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

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

In this supplemental material, we provide more discuss, details, results and analysis as follows:

- Discussion on the deficiency of establishing gaze-following dataset with manual annotations
- More details about the recording system
- Training and evaluation settings on GFIE dataset
- Further analysis on the Experiment section in main manuscript
- Limitation and future work

Dataset, model and demo are available on the project page: https://sites.google.com/view/gfie.

1. Discuss the Deficiency of Manual Annotation through Analysis of GazeFollow Dataset

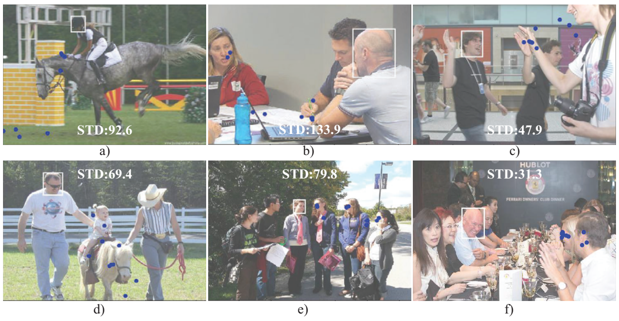

Figure 1. The example of manual annotation on test set of GazeFollow dataset. The white box is the target person whose gaze target needs to be annotated. The blue dots indicate annotations from different annotators while the numbers in the image indicate the average standard deviation of the annotated positions.

In this section, we take the GazeFollow dataset as an example to explain how manual annotation can introduce subjective bias in detail. As shown in Fig 1, test samples in the GazeFollow dataset are annotated by multiple annotators. The average standard deviation of the annotations over the entire test set is 46.3 pixel distance, which proves that different annotators have different opinions on the same example. Besides, in some complex scenes, as shown in Fig 1 b) e), it is difficult to figure out what the person is looking at. These all suggest that the subjectivity of annotators can deviate the annotation away from the ground truth.

2. More Details about Recording System

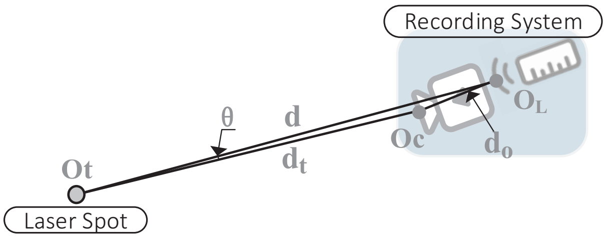

Figure 2. The geometric relationship between the laser spot and recording system

2.1. Reconstruct 3D Gaze Targets

In Section 3.3 of the main paper, we approximated the distance from the laser spot to the camera to be $d - d_0$, and here we explain why this approximation is possible. Considering that the laser rangefinder rotates around O_L, the distance between it and the fixed camera is always d_0, which is $0.071m$ meters. As shown in Fig. 2, we assume that the distance measured by the laser rangefinder is d, then the distance from the laser spot to the camera d_t can be expressed as follows:

$$d_t = d \cdot \cos\theta - \sqrt{d^2 \cdot (\cos^2\theta - 1) + d_0^2} \quad (1)$$

Since the laser spot in the recording process is far away from the recording system ($> 1m$), the angle θ between the two line segments O_LO_t and O_cO_t tends to zero $\theta \to 0$. In this case, $\lim_{\theta\to 0} d_t = d - d_0$.

In the main manuscript, we use $d - d_0$ to replace d_t, then we can get the reconstructed 3D gaze target (g_x, g_y, g_z) as follow:

$$\begin{aligned} g_x &= \frac{(d-d_o)(g_u-c_u)}{f_u\sqrt{\left(\frac{g_u-c_u}{f_u}\right)^2+\left(\frac{g_v-c_v}{f_v}\right)^2+1}} \\ g_y &= \frac{(d-d_o)(g_v-c_v)}{f_u\sqrt{\left(\frac{g_u-c_u}{f_v}\right)^2+\left(\frac{g_v-c_v}{f_v}\right)^2+1}} \\ g_z &= \frac{d-d_o}{\sqrt{\left(\frac{g_u-c_u}{f_u}\right)^2+\left(\frac{g_v-c_v}{f_v}\right)^2+1}} \end{aligned} \tag{2}$$

where (f_u, f_v, c_u, c_v) indicate the intrinsics of the RGB camera.

2.2. Technical Details of Laser Range Finder

Table 1. Some technical specifications of SK-Pro30

Measurement range↑	Measurement rate ↑	Measurement error↓
0.05m-30m	30Hz	$\leq$ 1mm

The laser rangefinder is SK-Pro30 from Shanghai shenji company. Some technical specifications are listed in Table 1. For more information, please refer to http://en.shsenky.com/index.php?c=show&id=2.

The following analyzes why the rangefinder was chosen to measure distance:

- For Azure Kinect, its standard measurement error is $\leq$ 17mm, which is much larger than that of the rangefinder.
- We move the laser spot randomly within 73s and apply these two devices to measure the distance to the laser spot. All the data obtained by the rangefinder are valid, but 8.7% of depth values acquired by Kinect are invalid and cause failures in calculating distances.

2.3. Flow Chart of Laser Spot Detection

The flow of laser spot inspection is visualized in Fig. 3, which illustrates the entire process.

3. Training and Evaluation Details

In this section, we provide more details about training and evaluation in the experiment section (Section 5). All methods in the experiments are implemented by Pytorch, and we use the early stopping strategy as the criterion for terminating training. For our proposed method, the batchsize is set to 32, Adam is used as the optimizer and the learning rate is 1×10^{-4}.

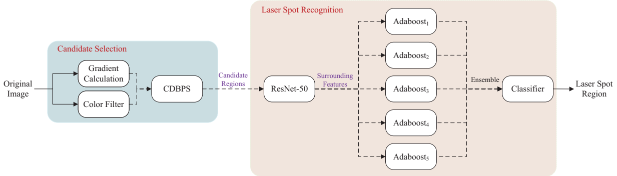

Figure 3. Flow chart of laser spot detection.

We have introduced in the experiment setup of the main paper that the 2D gaze-following method [1, 2, 4, 5] can be used to estimate the 3D gaze following with the help of the registered depth map, and the specific implementation is briefly described below:

Assume that the pixel location of the gaze target predicted by these methods in the RGB image is $(\hat{u}, \hat{v})$, we set a rectangular area $R \in w \times h$ in the registered depth map $\mathcal{D}$ with it as the center. This region is then cropped from the depth map and denoted as $\mathcal{D}_R$. After removing invalid values from it, we can get a set of depth values denoted as D'. Assuming that the size of the set D' is M, then the estimated 3D gaze target $(\hat{g}_x, \hat{g}_y, \hat{g}_z)$ in RGB camera coordinate system is as follow:

$$\begin{aligned} \hat{g}_z &= \tfrac{1}{M}\textstyle\sum_{i=1}^{M} D'(i) \\ \hat{g}_x &= \hat{g}_z\left(\hat{u}-c_u\right)/f_u \\ \hat{g}_y &= \hat{g}_z\left(\hat{v}-c_v\right)/f_v \end{aligned} \tag{3}$$

where (f_u, f_v, c_u, c_v) indicate the intrinsics of the RGB camera.

4. Further Analysis of the Experiment Section

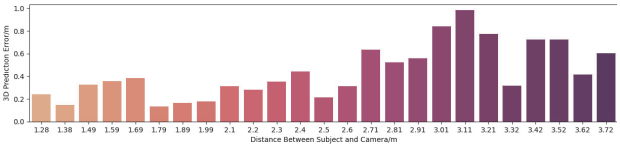

Figure 4. Relationship between the performance of the model and the distance between the camera and the subject.

4.1. Effect of distance on model performance

We made statistics on the relationship between performance of the model and the distance between the camera and subject in the test set of the GFIE dataset. Figure 4 shows that the prediction error of the model increases gradually with the distance. In the figure, the distance between the subject and the camera is divided into 25 intervals with a length of approximately 0.1m, the coordinates of the x-axis represent the midpoint of each interval, and the y-axis represents the average prediction error (with metric 3D Dist.) within this interval.

4.2. Evaluation on CAD-120 Dataset

Table 2. Performance comparison on the CAD-120 dataset

	2D		3D	
Method	AUC ↑	L^2 Dist. ↓	3D Dist. ↓	Angle Error ↓
Random	0.469	0.758	1.910	70.3°
Center	0.456	0.706	1.280	75.9°
GazeFollow [5]	0.862	0.196	1.030	44.1°
Lian [4]	0.871	0.180	0.813	34.8°
Chong [1]	0.891	0.152	0.812	31.9°
Rt-Gene [2]	0.463	0.492	0.483	26.5°
Gaze360 [3]	0.463	0.474	0.427	20.6°
GFIE (ours)	**0.921**	**0.114**	**0.365**	**19.8°**

In section 5.3 of the main paper, we describe the evaluation on the CAD-120 dataset, and provide a brief comparison between *Chong* [1] and our methods. For further analysis, we present the complete quantitative results evaluated on the CAD-120 dataset in Table 2.

Quantitative analysis in Table 2 shows that our proposed baseline method outperforms all comparison methods. In addition, we also draw the following conclusions:

- The performance of all methods in Table 1 degrades comparing to the evaluation on the GFIE dataset, which indicates that it is challenging to perform the test on the CAD-120 dataset due to different scenes, images sizes and camera parameters from the GFIE dataset.

- The excellent performance on cross-dataset generalization shows that our proposed method can effectively copy with unseen scenes and different camera settings.

5. Limitation and Future Work

Dataset: One of the limitations is that all scenes are located indoors. In the future. we will collect the gaze behavior of pedestrians outdoors.
Baseline method: Considering that the proposed baseline method requires multimodal input, it is unsuitable for 2D gaze-following based on a single RGB image. One solution is to use monocular depth estimation to generate the depth map. In the future, we will explore to achieve gaze following whether the input is RGB image or multimodal information.
Application: Another promising work is applying gaze following to human-robot interaction, which helps robots consider human intentions when providing services.

Institutional Review Board Statement

For the private data, all studies have been approved by the Nankai University Institutional Review Board (IRB).

References

[1] 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

[2] 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. 2, 3

[3] 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. 3

[4] 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. 2, 3

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