GFIE: A Dataset and Baseline for Gaze-Following from 2D to 3D in Indoor Environments

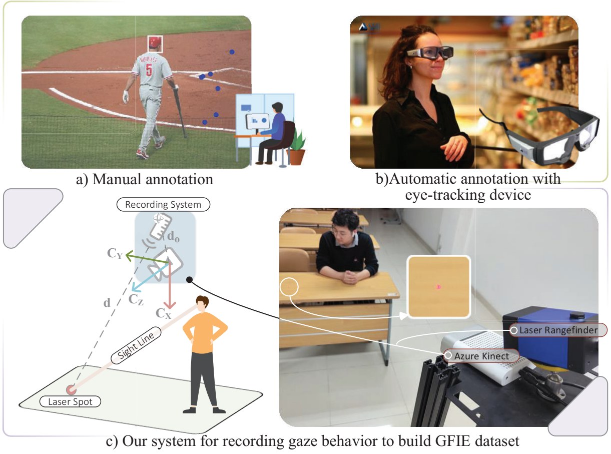
Zhengxi Hu[1,2,3], Yuxue Yang[1], Xiaolin Zhai[1,2,3], Dingye Yang[1,2,3], Bohan Zhou[1], Jingtai Liu[1,2,3]✉
[1]IRAIS, College of Artificial Intelligence, Nankai University
[2]tjKLIR, Nankai University [3]TBI center, Nankai University
{hzx,yyx,zhaixiaolin,abandon,zhoubohan}@mail.nankai.edu.cn liujt@nankai.edu.cn

Abstract

Gaze-following is a kind of research that requires locating where the person in the scene is looking automatically under the topic of gaze estimation. It is an important clue for understanding human intention, such as identifying objects or regions of interest to humans. However, a survey of datasets used for gaze-following tasks reveals defects in the way they collect gaze point labels. Manual labeling may introduce subjective bias and is labor-intensive, while automatic labeling with an eye-tracking device would alter the person's appearance. In this work, we introduce GFIE, a novel dataset recorded by a gaze data collection system we developed. The system is constructed with two devices, an Azure Kinect and a laser rangefinder, which generate the laser spot to steer the subject's attention as they perform in front of the camera. And an algorithm is developed to locate laser spots in images for annotating 2D/3D gaze targets and removing ground truth introduced by the spots. The whole procedure of collecting gaze behavior allows us to obtain unbiased labels in unconstrained environments semi-automatically. We also propose a baseline method with stereo field-of-view (FoV) perception for establishing a 2D/3D gaze-following benchmark on the GFIE dataset. Project page: https://sites.google.com/view/gfie.

1. Introduction

Gaze-following is a human skill that emerges in infancy [36] to learn about visual focus of other people, which helps to understand their personal thoughts and intentions [32]. For these reasons, detecting gaze targets automatically as humans do has great potential in some applications,


a) Manual annotation

b)Automatic annotation with eye-tracking device

c) Our system for recording gaze behavior to build GFIE dataset

Figure 1. The way of collecting gaze data in the existing gaze-following dataset and our proposed scheme. a) is a sample from the GazeFollow [28] dataset, the blue dots indicate the gaze targets annotated by the different annotators. b) indicate the case where annotations are collected with an eye-tracking device. c) is the system designed in this paper.

such as locating items of interest to a person in the retail environment [35] and judging the risk of driving by detecting whether the driver is distracted [9, 16]. In addition, gaze target detection can assist in action recognition [40], social relationship analysis [11, 41], autism diagnosis [7] and human-aware robot navigation [25].

As a device for monitoring gaze behavior, a wearable eye-tracking device was explored for gaze-following [23, 30]. [14, 23, 27] designed a custom system for tracking gaze. These methods are only applicable to constrained scenes due to extra burdens they bring, such as complex calibration and additional expense. This challenge has also attracted the attention of researchers in the computer vision community, and recent works [7, 22, 28, 38] have made an effort to establish datasets for inferring a person's gaze target from third-view image based on deep-learning methods.

✉Corresponding author

* This work is supported in part by National Key Research and Development Project of China under Grant 2019YFB1310604, in part by National Natural Science Foundation of China under Grant 62173189.

However, our survey of these datasets, which play an important role in this task, reveal deficiency in the way they gather gaze data. Most datasets are manually annotated, but the subjectivity of annotators may cause annotations to deviate from the actual gaze target. This is demonstrated by the sample in Figure 1 a) where each annotator has a different opinion on the gaze target of the same person. In addition, labor-intensive is another drawback. The eye-tracking device in Figure 1 b) can capture annotations automatically but alter subjects' appearance in the dataset, which brings the gap with the gaze-related behavior in the natural environment.

To address these problems, as shown in Figure 1 c), we propose a novel system for establishing our GFIE dataset that provides accurate annotations and clean training data recorded in natural environments. The system consists of a laser rangefinder and an RGB-D camera Azure Kinect, which allows us to manipulate the laser rangefinder to guide the subject's gaze target through the laser spot while recording their activities with the RGB-D camera. After detecting the laser spot in the image by our proposed algorithm, the gaze target of the person in the image can be located. Based on the distance to the laser spot measured by the laser rangefinder, the 3D gaze target can also be reconstructed. Considering that the laser spot introduces ground truth to the image, we employ an image inpainting algorithm to eliminate it for constructing the final dataset. Most of the processes are automated, alleviating the need for human resources. Our proposed GFIE dataset comprises rich activity clips with different subjects and diverse scenes. They are key to ensuring the diversity of gaze behaviors. Along with RGB-D images and 2D/3D gaze targets, we also provide camera parameters, head bounding boxes and 2D/3D eye locations.

Accompanying our proposed GFIE dataset, we design a novel baseline method that takes the stereo field of view (FoV) to estimate gaze targets into account. In this paper, FoV is defined as the extend to which a person can observe in 3D space. It is perceived based on the predicted gaze direction and transformed into a heatmap. Then the heatmap combined with scene saliency, helps the entire model localize 2D and 3D gaze targets more efficiently. State-of-the-art methods are introduced to establish 2D/3D gaze-following benchmarks on both GFIE and CAD-120 [20] datasets. Experiment results show that the GFIE dataset is reliable and the proposed baseline method achieves excellent performance in 2D images and 3D scenes.

In summary, our main contributions are as follows:

- We develop a system consisting of a laser rangefinder and RGB-D camera to guide and localize gaze target while recording gaze behavior.
- We release a new GFIE dataset for 2D/3D gaze-following that contains reliable annotations and diverse human activities in indoor environments.
- We introduce a stereo field of view (FoV) in the proposed baseline method for improving gaze-following.

Table 1. Comparison of GFIE with existing gaze-following datasts

Dataset	RGB/ RGB-D	Size	Gaze Target Localized by	3D Annot.	Data Source
GazeFollow [28]	RGB	122,143 frames, 130,339 people	Annotator	✗	MS COCO, SUN, PASCAL, etc
VideoAttentionTarget [7]	RGB	1331 tracks,164,541 frame-level anno.	Annotator	✗	YouTube
GazeFollow360 [21]	RGB	65 videos, 10,058 frames	Annotator	✗	YouTube
VideoGaze [29]	RGB	140 movies, 166,721 anno.	Annotator	✗	MovieQA
VideoCoAtt [10]	RGB	380 videos, 492,100 frames	Annotator	✗	TV shows
DL Gaze [22]	RGB	95,000 frames, 16 subjects	Annotator	✗	Recorded by iPhone
TIA [38]	RGB-D	330,000 frames, 14 subjects	Eye-tracking glasses	✓	Recorded by Kinect V2
GFIE (ours)	**RGB-D**	**71799 frames, 61 subjects**	**Laser spot**	✓	**Recorded by Azure Kinect**

2. Related Work

2.1. Gaze-following dataset

In Table 1, we present an analysis of existing datasets related to gaze-following. GazeFollow [28] is the first large-scale image dataset for gaze-following, which contains a total of $130,339$ people and $122,143$ images with hand-annotated ground truths. The VideoAttentionTarget [7] established by Chong *et al.* was proposed for a temporal task, which consists of $1,331$ tracks collected from YouTube. A team of 4 annotators provided $164,541$ frame-level annotations. VideoGaze [29] is built to predict what a person is looking at even if the gaze targets appear in other frames in the video, and it contains 140 movies and $166,721$ annotations from the crowdsourcing website. To address shared attention or gaze-following in social interaction, Fan *et al.* [10] established a VideoCoAtt dataset and the co-attention areas in $492,100$ frames are manually annotated. GazeFollow360 [21] focuses on gaze-following in 360-degree images and 4 knowledgeable annotators label gaze targets in $10,058$ frames collected from YouTube.

Several studies have explored building datasets by recording their own video. The video-based DL Gaze dataset [22] recorded $95,000$ frames and 16 volunteers in the video were asked to annotate where they were looking. External devices are also considered to collect ground truth in addition to manual annotation. TIA datasets [38] proposed by Wei *et al.* relies on eye-tracking glasses to locate a volunteer's gaze point and correlate it with video recorded from third-person view. 14 volunteers and $330,000$ frames formed this dataset. For an overview of all existing datasets in Table 1, the ground truth in most datasets is manually annotated, which may introduce subjective bias. And eye-tracking glasses changes the subject's appearance since it need to be worn during recording, which brings a gap from the natural environment.

2.2. Gaze-following method

Recasens *et al.* [28] first established a deep-learning-based framework including a saliency pathway and gaze pathway for learning to follow gaze from the image. On this basis, the estimation of gaze direction [6, 22] is considered to tackle this problem. Chong *et al.* [7] designed a spatial-temporal architecture with an attention mechanism to detect dynamic gaze targets. Gaze-related clues including human-object interaction [5] and scene depth [2, 12] are also explored and developed for gaze-following. Li's work [21] focused on gaze-following in 360-degree images and they proposed a dual-pathway to model the sightline in 3D sphere space to detect gaze targets in any region. There are also some researches who have focused on applications in special scenarios, such as inferring shared attention in social scene [10], following gaze in neighbor frames of a video [29] and tracking the attention of children in classrooms [1].

Some researchers extend the gaze following task to 3D space. Santner *et al.* [30] and Liu [23] *et al.* designed systems for 3D gaze tracking based on eye-tracking glasses and RGB-D camera, respectively, while Park *et al.* [24] utilized multiple head-mounted cameras to locate 3D joint attention in the social scene. Gaze-following in 3D scenes with only a single camera has also been proven to be feasible. Head poses [31] and geometric relationships [3, 34] are exploited for predicting joint attention. Based on the RGB-D camera, Wei *et al.* [38, 39] proposed a probabilistic graphical model to infer intent and attention jointly under a unified framework while Shi [33] proposed a Sequential Skeleton Based Attention Network under the LSTM framework to deal with attention inference in complex situations. In this paper, we designed a method for 2D/3D gaze-following simultaneously from RGB-D images.

3. GFIE Dataset Generation

In this section, we introduce the system setup for recording gaze behavior and then present the algorithm for detecting laser spots to acquire gaze targets. After constructing all 2D/3D annotations, we perform an analysis on the entire GFIE dataset. Figure 2 depicts the workflow for generating the GFIE dataset. The whole process is semi-automatic as human intervention is required only for recording the data and verifying the annotations.

3.1. System Setup

Considering the deficiency of existing datasets, we design a system that can locate gaze targets accurately while recording gaze behavior without changing the subjects' appearance. As shown in Figure 1, our designed system consists of a laser rangefinder and an Azure Kinect (RGB-D camera) mounted on a platform. The Azure Kinect is fixed on the platform to record the activities of the volunteers, while the laser rangefinder is placed on the universal ball joint so that the emitted laser spot can move smoothly. While recording gaze behavior, the Azure Kinect is set to capture RGB images and depth images with a resolution of 1920 × 1080, where the depth images have been registered into the RGB frames. We operate the laser rangefinder to guide the subject's attention target through the laser spot, which means that the subject is always staring at the laser spot while performing in front of the camera. At the same time, the distance measured by the laser rangefinder is recorded.

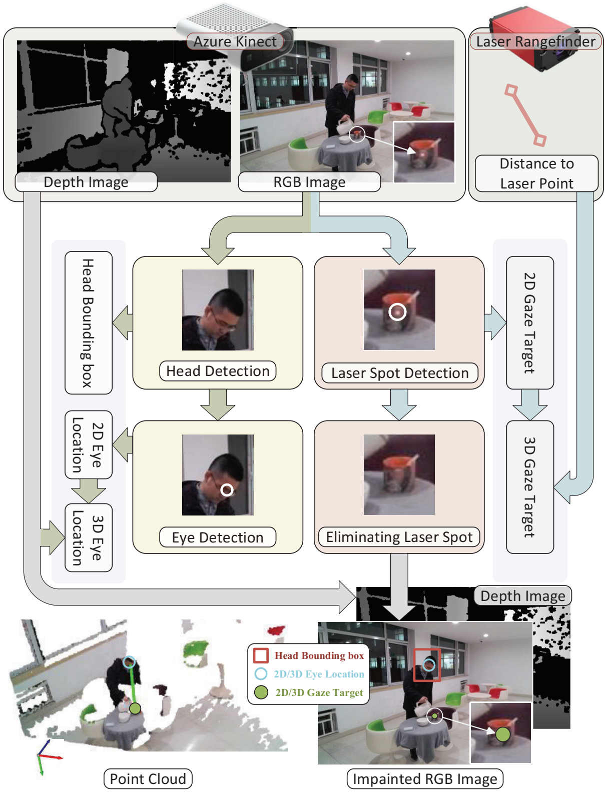


Figure 2. Workflow for GFIE dataset generation

3.2. Laser Spot Detection

In this paper, in order to detect the position of the laser spot in the RGB image to generate the gaze target annotation, we propose a laser spot detector that can be well applied to complex indoor scenes. To make the laser spots prominent, we preprocess the image by multiplying the saturation value and performing gamma correction on the lightness value in the image's HSL color space. After that, we extract the horizontal and vertical derivatives of the

Algorithm 1 CDBPS

```
Input: Boundary point sets $S$, Threshold $\eta$, Minimum radius $R_{\min}$, Maximum radius $R_{\max}$
Output: Boundary point sets of candidate regions $S_{\text{target}}$
for set $s$ in $S$ do
    $c, r \leftarrow$ find the minimum enclosing circle of set $s$
    if $R_{\min} \leq r \leq R_{\max}$ then
        for point $p_i$ in $s$ do
            $D_i \leftarrow \|p_i - c\|$
        end for
        $v \leftarrow$ compute the variance of $D$
        if $v \leq \eta$ then
            add $s$ in $S_{\text{target}}$
        end if
    end if
end for
```

grayscale image with the standard 3×3 Sobel operators to calculate the gradient magnitude. In addition to the gradient, we also need to select the candidate regions according to the color characteristics of the laser spot. Considering HSV color space is more suitable to screen out the laser spot matching the specified color ranges [19]. We filter out regions that fit the color ranges to extract the binary image in HSV color space and apply closing operation in morphology with an elliptical 7×7 structuring element, making the foreground regions smooth.

Based on preprocessing, we use the binary image as a mask for the gradient image to omit the regions that do not conform to the color features and retrieve boundary point sets of the remaining regions. Then the candidate regions where the laser spot may exist are selected via an algorithm named CDBPS (Circle Detection based on Boundary Point Set), which is summarized in Algorithm 1. The algorithm traverses the boundary point sets and calculates the center and radius of the minimum enclosing circle of each set. Among these sets, those whose enclosing circle is within the specified range will be preserved. After that, we calculate the variance of the distances between the boundary points of the reserved sets and the center of the enclosing circle, and discard the set with variance greater than the threshold. The selected sets are the boundary point set of the candidate regions.

Considering the laser spot moves smoothly in the recorded video sequence, we introduce the Kalman filter to locate a region most likely to have a laser spot from the candidate regions. Specifically, we first need to select a rectangular area containing laser spots in the initial frame of the video and then calculate the IoU between the rectangular area predicted by the Kalman filter and the enclosing rectangle of the candidate area in the next frame. The candidate region with the maximum IoU is regarded as the laser spot in the image and we update the estimated error covariance matrix for detecting in the next frame. We take the predicted rectangle as the region where the laser spot may exist if no candidate region is selected.

3.3. Annotation

We provide annotations including head bounding boxes, 2D/3D gaze targets and 2D/3D eye locations in our proposed GFIE dataset. In particular, all 3D coordinates are represented in the RGB camera coordinate system. Most of these annotations are generated automatically, with only a few failure cases need to be handled by the annotators.

2D gaze target: After performing the laser spot detection on all recorded RGB images, a team of 5 annotators was asked to check whether the laser spots in the images were detected correctly and correct the wrong cases. Then we take the verified position as the location of the gaze point in the image to form the final annotation.

3D gaze target: With the help of distance measured by the laser rangefinder in the recording system, 2D gaze targets in images can be transformed into 3D space. Then our established dataset can be extended for gaze-following in 3D scenes. Using the depth map directly to obtain the 3D gaze target is inappropriate because it contains invalid or noisy values.

In the recording system shown in Figure 1, we regard the laser rangefinder as a mass point and assume that the offset from it to the camera coordinate system is d_o, then the distance from the laser spot to the camera can be approximated as $d - d_o$, where d is the distance measured by the laser rangefinder. In addition, we use (g_u, g_v) to represent the coordinate of the detected laser spot in the image and $\mathcal{K} = (f_u, f_v, c_u, c_v)$ to indicate the intrinsics of the RGB camera. Then we need to calculate the coordinates of the 3D gaze point (g_x, g_y, g_z) in the RGB frame.

According to the unprojection principle of the pinhole camera and the measured distance, the following constraints can be established:

$$\begin{cases} \dfrac{(g_u - c_u)\, g_z}{f_u} = g_x \\ \dfrac{(g_v - c_v) g_z}{f_v} = g_y \\ \sqrt{g_x^2 + g_y^2 + g_z^2} = d - d_o \end{cases} \tag{1}$$

Then g_z is obtained by solving Equation 1 as follow:

$$g_z = \frac{d - d_o}{\sqrt{\left(\dfrac{g_u - c_u}{f_u}\right)^2 + \left(\dfrac{g_v - c_v}{f_v}\right)^2 + 1}} \tag{2}$$

we can calculate g_x, g_y according to Equation 1.

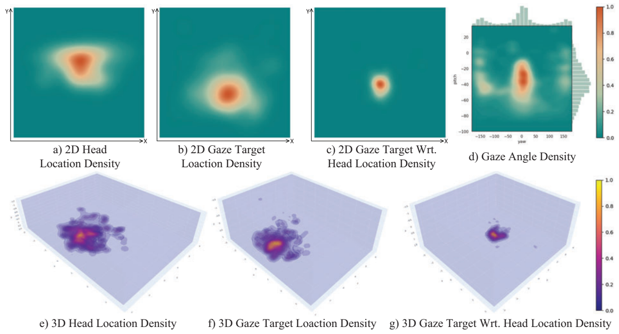


Figure 3. GFIE Dataset statics. Top left three: The distribution of annotations including head and gaze point location in the 2D plane. Top right one: The distribution of gaze angles. Bottom three: The distribution of annotations including head and gaze point location in the 3D space.

Head bounding box: Recent work [6, 22, 28] has demonstrated that the head is a crucial clue in gaze-following, so the head bounding box also need to be provided in the annotations. The robust face detector [42] is chosen for the coarse detection of the heads in the images. Then 5 annotators were asked to check the detected bounding box and correct the failed cases on the CVAT (Computer Vision Annotation Tool) platform.

2D/3D eye location: Based on the cropped head image, the facial landmark detector proposed by Bulat et al. [4] is used to detect the location of the left and right eye landmarks. The 2D eye position is at the center of these two landmarks, while the 3D eye position can be unprojected according to the values of all face landmarks in the depth map. The few failed case in detection are annotated by annotators.

3.4. Eliminating Laser Spot Process

Although the laser spot can guide and locate a person's gaze target, it also brings ground truth into the dataset by adding the spot in the image. As a technique to recover the missing region or remove some objects in the image, image inpainting is suitable for removing laser spots. In this paper, after setting the mask region around the laser spot, we adopt the generator network proposed by Ulyanov et al. [37] to inpaint the regions of laser spots in images. Since the laser spot is located in a small region, the algorithm can fill the target area effectively referring to the surrounding texture, which makes the inpainted image similar to the natural scene. Figure 2 shows the original images and the corresponding inpainted images.

3.5. Dataset Statics

The GFIE dataset includes diverse gaze behaviors of 61 subjects (27 male and 34 female), accompanied by a wide range of activities. The entire dataset consists of 71799 frames in total and each frame has annotations, which include head bounding box, eye location and gaze target in 2D plane and 3D space. In this paper, we divide the dataset into a training set with $59,217$ frames, a test set with $6,281$ frames and a validation set with $6,281$ frames. In addition, subjects and scenes that appeared in the training set were not included in the test and validation sets.

Statistics from 2D plane and 3D space for the entire GFIE dataset are shown in Figure 3. The probability densities of annotations in the image are placed in the top three subplots, indicating that the head locations are concentrated in the upper part of the image while the gaze points appear more in the lower part. Such a gaze related distribution is caused by these more common behaviors in daily life such as looking straight horizontally or looking down. The distribution of gaze angles shown at top right in Figure 3 is also consistent with our analysis. The probability distributions of 3D annotations in the RGB camera coordinate system are also placed in the bottom three subplots. The distance from the subject to the camera ranges from $1.04m$ to $6.48m$, with a mean distance of $2.41m$ and the gaze targets are distributed widely in space.

4. GFIE Model

4.1. Network Architecture

We design a baseline method to evaluate gaze-following in 2D images and 3D point clouds of our established GFIE dataset. The main idea of this proposed method is to infer human gaze targets on the basis of perceiving a person's field of view (FoV) in a stereo space. An overview is shown in Figure 4.

The architecture consists of three components: a module for estimating gaze direction, a module for perceiving stereo FoV, and a module for generating a gaze heatmap. We use ResNet50 [15] as the backbone to build the module for estimating gaze direction, which takes the cropped head image as an input and outputs a 3D gaze unit vector $\boldsymbol{v} = [a_x, a_y, a_z]$. The module that perceives stereo FoV is proposed to highlight regions that a person pays attention to in space according to gaze direction. With the help of camera intrinsics $\mathcal{K}$, we first unproject the registered depth map $\mathcal{D} \in m \times n$ into the RGB camera coordinate system and then subtract the eye coordinate E. The unprojected coordinates are represented by matrix $\mathcal{T} \in m \times n \times 3$ instead of a point set. The transformation from $\mathcal{D}$ to $\mathcal{T}$ is as follow, where u, v are the indices of the matrix:

$$\begin{aligned} (u - c_u)\,\mathcal{D}(u,v)/f_u - e_x &= \mathcal{T}(u,v,0) \\ (v - c_v)\,\mathcal{D}(u,v)/f_v - e_y &= \mathcal{T}(u,v,1) \\ \mathcal{D}(u,v) - e_z &= \mathcal{T}(u,v,2) \end{aligned} \tag{3}$$

After normalizing $\mathcal{T}$ to $\mathcal{T}'$ in the third dimension, heatmaps F and F' can be generated by:

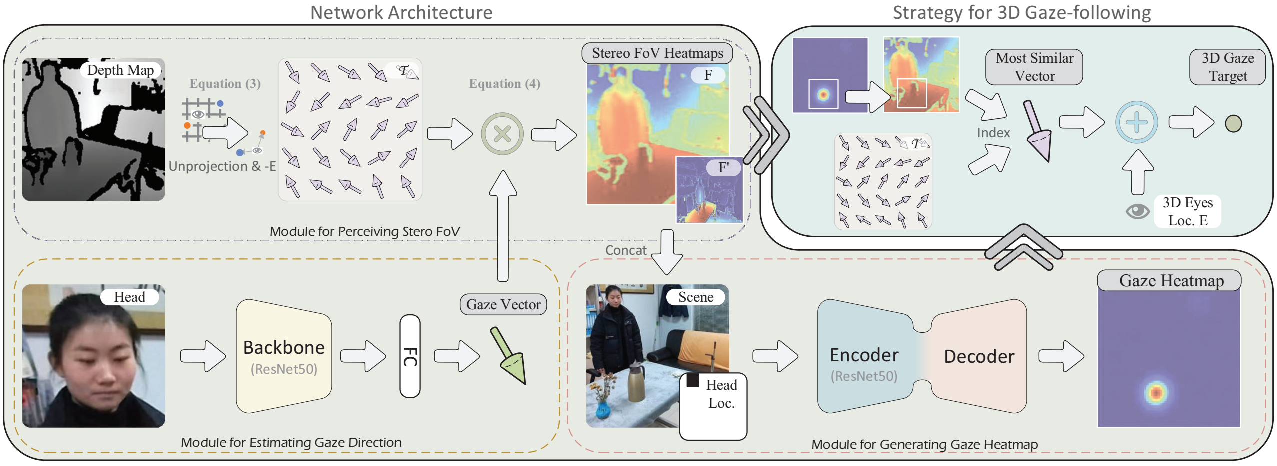


Figure 4. Network Architecture and Strategy for Gaze-following

$$F = \mathcal{T}' \cdot \boldsymbol{v}^T, F' = (\mathrm{Relu}(F))^{\alpha} \tag{4}$$

Assuming that gaze direction can be estimated accurately, these two heatmaps F and $F^{'}$ transformed from stereo FoV can indicate the region that the person is most likely to pay attention to. As shown in Figure 4, the activation function Relu and the exponent α can lead the model to focus more on regions with higher probability.

The stereo FoV heatmaps can be fed into the module for generating gaze heatmaps together with the scene images for estimating the gaze target. The final module is an encoder-decoder architecture, where the encoder contains all feature layer of ResNet50 and the decoder consist of 2 convolution layers and 3 deconvolution layers. The final output heatmap is H and the predicted gaze target is $(\hat{g}_u, \hat{g}_v) = \mathrm{argmax}(H)$.

4.2. Strategy for 3D Gaze-following

Our proposed baseline can be extended for 3D gaze-following based on predicted 2D gaze target $(\hat{g}_u, \hat{g}_v)$ and stereo FoV heatmap F. After transforming the 2D gaze target $(\hat{g}_u, \hat{g}_v)$ into the heatmap F, we set a rectangle area $R \in w \times h$ with it as the center and find the maximum value in R, which is at the (p, q) in the stereo FoV heatmap F. This maximum value means that the vector $\widetilde{\boldsymbol{v}} = (\mathcal{T}(p, q, 0), \mathcal{T}(p, q, 1), \mathcal{T}(p, q, 2))$ within the region R of the matrix $\mathcal{T}$ is most similar to the predicted vector $\boldsymbol{v}$. Then the point $\widetilde{\boldsymbol{v}} + E$ can represent the 3D gaze target, which is pointed by the vector $\widetilde{\boldsymbol{v}}$.

4.3. Implementation Details

Our proposed baseline is implemented by Pytorch and all inputs are resized to 224 × 224. The module for estimating gaze direction is pretrained on Gaze360 dataset [18] and the encoder is pretrained on the Imagenet [8]. The size of the two stereo FoV heatmaps and the final output heatmap is 64×64 and α is set to 3 in this paper.

We supervise both heatmap H and 3D gaze vector $\boldsymbol{v}$ for regression in the training process. Following [26], the ground truth heatmap is formed by generating a Gaussian centered on the gaze point. MSE loss l_h is chosen for heatmap regression and Cosine loss l_v is used for gaze vector regression. The total loss function is $l = \beta \cdot l_h + \gamma l_v$, where the β, γ are set to balance the two loss values. The training data is augmented with flipping, random cropping and color jittering for learning.

5. Experiment

5.1. Performance Comparison

Experiment setup: In this section, several methods are introduced to evaluate the performance on 2D/3D gaze-following in order to establish a benchmark as follows:

- *Random*: 2D and 3D gaze targets are selected randomly in the image and point cloud.
- *Center*: 2D and 3D gaze targets are located at the center of the image and the point cloud, respectively.
- *GazeFollow* [28], *Lian* [22] and *Chong* [7]: These methods all focus on gaze following in 2D planes, and we extend these methods for estimating 3D gaze targets by unprojecting the 2D gaze targets into 3D space from the registered depth map.
- *Gaze360* [18] and *RT-Gene* [13]: These methods for gaze estimation in an unconstrained environment are also introduced to perform 3D gaze-following by gaze direction. We find the vector most similar to the predicted gaze vector $\boldsymbol{v_s}$ in the set of vectors T, and the

Table 2. Performance comparison on the GFIE dataset

	2D		3D	
Method	AUC ↑	L^2 Dist. ↓	3D Dist. ↓	Angle Error ↓
Random	0.585	0.425	2.930	84.4°
Center	0.614	0.287	2.510	87.2°
GazeFollow [28]	0.941	0.131	0.856	41.5°
Lian [22]	0.962	0.091	0.542	26.7°
Chong [7]	0.972	0.069	0.455	20.8°
Rt-Gene [13]	0.823	0.123	0.552	21.0°
Gaze360 [21]	0.821	0.130	0.540	19.8°
GFIE (ours)	0.965	**0.065**	**0.311**	**17.7°**

Table 3. Quantitative results of ablation study on the GFIE dataset

	2D		3D	
Method	AUC ↑	L^2 Dist. ↓	3D Dist. ↓	Angle Error ↓
No encoder-decoder module	0.887	0.129	0.552	20.0°
No stereo FoV heatmap module	0.888	0.104	0.452	22.2°
One stereo FoV heatmap	0.945	0.079	0.391	20.8°
No supervision for the gaze vector	0.943	0.073	0.821	42.5°
3D gaze-following with only the predicted gaze vector	0.799	0.136	0.543	19.4°
3D gaze-following with only the predicted heatmap	0.965	0.065	0.333	18.7°
GFIE (ours)	**0.965**	**0.065**	**0.311**	**17.7°**

point $\boldsymbol{v_s} + E$ is taken as the 3D gaze target predicted by these methods.

The following metrics are introduced in the evaluation:

- *2D evaluation metrics*: *AUC*: The area under curve proposed by [17] is introduced to use the predicted heatmap as the confidence to draw the ROC curve. *L^2 Dist.*: The Euclidean distance between the predicted gaze point and the ground truth, we assume the size of the image is 1×1.
- *3D evaluation metrics*: *3D Dist.*: Similar to L^2 Dist., but for 3D scenes, its unit is m. *Angle Error*: The angular difference between predicted gaze direction and ground truth, in degrees.

Analysis: The quantitative results are shown in Table 2. Our proposed method achieves the best performance on the GFIE dataset both in the 2D and 3D scenes, which outperforms the other baseline methods. From the table, we also can draw the following conclusion: 1) All simple designed baseline methods perform poorly. 2) Extending the research on 2D gaze-following to 3D scenes may introduce errors, which may be attributed to invalid values or noise in the registered depth map. 3) The methods for gaze estimation still face challenges in locating gaze targets, although they perform well in predicting gaze direction.

We also show some examples predicted by our proposed method in Figure 5 a), which contains the results of gaze-following in the 2D image and 3D point cloud.

5.2. Ablation Study

We also design the following methods to demonstrate the effectiveness of the components, input setting, and strategies used for training and inference in our proposed method.

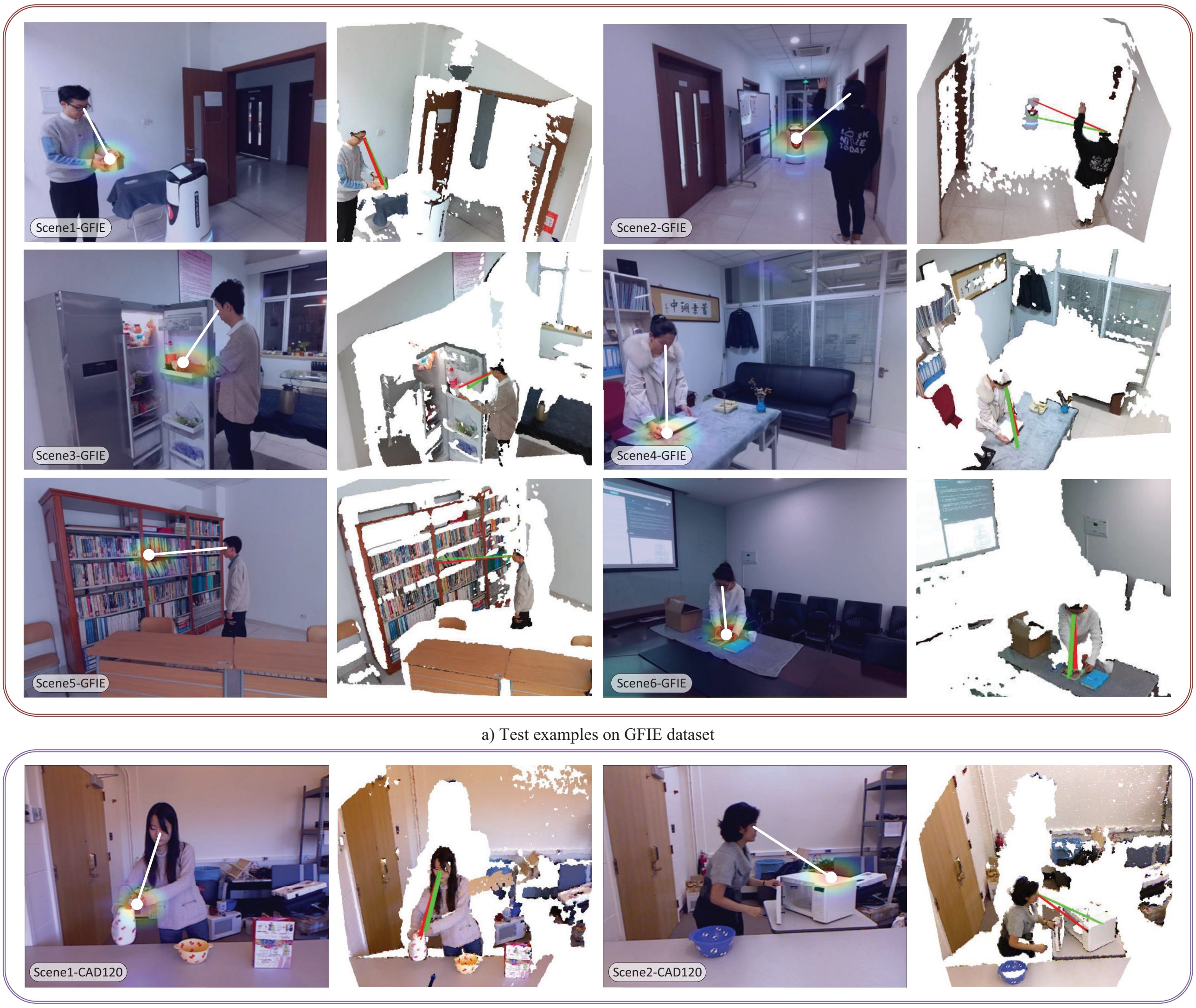
- *No encoder-decoder module*: The encoder-decoder module is removed, and we replace the final output heatmap with the stereo FoV heatmap F for inference. Only cosine loss is considered for training.
- *No stereo FoV heatmap module*: The module for perceiving stereo FoV is removed, and two other modules are used for heatmap generation and gaze estimation, respectively.
- *One stereo FoV heatmap*: Only a single stereo FoV heatmap F is used in the network and F' is ignored.
- *No supervision for the gaze vector*: The cosine loss l_v is removed and only the heatmap loss l_h is used for regression.
- *3D gaze-following with only the predicted gaze vector*: Only the predicted gaze vector is used to infer the 3D gaze target, and the implementation method is the same as the method for gaze estimation in the experiment setting.
- *3D gaze-following with only the predicted heatmap*: Only the predicted gaze heatmap is used to infer the 3D gaze target, and the implementation method is the same as the method for 2D gaze-following in the experiment setting.

The results of the ablation analysis are provided in Table 3. The first four rows in the table indicate that the encoder-decoder module, the two-layer FoV heatmap and the cosine loss settings in our designed network are necessary. The last two ablation methods in the table are designed to validate the effectiveness of the strategy for 3D gaze-following. We can learn that using predicted heatmaps or predicted gaze vectors alone has a certain performance gap compared with the complete approach, proving that our designed strategy for 3D gaze-following is effective.

5.3. Evaluation on CAD-120 Dataset

To explore whether the model trained on our GFIE dataset can be applied to other unseen scenes with different camera settings, we introduce the CAD-120 [20] dataset record with Kinect V2 for evaluation. The CAD-120 dataset is built for human activity analysis but has no gaze-related annotations. So we selected 1737 frames and asked 3 annotators to annotate the 3D gaze targets manually in the software *CloudCompare*. We selected 300 samples randomly and asked a volunteer to estimate the 3D attention target. The average distance error between the volunteer's estimated value and the annotated value was 0.113 meters.

After training on the GFIE dataset, our proposed method and the compared method *Chong* [7] perform testing on the



a) Test examples on GFIE dataset

b) Test examples on CAD-120 dataset

Figure 5. Visualization of test examples on GFIE and CAD-120 datasets. The endpoint of the white line indicates the gaze target in the image. The red and green lines in the point cloud represent the ground truth and sight line predicted by the GFIE model, respectively.

CAD-120 dataset. The evaluation results of our proposed method on this dataset are as follows: the 3D Dist. is 0.365 and the L^2 Dist. is 0.114, which outperforms the *Chong* (3D Dist. is 0.812 and L^2 Dist. is 0.152) in both 2D and 3D scenes. This demonstrates that the GFIE dataset contains sufficient information for gaze following task and the GFIE model has excellent generalization performance. The test samples is shown in Figure 5 b).

6. Conclusions

In this paper, we introduce a novel system and approach to guide, record and localize gaze targets efficiently while collecting gaze behaviors in indoor environments to generate GFIE dataset. This dataset covers reliable annotations and diverse activities which are suitable for gaze-following in 2D/3D scenes. We design a model that takes stereo FoV into account and present a strategy for 3D gaze-following, which constitutes our baseline method. Experiments for benchmarking on our dataset are also performed. Both quantitative results and qualitative analysis suggest that our proposed method can solve gaze-following problems well, whether under 2D or 3D circumstances. The evaluation on the CAD-120 dataset demonstrates that the dataset and baseline proposed in this paper can be used for other scenes with different camera settings.

References

[1] Arkar Min Aung, Anand Ramakrishnan, and Jacob R Whitehill. Who are they looking at? automatic eye gaze following for classroom observation video analysis. *International Educational Data Mining Society*, 2018. 3

[2] 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. 3

[3] Ernesto Brau, Jinyan Guan, Tanya Jeffries, and Kobus Barnard. Multiple-gaze geometry: Inferring novel 3d locations from gazes observed in monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 612–630, 2018. 3

[4] 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 International Conference on Computer Vision*, pages 1021–1030, 2017. 5

[5] Wenhe Chen, Hui Xu, Chao Zhu, Xiaoli Liu, Yinghua Lu, Caixia Zheng, and Jun Kong. Gaze estimation via the joint modeling of multiple cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 3

[6] 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 (ECCV)*, pages 383–398, 2018. 3, 5

[7] 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. 1, 2, 3, 6, 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[9] Tao Deng, Kaifu Yang, Yongjie Li, and Hongmei Yan. Where does the driver look? top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):2051–2062, 2016. 1

[10] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468, 2018. 2, 3

[11] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733, 2019. 1

[12] 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. 3

[13] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018. 6, 7

[14] Peter Hausamann, Christian Sinnott, and Paul R MacNeilage. Positional head-eye tracking outside the lab: an open-source solution. In *ACM Symposium on eye tracking research and applications*, pages 1–5, 2020. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Zhongxu Hu, Chen Lv, Peng Hang, Chao Huang, and Yang Xing. Data-driven estimation of driver attention using calibration-free eye gaze and scene features. *IEEE Transactions on Industrial Electronics*, 2021. 1

[17] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 7

[18] 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

[19] S Kolkur, D Kalbande, P Shimpi, C Bapat, and J Jatakia. Human skin detection using rgb, hsv and ycbcr color models. In *International Conference on Communication and Signal Processing 2016 (ICCASP 2016)*, pages 324–332. Atlantis Press, 2016. 4

[20] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013. 2, 7

[21] 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, 3, 7

[22] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018. 1, 2, 3, 5, 6, 7

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

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

[25] Remi Paulin, Thierry Fraichard, and Patrick Reignier. Using human attention to address human–robot motion. *IEEE Robotics and Automation Letters*, 4(2):2038–2045, 2019. 1

[26] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 1913–1921, 2015. 6

[27] Mahmoud Qodseya, Marta Sanzari, Valsamis Ntouskos, and Fiora Pirri. A3d: A device for studying gaze in 3d. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14*, pages 572–588. Springer, 2016. 1

[28] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015. 1, 2, 3, 5, 6, 7

[29] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017. 2, 3

[30] Katrin Santner, Gerald Fritz, Lucas Paletta, and Heinz Mayer. Visual recovery of saliency maps from human attention in 3d environments. In *2013 IEEE International Conference on Robotics and Automation*, pages 4297–4303. IEEE, 2013. 1, 3

[31] Samira Sheikhi and Jean-Marc Odobez. Recognizing the visual focus of attention for human robot interaction. In *International Workshop on Human Behavior Understanding*, pages 99–112. Springer, 2012. 3

[32] Stephen V Shepherd. Following gaze: gaze-following behavior as a window into social cognition. *Frontiers in integrative neuroscience*, 4:5, 2010. 1

[33] Xiang Shi, You Yang, and Qiong Liu. I understand you: Blind 3d human attention inference from the perspective of third-person. *IEEE Transactions on Image Processing*, 30:6212–6225, 2021. 3

[34] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2015. 3

[35] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3125–3133, 2021. 1

[36] Jochen Triesch, Christof Teuscher, Gedeon O Deák, and Eric Carlson. Gaze following: why (not) learn it? *Developmental science*, 9(2):125–147, 2006. 1

[37] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 5

[38] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6801–6809, 2018. 1, 2, 3

[39] Ping Wei, Dan Xie, Nanning Zheng, Song-Chun Zhu, et al. Inferring human attention by learning latent intentions. In *IJCAI*, pages 1297–1303, 2017. 3

[40] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, 22(6):1423–1432, 2019. 1

[41] Xingming Yang, Fei Xu, Kewei Wu, Zhao Xie, and Yongxuan Sun. Gaze-aware graph convolutional network for social relation recognition. *IEEE Access*, 9:99398–99408, 2021. 1

[42] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 5