processes a pair of images as its basic operation, with the potential to analyze more images by aggregating results from multiple pairs. For a pair with a primary view and a reference view, the method predicts the gaze target location and in/out probabilities for each view by leveraging information from the other view.

4.1. Processing pipeline

The processing pipeline of our framework is shown in Fig. 4. The input consists of a pair of images from the primary and reference views, $\mathbf{I}_1, \mathbf{I}_2 \in \mathbb{R}^{3\times H\times W}$, and the head bounding boxes of the subject in each view $\mathbf{x}_1^{box}, \mathbf{x}_2^{box} \in \mathbb{R}^4$. We assume that the camera intrinsic parameters, $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{3\times 3}$, and extrinsic parameters, $\mathbf{R}_1, \mathbf{R}_2 \in \mathbb{R}^{3\times 3}$ and $\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{R}^{1\times 3}$, are also known for the two views.

Since both images undergo similar processing steps, we will omit the view index for brevity unless specified otherwise. Given an image and the subject's head box, we crop out the head image $\mathbf{I}^h$, also creating a binary mask $\mathbf{M}^h \in \mathbb{R}^{1\times H\times W}$ of the subject head location in the image. The head image is first processed by the Head In-

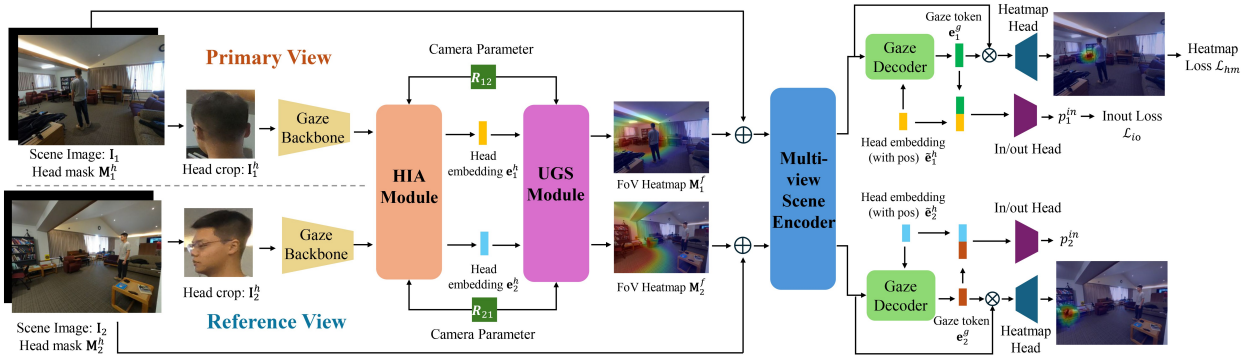


Figure 4. Overall framework. Our method takes images from a pair of camera views as input. The head images are processed by the HIA and UGS module to output enhanced head embeddings and generate FoV heatmaps from more accurately estimated gaze vectors benefiting from multi-view input. Camera parameters are provided as input to HIA and UGS to encode the geometric relationship and transform gaze vectors between views. The FoV heatmaps are input as priors to the multi-view scene encoder with the scene images. The output scene features and head embeddings are fed to a gaze decoder followed by output heads to output the gaze target heatmap and in/out probabilities.

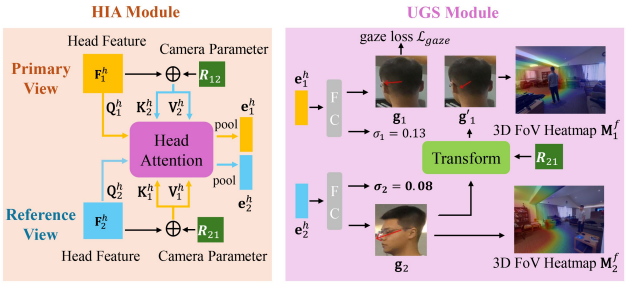


Figure 5. Structures of HIA and UGS. HIA aggregates head information from the other view using the head appearances and geometric relationships between views. UGS selects the more reliably predicted gaze vector based on the output uncertainty scores and transforms it to the other view using the camera parameters.

formation Aggregation (HIA) module, which outputs the head embedding $\mathbf{e}^h \in \mathbb{R}^d$ after interactions across views. The head embedding is then processed by the Uncertainty-based Gaze Selection (UGS) module to output a more accurate 3D gaze vector $\mathbf{g}$ benefiting from multiple views. Field-of-view (FoV) heatmaps $\mathbf{M}^f \in \mathbb{R}^{1\times H\times W}$ are computed from the predicted gaze vectors, serving as gaze target priors. These heatmaps are concatenated with the scene images and head masks for each view and are fed into a multi-view scene encoder that contains two Epipolar-based Scene Attention (ESA) modules, outputting the scene feature $\mathbf{F}^s \in \mathbb{R}^{c\times h\times w}$. The scene features and head embeddings are then input to a gaze decoder and the output heads, which produce the gaze target heatmap $\mathbf{H} \in \mathbb{R}^{1\times 64\times 64}$ and the probability p^{in} that the target is located in the image.

4.2. Head Information Aggregation Module

We propose the HIA module that leverages head images from both camera views to enhance the head embeddings and improve gaze vector estimations. First, the head feature $\mathbf{F}^h \in \mathbb{R}^{c_0\times h_0\times w_0}$ is extracted from the head image $\boldsymbol{I}^h$ using a ResNet-18 backbone. Then, $\mathbf{F}^h$ is flattened into tokens of $\mathbb{R}^{c_0\times h_0 w_0}$. The tokens enter the Head Attention module as queries, which is a cross-attention block to aggregate information from the keys and values of the other view. Take the primary view as an example:

$$\tilde{\mathbf{F}}_1^h = \mathbf{F}_1^h + CrossAtt(\mathbf{Q}_1^h, \mathbf{K}_1^h, \mathbf{V}_1^h), \quad (1)$$

where $\mathbf{Q}_1^h = W_q^h(\mathbf{F}_1^h), \mathbf{K}_1^h = W_k^h(\mathbf{F}_2^h \oplus \mathbf{R}_{21})$, $\mathbf{V}_1^h = W_v^h(\mathbf{F}_2^h \oplus \mathbf{R}_{21})$, and W_q^h, W_k^h, W_v^h are the linear projection layers. Following [41], we concatenate the keys and values with the relative camera rotation $\mathbf{R}_{21}$ to incorporate the geometric relationship information between views, where $\mathbf{R}_{21} = \mathbf{R}_1\mathbf{R}_2^{-1}$. As shown in the Supplementary, both the relative rotation and the head appearance from the other view are vital for performance improvement. $\tilde{\mathbf{F}}_1^h$ are average-pooled to produce the head embedding $\mathbf{e}_1^h$. The reference view undergoes a similar process to yield $\mathbf{e}_2^h$.

4.3. Uncertainty-based Gaze Selection Module

Although the HIA module helped the information propagation, the two input views can still predict gaze vectors of different qualities. Therefore, we propose the UGS module which picks the more reliably predicted gaze vector from the two views to generate better-quality FoV heatmaps. To find out the more reliable view, we extend the gaze estimator in GTE models to predict an uncertainty score σ along with the gaze vector $\mathbf{g}$. We make the model learn the *Aleatoric Uncertainty* [35] of the input image with the uncertainty-aware loss $\mathcal{L}_{gaze}$ between the predicted gaze vector $\mathbf{g}$ and the ground-truth $\hat{\mathbf{g}}$, similar to [10]:

$$\mathcal{L}_{gaze}(\mathbf{g}, \hat{\mathbf{g}}) = \frac{1}{2\sigma^2}\left(1 - \frac{\mathbf{g}\cdot\hat{\mathbf{g}}}{||\mathbf{g}||_2||\hat{\mathbf{g}}||_2}\right) + \frac{1}{2}log(\sigma^2), \quad (2)$$

where the first term is the cosine angular loss suppressed by the uncertainty score, and the second is a regularization

term. The ground truth gaze vector $\hat{\mathbf{g}}$ can be obtained from the pseudo point cloud using the ground truth gaze coordinate, which we will show later. As shown in previous works [10, 35], the model trained with this form of uncertainty loss tends to predict a larger σ for the predictions with larger errors. Therefore, we select the view predicted with lower σ and replace the other view's prediction with the one with lower uncertainty using camera transformation:

$$\mathbf{g}'_j = \mathbf{R}_j \mathbf{R}_i^{-1} \mathbf{g}_i, \quad i, j \in \{1, 2\}, \sigma_i < \sigma_j \tag{3}$$

The gaze vectors are used to generate the FoV heatmaps with the monocular depth maps $\boldsymbol{D}$. The camera intrinsic matrix $\mathbf{K}$ is represented as:

$$\mathbf{K} = \begin{bmatrix} f^x & 0 & c^x \\ 0 & f^y & c^y \\ 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

Given a pixel coordinate (u, v) in the image, the 3D point cloud $\boldsymbol{P}^{(u,v)} = [P^x, P^y, P^z]$ in the camera coordinate system of the input view is represented as:

$$\begin{aligned} P^x &= (u - c^x)/f^x * \boldsymbol{D}(u, v), \\ P^y &= (v - c^y)/f^y * \boldsymbol{D}(u, v), \\ P^z &= \boldsymbol{D}(u, v). \end{aligned} \tag{5}$$

Therefore, the 3D vector from any pixel (u, v) to the subject's eye (e_x, e_y) can be obtained as $\mathbf{V}^{(u,v)} = \boldsymbol{P}^{(u,v)} - \boldsymbol{P}^{(e_x,e_y)}$, and the ground truth gaze vector $\hat{\mathbf{g}}$ is $\mathbf{V}^{(gt_x,gt_y)}$. Similar to [54], they are computed from "pseudo" point clouds, as $\boldsymbol{D}$ contains relative depth values from a monocular depth estimation model, which is up to a scale and shift factor to the absolute depth $\boldsymbol{D}^*$. However, when we use a depth estimation model that has low depth distortion and shift [69, 70], we can ignore the shift term, and $\mathbf{V}^{(u,v)}$ will only be the same due to the elimination of the scale factor based on Eq. (5). Based on the pseudo point clouds, we obtain the value at (u, v) in the FoV heatmap:

$$\mathbf{M}^f(u, v) = max(0, \frac{\mathbf{V}^{(u,v)} \cdot \mathbf{g}}{||\mathbf{V}^{(u,v)}||_2 ||\mathbf{g}||_2}). \tag{6}$$

We adopt an exponential decay scheme if the value on the FoV heatmap is lower than 0.9, as in [54]. As shown in Supplementary, we observe larger σ values for gaze vectors with larger errors, and by selecting the more reliable view, UGS improves the FoV heatmap and final prediction, especially when the original predicted vector has a large error.

4.4. Multi-view Scene Encoder

The FoV heatmap $\mathbf{M}^f$ is used as priors and concatenated with the scene image $\mathbf{I}$ and the head mask $\mathbf{M}^h$ for both views, and input to the multi-view scene encoder. As shown in Fig. 6, the multi-view scene encoder consists of a ViT-base [12] encoder along with two Epipolar-based Scene Attention (ESA) modules for propagating scene information between views. The transformer blocks are shared between views. In the ESA module, each token in one view attends to the feature tokens uniformly sampled along the epipolar line of the other view. Take a token $\mathbf{p}$ with a coordinate of (u, v) in the primary view feature $\mathbf{F}_1^s$ as an example:

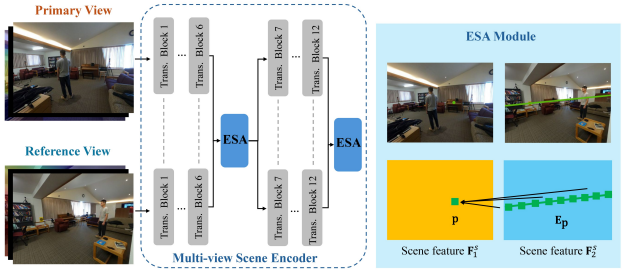


Figure 6. Structure of the multi-view scene encoder and ESA module. The transformer blocks are shared between the two input views. Two ESA modules are inserted in the transformer encoder. In ESA, each feature token attends to multiple tokens sampled along the epipolar line in the other view.

$$\mathbf{p}' = \mathbf{p} + CrossAtt(W_q^s(\mathbf{p}), W_k^s(\mathbf{E_p}), W_v^s(\mathbf{E_p})), \tag{7}$$

where $\mathbf{E_p} = \{\mathbf{q}_l\}_{l=1}^N$ are the feature vectors sampled along the epipolar line of $\mathbf{p}$ in $\mathbf{F}_2^s$. The epipolar line is computed from the fundamental matrix: $l^{epi} = \mathbb{F}\mathbf{x}$, where $\mathbf{x} = [u, v, 1]^T$ and $\mathbb{F}$ is computed from the camera parameters using multi-view geometry [22]. Epipolar attention has been used in multi-view tasks including 3D reconstruction [30, 67] and pose estimation [24], where it enhances the feature with another view's appearance associated in 3D, especially in the case of occlusion in the primary view. In our case, it also saves computation and memory compared to dense cross-attention on the higher-resolution scene features. We observed that ESA improves GTE performance when the other view contains the gaze target.

4.5. Output and Losses

In the final stage, the head embedding $\tilde{\mathbf{e}}^h$ and the scene feature $\mathbf{F}^s$ from the scene encoder is fed to a gaze decoder. The head embedding $\tilde{\mathbf{e}}^h$ is $\mathbf{e}^h$ from the HIA module added with a positional encoding $\mathbf{e}^{pos}$, which is mapped from the normalized head center coordinate with an MLP. The head embedding $\tilde{\mathbf{e}}^h$ is used as a query while $\mathbf{F}^s$ serves as the key and value. The gaze decoder outputs a gaze token $\mathbf{e}^g$. For gaze target estimation, $\mathbf{e}^g$ is element-wise multiplied with each token in $\mathbf{F}^s$, and fed into a heatmap head to output the gaze target heatmap $\mathbf{H}$. On the other hand, the head embedding $\tilde{\mathbf{e}}^h$ is also concatenated with the gaze embedding $\mathbf{e}^g$ to output the probability of the target located in the frame p^{in}.

The overall training loss is formulated as:

$$\mathcal{L} = \alpha \mathcal{L}_{hm} + \beta \mathcal{L}_{io} + \lambda \mathcal{L}_{gaze}, \tag{8}$$

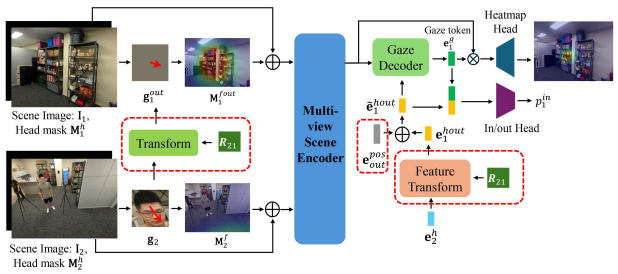


Figure 7. The modified model structure for cross-view GTE. The parts encircled in red are the modules we modified/added for the cross-view gaze estimation cases. Modules for estimating the gaze vectors are the same as above and omitted.

where $\mathcal{L}_{hm}$ is the Mean-Square-Error (MSE) loss between the predicted heatmap $\mathbf{H}$ and the ground truth heatmap $\hat{\mathbf{H}}$ which is a Gaussian centered at the ground truth gaze target coordinate. $\mathcal{L}_{io}$ is a binary cross-entropy loss between p^{in} and the ground truth in/out label. For gaze targets labeled with "occlusion", we assign them to the "inside" class for the in/out task, as these targets remain within the frame but are merely obscured by other objects.

4.6. Cross-view Estimation

Our method can be extended to address cross-view estimation, where the primary view only contains the gaze target, and the subject is only visible in the reference view. In this case, we need to generate the FoV heatmap $\mathbf{M}_1^f$ using the eye location $\boldsymbol{P}_2^{(e_x,e_y)}$ and the predicted gaze vector $\mathbf{g}_2$ from the reference view. However, as mentioned in Sec. 4.3, the point clouds were "pseudo" point clouds of which the depth values are up to a scale and shift to the real absolute depth, so the actual 3D location of $\boldsymbol{P}_2^{*(e_x,e_y)}$ is not known. Therefore, the eye location $\boldsymbol{P}_1^{*e}$ in the primary view's camera coordinate system cannot be directly obtained.

We use a Multi-view Stereo model to reconstruct the scene in 3D to obtain the absolute depth values. As in the cross-view cases, the input pair of view usually has little or no overlap (Fig. 7), we assume that a set of images from all six cameras have been obtained before the capturing of input data, so that the 3D scene can be reconstructed. We use, e.g., six images collected in extrinsic parameters calibration to reconstruct the 3D scene with a SOTA multi-view reconstruction model, Dust3R [64]. When inputting the camera parameters calibrated in real-world metrics, Dust3R can generate depth estimations that are very close to the absolute depth values by optimizing a reconstruction loss [64]. After obtaining the absolute depth values for both views, for each image, we estimate the scale and shift between the monocular depth map in each new input image and the absolute depth in the reconstructed scene. In this way, $\boldsymbol{P}_2^{*(e_x,e_y)}$ can be obtained, along with $\boldsymbol{P}_1^{*e}$ via camera transformation, from which $\mathbf{M}_1^f$ can be generated. We provided the detailed procedure for scale and shift estimation and some reconstruction examples in Supplementary.

On the other hand, we also updated the later part of the model for cross-view GTE (Fig. 7). As the primary view does not contain the subject's head, we add a feature transform module to generate the head embedding $\mathbf{e}_1^{hout}$ from the reference view. The feature transform module is a two-layer MLP, which takes the concatenated input of the reference view's head embedding $\mathbf{e}_2^h$ and the relative camera rotation R_{21}. Meanwhile, we add a learnable "outside embedding" $\mathbf{e}_{out}^{pos}$ to $\mathbf{e}_1^{hout}$ as the positional embedding to get the final head embedding for primary view. In the experiments, we fine-tuned our model trained for the ordinary multi-view setting above on the view pairs that fall in the cross-view GTE category. As will be seen, our method can predict the cross-view gaze targets well.

5. Experiments

5.1. Experimental setups

Implementation details. We process the scene image at a resolution of 512×384 and the head crop at a resolution of 224×224. The gaze backbone is a ResNet-18 [23] pre-trained on Gaze360 [34], while the transformer part of the multi-view scene encoder is a ViT-base [12] model pre-trained with MultiMAE [2]. We used Metric3D [70] for monocular depth estimation. In ESA, we sampled 48 feature vectors along the epipolar line for each query token.

Training and evaluation. To simulate the real-world application of applying a trained model to a new scene with multiple camera setups, we performed leave-one-scene-out cross-validation in our experiments. In the experiments, all models are first trained on the GazeFollow dataset [49], fine-tuned on three scenes of our MVGT dataset, and then validated on the left-out scene. Each of the four scenes was left out for validation, and the results averaged across scenes are reported. For our method, we train the single-view version of the model on GazeFollow and fine-tune the whole model on the MVGT dataset. We used a batch size of 40 pairs of views. The specific parameter settings for validating each scene are described in the Supplementary.

Evaluation metrics. We use the normalized L_2 Distance (**Dist.**) between the predicted gaze target coordinates and the ground truth gaze target annotations to evaluate GTE performance. AUC is not used because it is more suited for evaluating the alignment of predicted heatmaps with group-level annotations, making it less suitable for datasets with single-point annotations, as explained in [55]. We use **AP** to evaluate the performance on in/out classification.

Method	Head Visible				Head Not Visible			
	Target Visible		Target Not Visible		Target Visible		Target Not Visible	
	Dist. ↓	AP ↑	Dist. ↓	AP ↑	Dist. ↓	AP ↑	Dist. ↓	AP ↑
Random	0.451	0.555	0.456	0.546	0.464	0.502	0.457	0.564
Center	0.259	/	0.261	/	0.272	/	0.253	/
Chong [9]	0.159	0.855	0.157	0.862	0.191	0.792	0.174	0.866
Miao [42]	0.141	0.886	0.140	0.892	0.164	0.831	0.163	**0.886**
Tafasca* [54]	0.149	0.893	0.148	0.895	0.166	0.831	0.154	0.884
Ours-Single	0.151	0.877	0.148	0.878	0.179	0.758	0.154	0.855
Ours	**0.129**	**0.909**	**0.122**	**0.912**	**0.161**	**0.836**	**0.152**	0.868

Table 1. Comparing with single-view GTE methods on test data divided based on the head and target visibility in the reference view. Best numbers are marked as bold and 2nd best are underlined. Our method shows large improvement when the subject head is visible, while maintaining comparable performance when the head is not visible.

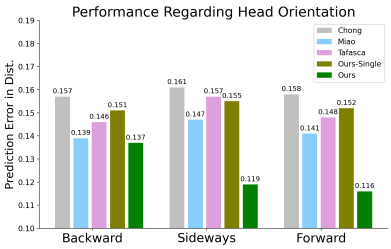


Figure 8. Results regarding different head orientations in the reference view. Our method shows a much larger improvement when the subject face is half/fully visible.

5.2. Comparison with Single-View Methods

We compare our multi-view GTE method with SOTA single-view GTE methods, and the single-view baseline version of our method (*Ours-Single*). In *Ours-Single*, we exclude the HIA, UGS, and ESA modules to eliminate interaction between views, and the gaze estimator only outputs a single gaze vector without the uncertainty score. We evaluate the model on primary views, treating the reference view as additional input. To ensure a fair comparison with single-view methods, when a primary view image (e.g., Camera1) is paired with different cameras (Camera2, Camera3, etc.) in multiple pairs, we evaluate on the primary view for each pair and average the scores for the same primary view image. This ensures the same total number of testing samples, enabling direct comparison between methods.

We experimented with the following baseline methods: *Random* generates heatmap response and in/out probability randomly in a [0,1] uniform distribution. *Center* generates a heatmap always at the image center. *Chong* [9] is a popular GTE model that only uses RGB input. *Miao* [42] uses monocular depth maps as direct input for GTE. *Tafasca* [54] is a recent model that generates a FoV heatmap from a predicted 3D gaze vector and a monocular depth map. We re-implemented the model and were able to reproduce its performance on the GazeFollow dataset (See Supplementary).

To better understand the benefits of using an additional view, we divide the test cases into four categories based on the visibility of the subject's head and the gaze target in the reference view. The number of samples for each category is shown in Supplementary. In Tab.1, our method shows significant improvement when the reference view contains the head, while maintaining comparable performance with the best baselines when it does not. The improvement is most pronounced when the reference view includes the head but not the gaze target, which typically occurs when the subject is facing toward the camera with a clear face appearance.

The benefits of the reference view are further highlighted by comparing to our method without the reference view and multi-view processing (Ours-single). The advantage of the reference view is clear, except when neither the head nor the target is visible in it, as expected, since little additional information can be gained in this case. As shown in Fig. 9, our methods can leverage the useful face information from the reference view and obtain higher-quality FoV heatmaps (Rows 1-2). It also benefits from the appearance of the gazed objects in a different perspective (Row 3). This demonstrates the effectiveness of our method in leveraging two-view input. Using more than two views can further improve performance, as shown in the Supplementary.

Performance Regarding Head Orientation. We further investigate the effect of the face visibility when the reference view contains the subject's head, by analyzing the head orientation. We use a head pose estimation model [51] to obtain the yaw angle from the head image. The head pose estimator does not perform well when the head is facing away from the camera, so we combine it with a face keypoint estimator [5] to divide the head orientations into three categories: backward: $<$ 30 face keypoints are detected; sideways: $\geqslant$ 30 face keypoints are detected and the head pose yaw angle $\geqslant 55°$; forward: the remaining images.

The results are shown in Fig. 8. As illustrated, having the head visible in the reference view consistently provides benefits, reducing the distance error. The error reduction rate is 23.7% and 23.2% for forward and sideways orientations, respectively—significantly greater than the 9.3% error reduction rate when the person is facing backward.

5.3. Ablation Study

Tab. 2 shows the ablation study results. The first row shows the single-view baseline version of our method. When the gaze estimator is extended to predict an uncertainty score σ, the model shows a small improvement due to better gaze vector prediction [10]. The HIA module significantly improves the performance by incorporating head information from the reference view and the geometry relations, which lead to more accurate gaze vectors and enhanced head embeddings when input to the output heads. The UGS module further improves the performance by selecting the more

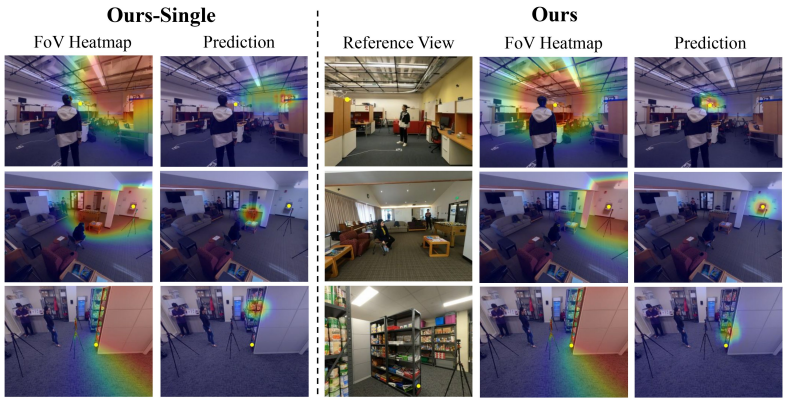


Figure 9. Qualitative comparisons of our method with and without multi-view processing. Yellow dots indicate ground truth. Our method can leverage the head appearance in the reference view to obtain better FoV heatmaps (Rows 1&2), and get enhanced performance using the scene appearance from the other view (Row3).

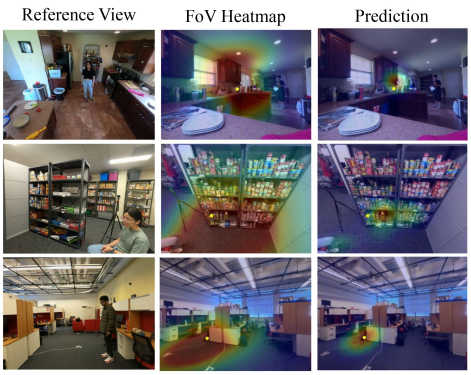


Figure 10. Qualitative examples for cross-view GTE. The reference view input, and the primary view's FoV heatmaps and predicted heatmaps are shown. Ground truth is shown as yellow dots.

σ	HIA	UGS	ESA	Head Visible				Head Not Visible			
				Target Visible		Target Not Visible		Target Visible		Target Not Visible	
				Dist. ↓	AP ↑	Dist. ↓	AP ↑	Dist. ↓	AP ↑	Dist. ↓	AP ↑
				0.151	0.877	0.148	0.878	0.179	0.758	0.154	0.855
✓				0.145	0.874	0.147	0.874	0.177	0.756	**0.151**	0.855
✓	✓			0.135	0.896	0.133	0.897	0.174	0.821	0.153	**0.873**
✓	✓	✓		0.130	0.902	0.123	0.908	0.170	0.810	0.152	0.863
✓	✓	✓	✓	**0.129**	**0.909**	**0.122**	**0.912**	**0.161**	**0.836**	0.152	0.868

Table 2. Ablation Study Results. All three modules demonstrate their effectiveness in leveraging multi-view information to improve GTE accuracy.

Method	Dist. ↓	AP ↑
Random	0.446	0.462
Center	0.245	/
DeepGazeIIE [39]	0.248	/
Recasens [50]	0.271	0.542
Ours	**0.188**	**0.820**

Table 3. Cross-view GTE results.

accurate gaze vector in the two views. The ESA modules provide further improvement when the gaze target is visible (1st and 3rd columns), suggesting that they add helpful scene context for GTE. The improvement is relatively less when the head is visible in reference because the gaze vectors and the corresponding FoV heatmaps determine the priors for the potential attended area, instead of the scene background. See the Supplementary for more detailed analyses of the proposed modules. Notably, the benefits of incorporating camera parameters as input in the HIA module for the in/out prediction task, are evident.

5.4. Cross-View Estimation

In this section, we train the model to predict the gaze target in the primary view which contains the target but not the subject, by using the person's appearance in the reference view. We fine-tune the model trained in the ordinary multi-view setting above on these cross-view camera pairs (around 7000 pairs). In this case, none of the single-view GTE models can predict the targets or serve as baseline methods. The strongest baseline we propose is adapting a method for predicting gaze targets in future video frames from the current observed frame [50] (outside of the current observed frame). We fine-tune it on the cross-view samples in our dataset, treating the reference view as the "current frame" and the primary view as the "future frame." We also evaluated DeepGazeIIE [39] as a representative baseline of the saliency prediction model. Table 3 shows that our method outperforms the other approaches by a wide margin. The qualitative examples in Fig. 10 demonstrate that our method generates reasonable FoV heatmaps in the primary view based on the person's appearance in the reference view, and predict target locations reasonably well.

6. Conclusions

This paper proposed the first method for multi-view gaze target estimation (GTE). The model incorporates a Head Information Aggregation (HIA) module to aggregate head information, an Uncertainty-based Gaze Selection (UGS) module to select the more reliable gaze vector predicted, and Epipolar-based Scene Attention (ESA) module for integrating scene background information. Our method shows large improvements when the reference view contains the person's head, and can be extended to cross-view GTE that single-view methods cannot handle. In addition, we introduced the MVGT dataset, the first dataset for multi-view GTE with calibrated camera parameters and precisely annotated targets. Future work could explore learning geometric-aware features without inputting camera parameters, or address the cross-view task without access to the reconstructed 3D scene. We expect our work can draw more attention to using multi-view input in the GTE domain.

Acknowledgments. This project was partially supported by NSF award IIS-2123920 and the Department of Surgery at Stony Brook University. Minh Hoai was initially supported by NSF award DUE-2055406, and later in part by the Australian Institute for Machine Learning (University of Adelaide) and the Centre for Augmented Reasoning, an initiative of the Australian Government's Department of Education. The authors thank Haoyu Wu for helpful discussions.

References

[1] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017. 1

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Proceedings of the European Conference on Computer Vision*, pages 348–367. Springer, 2022. 6

[3] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135, 2022. 2

[4] Yiwei Bao and Feng Lu. Unsupervised gaze representation learning from multi-view face images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1419–1428, 2024. 2

[5] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 7

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2

[7] Yihua Cheng and Feng Lu. Dvgaze: Dual-view gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20632–20641, 2023. 2

[8] 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 *Proceedings of the European Conference on Computer Vision*, pages 383–398, 2018. 2

[9] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020. 2, 3, 7

[10] Philipe Ambrozio Dias, Damiano Malafronte, Henry Medeiros, and Francesca Odone. Gaze estimation for assisted living environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 290–299, 2020. 4, 5, 7

[11] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning and Representation*, 2021. 5, 6

[13] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000. 1

[14] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11390–11399, 2021. 2

[15] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012. 1

[16] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the European Conference on Computer Vision*, pages 314–327. Springer, 2012. 1

[17] Thomas W Frazier, Mark Strauss, Eric W Klingemier, Emily E Zetzer, Antonio Y Hardan, Charis Eng, and Eric A Youngstrom. A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(7):546–555, 2017. 1

[18] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *pami*, 32(8):1362–1376, 2009. 2

[19] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5041–5050, 2022. 2

[20] Anshul Gupta, Samy Tafasca, Naravich Chutisilp, and Jean-Marc Odobez. A unified model for gaze following and social gaze prediction. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 1–9. IEEE, 2024. 2

[21] Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean marc Odobez. MTGS: A novel framework for multi-person temporal gaze following and social gaze prediction. In *Advances in Neural Information Processing Systems*, 2024. 2

[22] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 6

[24] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 7779–7788, 2020. 2, 5

[25] Yoichiro Hisadome, Tianyi Wu, Jiawei Qin, and Yusuke Sugano. Rotation-constrained cross-view feature fusion for multi-view appearance-based gaze estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5985–5994, 2024. 2

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[27] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. We know where they are looking at from the rgb-d camera: Gaze following in 3d. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 2

[28] Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8907–8916, 2023. 2, 3

[29] Xianjun Huang. Smart construction. https://github.com/PeterH0323/Smart_Construction. 3

[30] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 5

[31] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 2

[32] Robert JK Jacob and Keith S Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*, pages 573–605. Elsevier, 2003. 1

[33] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[34] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019. 6

[35] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017. 4, 5

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. 1

[37] Yunhao Li, Wei Shen, Zhongpai Gao, Yucheng Zhu, Guangtao Zhai, and Guodong Guo. Looking here or there? gaze following in 360-degree images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3742–3751, 2021. 2

[38] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Proceedings of the Asian Conference on Computer Vision*, pages 35–50. Springer, 2018. 2

[39] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021. 8

[40] Meng Liu, You Fu Li, and Hai Liu. 3d gaze estimation for head-mounted devices based on visual saliency. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems*, pages 10611–10616. IEEE, 2020. 1

[41] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 4

[42] Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 880–889, 2023. 2, 7

[43] Qiaomu Miao, Alexandros Graikos, Jingwei Zhang, Sounak Mondal, Minh Hoai, and Dimitris Samaras. Diffusion-refined vqa annotations for semi-supervised gaze following. In *Proceedings of the European Conference on Computer Vision*, pages 439–457. Springer, 2024. 2

[44] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[45] Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2192–2201, 2022. 2

[46] Hyun Park, Eakta Jain, and Yaser Sheikh. 3d social saliency from head-mounted cameras. *Advances in Neural Information Processing Systems*, 25, 2012. 1

[47] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *Proceedings of the European Conference on Computer Vision*, pages 747–763. Springer, 2020. 2

[48] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4342–4351, 2019. 2

[49] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 2, 3, 6

[50] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1435–1443, 2017. 2, 8

[51] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 7

[52] Kenji Sakita, Koichi Ogawara, Shinji Murakami, Kentaro Kawamura, and Katsushi Ikeuchi. Flexible cooperation between human and robot by interpreting human intention from gaze information. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems*, pages 846–851. IEEE, 2004. 1

[53] Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: Gaze following with interaction features in vision transformers. *arXiv preprint arXiv:2403.12778*, 2024. 2

[54] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946, 2023. 2, 5, 7

[55] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2017, 2024. 2, 6

[56] Francesco Tonini, Cigdem Beyan, and Elisa Ricci. Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 420–431, 2022. 2

[57] Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Object-aware gaze target detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21860–21869, 2023. 2

[58] Francesco Tonini, Nicola Dall'Asen, Lorenzo Vaquero, Cigdem Beyan, and Elisa Ricci. Al-gtd: Deep active learning for gaze target detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2360–2369, 2024. 2

[59] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2192–2200. IEEE, 2022. 2

[60] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *Proceedings of the European Conference on Computer Vision*, pages 197–212. Springer, 2020. 2

[61] Ultralytics. YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com, 2021. 3

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[63] Fred R Volkmar and Linda C Mayes. Gaze behavior in autism. *Development and Psychopathology*, 2(1):61–69, 1990. 1

[64] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 6

[65] Yaokun Yang and Feng Lu. Gaze target detection based on head-local-global coordination. In *Proceedings of the European Conference on Computer Vision*, pages 305–322. Springer, 2024. 2

[66] Yaokun Yang, Yihan Yin, and Feng Lu. Gaze target detection by merging human attention and activity cues. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 6585–6593, 2024. 2

[67] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8574–8584, 2022. 5

[68] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, pages 767–783, 2018. 2

[69] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 5

[70] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 5, 6

[71] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1358–1366, 2019. 2

[72] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Proceedings of the European Conference on Computer Vision*, pages 365–381. Springer, 2020. 2