The Visual Computer (2021) 37:1731–1741
https://doi.org/10.1007/s00371-020-01934-1

ORIGINAL ARTICLE

RGB-D-based gaze point estimation via multi-column CNNs and facial landmarks global optimization

Ziheng Zhang[1,2] · Dongze Lian[1,2] · Shenghua Gao[1,2]

Published online: 30 October 2020

Abstract

In this work, we utilize a multi-column CNNs framework to estimate the gaze point of a person sitting in front of a display from an RGB-D image of the person. Given that gaze points are determined by head poses, eyeball poses, and 3D eye positions, we propose to infer the three components separately and then integrate them for gaze point estimation. The captured depth images, however, usually contain noises and black holes which prevent us from acquiring reliable head pose and 3D eye position estimation. Therefore, we propose to refine the raw depth for 68 facial keypoints by first estimating their relative depths from RGB face images, which along with the captured raw depths are then used to solve the absolute depth for all facial keypoints through global optimization. The refined depths will provide us reliable estimation for both head pose and 3D eye position. Given that existing publicly available RGB-D gaze tracking datasets are small, we also build a new dataset for training and validating our method. To the best of our knowledge, it is the largest RGB-D gaze tracking dataset in terms of the number of participants. Comprehensive experiments demonstrate that our method outperforms existing methods by a large margin on both our dataset and the Eyediap dataset.

Keywords Gaze tracking · Human–computer interaction · Multi-column CNNs

1 Introduction

Gaze estimation[1] is an important task and has wide applications in human–computer interaction [17], visual behavior analysis [22], and psychological studies [24]. Traditional methods usually utilize dedicated hardware, such as electrodes attached to the skin around the eye or radar range finder, to estimate eyeball pose and head pose [13,14, 30]. In contrast, recent researches switch the attention to appearance-based methods, where gaze direction or gaze point is predicted using only cameras. The appearance-based methods free the gaze tracking systems from the dependence of special hardware while still retaining satisfactory performance and thus become a desirable approach for gaze estimation. However, existing appearance-based approaches often suffer from varying illuminations, occlusions, image qualities, and head poses. Besides, the cross-subject performance of existing methods still remains unsatisfactory.

On the other hand, the convolutional neural network (CNN) has shown its excellent capability of learning complex visual concepts from large datasets [7,12,25,28]. In particular, CNN is able to learn robust intermediate representation from diverse samples in a large dataset, thereby achieving excellent generalization performance. On this account, many researchers [11,23,38,39,41] have sought to leverage CNN for appearance-based gaze estimation, where superior performance has been achieved compared to traditional approaches. During these attempts, the researchers also observed that CNN would have poor generalizability among different subjects if gaze points are regressed directly from face images, while good cross-subject performance could be achieved by adding eye images or attention maps/masks indicating eye/face positions. We identify the reasons behind

[1] Gaze estimation is also referred to as eye-gaze estimation, eye tracking, and gaze tracking in some literature.

✉ Ziheng Zhang
zhangzh@shanghaitech.edu.cn

Dongze Lian
liandz@shanghaitech.edu.cn

Shenghua Gao
gaoshh@shanghaitech.edu.cn

[1] ShanghaiTech University, Shanghai, China

[2] Shanghai Institute of Microsystem and Information Technology, University of Chinese Academy of Sciences, Shanghai, China

these observations as follows: (1) The models used in previous work were big, while the existing datasets are not large enough in terms of the number of participants, which makes the gaze estimators easy to overfit few subjects in the training data. (2) The cross-subject performance can be improved by explicitly highlighting critical regions such as the eye regions. Therefore, in this work, we introduce a larger RGB-D gaze dataset in terms of the number of participants. In addition, we also propose to explicitly decompose the gaze estimation problem into an eyeball pose, a head pose, and a 3D eye position estimation problem and use only necessary face regions to solve each. To further boost the cross-subject generalization performance, we propose a multi-column CNNs [2] architecture for this task, in which gaze points are estimated from different data sources with multiple small CNNs, and the final results are acquired by averaging these estimations.

Another problem in gaze estimation is whether depth data should be used. In fact, researchers have tried both settings and each seems to be able to produce reasonable results. To make the decision, we first need to analyze the problem itself. As we know, the gaze point of a person is determined by the gaze direction and the distance between the eyes and the display. And the gaze direction is further determined by the head pose and the the eyeball pose [39].[2] Because estimating the gaze point requires the distance between the eye and the display, depth data should be incorporated to provide this distance information. Although in some cases, where the distance between the display and the subject does not change greatly, reasonable performance could be achieved without depth data. On the other hand, even if we only need the gaze direction, depth data also help to provide accurate head pose estimation. Although there are approaches which directly estimate depth from single RGB images [3,32,35,37], as discussed in [40], estimating absolute depth from a monocular color image is highly ill-posed and difficult even for people. Hence, we propose to introduce depth images to facilitate both 3D eye position and head pose estimation.

In practice, the captured raw depth images often contain noises and black holes caused by occlusions, specularity of eyeglasses, and depth range limitations, which is unfavorable for reliable 3D eye position and head pose estimation. Inspired by recent researches on depth completion [1,40], which suggests that holes in the raw depth images can be completed with the help of RGB images, we further introduce a semi-supervised convolutional neural network to complete depth using both RGB and raw depth image. Similar as [40], we choose to use the CNN only for relative depth estimation and solve the absolute depth from the estimated relative depth and the raw depth map via global optimization. Given the fact that our purpose for depth completion is to acquire 3D eye position and head pose, which could be well characterized by 68 3D facial keypoints (see Fig. 4), we only optimize the depths for these facial keypoints.

So far, there are only two gaze tracking datasets, i.e., the Eyediap [20] and the UT-Multiview [31] datasets, containing depth information of faces or eye regions. But both of them are small in terms of the number of participants. The Eyediap dataset provides 86,222 labeled RGB-D video frames collected from only 16 subjects with 12 males and 4 females. On the other hand, though not directly containing depth maps, the UT-Multiview dataset provides 3D face models reconstructed by 8 cameras, from which it is possible to acquire both RGB face images and depth maps. But there are only 50 participants in it, which is still very limited for study DNN-based gaze tracking methods. The study by Krafka et al. shows that more participants will improve the gaze tracking performance [11]. Therefore, we introduce a larger RGB-D gaze tracking dataset in terms of the number of participants. Our dataset consists of over 165K RGB-D images from 218 participants, probably the largest RGB-D gaze dataset readily available to the research community.

The contributions of this paper can be summarized as follows: (1) We build a large-scale RGB-D gaze tracking dataset to facilitate the exploration of data-driven approaches for gaze tracking; (2) we propose to decompose gaze point estimation into eyeball pose, head pose, and 3D eye position estimation and design a multi-column CNNs architecture to complete depth for gaze prediction; and (3) we conduct extensive experiments to show that our new technique outperforms state-of-the-art methods by a large margin.

2 Related work

2.1 RGB image-based gaze tracking

Generally, gaze estimation can be categorized into model-based and appearance-based methods [6]. Model-based methods [42] utilize geometric eyeball models and features for gaze estimation, while the appearance-based approaches directly predict gaze direction or gaze point with face or eye images. We refer readers to [10] for a comprehensive review of model-based methods and early appearance-based methods where hand-crafted features are used. Recent state-of-the-art appearance-based methods usually utilized CNNs to regress gaze direction or gaze points from face or eye images. Zhang et al. [38] introduced an in-the-wild RGB gaze dataset and learned the mapping from the 2D head angle and eye images to gaze angles with multimodal CNNs. Krafka et al. [11] collected a mobile-based RGB eye tracking dataset and designed a CNN-based architecture where face images, two eye images, and a binary mask indicating the location

[2] The eyeball pose has also been denoted as the eyeball movement in [41].

and size of the head are used to regress the gaze point on a mobile phone. In [39], a spatial weights CNN was proposed to generate a weight map from a face image, which was applied to the feature tensor extracted from the face image. Then, the weighted feature tensor was fed into a few fully connected layers to output the final 2D or 3D gaze point. Recent researchers have noticed the effect of head-gaze correlation overfitting and started to explicitly decompose gaze estimation into head pose and eyeball pose estimation. Zhu and Deng [41] proposed to use two CNNs for head pose and eyeball pose estimation, respectively. Then, a gaze transform layer was introduced to aggregate them into the final gaze prediction. Ranjan et al. [23] made a further attempt toward head pose independent gaze estimation by introducing a branched CNN, of which each branch shared the same backbone network and accounted for estimating the gaze direction for the corresponding head pose cluster.

As one might have noticed, most of the methods above are person-agnostic, which means that once trained, the gaze estimator can be directly applied to any person without extra fine-tuning or adaption steps. In fact, there are plenty of researches targeting person-specific gaze estimation by learning person-specific models,[3] fine-tuning pre-trained models for individuals [18], incorporating adaption mechanisms into existing models [11,33], or learning a model to predict the gaze differences between input and reference eye images [15,16]. Although these person-specific methods often show superior performance compared with person-agnostic ones, they require annotated samples for new users, which introduces inconvenience. On the other hand, researchers [11,15,16] have shown that the performance of person-agnostic methods can be significantly improved by directly applying additional adaption components to existing models, and the final performance is determined by both the backbone models and the design of adaption components. In this work, we focus on person-agnostic gaze estimation, which can either be used alone or serve as a strong backbone for extra adaption components.

2.2 RGB-D image-based gaze tracking

There are relatively fewer researches on RGB-D-based gaze tracking than RGB-based ones. Mora et al. [21] used the sparse representation technique to reconstruct new eye images from an existing set of eye image/gaze pairs, and acquired gaze point accordingly. Xiong et al. [36] applied a personalized 3D face model, which was calibrated for each subject, and tracked six 2D facial landmarks whose 3D locations were provided by an RGB-D camera, to detect the eye gaze. Sugano et al. [31] collected a large multi-view gaze estimation dataset and synthesized eye images from dense viewing angles, which are used to learn an appearance-based gaze estimator. Mora et al. [4] leveraged RGB-D images and fitted a 3DMM model to obtain the head pose for each subject. Then, the textured mesh for each subject was rendered in a frontal pose, which provided the head pose invariant eye region, and the gaze direction was estimated from the rendered eye region and the head pose. Recently, Ghiass et al. [27] proposed a 3D pose estimator, which utilized low quality depth data and RGB images for face modeling, and then tracked face poses with depth frames only. Wang et al. [34] incorporated ICP-based head pose tracking and appearance-based gaze eyeball pose estimation with a neighbor selection strategy to estimate eye gaze using RGB-D cameras. Different from these methods, we use 68 3D facial landmarks rather than a single head pose vector to incorporate both head poses and eye positions into gaze estimation and utilize a global optimization procedure to acquire reliable facial landmarks from RGB and noisy depth images.

2.3 Gaze tracking dataset

Several datasets [8,9,11,20,31,38] have been made publicly available to the community. For instance, in [20], an RGB-D gaze tracking dataset, namely Eyediap, was proposed, which consists of videos from 16 participants. Sugano et al. [31] introduced a large and fully calibrated multi-view gaze dataset UT-Multiview, which contains 64k samples collected from 50 participants. Zheng et al. [38] created the MPIIGaze dataset, which is an in-the-wild RGB gaze dataset collected from 15 participants during natural everyday laptop use over several months. In [11], the researchers built a larger RGB gaze dataset containing about 2.5M video frames from over 1450 people. They also demonstrated that dataset with more participants could apparently improve the accuracy of gaze tracking systems. Therefore, in this work, we build the largest RGB-D gaze tracking dataset so far collected from 218 participants containing over 165k images. The dataset will be made publicly available for all researchers to facilitate the study of data-driven approaches for gaze tracking.

3 Our RGB-D gaze tracking dataset

In this section, we will detail how we collect data and show some important statistics of our proposed RGB-D gaze tracking dataset.

3.1 Data acquisition procedure

An illustration of our data collecting system can be found in Fig. 1. In the beginning, each participant was asked to sit

[3] In fact, most learning-based methods, including ours, can be person-specific if they are trained or fine-tuned with samples from each specific person.

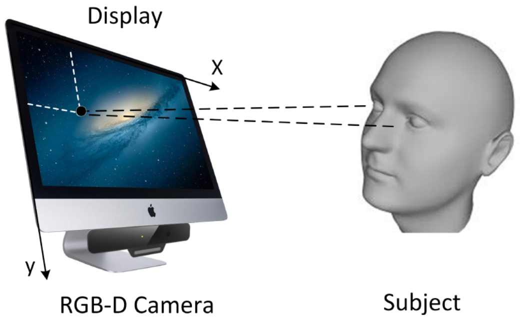


Fig. 1 Our data acquisition system

in front of a 27-inch iMac, with an Intel RealSense SR300 camera attached on it, at a comfortable distance ranging from 50 to 100 cm. When the participant was ready, he or she could enter his or her name in our data collecting software and start a collecting session. During each session, 50 white dots with a radius of 30 pixels (about 0.93 cm at the screen resolution of 1920 × 1080 pixels) successively appeared on the screen at random locations, and the participant was required to click the center of the white dots one after another. At the same time, the data collecting software recorded the location of each white dot, the location of each click, and the RGB-D image of the participant. After one session was finished, the participant could take a short break before starting the next session until all 16 sessions were finished.

In the data collection process, we assume that the participants should stare at the center of each white dot. However, it is not always true, especially when a participant is distracted while clicking the dots. To avoid this case, we also showed the participants the location of each click with a blue dot. In addition, we identify a bad sample by measuring the distance between the center of the white dot and the click location. If the distance is further than a given threshold, which is 10 pixels (about 0.31 cm at the screen resolution of 1920 × 1080 pixels), we simply exclude the sample from the dataset. All of the RGB and the depth images were re-sampled into a fixed size of 1920 × 1080 pixels. Figure 2 shows some RGB and depth images in our dataset. We further split the dataset into a training set and a validation set, of which the former contains 119,318 samples from 159 participants, and the later contains 45,913 samples from the remaining 59 participants.

3.2 Dataset statistics

There are 218 participants, including 141 males and 77 females between 19 and 37 years old, in our dataset. All participants have a normal or corrected-to-normal vision. The total number of samples (RGB-D images with gaze point ground truths) is 156,231, and each participant has about 600-800 samples. Because we do not fix the distance between the participants and the display, the depth of the left and the right eye of each participant varies in wide ranges,[4] as shown in Fig. 3. In addition, over half the participants wear glasses in the experiment, which causes their depth images to contain large black holes within the eye region, which can be observed in Fig. 2. The comparison between our dataset and existing RGB and RGB-D gaze tracking datasets is listed in Table 1.

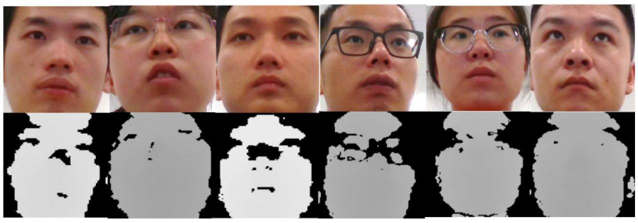


Fig. 2 Some examples of color and depth images in our RGB-D gaze dataset

4 Method

4.1 Overview

As aforementioned, we decompose the gaze estimation problem into eyeball pose, head pose, and eye position estimation. The first problem is solved via multi-column CNNs, of which the three CNN columns take left eye images, right eye images, and full-face images, respectively, as inputs and output three feature vectors containing the eyeball pose information. The last two problems are tackled by estimating the 68 3D facial landmarks (see Fig. 4) from RGB and depth images in a two-step procedure. In the first step, an ResNet-18 is used to regress the relative depth of every two facial landmarks from an RGB image. The second step utilizes global optimization to find the absolute depths of all facial landmarks consistent with the estimated relative depths and the raw depth images. The gaze point is estimated by two fully connected layers in each CNN column from eyeball pose features and 3D facial landmarks. Finally, the three gaze predictions from three CNN columns are fed into a fully connected layer for the final gaze prediction. Our method is demonstrated in Fig. 5.

4.2 3D facial landmarks estimation

For 3D facial landmarks estimation, we first extract the 2D facial landmarks with the *Dlib* toolkit.[5] Then, the only problem that remains is acquiring the depths of those landmarks.

[4] The depth of the eye is calculated by taking the median of depth values within a square region in the depth image.

[5] http://dlib.net/.

Table 1 Comparison of our dataset with popular publicly available datasets

Dataset	#Participants	#Poses	#Targets	Illum.	#Amount of data	#Modality
Eyediap [20]	16	cont.	cont.	2	Videos	RGB + depth
MPIIGaze [38]	15	cont.	cont.	cont.	213,659	RGB
PoG [19]	20	1	16	1	97 min	RGB
OMEG [8]	50	cont.	10	1	333 min	RGB
TabletGaze [9]	51	cont.	35	cont.	Videos	RGB
iTracker [11]	1474	cont.	cont.	cont.	2,445,504	RGB
UT-Multiview [31]	50	8 + synth.	160	1	64,000	3D face model
Free-head [41]	200	cont.	cont.	cont.	240,000	RGB
Ours	218	cont.	cont.	cont.	165,231	RGB + depth

We use the following abbreviations: *cont.* for continuous, *illum.* for illumination and *modality* for containing modality

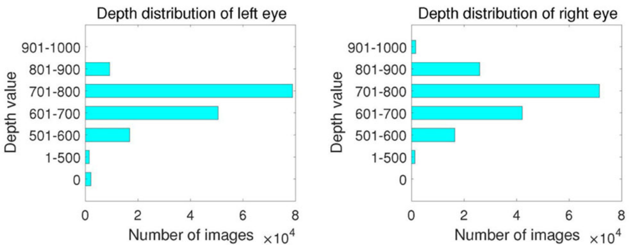


Fig. 3 Depth distribution of left and right eyes in our dataset

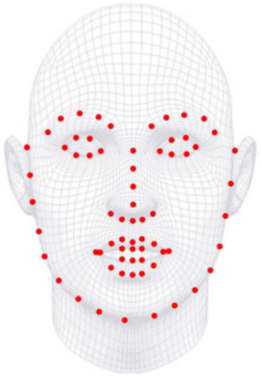


Fig. 4 The illustration of 68 facial landmarks used in our framework. The location of each landmark is defined in the iBUG 300-W dataset [26]

Though the raw depth images are already available, they contain too much noises and black holes to provide accurate and reliable depths of facial landmarks (see the first row in Fig. 6). Fortunately, we also have color images in hand. Previous work has shown the practicability of single RGB image-based depth image generation [3], which implies that RGB images help to infer some depth information. However, as discussed in [40], estimating absolute depths from a monocular color image is highly ill-posed and difficult even for people. Inspired by [40], we propose to refine the raw depths of facial landmarks with the help of the corresponding RGB images. Specifically, we use the similar strategy as that in [40], where RGB image is used to predict relative depths of facial landmark pairs and the absolute depths are acquired through global optimization using both the estimated relative depths and the raw depths. Note that because we rely on an out-of-box 2D facial landmark detector to estimate 3D facial landmarks, the accuracy of our method is affected by the performance of the facial landmark detector. In fact, in the cases where we fed raw 3D facial landmarks or random vectors instead of the optimized 3D facial landmarks into our network, we observed a significant performance degradation in both gaze point estimation (see Table 5). Furthermore, we do find that Dlib facial landmark extractor fails in some samples on both our dataset and the Eyediap dataset. Therefore, one can expect that the accuracy of our method will increase by substituting the Dlib detector with some more powerful facial landmark detectors. Despite some failure cases, Dlib is still a fairly good choice for research, which provides sound accuracy, efficiency, and reliability.

4.2.1 Relative depth estimation

The first step to solving the absolute depths for facial landmarks is estimating the relative depth. We use only RGB image to estimate relative depths for facial landmarks, which according to [40] achieves better performance compared with the alternatives using depth images or both. The relative depth is a 68 × 68 antisymmetry matrix, where the element at ith row and jth column is the depth difference between the ith and the jth facial landmarks, i.e., $\boldsymbol{R}_{ij} = \boldsymbol{d}_i - \boldsymbol{d}_j$. The input RGB images are fed into a ResNet18 network, of which we change the dimension of the last fully connected layer to regress only the lower part of the relative depth matrix, i.e., 2278 elements below the main diagonal, due to the antisymmetry property.

4.2.2 Global depth optimization

After predicting the relative depth $\hat{\boldsymbol{R}}$ for facial landmarks, we solve a system of equations to get the estimation of absolute depth $\hat{\boldsymbol{d}}$. The objective function is defined as the weighted

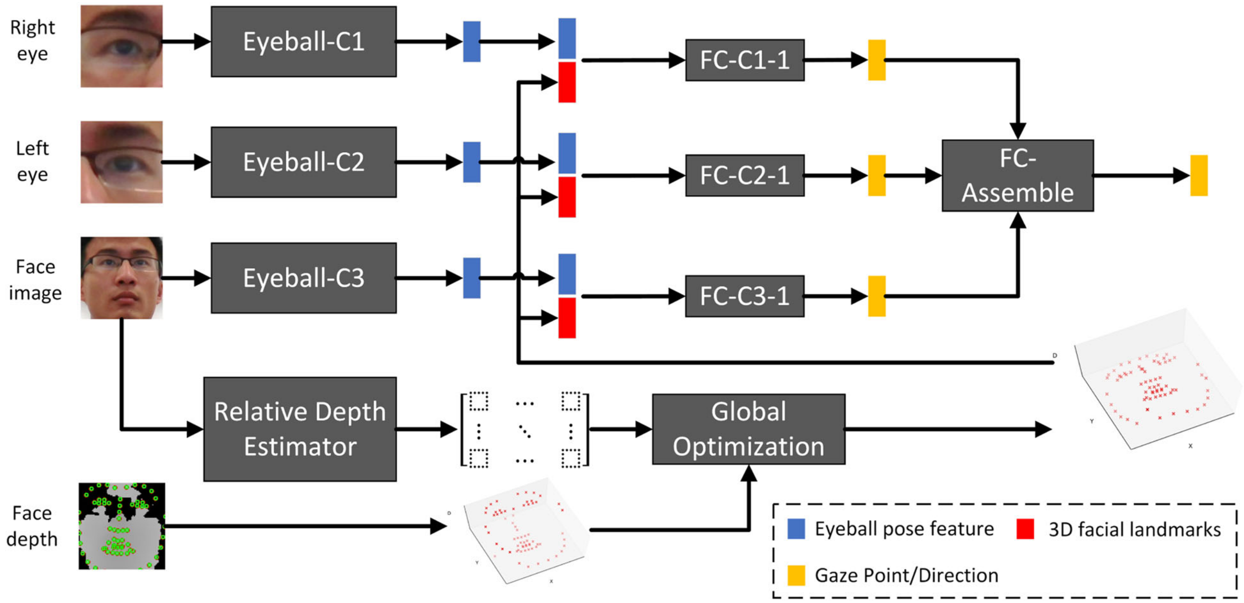


Fig. 5 Our framework for gaze point estimation. Eyeball pose features are extracted from two single-eye images and face images with three independent feature extractors (i.e., Eyeball-C1, Eyeball-C2, and Eyeball-C3, which are three independent ResNet-18 in our implementation). Relative depths for facial landmarks are estimated by the relative depth estimator, which is another ResNet-18, from face images. Then, the estimated relative depths along with the raw depths are used to acquire reliable 3D facial landmarks through global optimization. The eyeball pose features from the three CNN columns and the optimal 3D facial landmarks are fed into three sets of fully connected layers (i.e., FC-C1-1, FC-C2-1, and FC-C3-1 in the figure, of which each consists of two cascaded fully connected layers in our implementation) for gaze point regression. Finally, the three independent gaze predictions from the three CNN columns are collected by the last fully connected layer for the final prediction. (Best viewed in color)

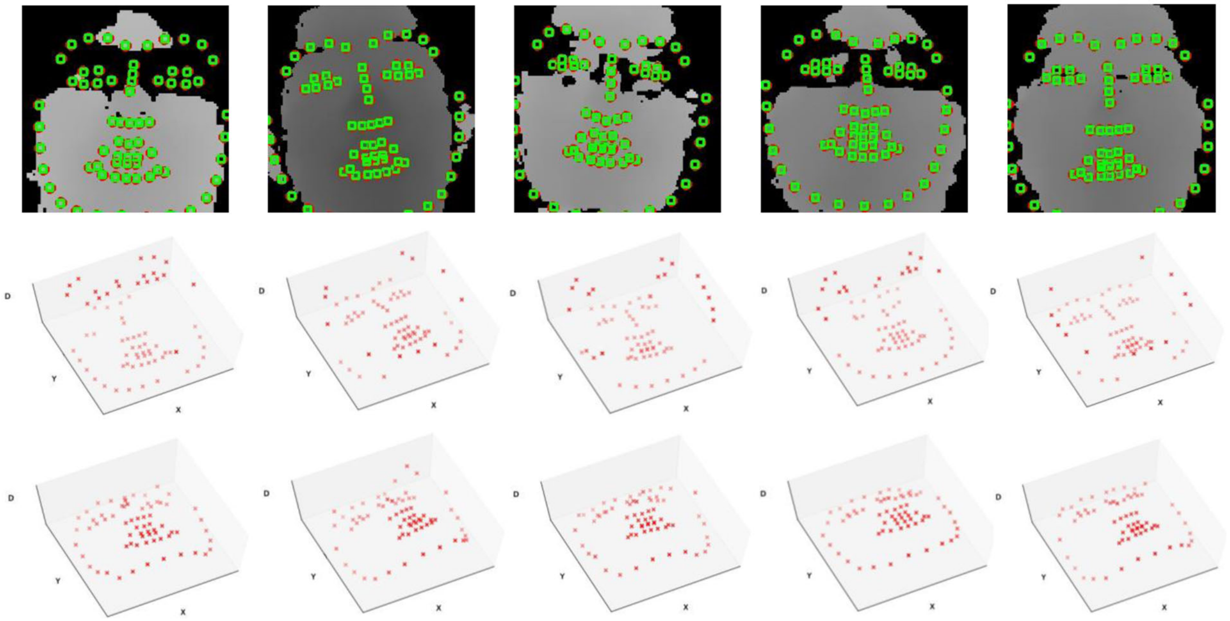


Fig. 6 Illustration of depth refinement for facial landmarks. The 1st row: the raw depth images; the 2nd row: the 3D facial landmarks before global optimization; and the 3rd row: the 3D landmarks after global optimization

sum of three terms

$$E = \lambda_{\mathrm{D}} E_{\mathrm{D}} + \lambda_{\mathrm{R}} E_{\mathrm{R}} + \lambda_{\mathrm{S}} E_{\mathrm{S}} \tag{1}$$

where

$$E_{\mathrm{D}} = \sum_{i \in L_{\mathrm{obs}}} \|\hat{\boldsymbol{d}}_i - \boldsymbol{d}_i\|^2 \tag{2}$$

$$E_{\mathrm{R}} = \sum_{i,j \in L} \|\hat{\boldsymbol{R}}_{ij} - \left(\hat{\boldsymbol{d}}_i - \hat{\boldsymbol{d}}_j\right)\|^2 \tag{3}$$

$$E_{\mathrm{S}} = \sum_{i,j \in L} \omega\,(i,j)\,\|\hat{\boldsymbol{d}}_i - \hat{\boldsymbol{d}}_j\|^2 \tag{4}$$

The term E_{D} measures the difference between $\hat{\boldsymbol{d}}$ and the raw depth $\boldsymbol{d}$, E_{R} measures the consistency between the estimated relative depth $\hat{\boldsymbol{R}}$ and the predicted depth $\hat{\boldsymbol{d}}$, and E_{S} encourages close landmarks to have the similar depths. Intuitively, the first term E_{D} (absolute depth term) ensures the optimized depths are close to the measured depth; the second term E_{R} (relative depth term) keeps the consistency between the optimized and the estimated relative depths; and the third them E_{S} (smoothness term) encourages the depths of neighborhood facial landmarks to be close to each other. In all terms, L represents the index of all 68 facial landmarks, and L_{obs} represents the set of the index of landmarks whose depths are valid, i.e., greater than zero. In the smoothness term E_{S}, $\omega(i, j)$ takes the form of

$$\omega(i,j) = \exp\left[-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma^2}\right], \qquad i, j \in L, \tag{5}$$

where $\hat{\boldsymbol{x}}$ stands for the location of a facial landmark specified by the subscript, and σ is a hyper-parameter which controls the locality of the smoothness term.

Since minimizing the objective function with respect to $\hat{\boldsymbol{d}}$ is a quadratic programming (QP) problem, we can solve it efficiently with existing QP solvers like OSQP [29], and the final solution is the global minimum of the objective function. Figure 6 shows some examples of 3D facial landmarks before and after global optimization.

4.3 Gaze point prediction network

The gaze point prediction network contains three independent CNN columns. Two of them regress gaze points from the left and right eye images, respectively, and the other one regresses gaze points from the full-face images. In each CNN column, the eyeball pose feature is first extracted by a ResNet-18, of which the last fully connected layer is replaced to output a 128-dimension feature vector. Then, we concatenate the eyeball pose feature and the coordinates of 68 3D facial landmarks to form a combined feature vector and feed it into two fully connected layers (332×64–64×2) to make the gaze prediction. Finally, the three independent gaze predictions from the three CNN columns are concatenated and fed into the last fully connected layer (6×2) to make the final gaze prediction. In principle, the eyeball pose features should be extracted using only eye regions to decouple the eyeball pose from other factors that may affect the gaze point. However, we empirically found that although using face images only gives us inferior performance than using two eye images due to the disturbance of head pose, the face images can boost the gaze prediction accuracy in conjunction with the eye images.

The loss function of the network is the mean square error between the ground truth and the predicted gaze points for all training samples, i.e.,

$$\ell_{\mathrm{gp}} = \frac{1}{M} \sum_{i}^{M} \|\hat{\boldsymbol{p}}^{\boldsymbol{i}} - \boldsymbol{p}^{\boldsymbol{i}}\|_2^2 \tag{6}$$

where p_i and $\hat{p}_i$ denote the ground truth and the predicted gaze point for the ith training sample ($i = 1, \ldots, M$). Note that the three intermediate and the final gaze point predictions are supervised by the same loss function.

4.4 Implementation details

Data preparation We first use the *Dlib* library to detect faces and 2D facial landmarks over our dataset. Then, we crop face regions and eye regions from the original RGB images and depth images, where the eye region is a square patch of which the length of four sides is equal to 1.5 times of distance between eye corners. We use the median of nonzero values within a square patch of size 7 × 7 centered around each facial landmark on the depth map as the raw absolute depth for each landmark, and the raw depth for a landmark is considered to be unobserved if there is no nonzero value in the square patch.

Training phase The CNN used for relative depth estimation is first trained with RGB face images and the observed ground truth landmark depths. Then, we fix the parameters in the relative depth estimator and train the three CNN columns independently. The refined absolute depths for facial landmarks are acquired from the estimated relative depth and the raw landmark depth using global optimization. The 2D facial landmarks and the inferred absolute depths constitute the 3D facial landmarks, which are also fed into the three CNN columns. Finally, we train the last fully connected layer with all other parameters fixed.

Testing phase During testing, we first predict the 3D facial landmarks using the relative depth estimator and global optimization. Then, the three CNN columns in our network are executed one by one to generate the three independent gaze point estimations from their respective input images. Finally, all of the three gaze point estimations (i.e., three 2D coordinates) are concatenated to constitute a single 6D vector, which is fed into the last fully connected layer to make the final gaze point prediction.

5 Experiments

5.1 Experimental setup

Training setup We implement our method with the PyTorch framework.[6] The batch size for all of our experiments is 100 in training. In the testing phase, we use a larger batch size of 200, which can speed up the testing phase without affecting the accuracy. We use 8 NVIDIA Tesla k40m GPUs to train our network. The stochastic gradient descent (SGD) optimization algorithm is adopted with learning rate equal to 0.1 and momentum equal to 0.9.

Datasets We evaluate our method on both our RGB-D gaze dataset and the Eyediap dataset. Our RGB-D gaze dataset is used for evaluating the gaze point prediction, and we choose images corresponding 159 subjects from the total 218 subjects as the training set, with the remaining samples used for testing. The Eyediap dataset is used for evaluating gaze direction prediction. As our method is mainly for gaze point prediction, we slightly modify the last fully connected layer of our network to regress the gaze direction rather than the gaze point. We follow the same strategy as that in [39] to choose frame images and gaze points. After that, we divide the 14 participants into five groups and perform cross-validation. Note that because our dataset is about two times larger than the Eyediap dataset, and the k-fold validation for all SOTA methods and our method is quite time consuming, we only compare our method with SOTA methods on the fixed validation set for our dataset.

Evaluation metrics For gaze point estimation on our RGB-D gaze dataset, we use the Euclidean distance to measure the error between the predicted and the ground truth gaze points, i.e.,

$$d_e = \frac{1}{M}\sum_{i}^{M} \|\boldsymbol{p}^{\boldsymbol{i}} - \hat{\boldsymbol{p}}^{\boldsymbol{i}}\|_2 \tag{7}$$

where M is the total number of images in our dataset. $\boldsymbol{p}^{\boldsymbol{i}}$ and $\hat{\boldsymbol{p}}^{\boldsymbol{i}}$ are the ground truth and the predicted gaze point, respectively, for the ith image.

For gaze direction estimation on the Eyediap dataset, the angle deviation between the estimation and the ground truth gaze direction is used for the performance evaluation, i.e.,

$$a_e = \frac{1}{N}\sum_{i}^{N} \arccos \frac{\langle \boldsymbol{a}^{\boldsymbol{i}}, \hat{\boldsymbol{a}}^{\boldsymbol{i}} \rangle}{|\boldsymbol{a}^{\boldsymbol{i}}||\hat{\boldsymbol{a}}^{\boldsymbol{i}}|} \tag{8}$$

where N is the total number of images in the Eyediap dataset. $\boldsymbol{a}^{\boldsymbol{i}}$ is the ground truth gaze direction of the ith image, and its prediction is $\hat{\boldsymbol{a}}^{\boldsymbol{i}}$. $\langle \boldsymbol{a}^{\boldsymbol{i}}, \hat{\boldsymbol{a}}^{\boldsymbol{i}} \rangle$ refers to the inner product between $\boldsymbol{a}^{\boldsymbol{i}}$ and $\hat{\boldsymbol{a}}^{\boldsymbol{i}}$.

To keep consistent with [39], Eq. (7) measures the gaze point error in millimeters (mm), and Eq. (8) evaluates the gaze direction error in degrees.

5.2 Performance comparison

We compare our proposed method with state-of-the-art deep learning-based methods for gaze point estimation on our dataset and gaze direction estimation on the Eyediap dataset, including:

- Multimodal CNN [38]: Normalized eye images and 3D head poses are fed into a CNN consisting of a LeNet feature extractor and a few fully connected layers to predict gaze direction.
- iTracker [11]: Eye images, faces, and face grids are fed into a multi-region CNN architecture. In the fully connected layer, all features are combined to predict gaze points.
- iTracker* [11]: It is an another version of iTracker, where we substitute the original feature extractor in iTracker with a ResNet-18 for a fair comparison. All other parts are the same as the iTracker [11].
- Spatial weights CNN [39]: A spatial weight CNN is used to generate weight maps from face images, which are applied to the face features. Then, the weighted face features are fed into a few fully connected layers for gaze point and gaze direction prediction.
- Ghiass et al. [27]: First, RGB-D data are used to reconstruct 3D face model for each subject. Then, the subject's face is tracked from depth frames with ICP registration.

We list the results of different methods on our dataset and the Eyediap dataset in Tables 2 and 3, respectively. From the results, one can observe that our method achieves superior performance compared to all existing methods. Specifically, as a state-of-the-art method, the iTracker can predict gaze point and gaze direction in a much higher accuracy only by changing the original feature extractor to the ResNet-18. It suggests that the architecture of the feature extractor is a key factor that affects the performance. Also, with the same feature extractor, our network achieves even better accuracy than the iTracker. The performance improvement comes from:

[6] https://pytorch.org/.

Table 2 Performance comparison of gaze point estimation on our dataset (unit: mm)

Methods	d_e
Multimodal CNN [38]	67.2
iTracker [11]	55.5
iTracker* [11]	47.5
Spatial weights CNN [39]	60.6
Our method	**32.3**

Bold value is used to highlight the best results

Table 3 Performance comparison of gaze direction estimation on Eyediap (unit: degree)

Methods	a_e
Multimodal CNN [38]	10.2 (2.9)
iTracker [11]	8.3 (1.7)
iTracker* [11]	5.7 (1.1)
Spatial weights CNN [39]	6.0 (1.2)
Ghiass et al. [5]	7.2 (1.3)
Ghiass et al. [27]	7.2 (0.4)
Our method	**4.6 (0.8)**

Bold value is used to highlight the best results

(1) the decomposition strategy, where the original problem is explicitly decomposed into eyeball pose, head pose, and eye position estimation; (2) the multi-column CNNs architecture, which extracts eyeball pose features from different data sources independently and thus improves the robustness and generalizability of our model; and (3) the introduction of the global optimization for depth refinement, which provides more accurate and reliable depths for facial landmarks.

5.3 Ablation studies

In order to explore the effectiveness of different modules and their possible alternatives for gaze tracking, we conducted extra experiments in which some components or intermediate variables of our method were replaced or removed. The results of these experiments are listed in Tables 4 and 5.

In the first three experiments, we studied the effect of different input settings for eyeball pose estimation, where *Face*, *Eyes* and *Face* + *Eyes* mean that we used only face images, only eye images, and both for eyeball pose estimation, respectively. We can see that even though face images contain information of two eyes, using only eye images achieves better performance than using only face images, and using both gives us the best performance.

In the following experiments, we also studied the effect of different terms in (1) on depth optimization. Using only the E_{D} is equivalent to using raw depth directly, which gave us the worst result because of the holes and noises. The introduction of the second term E_{R} greatly improved the performance, which demonstrates the effectiveness of our relative depth estimation network. The third smoothness term E_{S} further boosted the performance slightly, which shows that the estimated relative depth may still contain errors and the smoothness term helps ease the problem.

Table 4 Network architecture evaluation on our dataset (unit: mm)

Baselines	d_e
Face	41.1
Eyes	35.4
Face + Eyes	32.3
E_{D}	45.0
$E_{\mathrm{D}} + E_{\mathrm{R}}$	33.1
$E_{\mathrm{D}} + E_{\mathrm{R}} + E_{\mathrm{S}}$	32.3

Table 5 The exploration of how 3D facial landmarks quality affects the accuracy of our framework on our dataset (unit: mm)

Settings	d_e
Raw landmarks	45.0
Random landmarks	64.7
Optimized landmarks	32.3
No landmark	46.7

In Table 5, we performed three more experiments to explore how the quality of 3D facial landmarks affects the gaze estimation accuracy. In the first experiment, we used the raw depths of facial landmarks when constituting 3D facial landmarks. In the second experiment, we first calculated the mean and the variance of the ground truth 3D facial landmarks[7] and fed random vectors sampled from the normal distribution with the same mean and variance into the network. In the third experiment, we used the optimized 3D facial landmarks as normal. And in the final experiment, we removed 3D facial landmarks from our model.

As one can see, either using the raw facial landmarks or using random variables harms the performance of our method. The first result is the same as that of the experiment in which only the E_{D} term was included in the optimization step. This is reasonable because using only the E_{D} term is equivalent to using the raw depths directly, just as we mentioned before. Moreover, the performance got even worth in the second experiment, which proves the importance of accurate 3D facial landmark for our framework to produce reliable gaze prediction. The last result demonstrated the effectiveness of 3D facial landmarks for gaze estimation.

[7] Only the landmarks having nonzero depths were used. For simplicity, we assumed that all of the elements of 3D facial landmark vectors are independent random variables.

6 Conclusion and discussion

In this paper, we introduced a new RGB-D gaze dataset and a novel framework to tackle the gaze point estimation problem. Our proposed dataset consists of over 165K RGB-D images collected from 218 participants, which makes it the largest RGB-D gaze dataset in terms of the number of participants so far. As for our proposed gaze estimation framework, we first utilized the multi-column CNNs to extract the eyeball pose features from face images and two eye images. Then, the relative depths for the 68 facial landmarks are estimated from face images. Global optimization is applied to acquire the refined absolute depths for facial landmarks that are consistent with both the estimated relative depths and the raw depths. The final gaze point is inferred from eyeball pose features and the 3D facial landmarks. Extensive experiments on our dataset and the Eyediap dataset show that our gaze tracking method outperforms existing methods by a large margin.

On the other hand, we would like to point out that better gaze tracking accuracy could be achieved by applying extra person-specific calibration to our method. In fact, as we have mentioned in Sect. 2, person-specific often shows superior performance than person-agnostic ones. Our method, as an general gaze estimator, can either be used alone or serve as a strong backbone for extra adaption components in person-specific setting. Moreover, our model cannot achieve real-time prediction. A single global optimization step took about 0.3 s, which severely slows down the overall speed to about 0.5 s per sample. Further efforts still need to be taken to boost the speed of our method.

Compliance with ethical standards

Conflict of interest The author declares that they have no conflict of interest.

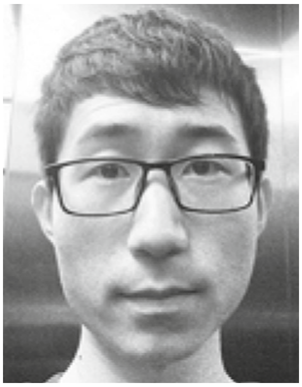
Ethical approval Ziheng Zhang declares that all procedures performed in studies involving human participants were in accordance with the ethical standards of the ShanghaiTech Ethics Committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards, and that no study with animals was performed by any of the authors. Ziheng Zhang declares that informed consent was obtained from all individual participants included in the study.

References

1. Barron, J.T., Malik, J.: Intrinsic scene properties from a single RGB-D image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 17–24 (2013)
2. Cireşan, D., Meier, U., Schmidhuber, J.: Multi-column Deep Neural Networks for Image Classification (2012). arXiv:1202.2745
3. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems, pp. 2366–2374 (2014)
4. Funes-Mora, K.A., Odobez, J.M.: Gaze estimation in the 3d space using RGB-D sensors. Int. J. Comput. Vision **118**(2), 194–216 (2016)
5. Ghiass, R.S., Arandjelovic, O.: Highly accurate gaze estimation using a consumer RGB-D sensor. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 3368–3374. AAAI Press (2016)
6. Hansen, D.W., Ji, Q.: In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 478–500 (2010)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. He, Q., Hong, X., Chai, X., Holappa, J., Zhao, G., Chen, X., Pietikäinen, M.: Omeg: oulu multi-pose eye gaze dataset. In: Scandinavian Conference on Image Analysis, pp. 418–427. Springer (2015)
9. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: Unconstrained Appearance-based Gaze Estimation in Mobile Tablets (2015). arXiv:1508.01244
10. Kar, A., Corcoran, P.: A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. IEEE Access **5**, 16495–16519 (2017)
11. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2176–2184 (2016)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
13. Kuno, Y., Yagi, T., Uchikawa, Y.: Development of a fish-eye VR system with human visual functioning and biological signals. In: 1996 IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems (Cat. No. 96TH8242), pp. 389–394. IEEE (1996)
14. Kuno, Y., Yagi, T., Uchikawa, Y.: Development of eye-gaze input interface. In: Proceedings of 7th International Conference on Human Computer Interaction. Volumen 1, vol. 44 (1997)
15. Liu, G., Yu, Y., Funes-Mora, K.A., Odobez, J.M.: A differential approach for gaze estimation with calibration. In: 29th British Machine Vision Conference 2018 (2018)
16. Liu, G., Yu, Y., Mora, K.A.F., Odobez, J.M.: A differential approach for gaze estimation (2019). arXiv:1904.09459
17. Majaranta, P., Bulling, A.: Eye tracking and eye-based human–computer interaction. In: Advances in Physiological Computing, pp. 39–65. Springer (2014)
18. Masko, D.: Calibration in eye tracking using transfer learning. Master thesis, KTH, School of Computer Science and Communication (CSC) (2017)
19. McMurrough, C.D., Metsis, V., Rich, J., Makedon, F.: An eye tracking dataset for point of gaze detection. In: Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA'12, pp. 305–308. ACM, New York, NY, USA (2012). https://doi.org/10.1145/2168556.2168622
20. Mora, K.A.F., Monay, F., Odobez, J.M.: Eyediap: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 255–258. ACM (2014)
21. Mora, K.A.F., Odobez, J.M.: Gaze estimation from multimodal kinect data. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 25–30. IEEE (2012)
22. Morimoto, C.H., Mimica, M.R.: Eye gaze tracking techniques for interactive applications. Comput. Vis. Image Underst. **98**(1), 4–24 (2005)

23. Ranjan, R., De Mello, S., Kautz, J.: Light-weight head pose invariant gaze tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2156–2164 (2018)
24. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychol. Bull. **124**(3), 372 (1998)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
26. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: database and results. Image Vis. Comput. **47**, 3–18 (2016)
27. Shoja Ghiass, R., Arandjelović, O., Laurendeau, D.: Highly accurate and fully automatic 3d head pose estimation and eye gaze estimation using rgb-d sensors and 3d morphable models. Sensors **18**(12), 4280 (2018)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
29. Stellato, B., Banjac, G., Goulart, P., Bemporad, A., Boyd, S.: OSQP: an operator splitting solver for quadratic programs. Math. Program. Comput. **12**, 637–672 (2020). https://doi.org/10.1007/s12532-020-00179-2
30. Stiefelhagen, R., Yang, J., Waibel, A.: A model-based gaze tracking system. Int. J. Artif. Intell. Tools **6**(02), 193–209 (1997)
31. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1821–1828 (2014)
32. Suwajanakorn, S., Hernandez, C., Seitz, S.M.: Depth from focus with your mobile phone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3497–3506 (2015)
33. Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with Bayesian adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11907–11916 (2019)
34. Wang, Y., Yuan, G., Mi, Z., Peng, J., Ding, X., Liang, Z., Fu, X.: Continuous driver's gaze zone estimation using RGB-D camera. Sensors **19**(6), 1287 (2019)
35. Xie, J., Girshick, R., Farhadi, A.: Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In: European Conference on Computer Vision, pp. 842–857. Springer (2016)
36. Xiong, X., Liu, Z., Cai, Q., Zhang, Z.: Eye gaze tracking using an RGBD camera: a comparison with a RGB solution. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pp. 1113–1121. ACM (2014)
37. Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M.: Shape-from-shading: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **21**(8), 690–706 (1999)
38. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4511–4520 (2015)
39. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: full-face appearance-based gaze estimation. In: CVPRW (2017)
40. Zhang, Y., Funkhouser, T.: Deep depth completion of a single RGB-D image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 175–185 (2018)
41. Zhu, W., Deng, H.: Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3143–3152 (2017)
42. Zhu, Z., Ji, Q., Bennett, K.P.: Nonlinear eye gaze mapping function estimation via support vector regression. In: 18th International Conference on Pattern Recognition, 2006. ICPR 2006, vol. 1, pp. 1132–1135. IEEE (2006)

Ziheng Zhang received the BSc degree from Xidian University, Xi'an, China, in 2016. He is currently pursuing the MSc degree with the ShanghaiTech University, Shanghai, China, supervised by Prof. Shenghua Gao. His research interests include semantic segmentation, 3D vision, saliency detection, and gaze estimation.



Dongze Lian received the BSc degree from Dalian University of Technology, Dalian, China, in 2016. He is currently pursuing the PhD degree with the ShanghaiTech University supervised by Prof. Shenghua Gao. His research interests include gaze estimation, action recognition, video understanding, and crowd counting.

Shenghua Gao is an assistant professor at ShanghaiTech University, China. He received the BE degree from the University of Science and Technology of China in 2008 and received the PhD degree from the Nanyang Technological University in 2012. From June 2012 to August 2014, he worked as a research scientist in UIUC Advanced Digital Sciences Center in Prof Yi Ma's group, Singapore. From January 2015 to June 2015, he visited UC Berkeley as a visiting professor, hosted by Prof Jitendra Malik. His research interests include computer vision and machine learning. He has published more than 50 papers on image and video understanding in many top-tier international conferences and journals. He also served as a chair for some workshops in CVPR2017, ACCV2014, ACCV2016, and area chair in ICCV2019. He also served as the Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology (IF:3.558) and Neurocomputing (IF:3.224). His work on personalized saliency detection was nominated as outstanding student award (runner-up) in IJCAI 2017. He was awarded the Microsoft Research Fellowship in 2010 and ACM Shanghai Young Research Scientist in 2015.