GazeOnce: Real-Time Multi-Person Gaze Estimation

Mingfang Zhang[1,2], Yunfei Liu[3], Feng Lu[1,3,*]
[1]Peng Cheng Laboratory, [2]The University of Tokyo,
[3]State Key Lab. of VR Technology and Systems, School of CSE, Beihang University
mfzhang@iis.u-tokyo.ac.jp, lyunfei@buaa.edu.cn, lufeng@buaa.edu.cn

Abstract

Appearance-based gaze estimation aims to predict the 3D eye gaze direction from a single image. While recent deep learning-based approaches have demonstrated excellent performance, they usually assume one calibrated face in each input image and cannot output multi-person gaze in real time. However, simultaneous gaze estimation for multiple people in the wild is necessary for real-world applications. In this paper, we propose the first one-stage end-to-end gaze estimation method, GazeOnce, which is capable of simultaneously predicting gaze directions for multiple faces (>10) in an image. In addition, we design a sophisticated data generation pipeline and propose a new dataset, MPSGaze, which contains full images of multiple people with 3D gaze ground truth. Experimental results demonstrate that our unified framework not only offers a faster speed, but also provides a lower gaze estimation error compared with state-of-the-art methods. This technique can be useful in real-time applications with multiple users.

1. Introduction

Eye gaze is one of the important channels in revealing human intentions. It has been adopted for a wide range of applications such as human-computer interaction [23], virtual/augmented reality [3, 26], medical diagnostics [4], and surveillance systems [17]. To estimate the gaze direction, various systems have been developed. However, fast and accurate calculation of gaze direction in a large range of environment remains challenging.

With the development of deep learning, appearance-based gaze estimation has attracted more and more attention, *i.e.*, gaze estimation using face images captured by common cameras. The main drawbacks of existing methods are: 1) they usually only support the gaze estimation for a single person, while multi-person with different head poses have been less explored; 2) they need to pre-process the images, *i.e.*, cropping and calibration of the face images, resulting in a longer computation time. Fig. 2 illustrates the typical flow of existing systems. It first extracts face ROI using a face detector, calibrates each face using the detected facial landmarks, and then the normalized face is fed into the gaze direction estimation system. It can be seen that the system errors accumulated after these steps. Moreover, their computational complexity is proportional to the number of people in the image, and they normally cannot operate in real time when there are more than 5 faces in each frame, as shown in Fig. 1.

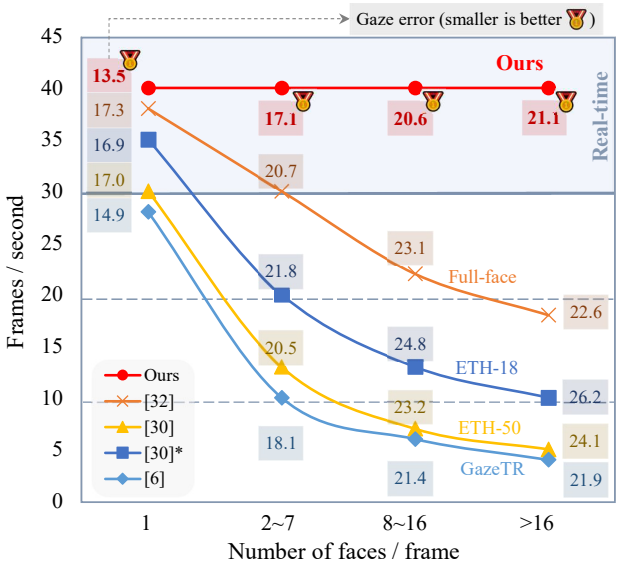

Figure 1. Compared with previous appearance-based gaze estimation methods [6, 30, 32], our method is the only one that can maintain the real-time speed as the number of faces increases in the input image. Consider the average gaze accuracy across different face resolutions, our method also achieves the best performance. The experiment setting is the same as Tab. 3.

In this paper, we reframe the multi-person gaze estimation as a single-stage regression task, which directly maps image pixels to multiple gaze directions. Specifically, we propose the first one-stage gaze estimation method,

*Corresponding author.

This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant 61972012.
This work is done during M. Zhang's internship at Pengcheng Lab.

Figure 2. Comparison between existing appearance-based gaze estimation (AGE) methods and our method. AGE methods usually conduct localization, normalization, and gaze estimation for each face one by one. We present the first one-stage method to simultaneously estimate gaze directions for multiple people in one pass.

i.e., GazeOnce, which estimates all human gaze directions within one pass. The proposed method not only estimates gaze directions but also predicts auxiliary face information including bounding box and facial landmarks. In addition, we carefully design a projection-based self-supervision loss for 3D gaze estimation.

Another difficulty to overcome is about the dataset. Appearance-based gaze estimation relies on high-quality datasets with face images and ground truth gaze directions. However, obtaining gaze ground truth is very challenging. Many gaze datasets have been released [30, 32], while they usually record data for each single person in a strictly controlled environment, leading to limited image styles and body poses. On the other hand, manual annotation of 3D gaze directions is time-consuming and error-prone.

In order to train our GazeOnce method, a new high quality dataset is needed with multiple people and their gaze ground truth in every image. To this end, we propose a sophisticated gaze swapping method for generating a high-quality multi-person gaze dataset. The proposed MPSGaze dataset has no restrictions on the number of people and scenes, and is also easily extensible. This makes the training and evaluation of multi-person gaze estimation possible.

Based on the proposed dataset, our method not only achieves real-time multi-person gaze estimation, but also outperforms state-of-the-art methods in terms of estimation error and running time, as shown in Fig. 1.

In summary, our main contributions are as follows:

- We propose the first one-stage gaze estimation method, *i.e.*, GazeOnce, which can estimate multi-user gaze directions simultaneously in a single image. In addition, we design a projection-based self-supervised strategy that can further improve the gaze accuracy.
- We provide a new gaze dataset MPSGaze, which enables one-stage gaze estimation training and evaluation. This dataset is generated by a sophisticated swap-gaze procedure to produce full images of multi-person with their gaze ground truth.
- Our method outperforms state-of-the-art methods in terms of gaze accuracy and speed, especially in the cases of a large number of faces.

2. Related Works

Appearance-based gaze estimation (AGE). AGE has been a long-standing computer vision problem [15, 16]. Recent deep learning-based AGE methods [7, 9, 31, 32] have significantly improved the accuracy using various strategies, such as a coarse-to-fine approach [5], an adversarial learning approach [24], a self-attention approach [2], *etc*. At the same time, large-scale gaze datasets [9, 13, 22, 30] have been proposed. Most of them are collected in laboratory environments with a strict setting of multi-view cameras, 3D positions of human participants and gaze targets, *etc*. This procedure always results in these datasets containing only single face images in a limited number of scenes. Correspondingly, current AGE methods all assume that there is only one calibrated face in the input image. However, this will lead to a disadvantage that the speed of current AGE methods depends on the number of faces in the input image. Most methods cannot achieve real-time performance when there are multiple people in the image.

Real-time multi-face process. Face understanding receives keen attention because of its wide range of applications. Many methods have been proposed for face localization [18], facial expression recognition [27], head pose estimation [1], *etc*. With the development of object detection methods [14], one-stage methods for multi-face understanding are favored by real-time applications because of their lightweight design and high accuracy. For example, face detection methods [8, 18] apply the one-stage architecture and design more efficient modules for face characteristics. Correspondingly, large-scale face datasets [28] have been constructed by employing a large number of manual annotations. In addition, researchers find that it is an efficient method [20, 29] to conduct multi-task (landmark, head pose, gender, *etc*.) learning alongside face detection because

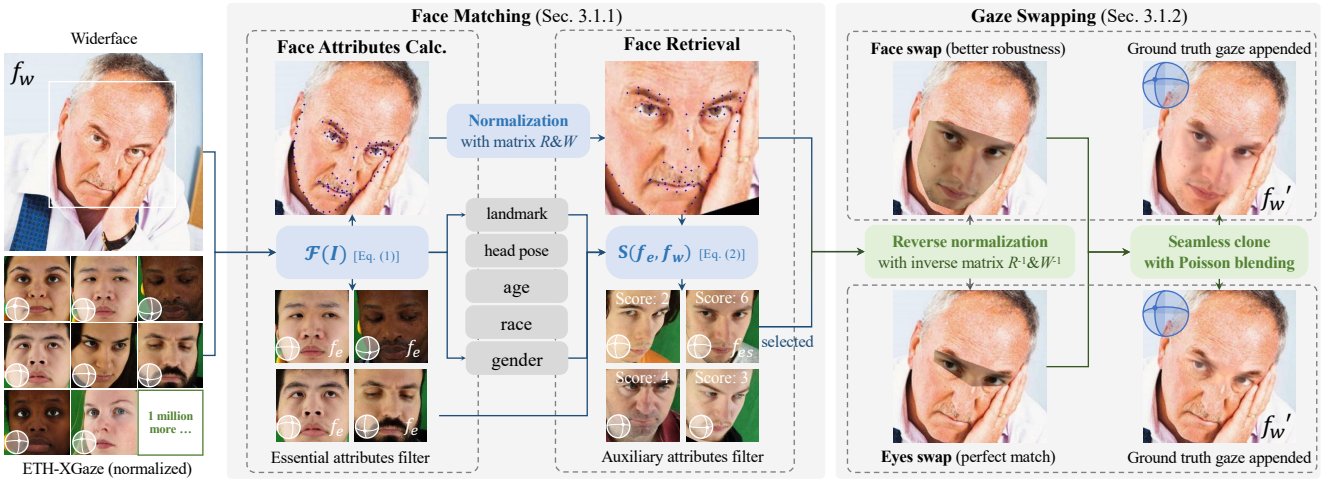

Figure 3. Generation of the MPSGaze. To create a dataset with full images of multi-person with gaze ground truth, we conduct gaze swapping between the Widerface [28] dataset (with face bounding box labels) and the ETH-XGaze [30] dataset (with gaze labels). The pipeline consists of 2 phases, matching and swapping. For each qualified face in Widerface, we retrieve the nearest face from ETH-XGaze by filtering various face attributes. Based on the matching result, we design 2 strategies to swap gaze, *i.e.*, face exchange and eyes exchange.

these tasks share common facial features. Inspired by these works, we propose to develop a one-stage gaze estimation method.

3. Multi-Person Swap Gaze Dataset

We propose a new Multi-Person Swap Gaze Dataset, MPSGaze, for our task of gaze estimation for multiple people in one stage. To the best of our knowledge, existing datasets either only contains face information (*e.g.*, bounding box, landmark, *etc.*) or contains normalized single faces with gaze labels. Therefore, our first obstacle is to construct a dataset that contains both multiple people in the wild and corresponding gaze ground truth. To this end, we propose to merge the advantages of face datasets and gaze datasets to enable the training and evaluation of multi-person gaze estimation in one stage. In the following, we first introduce the pipeline of generating the MPSGaze dataset, then show the details of the dataset.

3.1. Generation Pipeline

We choose the largest and most common gaze dataset available, ETH-XGaze [30], and the face detection dataset, Widerface [28], for our task. The proposed approach consists of two phases, face matching and gaze swapping, as shown in Fig. 3.

3.1.1 Face Matching

The left part of Fig. 3 shows the process of face matching between the two datasets. First, we conduct face attributes calculation for each qualified face in Widerface [28] and ETH-XGaze [30] by

$$\mathbf{A} = \mathcal{F}(I), \tag{1}$$

where $\mathbf{A}$ is the attributes extracted from a single-face image I. Here $\mathbf{A} = \{\mathbf{a}_{lmk}, \mathbf{a}_{pose}, \mathbf{a}_{age}, \mathbf{a}_{race}, \mathbf{a}_{gender}\}$ and $\mathbf{a}_{lmk} \in \mathbb{R}^{68\times2}$, $\mathbf{a}_{pose} \in \mathbb{R}^2$, $\mathbf{a}_{age} \in \mathbb{R}^9$, $\mathbf{a}_{race} \in \mathbb{R}^7$, $\mathbf{a}_{gender} \in \mathbb{R}^2$. The function $\mathcal{F}$ is implemented by state-of-the-art methods [12, 25].

Next, for each qualified face f_w in the Widerface [28], we retrieve the nearest face f_{es} from the ETH-XGaze [30]. The implementation of our retrieval is as follows. 1) We first conduct an essential-attribute (we choose gender) filter for faces in ETH-XGaze [30] to match with f_w. The chosen faces are called f_e. 2) f_w is normalized according to its landmarks and the normalization steps are consistent with ETH-XGaze [30]. 3) We save the image warp matrix W and the head pose rotation matrix R. 4) We calculate the difference in landmarks and head poses between f_w and faces in f_e. Jointly we compute a matching score for each face in f_e by a scoring function

$$\mathcal{S}(f_e, f_w) = \sum_{\tau\in\{lmk,pose\}} \alpha_\tau * |\mathbf{a}_{\tau,w} - \mathbf{a}_{\tau,e}|, \tag{2}$$

where α_τ is determined by the experience of comparing the matching results. 5) We keep n highest-scoring faces for the final filtering, auxiliary-attribute filtering, where we penalize the score of the left n faces by a joint measurement of age and race differences. 6) Finally, we choose the face f_{es} with the final highest score in f_e as a match for f_w from Widerface [28].

3.1.2 Ground Truth Available Gaze Swapping

We propose a gaze swapping method to produce synthetic face images f'_w with ground truth gaze g'. Through affine

Figure 4. Gaze swap results (second row) with ground truth gaze labels in our MPSGaze dataset. Please zoom in for details.

transformations, the ground truth gaze of f_{es} can be preserved. The procedure is shown in the right part of Fig. 3.

Given a pair of matched faces f_w and f_{es}, we swap the gaze as follows. 1) We first warp f_{es} to match with the original f_w through W^{-1} and R^{-1}, where W and R are image warp matrix and head pose rotation matrix. These two matrices are pre-computed in the step of Face Matching. 2) Then we load the matching error computed by Eq. (2) according to the landmarks and head poses difference between the face pair. 3) When the error is under a given threshold, we only swap the eye region by replacing f_w with f_{es}. 4) Otherwise, we keep the whole face region of f_{es} to replace f_w. This operation produces more robust results because the head pose information of face f_{es} is preserved. 5) Next, we employ the Poisson blending method [19] to seamlessly fuse the two overlapped faces for the sake of reality. 6) Finally, we apply the inverse rotation matrix R^{-1} on the original gaze label of f_{es} as the ground truth gaze g' for the final swapped image f'_w.

3.2. Details of MPSGaze

In the Widerface [28] training dataset, 24282 faces are swapped leaving faces that are too small, too blurry, and with too much occlusion unswapped. Some examples are shown in Fig. 4. As shown in Tab. 1, compared to other gaze datasets, the MPSGaze contains images of ∼20 thousand people subjects with multiple faces per image. In addition, MPSGaze shares the advantages of [28] that it contains images in a large variety of wild scenes which are specially designed.

4. Method

After acquiring the MPSGaze dataset which enables the training of one-stage multi-person gaze estimation, we propose the GazeOnce framework. The architecture overview is illustrated in Fig. 5. The input to our model is a full image containing any number of faces and the output is multi-user gaze directions. Instead of processing every face one by one, we propose the first model to estimate gaze for multiple people in one stage.

Gaze dataset	# people	# faces/image	Unconstrained
MPIIGaze [32]	15	1	×
ETH-XGaze [30]	110	1	×
Gaze360 [13]	238	1	×
MPSGaze (ours)	$\sim 10^4$	1∼30	✓

Table 1. Comparison with other gaze datasets. In existing datasets, limited number of *people* subjects are requested to look at preset targets to collect *constrained* gaze data.

4.1. Multi-task Learning for Face Detection and Gaze Estimation

We equip the proposed GazeOnce with a multi-task learning strategy, *i.e.*, jointly optimizing face localization and gaze estimation. Inspired by the RetinaFace [8], we employ a similar architecture for our task. The proposed GazeOnce mainly consists of two components: *feature extractor* and *downstream heads*. The *feature extractor* aims to encode different faces from the input image I into latent codes. To get a rich embedding in which faces with different scales can be treated equally, we adopt the feature pyramids and the context modules from [8] as the feature extractor. Specifically, different levels of the feature pyramid produce features of different scales computed from the output of the corresponding stages of MobileNet [21] or ResNet [11] using top-down and lateral connections. Next, for each feature level, context modules [18] are implemented to increase the receptive field. The feature extraction is effective for both the face detection task and the gaze estimation task because they share all-face information, such as the head pose, besides the eye-region information. Similar conclusions have also been made in [32].

After the feature extraction, 1×1 convolutions are used as *downstream heads* for different tasks. For the face detection task, we employ three heads, namely classification head, localization head, and landmarks head. These three heads are used to predict probabilities of existence y_p, bounding boxes y_b, and positions of landmarks y_l, respectively. We design a 3D gaze head and three auxiliary projection heads for the gaze estimation task. These heads estimate the 3D gaze y_g and three 2D gaze projections y_F, y_T, and y_S. For each training anchor i, we minimise a multi-task loss:

$$\mathcal{L} = \alpha \mathcal{L}_{face} + \beta \mathcal{L}_{gaze}, \tag{3}$$

where:

$$\begin{aligned} \mathcal{L}_{face} =& \mathcal{L}_{class}(y_p^i, y_p^{i*}) + \lambda_1 y_p^{i*} \mathcal{L}_{box}(y_b^i, y_b^{i*}) \\ &+ \lambda_2 y_p^{i*} \mathcal{L}_{landmark}(y_l^i, y_l^{i*}), \end{aligned} \tag{4}$$

where notations with $*$ are the corresponding ground truth. Landmark regression is an auxiliary task to benefit face detection which is proven in [8, 29]. We will introduce $\mathcal{L}_{gaze}$ in the next section.

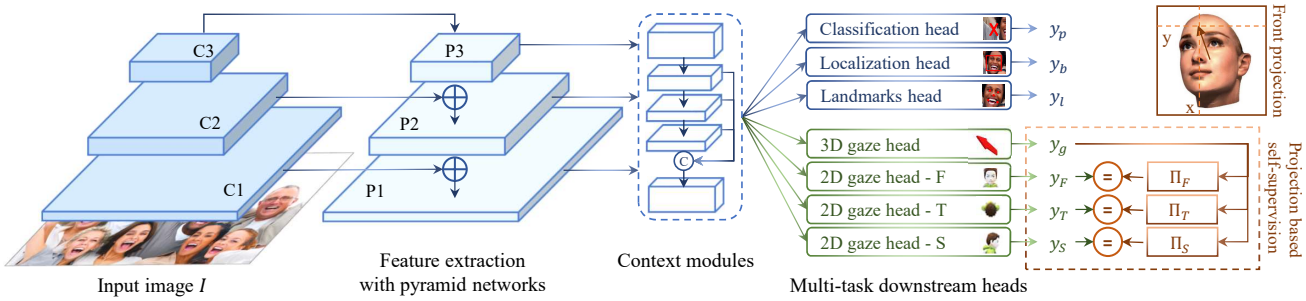

Figure 5. Overview of the GazeOnce framework. The feature extraction is based on feature pyramids followed by context modules [18], which is adopted from RetinaFace [8]. Next, we calculate a joint loss for gaze estimation and face localization for each positive anchor. To achieve higher gaze accuracy, we propose to project the 3D gaze from 3 directions as an auxiliary supervision signal and design a self-supervision loss function to constrain the predictions from different views to be equal.

4.2. 2D Projection-based Self-supervision for 3D Gaze Estimation

We propose a projection-based self-supervision technique for our method. The idea is to project the 3D gaze onto three planes to form three 2D gaze estimation subtasks, which constrains the original 3D gaze estimation task in a self-supervised manner.

Inspired by 3D head pose estimation work [1] which projects the 3D pose onto the image plane for supervision, we apply the projection operation in the 3D gaze estimation task. However, we notice that projecting a 3D gaze onto different planes may result in different performances.

Fig. 6 shows three projections of a 3D gaze onto three planes, namely the front plane (image plane), the side plane, and the top plane. When there is a certain variation dg in the 3D gaze, its projection variations in pixels are different on the three planes. For instance, the 2D projection point on the side plane falls near the origin, and thus the corresponding pixel variation is larger.

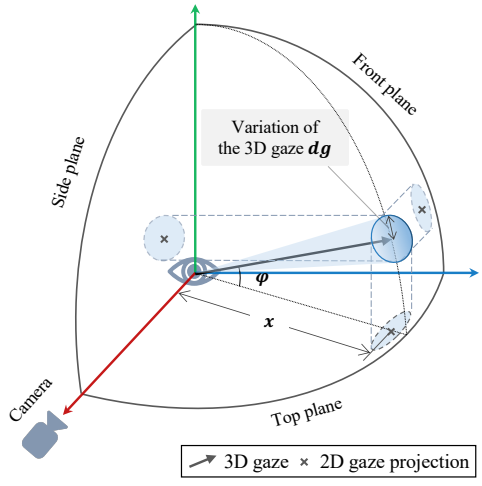

Figure 6. Visualization of 2D projections of a 3D gaze onto 3 planes. When there is a variation dg in the 3D gaze, its projection variations in pixels are different on the three planes.

To mathematically model such differences, we introduce the concept of *2D Gaze Sensitivity (GS)*:

$$GS = \frac{dg}{dx} = \frac{d\varphi}{dx} = \frac{r}{\sqrt{r^2 - x^2}}. \tag{5}$$

GS defines the consequential variation dg of the 3D gaze angle with respect to a change dx at the position x on the projection plane. According to Eq. (5), the further x is apart from the origin, the larger the 2D GS is.

Clearly, lower 2D GS means lower uncertainty in the corresponding 3D gaze direction given the fixed pixel resolution, which is good for our task. However, as shown in Fig. 6, the 2D GS can be large in two projection planes with large x values. Therefore, we propose to use all of the three projections on different planes to ensure the existence of at least one lower 2D GS.

We implement this idea in our network. As shown in Fig. 5, besides the 3D gaze task, we introduce three additional sub-tasks to estimate three sets of 2D gaze points on the projection planes, namely y_F, y_T and y_S, where 'F', 'T' and 'S' stand for front, top, and side.

Then, a self-supervised mechanism can be constructed by checking the consistency between each of the three 2D gaze outputs and three projections Π_F, Π_T and Π_S of the 3D gaze output y_g respectively as shown in Fig. 5. Specifically, the three projections of y_g follow these equations:

$$\begin{aligned}
\Pi_F(\theta, \phi) &= [-r * \sin\phi * cos\theta, \quad -r * \sin\theta], \\
\Pi_T(\theta, \phi) &= [r * \cos\phi * cos\theta, \quad -r * \sin\phi * cos\theta], \\
\Pi_S(\theta, \phi) &= [-r * \cos\phi * cos\theta, \quad -r * \sin\theta],
\end{aligned} \tag{6}$$

where Π is the projection function, r is the half face width, and $\{\theta, \phi\}$ are the *pitch* and *yaw* components of y_g.

As explained above, by projecting the 3D gaze onto three planes simultaneously, there is always at least one projection with low 2D GS, which is favorable for the estimation. This is supported by the results in Tab. 4.

Table 2. Running speed comparison on NVIDIA GeForce RTX 2080. Our one-stage gaze estimation method can run at almost the same speed as RetinaFace [8] (SOTA face detection method). Assuming that existing AGE methods [6, 30, 32] employ RetinaFace [8] for face detection costing time T (25ms on average), their average values of running speed tested on the Widerface [28] validation set are shown in the table.

Method	Ours (MobileNet)	Full-face [32]	ETH-18 [30]	ETH-50 [30]	GazeTR [6]
Time/image (ms)	24.93	$T(\approx 25)$+1.21×*#face*	T+3.15×*#face*	T+6.64×*#face*	T+9.98×*#face*

Table 3. Gaze error evaluated on the MPSGaze. It shows comparison between our method and existing AGE methods, including Full-face [32], ETH-18 [30], ETH-50 [30], and GazeTR [6] (trained on ETH-XGaze dataset [30]). Our method shows higher accuracy in grading comparisons of faces with various scales, even when compared to the transformer-based method [6].

Method	Backbone	Input	Gaze error (lower is better) w.r.t. **the width of faces**							
			30-60	60-90	90-120	120-150	150-180	180-210	210-240	>240
Full-face	AlexNet	1 normalized face	24.99	20.00	17.56	17.03	16.47	14.74	13.43	12.31
ETH-18	ResNet18	1 normalized face	28.89	21.93	16.66	14.90	14.33	12.44	11.68	10.32
ETH-50	ResNet50	1 normalized face	29.82	21.87	16.93	14.76	13.87	11.79	11.13	9.98
GazeTR	ResNet18	1 normalized face	24.51	16.84	14.59	13.37	13.65	11.72	10.71	9.96
Ours	MobileNet0.25	1 full image	22.94	17.55	13.69	11.08	11.13	9.41	8.17	7.74
Ours	ResNet50	1 full image	**21.17**	**13.77**	**10.58**	**7.9**	**8.57**	**6.68**	**6.01**	**5.56**

Loss Design. By constraining the three 2D gaze estimation points y_F, y_T, y_S on the three planes to be equal to the three projections of the 3D gaze estimation direction y_g, our self-supervision improves the final gaze estimation accuracy. The self-supervision is implemented by using the following loss function:

$$\mathcal{L}_{self} = \sum_{\tau \in \{F,T,S\}} |y_\tau - \Pi_\tau(y_g)|_1 * e^{-p_\tau} + p_\tau, \quad (7)$$

where F, T, S represents front, top, and side, Π functions are from Eq. (6), and p is trainable weights [10] associated with each projection plane. Finally, the total loss is $\mathcal{L} = \alpha\mathcal{L}_{face} + \beta\mathcal{L}_{gaze}$, where $\mathcal{L}_{face}$ is defined in Eq. (4) and $\mathcal{L}_{gaze}$ is defined as:

$$\mathcal{L}_{gaze} = \lambda_1\mathcal{L}_{self} + \lambda_2|y_g - y_g^*|_1 + \lambda_3 \sum_{\tau \in \{F,T,S\}} |y_\tau - \Pi_\tau(y_g^*)|_1, \quad (8)$$

where $*$ represents the ground truth and other expressions are the same as in Eq. (7).

5. Experiments

5.1. Experimental Setup

Our evaluation is mainly conducted on the test set of MPSGaze which is based on the validation set of Widerface [28] and ETH-XGaze [30], and we conduct gaze swap on 6277 faces of different resolutions. To match the input format of existing AGE methods, we conduct cropping and normalization on these faces. The experiments show that our method not only achieves higher accuracy and speed on synthetic data in MPSGaze, but we also test on real full images with gaze annotation from human experts and our method still performs better than existing AGE methods.

5.2. Comparison with Existing AGE Methods

We compare our method with 4 full-face appearance based gaze estimation methods [6, 30, 32]. They are all trained on the ETH-XGaze [30] dataset to match the source of the gaze-swap of our test data which is also ETH-XGaze [30]. As shown in Tab. 2 and Tab. 3, Full-face [32] is the earliest to be proposed and its speed is relatively high but it shows the worst accuracy. ETH-18 and ETH-50 [30] are the models trained with ResNet18 and ResNet50 [11] as backbones, where ETH-50 is used as the baseline method published in [30]. GazeTR [6] is the latest method based on transformer design which achieves the highest accuracy among the four methods. However, it runs the slowest and cannot achieve real-time performance if there is more than one face in an image.

We test our method with MobileNet [21] and ResNet50 [11] as backbones. The speed evaluation in Tab. 2 shows that our method performs the best because our method can achieve similar speed as the RetinaFace [8], the SOTA face detection method. Assuming that the four existing AGE methods mentioned above use RetinaFace [8] for face detection, their speed comparison with our method is clearly lagging behind even without considering the time spent on the normalization based on facial landmarks. In Tab. 2, the

Table 4. Ablation studies on the MPSGaze test set. Constraining **the width of faces** and **the angle of ground truth gazes**, two comparisons are conducted. It is worthy of attention that, compared to the model that only predicts 3D gaze (✓,×,×), the model that only predicts front-projection 2D gaze (×,F,×) achieves higher accuracy in the small-angle range (0-60°) and lower accuracy in the large-angle range due to the uneven distribution of the *Gaze Sensitivity* (Eq. (5)). This table shows the advantage of our full model with F, T, S (Eq. (6)) and $\mathcal{L}_{self}$ (Eq. (7)).

3D gaze task	2D gaze task	$\mathcal{L}_{self}$	Gaze error (lower is better) w.r.t. **the width of faces**							
			30-60	60-90	90-120	120-150	150-180	180-210	210-240	>240
✓	×	×	24.51	19.51	15.59	13.23	12.84	11.34	10.68	9.76
×	F	×	24.00	18.73	14.56	12.6	12.22	10.83	8.77	8.99
✓	F	F	23.44	18.10	14.16	11.76	11.09	10.01	9.15	8.09
✓	F, T, S	F, T, S	**22.94**	**17.55**	**13.69**	**11.08**	**11.13**	**9.41**	**8.17**	**7.74**
3D gaze task	2D gaze task	$\mathcal{L}_{self}$	Gaze error (lower is better) w.r.t. **the angle of GT gaze**							
			0-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
✓	×	×	9.12	9.94	10.83	11.59	14.3	18.67	26.86	**43.04**
×	F	×	8.56	8.49	9.75	10.77	12.64	18.37	30.2	54.35
✓	F	F	8.01	8.66	9.21	10.04	11.52	17.12	26.57	43.75
✓	F, T, S	F, T, S	**7.97**	**8.45**	**8.99**	**9.51**	**10.53**	**16.36**	**23.31**	43.4

larger the *#face*, the slower the speed is. With the absolute speed advantage, Tab. 3 also shows that our method has the highest accuracy.

5.3. Ablation study

Tab. 4 shows the ablation study of our method. Our full model contains 4 gaze tasks: one 3D gaze task (*pitch*, *yaw*) and three projected 2D gaze tasks (x, y) in which the projection in the front direction has a one-to-one correspondence with the 3D gaze, thus it can be easily transformed to *pitch* and *yaw* and compared with the ground truth. In addition, according to Eq. (7), we also add equal-loss to the 4 gaze predictions, which is also proven to be effective in Tab. 4.

There are two parts in Tab. 4 constraining the target face size and ground truth gaze direction, respectively. For each table there are 4 rows of data. The first row indicates that the model only regresses 3D gaze; the second row indicates that the model only regresses 2D gaze projection in the front direction which is later converted to 3D gaze; the third row indicates that the model regresses both 3D gaze and 2D gaze projection (front) with a training loss restricting them to be equal; and the fourth row indicates that the model regresses all 4 gaze values and restricts them to be equal. The results in the table show the advantages of our full model. It is worthy of attention that, as we described in Method Sec. 4.2, compared to the model that only regresses 3D gaze, the model that only regresses 2D gaze in the front direction has higher accuracy in the small-angle range (0-60°) and lower accuracy in the large-angle range due to the uneven distribution of the *Gaze Sensitivity* (Eq. (5)). To solve this problem we propose to project 3D gaze from different directions, which solves the problem theoretically and achieves better results.

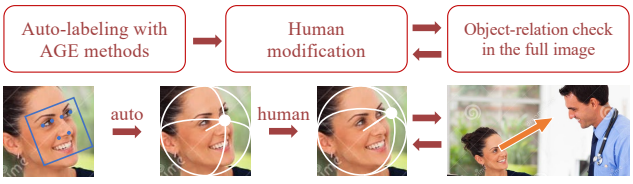

Figure 7. Human expert annotation pipeline. First, each face is cropped out, normalized, and then [30] is employed to generate preliminary gaze labels. Second, human experts are asked to modify the gaze labels. Third, the gaze labels are shown in the original full image, and experts can modify them again according to object relations to get the final annotation.

5.4. Evaluation on Human Annotation Dataset

Besides testing on swap-gaze synthetic data, we also conduct evaluation on real full images. To get corresponding gaze labels, we ask some experts to conduct a subjective gaze annotation and finally acquire gaze annotation for 2719 faces in the Widerface [28] validation set. The annotation process is shown in Fig. 7. First, each face is cropped out for existing AGE methods to automatically generate preliminary gaze labels. Next, human experts are asked to modify the gaze labels and the gaze is then shown in the original full image so experts can modify them again according to object relations to get the final annotation. We also develop a GUI software to facilitate human experts to

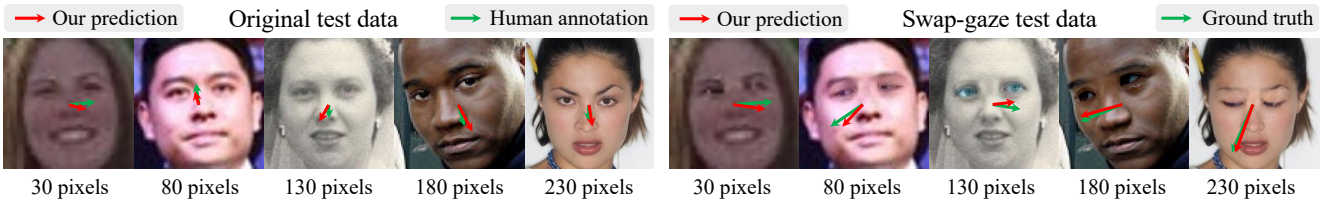

Figure 8. Visualization of predicted gaze on cropped single faces. Faces with various resolutions in human annotation dataset (left) and MPSGaze test set (right) are cropped from the whole images and resized for better visualization. Note that these results are generated by the MobileNet version of our model.

Figure 9. Visualization of multi-person gaze estimation on full images. The gaze directions of different people are estimated at the same time by the MobileNet version of our model.

conduct the annotation.

Tab. 5 shows the comparison between our method and existing AGE methods. Although humans cannot perform well in 3D annotation tasks, we can consider this experiment as a subjective test. The clear leading rank of our method shows our advantage. From another perspective, the relatively large test error shows the necessity of our proposed MPSGaze dataset with ground truth labels. We also show the visualization results in Fig. 8 and Fig. 9.

Table 5. Comparison on the human annotation dataset. Our methods (m: MobileNet0.25 backbone, r: ResNet50 backbone) show higher accuracy than existing AGE methods [6, 30, 32] (trained on the ETH-XGaze [30] dataset).

Method	Gaze error w.r.t. the width of faces				
	0-60	60-120	120-180	180-240	>240
Full-face	36.06	35.7	33.38	25.37	21.67
ETH-18	31.95	31.38	28.5	22.85	19.93
ETH-50	30.43	31.24	28.31	22.11	18.79
GazeTR	36.00	33.53	31.10	26.81	23.59
Ours-m	26.06	25.88	24.02	22.27	18.41
Ours-r	**25.90**	**22.69**	**22.02**	**19.92**	**15.54**

6. Limitation and Future Work

First, our method cannot produce estimates for people who show their backs to the camera or look towards the back side of the scene. This is a common problem for existing appearance-based gaze estimation methods while it is inevitable in the real world. In future research, such back-to-camera situations can be further considered and tried to handle. Second, although we propose an effective method to synthesize full images with multi-person gaze ground truth, it could be still worth considering to try to collect real data directly with accurate gaze directions of multiple people in the wild.

7. Conclusion

We propose the first one-stage gaze estimation method, *i.e.*, GazeOnce, which can estimate multi-user gaze directions simultaneously in a full image. In addition, we design a projection-based self-supervised strategy that can further improve the gaze accuracy. To enable one-stage gaze estimation training and evaluation, we provide a new gaze dataset, MPSGaze, which is generated by a sophisticated swap-gaze procedure to produce full images of multi-person with gaze ground truth data. Finally, our method outperforms state-of-the-art methods in terms of gaze accuracy and speed.

References

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7617–7627, 2021. 2, 5

[2] Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. Adaptive feature fusion network for gaze tracking in mobile tablets. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9936–9943. IEEE, 2021. 2

[3] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. 1

[4] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–10, 2020. 1

[5] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10623–10630, 2020. 2

[6] Yihua Cheng and Feng Lu. Gaze estimation using transformer. *arXiv preprint arXiv:2105.14424*, 2021. 1, 6, 8

[7] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29, 2020. 2

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 2, 4, 5, 6

[9] 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

[10] Nicola Garau, Niccolo Bisagno, Piotr Bródka, and Nicola Conci. Deca: Deep viewpoint-equivariant human pose estimation using capsule autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11677–11686, 2021. 6

[11] 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. 4, 6

[12] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 3

[13] 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. 2, 4

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[15] Feng Lu, Takahiro Okabe, Yusuke Sugano, and Yoichi Sato. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing*, 32(3):169–179, 2014. 2

[16] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2033–2046, 2014. 2

[17] Alexandre Marois, Laura Salvan, Daniel Lafond, Alexandre Williot, Noémie Lemaire, and Sébastien Tremblay. Improving usability of a gaze-based surveillance support tool through user-centered design. In *International Conference on Applied Human Factors and Ergonomics*, pages 732–740. Springer, 2021. 1

[18] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 4875–4884, 2017. 2, 4, 5

[19] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003. 4

[20] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017. 2

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4, 6

[22] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014. 2

[23] Haofei Wang, Xujiong Dong, Zhaokang Chen, and Bertram E Shi. Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1476–1479. IEEE, 2015. 1

[24] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11907–11916, 2019. 2

[25] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[26] Zhimin Wang, Huangyue Yu, Haofei Wang, Zongji Wang, and Feng Lu. Comparing single-modal and multimodal interaction in an augmented reality system. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 165–166. IEEE, 2020. 1

[27] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2168–2177, 2018. 2

[28] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 2, 3, 4, 6, 7

[29] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2, 4

[30] 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 *European Conference on Computer Vision*, pages 365–381. Springer, 2020. 1, 2, 3, 4, 6, 7, 8

[31] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 2

[32] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–60, 2017. 1, 2, 4, 6, 8