Gaze Estimation via Modulation-Based Adaptive Network With Auxiliary Self-Learning

Yong Wu, Gongyang Li, Zhi Liu, *Senior Member, IEEE*, Mengke Huang, and Yang Wang

***Abstract*— Given a face image, most of previous works in gaze estimation infer the gaze via a well-trained model with supervised training. However, the distribution of test data may be very different compared to that of training data since samples might be corrupted in real-world scenarios (*e.g.*, taking a photo in strong light). This will lead to a gap between source domain (*i.e.*, training data) and target domain (*i.e.*, test data). In this paper, we first introduce self-supervised learning into our method for addressing challenging situations in gaze estimation. Moreover, existing appearance-based gaze estimation methods focus on directing towards the development of powerful regressors, which mainly utilize face and eye images simultaneously or face (eye) images only. However, the problem of inter cues between face and eye features has been largely overlooked. To this end, we propose a novel Modulation-based Adaptive Network (MANet) for gaze estimation, which uses high-level knowledge to filter the distractive information and bridges the intrinsic relationship between face and eye features. Further, we combine self-supervised learning and MANet to learn to adapt to challenging cases, such as abnormal lighting conditions and poor-quality images, by minimizing a self-supervised loss and a supervised loss jointly. The experimental results on several datasets demonstrate the effectiveness of our proposed approach with a real-time speed of 900 *fps* on a PC with an NVIDIA Titan RTX GPU.**

***Index Terms*— Gaze estimation, self-supervised learning, corrupted images, inter cues, high-level knowledge.**

I. Introduction

GAZE is a non-verbal cue for understanding internal states of humans. It is widely applied in various applications, such as saliency detection [1]–[5], human-computer interaction [6], [7], brain–computer interface (BCI) [8] and virtual reality industry [9]. Recently, deep neural networks have dramatically improved the performance in gaze estimation [10]–[14]. However, after existing methods are well trained, the distribution of test data may be very different compared to that of training data. This is because, during testing, samples may be corrupted by noise, different lighting conditions, and environmental changes (shown in Fig. 1). These corruptions and the resulting distribution shifts cause a dramatic drop in performance [15], [16].

(a) (b) (c) (d) (e)

Fig. 1. Examples of a face image with different corrupted versions of the most severe level. (a) original face image; (b) Gaussian blur; (c) adjusting brightness; (d) Gaussian noise; and (e) glass blur.

To address the above problem, recent work mainly focuses on personal calibration from the same dataset to adversarial examples [17] or few-shot learning [11], [18]. Both areas aim to train a model to be robust against various types of distribution shifts or domain shifts during testing. In this paper, we introduce the *self-supervised* auxiliary task [19]–[22] into our method to handle distribution/domain shifts. Self-supervised learning can force the predictions of the two augmented views (*i.e.*, normal images and corrupted images) to be as similar as possible. Our first intuition is to impose a constraint over the feature space during training, so that the feature distribution of training examples can remain close to that of the real-world domain. In general, our self-supervised learning based framework consists of two neural networks, that is, an online network and a target network. By minimizing the mean squared euclidean distance of both l_2-normalized predictions, our proposed method can obtain robust parameters to adapt to various corrupted images.

Some appearance-based models use CNNs for gaze estimation from either a single eye patch (Fig. 2(a)) [11], [23], [24] or both eye patches (Fig. 2(b)) [18], [25], [26]. There are also some works [27], [28] on gaze estimation from full face images (Fig. 2(a)). Most state-of-the-art approaches [29]–[32] use both face and eye images (Fig. 2(b)). However, these methods use simple techniques to fuse the information from the face and eye images, *e.g.*, by simple concatenation or fully-connected layers. Since gaze estimation is inherently a challenging task that requires high-level understanding of the gaze, simple feature concatenation operation is fundamentally limited, which is not conducive to modeling the interaction

Manuscript received 5 January 2022; revised 14 February 2022; accepted 16 February 2022. Date of publication 18 February 2022; date of current version 4 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62171269 and in part by the China Scholarship Council under Grant 202006890081 and Grant 202006890079. This article was recommended by Associate Editor J. Meng. *(Yong Wu and Gongyang Li contributed equally to this work.) (Corresponding authors: Zhi Liu; Yang Wang.)*

Yong Wu, Gongyang Li, Zhi Liu, and Mengke Huang are with the School of Communication and Information Engineering, and Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China (e-mail: yong_wu@shu.edu.cn; ligongyang@shu.edu.cn; liuzhisjtu@163.com; huangmengke@shu.edu.cn).

Yang Wang is with the Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada, and also with Huawei Technologies Canada, Markham, ON L3R 5A4, Canada (e-mail: ywang@cs.umanitoba.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3152800.

Digital Object Identifier 10.1109/TCSVT.2022.3152800

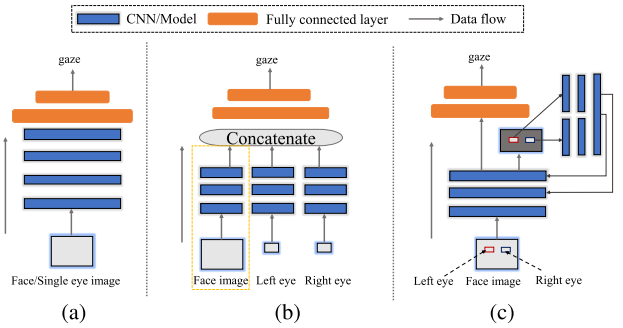


Fig. 2. Two typical gaze estimation architectures to explore the correlation between face and eye images. (a) full face or single eye input [11], [18], [27], [28]; (b) face and eye images simultaneously or both eyes input [14], [29], [30], [33]; and (c) our unique architecture for gaze estimation.

between face and eye images. We believe that modeling the proper interaction between face and eye images is very important for gaze estimation. Besides, the capacity of gaze cues between eyes and face is different that means we cannot address them by simple concatenation. Eyes can provide more accurate gaze cues, instead face contains richer gaze cues, including more noises. Therefore, we use more accurate eye features to filter distractive information of face features.

To this end, in this paper, we propose a novel modulation-based adaptive architecture named MANet (Fig. 2(c)). Different from existing architectures (Fig. 2 (a&b)) that simply fuse face and eye features or just input a single patch, our model uses the eye features to modulate the face features for gaze estimation. The advantage is that our model effectively models the proper interaction between face and eye features and successfully captures the complementary information between them. Concretely, our MANet consists of three main parts: a feature extraction network, an eye-guiding network and a gaze prediction network. The eye-guiding network is based on a two-stream structure that encodes precise information from the bounding boxes for the two eyes. The gaze prediction network encodes the visual feature from the full face for gaze estimation. The output from the eye-guiding network is used to modulate the face features in the gaze prediction network. This feature modulation enables to learn rich interactions between face and eye features.

We combine a self-supervised auxiliary task and MANet to cope with domain shifts when test data is corrupted. Our method mitigates the harmful effect of data distribution shifts between training data and test data, which can tackle extreme conditions, such as low-quality images, blurred images and abnormal illumination.

Our major contributions are summarized as follows:

- First, previous works in gaze estimation have largely overlooked the issue of the covariate shift between training and testing. To solve it, we introduce self-supervised learning to promote the robustness and generalization of our gaze model. Our core idea is to impose a constraint over the feature space during training so that the feature distribution of training examples remains close to that of the test domain.
- Second, we propose a novel modulation-based adaptive architecture called MANet for gaze estimation. To the best of our knowledge, this is the first method that utilizes high-level cues of eye regions to modulate face features in gaze estimation.
- Finally, we combine the self-supervised learning and MANet at the training phase which is able to adapt to corrupted images at test time. Our proposed model with joint training shows competitive performance compared with other state-of-the-art methods on several datasets.

II. Related Work

In this section, we briefly introduce self-supervised learning and review existing representative works on gaze estimation, including model-based methods and appearance-based methods.

A. Self-Supervised Learning

The goal of self-supervised learning is to learn general representations with unlabeled data. Recent state-of-the-art approaches for representation learning rely on contrastive learning [20], [34], [35]. The key idea of contrastive learning is to jointly maximize the similarity of representations of augmented views of the same image, while minimizing the similarity of representations of other samples, *i.e.*, the so-called negatives. Our paper uses another state-of-the-art method, BYOL [19], which shows that the self-supervised learning on only a single image can surprisingly produce representations that generalize well. Fu *et al.* [21] present a novel self-supervised synthesis ranking auxiliary framework for better metric learning. Guo *et al.* [22] use 2D spatial relationships and 3D geometric knowledge to build a self-supervised module to eliminate domain gaps between 2D and 3D space in 3D hand pose estimation. In gaze estimation, for the first time, we introduce self-learning to adapt domain shifts when test data is corrupted. In addition to self-supervised learning methods, many outstanding semi-supervised works are published. Chen *et al.* [36] propose a semi-supervised deep model for imbalanced activity recognition from multimodal wearable sensory data. DML [37] presents a deep mutual learning strategy to transfer knowledge to meet the low-memory or fast execution requirements. Then the proposed DML is extended straightforwardly to semi-supervised learning. Laine and Aila [38] present a simple and efficient method for training deep neural networks in a semi-supervised setting. Zheng and Yang [39] propose an orthogonal method to exploit the intra-domain knowledge and regularize the model training.

B. Model-Based Methods

Much attention has been spent on investigating model-based methods for gaze estimation, and they use a geometric model of eyes, usually requiring either high-resolution images or a person-specific calibration stage to estimate personal eye parameters [40]–[46]. Although these model-based methods have achieved promising performance, most of them require special devices such as infrared lights or RGB-D cameras.

Meanwhile, model-based methods are prone to noise or illumination perturbations, and cannot handle well with head orientation variabilities, which limits their practical application. Besides, the users usually need to provide a strict controlled environment such as a laboratory, because all of the present model-based methods have limited working distance. Different from the model-based methods, we propose a novel method based on appearance.

C. Appearance-Based Methods

Because of the above-mentioned limitations of model-based methods, recent research works focus more on appearance-based methods as they can learn a direct mapping from an image, or extract eye features to estimate gaze direction, thus being potentially applicable to relatively low-resolution images and mid-distance scenarios. The appearance-based methods can roughly be classified into three major categories based on the ways of input: single eye-patch input, full face input and multiple regions input.

1) Single Eye-Patch Input: Many appearance-based methods [13], [17], [24], [47]–[54] take a single eye region as input (Fig. 2(a)), and these methods can be inputted the left or right eye of a person separately. Zhang *et al*. first propose a CNN-based method to map eye images to gaze directions [55]. Because existing appearance-based methods assume person-specific training data, Sugano *et al*. [50] use a large amount of cross-subject training data to train a 3D gaze estimator. Zhang *et al*. [26] first present an in-the-wild dataset for gaze estimation to evaluate gaze estimation methods. Wang *et al*. [17] propose to incorporate adversarial learning and Bayesian inference into a unified framework to overcome poor generalization performance. Although their inputs to the network are relatively straightforward, the performance is significantly improved.

2) Full Face Input: Because the regions need to be chosen by handling for above-mentioned methods, previous methods also use full face as input (Fig. 2(a)) [27], [28], which can be simple in real-world applications. FullFaze [27] encodes the face image using a convolutional neural network with spatial weights applied on the feature maps to flexibly suppress or enhance information in different facial regions. Zhang *et al*. [56] leverage a standard CNN architecture, trained with the task of estimating gaze from a monocular face patch. Generally, full face image can provide more information, and the CNNs can avoid over-fitting when the input is a full face image. But the performance of the gaze estimator is still not sufficient for high-accuracy applications.

3) Multiple Regions Input: Performance generally improves when considering both eye regions simultaneously or using multiple input regions (Fig. 2(b)), such as the two eyes alongside the face [14], [18], [25], [29], [30], [32], [33], [57]. To further improve the accuracy of the gaze estimator, Cheng *et al*. propose an Asymmetric Regression-Evaluation Network by utilizing asymmetry of two eyes. Recently, a lot of works focus on few-shot learning [58]–[64]. For example, Park *et al*. [18] lower the angular errors by using a few-shot adaptive network for learning person-specific gaze estimation networks. From scientific research to commercial applications, Krafka *et al*. [29] present the first large-scale dataset to build an eye tracking software that works on commodity hardware. Fischer *et al*. [25] address the gaze estimation task by measuring head pose using a motion capture system and eye gaze using mobile eye-tracking glasses. AR-Net [57] and ARE-Net [57] try to improve the gaze estimation performance using the property of "two eyes asymmetry". Based on AR-Net [57] and ARE-Net [57], FAR-Net [30] and FARE-Net [30] present the face-based asymmetric regression-evaluation network to optimize the gaze estimation results. These works have obtained promising results, however, the intrinsic regularities between face and eye features are largely ignored because face and eye images only serve as independent or parallel feature sources in these works.

Existing appearance-based gaze estimation approaches with CNNs have poor generalization performance, due to three issues in our opinion, *i.e*., lacking the intrinsic regularities correlation between face and eyes, inputting low-quality or occluded image, and over-fitting issue with point estimation. Some works have tried to tackle these issues. Cheng *et al*. [31] proposed a coarse-to-fine adaptive network to address the intrinsic correlation between face and eyes. Further, Zhang *et al*. [32] take a dynamic region selection approach to overcome the problems of illumination conditions, low-quality images and occlusions to some extent, but this method requires to undergo a complex training process due to the non-end-to-end network. Inspired by the previous studies mentioned above, compared to [32], in our method, we focus on making the model adapt to various situations. Besides, we would expect to extract more meaningful adaptive eye features to filter the distractive information, and learn inter cues between face and eyes with a concise architecture.

Thus, based on the above analysis, we first introduce self-supervision to address corrupted images, and we also design a modulation-based structure called MANet, to learn inter cues between face and eye images. Our model can eliminate distractive information and avoid performance degradation when the input is with low-quality, abnormal illumination and occlusion.

III. Our Approach

In this section, we elaborate our proposed model. In Sec. III-A, we present the architecture of our model. In Sec. III-B, we give the details of the proposed Modulation-Based Adaptive Network (MANet), which consists of feature extraction, gaze prediction network and eye-guiding network.

A. Architecture Overview

Similar to previous work in self-supervised learning (or representation learning) [19], [20], our proposed model learns representations by maximizing agreement between different views of the same data example via a self-supervised loss in the latent space. The overall architecture as shown in Fig. 3 consists of two neural networks: the *online* and *target* ones. The online network is defined by a set of weights θ and is comprised of three stages: an encoder f_θ, a projector g_θ and a predictor. The target network has the same architecture as

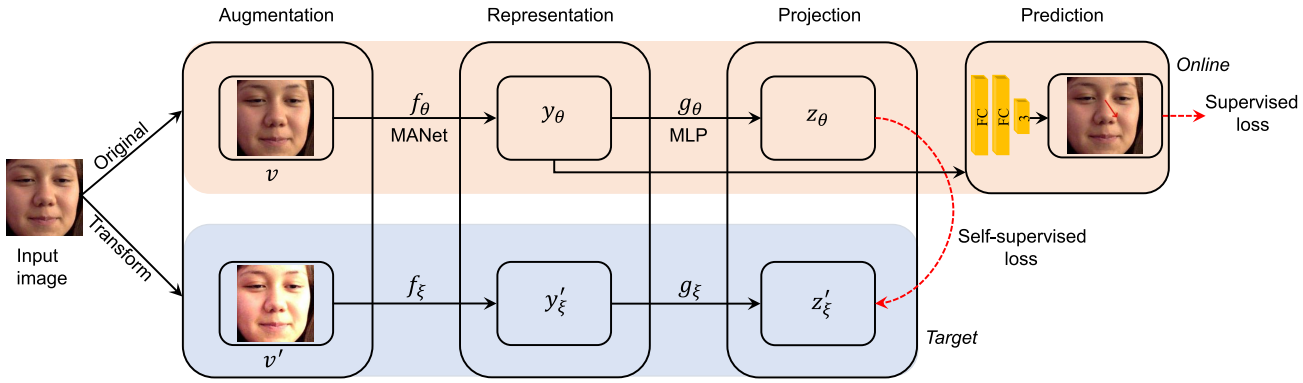


Fig. 3. Overview of our approach. Similar to BYOL [19], our approach has a online network and a target network. For the *online* network, we use the original image into *augmentation*, which is then passed through f_θ (MANet, see Fig. 4) to obtain y_θ. After that, y_θ is passed through g_θ (MLP) to obtain z_θ, where θ are the trained weights. The target branch has same the architecture as the online branch except *prediction*. We first augment (simulating image corruptions) the original image into *augmentation*, then the image undergoes the same process as that of *online*. ξ are an exponential moving average of θ that means the *target* never calculates gradients. We minimize a similarity loss (self-supervised loss) between z_θ and z'_ξ. Meanwhile, we minimize the supervised loss via *prediction*. At the end of training, everything is discarded but f_θ (MANet).

the online network except predictor (target has no predictor). The target network provides the regression targets to train the online network, and its parameters ξ are an exponential moving average of the online parameters θ [65]. Given a target decay rate $\tau \in [0, 1]$, after each training step we perform the following update:

$$\xi \leftarrow \tau\xi + (1-\tau)\theta, \tag{1}$$

where we set τ to 0.99. In addition, we elaborate major components in Fig. 3 as follows:

1) Augmentation: We use stochastic data augmentation to transform any given data example randomly resulting in a correlated view of the input image v, denoted as v'. In this work, we sequentially apply several simple augmentations, including color distortion, random sequence of brightness, contrast, saturation, hue adjustments, Gaussian blur and Gaussian noise, to the face image.

2) Representation: We apply the proposed MANet (refer to Sec. III-B for details) as the base encoder $f(\cdot)$ that extracts the representation vectors y_θ and y'_ξ from the two samples v and v', respectively.

3) Projection: We use a Multi-Layer Perception (MLP) as *projection* $g(\cdot)$ that maps representations to the space where self-supervised loss is applied. We obtain z_θ via $g(\cdot)$. The MLP consists of a linear layer with output dimension 1024 followed by batch normalization, and a linear layer with output dimension 512 followed by ReLU. Additional details of the MLP are given in Tab. V.

4) Objective: Our proposed model has two loss functions, *i.e.*, a supervised loss and a self-supervised loss. We obtain the supervised loss via *prediction*. We use the angular gaze estimation error as the supervised loss and evaluation metric. To calculate this error, we first convert the yaw and pitch angles, *i.e.*, $(\boldsymbol{\phi}, \boldsymbol{\psi})$, into three-dimensional representation in the Cartesian coordinate system as $p = (cos\boldsymbol{\phi}cos\boldsymbol{\psi}, -sin\boldsymbol{\phi}, cos\boldsymbol{\phi}sin\boldsymbol{\psi})$. Given the ground-truth gaze angle p and the predicted gaze angle $\hat{p}$, the angular error $\mathcal{L}(\hat{p}, p)$ is defined as:

$$\mathcal{L}(\hat{p}, p) = arccos\left(\frac{\hat{p} \cdot p}{\|\hat{p}\| \cdot \|p\|}\right). \tag{2}$$

To compute the self-supervised loss, we first use ℓ_2-normalize to output z_θ and z'_ξ, then obtain $\bar{z}_\theta = z_\theta / \|z_\theta\|_2$ and $\bar{z}'_\xi = z'_\xi / \|z'_\xi\|_2$. Finally we use the following mean squared error between the normalized predictions like [19],

$$\mathcal{L}_{self} = \left\|\bar{z}_\theta - \bar{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\langle z_\theta, z'_\xi \rangle}{\|z_\theta\|_2 \cdot \|z'_\xi\|_2}. \tag{3}$$

Note that we only compute gradient with respect to θ in Eq. 3, but not ξ (refer to Eq. 1).

We symmetrize the loss $\mathcal{L}(\hat{p}, p)$ and $\mathcal{L}_{self}$ in Eq. 2 and Eq. 3, by separately feeding v' to the online network and v to the target network, to compute $\tilde{\mathcal{L}}(\hat{p}, p)$ and $\tilde{\mathcal{L}}_{self}$, respectively. The final loss can be summarized as:

$$\mathcal{L} = \mathcal{L}(\hat{p}, p) + \tilde{\mathcal{L}}(\hat{p}, p) + \mathcal{L}_{self} + \tilde{\mathcal{L}}_{self}. \tag{4}$$

B. Modulation-Based Adaptive Network

Most of the appearance-based methods handle eyes and face by simply concatenating, however, inter cues between eyes and face are vital for regressing gaze. In this paper, we propose a modulation-based adaptive network (MANet) to address this issue. As illustrated in Fig. 4, MANet consists of a feature extraction backbone network, an eye-guiding network and a gaze prediction network. The eye-guiding network extracts visual information from the two eye images, *i.e.*, the bounding boxes of two eyes, and outputs modulator vectors for the gaze prediction network. The gaze prediction network predicts gaze based on the modulated features of the face image.

1) Feature Extraction: We use the relatively shallow ResNet-18 [66] as our backbone network for feature extraction. We remove the last fully-connected layer of ResNet-18, and retain its original five blocks, *i.e.*, `Block1` to `Block5`. We define the extracted features of `Block-i` as $\boldsymbol{F}^i$. Notably, the resolution of the input face image is set to $224 \times 224 \times 3$.

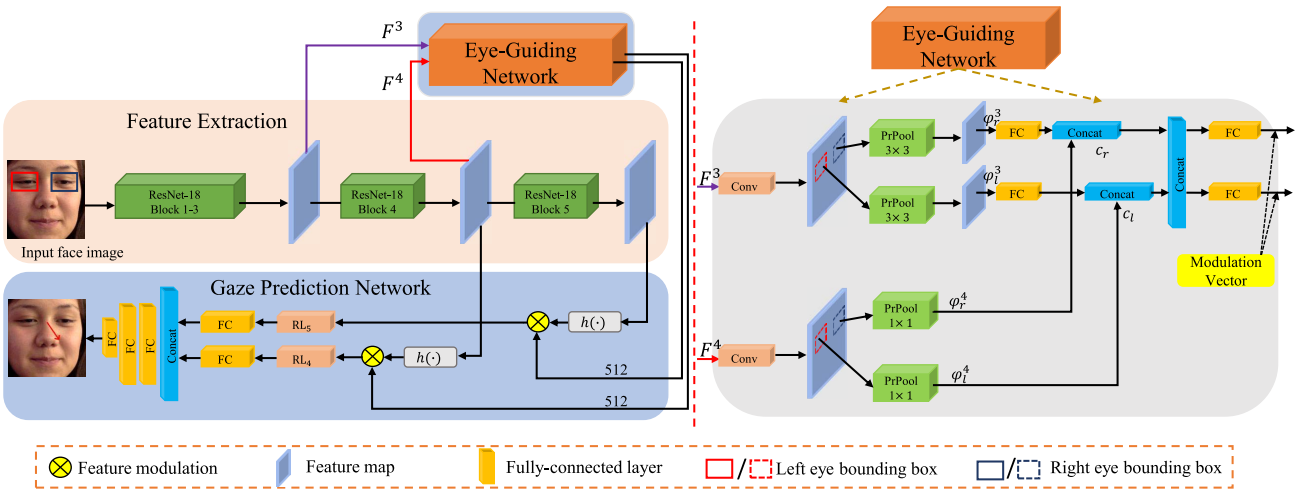


Fig. 4. The overall framework of MANet, which is composed of three components: feature extraction, eye-guiding network, and gaze prediction network. First, the feature extraction component extracts features from the face image by ResNet-18 `Block3` and `Block4`. These features are passed into the eye-guiding network. The eye-guiding network first pools two eyes' regions to a fixed size using `PrPool` layers. Then, ResNet-18 `Block4` and `Block5` features extracted from the face image pass through $h(\cdot)$ which is consisted of two `Conv` layers. The generated features are modulated by channel-wise multiplication with the modulator vectors returned by the eye-guiding network. Finally, the gaze prediction network predicts the gaze directions.

2) Eye-Guiding Network: As shown in Fig. 4, the eye-guiding network is a key component of MANet. Its goal is to encode the visual information of the two eyes in the image and to produce the modulator vectors for face features modulation in the gaze prediction network. The eye-guiding network takes the visual information extracted from the two eyes as inputs. It then produces the modulator vectors, which modulator vectors allow us to capture the rich interaction between face and eye images.

To simplify, we take $\boldsymbol{F} \in \mathbb{R}^{K\times K\times D}$ as an input example to introduce our eye-guiding network. We use $\{B_l, B_r\} \in \mathbb{R}^4$ to denote the two bounding boxes of both eyes. Here $B_{l/r}$ indicates continuous coordinates of the top-left and bottom-right points of the bounding box for the left/right eye. The eye-guiding network first feeds $\boldsymbol{F}$ through a `Conv` layer followed by a pooling layer on the two eye bounding boxes B_l and B_r. Here, we use the precise ROI pooling layer (`PrPool`)[1] [67], which is a continuous variant of adaptive average pooling. The key advantage of `PrPool` is that it is differentiable with respect to the bounding box coordinates, and can better extract features of the bounding box. The output from each eye bounding box after `PrPool` is a feature map of size $K \times K \times D$, denoted to $\boldsymbol{\varphi}_{l/r}$ for left or right eye. We formulate the above operations as follows:

$$\text{Left eye: } \boldsymbol{\varphi}_l = \texttt{PrPool}(B_l, \texttt{Conv}(\boldsymbol{F})), \tag{5a}$$

$$\text{Right eye: } \boldsymbol{\varphi}_r = \texttt{PrPool}(B_r, \texttt{Conv}(\boldsymbol{F})), \tag{5b}$$

where `Conv` consists of a convolutional layer, a batch normalization layer, and a ReLU activation function.

In this work, we perform the operations in Eq. 5b to $\boldsymbol{F}^3$ and $\boldsymbol{F}^4$. This two-stream interaction is in charge of capturing local and global information. The output size of `PrPool` to $\boldsymbol{F}^3$ and $\boldsymbol{F}^4$ are set to 3×3 and 1×1, respectively. We use $\boldsymbol{\varphi}_l^3$ and $\boldsymbol{\varphi}_l^4$ to denote the two outputs for the left eye. We then apply a fully-connected layer on $\boldsymbol{\varphi}_l^3$ to match the feature dimension of $\boldsymbol{\varphi}_l^4$, and concatenate them together to form a feature vector corresponding to the left eye, *i.e.*,

$$c_l = \texttt{FC}(\boldsymbol{\varphi}_l^3) \circledcirc \boldsymbol{\varphi}_l^4, \tag{6}$$

where $\texttt{FC}(\cdot)$ is the fully-connected layer. The feature vector c_r for the right eye is similarly defined.

Finally, we concatenate c_l and c_r, and pass through the fully-connected layers to obtain two modulation vectors $\{\nu_1, \nu_2\} \in \mathbb{R}^{1\times 1\times 512}$. Here, we give the implementation details of operations to $\boldsymbol{F}^3$ and $\boldsymbol{F}^4$ in Tab. VI.

3) Gaze Prediction Network: The goal of the gaze prediction network is to perform gaze estimation using full face features. We use features extracted from `Block4` and `Block5` of the backbone network, *i.e.*, $\boldsymbol{F}^4$ and $\boldsymbol{F}^5$, as the input to the gaze prediction network. Note that gaze estimation from full face image is quite challenging, since the underlying factors (*e.g.*, gaze and head orientation) that we wish to precisely encode gaze estimation entangled with many other extraneous factors (*e.g.*, lighting, hue, blur, *etc.*). In order to address this challenge, we introduce the modulation mechanism, where the face features are modulated by the information extracted from the eye-guiding network of the two eyes. This modulation mechanism not only disentangles the gaze from distractive information, but also models the rich interaction between face and eye features.

Therefore, the modulator vectors from the eye-guiding network are also sent to the gaze prediction network, playing a role in modulating the visual features extracted from the face image (*i.e.*, $\boldsymbol{F}^4$ and $\boldsymbol{F}^5$). Concretely, as shown in Fig. 4, the features of the face image first pass through two `Conv` layers. Then the generated features are modulated by the modulator vectors ν_1 and ν_2 via the channel-wise multiplication. The two modulated representations pass through $\texttt{RL}_4$ and $\texttt{RL}_5$,

[1]Given a deep feature representation x of an image and a bounding box B of an object in the image, a precise ROI pooling layer performs the pooling operation in x over the region given by B, resulting in a feature map of a pre-determined size.

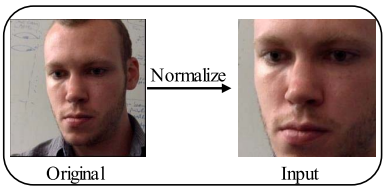


Fig. 5. The process of data normalization.

respectively, for reshaping the feature maps. Finally, the feature maps are concatenated then fed into the gaze predictor module $u(\cdot)$, consisting of three fully-connected layers (512D, 256D, 3D). The predicted gaze is hence given by

$$\hat{p} = u\Big(\mathtt{RL}_4\big(h(\boldsymbol{F}^4) \otimes \nu_1\big) \circledcirc \mathtt{RL}_5\big(h(\boldsymbol{F}^5) \otimes \nu_2\big)\Big), \tag{7}$$

where $h(\cdot)$ stands for the two `Conv` layers, $\mathtt{RL}_{4/5}$ is reshaping layer, $\otimes$ is channel-wise multiplication, and $\circledcirc$ is concatenation. We will describe the details of $h(\cdot)$ and reshaping layers in Tab. VII.

IV. Experiments

In this section, we present experimental results to assess the effectiveness of the proposed method. We first give the details of our datasets and implementation details in Sec. IV-A. We then perform ablation experiments to study the effect of different components of our model in Sec. IV-B. Finally, we compare our method with the current state-of-the-arts in Sec. IV-C.

A. Datasets and Implementation Details

1) Datasets: We use the GazeCapture [29], MPIIGaze [26], EyeDiap [23] and Gaze360 [28] datasets in our experiments. All are widely used benchmark datasets in 3D gaze estimation.

- **GazeCapture**[2]$(\mathcal{D}_{GC})$ [29] is the largest in-the-wild dataset for gaze estimation. It contains data over 1,450 people consisting of almost 2.5M frames. Since the original GazeCapture dataset only provides gaze labels on a 2D screen, we use the pre-processing pipeline [18] to attain 3D head pose from GazeCapture. Note that the same ground-truth gaze vectors in the normalized face coordinate system are always used in the following experiments. We use both phone and tablet sessions in GazeCapture, and adopt the same training and test sets as [29]. All face images are cropped to 224×224, and we roughly set the size of eye regions to be 0.3 times of the face image size, *i.e.*, about 68×68 pixels.
- **MPIIGaze**[3]$(\mathcal{D}_M)$ [26] is another established benchmark dataset for in-the-wild gaze estimation. Compared with GazeCapture, this dataset has higher within-person variations in appearance including illumination, make-up and facial hair changes. These factors potentially make this dataset more challenging. The MPIIGaze dataset provides a standard subset for evaluation, which contains 1,500 left eye images and 1,500 right eye images independently selected from each participant. So we use the images specified in the MPIIFaceGaze subset [27] only for evaluation purposes. The MPIIFaceGaze subset consists of 15 subjects each with 2,500 samples on average.
- **EyeDiap**[4]$(\mathcal{D}_E)$ [23] contains a set of video clips of 16 participants. Since EyeDiap dataset does not provide a evaluative standard, we sample one image per 15 frames [30] from VGA videos. We obtain the video clip from 14 participants since the other two participants are lack of screen target session videos.
- **Gaze360**[5]$(\mathcal{D}_G)$ [28] is a large-scale gaze-tracking dataset for robust 3D gaze estimation in unconstrained images. The dataset consists of 238 subjects in indoor and outdoor environments with labeled 3D gaze across a wide range of head poses and distances. It is the largest publicly available dataset of its kind in term of both subject and variety. This dataset is split into training, validation and test sets.

2) Data Normalization: We pre-process GazeCapture [29], MPIIGaze [26], EyeDiap [23], and Gaze360 [28] datasets following the data normalization procedure described in [68] to extract the face images and the corresponding gaze direction labels. As shown in Fig. 5, the data normalization procedure places a virtual camera to re-render the eye image from a reference point with the head upright, which results in a normalized face image without any in-plane rotation.

3) Implementation Details: We implement the proposed model in PyTorch on two NVIDIA Titan RTX GPUs with approximately 24 hours for training. The ADAM [69] optimizer is employed with initial learning rates of 2e-5 and 1e-4 for the backbone network (*i.e.*, ResNet-18) and the others, respectively. The batch size is set to 128. On GazeCapture and Gaze360 datasets, we train our network for 20 epochs; and on MPIIGaze and EyeDiap datasets, we train it for 200 epochs with a leave-one-person-out strategy. The inference time of one sample is around 1.1ms.

B. Ablation Study

In this section, we provide comprehensive ablation studies to evaluate the contribution of each key component in our method. We follow [32] and use the following settings for the ablation study: (1) training and testing on GazeCapture [29]; (2) cross-dataset testing by training on GazeCapture and testing on MPIIGaze [26], EyeDipa [23] and Gaze360 [28]; (3) within dataset evaluations on MPIIGaze [26], EyeDiap [23] and Gaze360 [28].

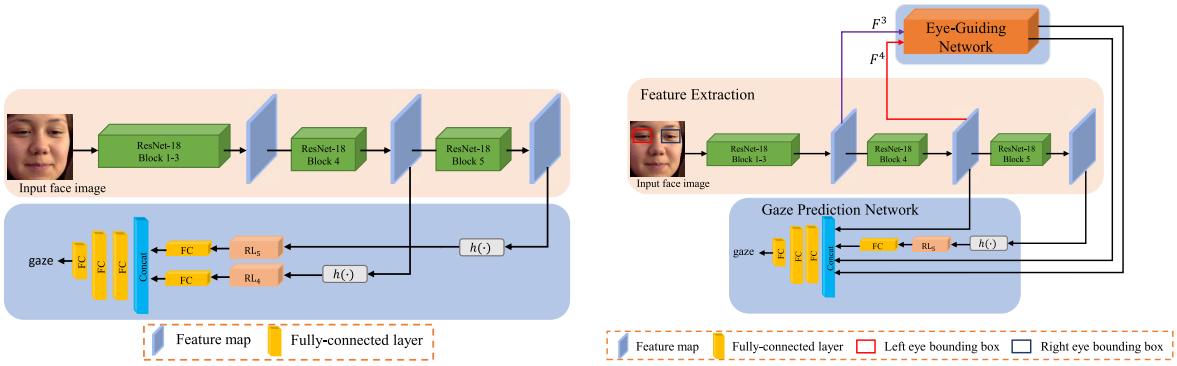
1) Importance of Self-Supervised Learning: To show the ability of self-supervised learning, we provide a variant without the self-supervised learning strategy, named *w/o self-supervision*. As shown in Tab. I, the performance of our full model is better than *w/o self-supervision*, especially on $\mathcal{D}_{GC} \rightarrow \mathcal{D}_M$. The angular error descends as shown by the large gap with and without the self-supervised learning strategy.

[2]https://gazecapture.csail.mit.edu/

[3]https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/gaze-based-human-computer-interaction/appearance-based-gaze-estimation-in-the-wild/

[4]https://www.idiap.ch/en/dataset/eyediap/

[5]http://gaze360.csail.mit.edu/



(a) (b)

Fig. 6. Two variant architectures. (a) *w/o eye-guiding networ*: discarding the eye-guiding network; (b) *concatenation*: discarding the modulation mechanism.

TABLE I

ABLATION STUDY RESULTS ON GAZECAPTURE ($\mathcal{D}_{GC}$) AND MPIIGAZE ($\mathcal{D}_M$). HERE "*Ang error*" MEANS ANGULAR ERROR WHICH IS USED AS EVALUATION METRIC. THE BEST RESULT IN EACH COLUMN IS **BOLD**

Models	$\mathcal{D}_{GC}$ [29]	$\mathcal{D}_{GC}\rightarrow\mathcal{D}_M$ [26]	FLOPs (G)↓	Params (M)↓
	Ang error(∘) ↓	*Ang error*(∘) ↓		
Ours (*w/o self-supervision*)	3.53°	4.64°	2.7	29.5
w/o eye-guiding network	3.91°	5.00°	**2.4**	**26.5**
w/o PrPool	3.63°	4.92°	2.8	30.1
ROI Pooling	3.61°	4.70°	2.7	29.3
concatenation	3.85°	4.90°	2.6	29.2
single branch w/ Block4	3.61°	4.77°	2.6	29.3
single branch w/ Block5	3.58°	4.73°	2.6	29.3
Ours	**3.51°**	**4.37°**	2.7	29.5

This clearly shows that the learned parameters are able to address lots of corrupted situations.

2) Significance of Eye-Guiding Network: In this experiment, we investigate the significance of the eye-guiding network in the proposed architecture. We compare our model with a baseline without the eye-guiding network, named *w/o eye-guiding network*, that is, directly predicting the gaze from the full face image, as shown in Fig. 6a. The performance of this baseline is reported in the second row in Tab. I. We can see that our method achieves better performance (3.53° and 4.64°) than that of *w/o eye-guiding network* (3.91° and 5.00°) with a significant margin (9.7% and 7.2%) in both settings. This demonstrates the effectiveness of adapting the gaze prediction network using the information from the eye-guiding network.

To further prove the effectiveness of our model, we give quantitative results for poor-quality images in Tab. II. Because the original dataset doesn't provide the files about poor-quality images, we pick them manually. We observe that the quantitative results of three subjects' poor-quality images (P00, P02 and P11) are incremental in terms of angular error (*e.g.*, P00: 3.56°→3.02°→3.00°, P02: 4.11°→3.96°→3.85° and P11: 4.60°→4.51°→4.32°). This shows that the eye-guiding network and self-supervised learning improve performance even in bad conditions.

3) Effectiveness of PrPool in Eye-Guiding Network: In the eye-guiding network, we have used `PrPool` to extract precise semantic cues from two eyes in order to generate the two modulation vectors as shown in Fig. 4. To investigate the effectiveness of `PrPool`, we compare with two alternative baselines in Tab. I. The first one, named *w/o PrPool*, removes `PrPool` in the eye-guiding network, and directly extracts eye features from $\boldsymbol{F}^3$ and $\boldsymbol{F}^4$. The second one, named *ROI Pooling*, uses `ROI Pooling` to replace `PrPool` in the eye-guiding network. Our model outperforms these two baselines, which demonstrates the benefit of `PrPool` in extracting effective information from the eye regions.

4) Effectiveness of Modulation Mechanism: In our model, we use the feature representations from the eye images to modulate the features from the face image. To validate the effectiveness of this modulation mechanism, we compare with a baseline, named *concatenation*, that simply concatenates the features from the face and two eyes (see Fig. 6b). This baseline is similar to previous work, *e.g.*, [30], [57]. Note that the simple concatenation does not model rich interactions between face and eye images. As shown in Tab. I, we can see that *concatenation* performs inferior.

5) Usefulness of Two Branches in Gaze Prediction Network: In the gaze prediction network, we use two modulation vectors to adjust the generated two groups of feature maps from `Block4` and `Block5`. This two-branch structure is in charge of capturing local and global information. To investigate the effectiveness of two branches in gaze prediction network, we design two variants: using a single modulation vector to adjust the generated features from `Block4` and `Block5`,

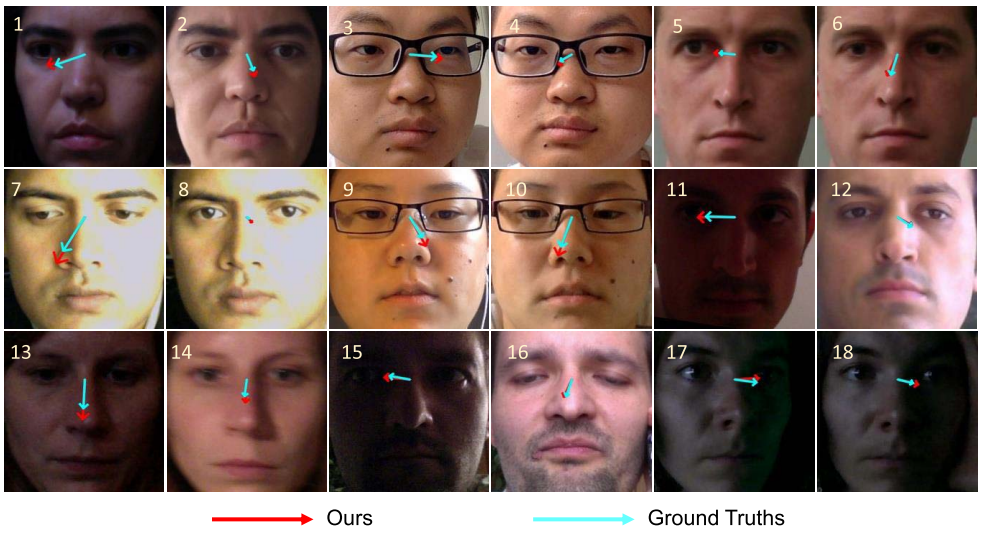


Fig. 7. Some visual results of estimated 3D gaze.

TABLE II

THE QUANTITATIVE RESULTS OF ABLATION STUDY ABOUT EYE-GUIDING NETWORK. WE RANDOMLY CHOOSE 3 SUBJECTS ON MPIIGAZE DATASET AND MANUALLY PICK POOR-QUALITY IMAGES

Subjects	Poor-quality images	*w/o eye-guiding network* $Ang\ error(\circ)\downarrow$	Ours (*w/o self-supervision*) $Ang\ error(\circ)\downarrow$	**Ours** $Ang\ error(\circ)\downarrow$	$\nabla(\%)$
P00	206	3.56°	3.02°	**3.00**°	15.7
P02	829	4.11°	3.96°	**3.85**°	6.3
P11	1453	4.60°	4.51°	**4.32**°	6.0

TABLE III

COMPARISON WITH THE STATE-OF-THE-ARTS. NOTE THAT OUR METHOD PERFORMS PARTICULARLY WELL IN THE MOST CHALLENGING CROSS-DATASET SETTING WHEN TRAINED ON THE LARGE GAZECAPTURE ($\mathcal{D}_{GC}$) DATASET EVEN IF TESTED ON MPIIGAZE ($\mathcal{D}_M$), EYEDIAP ($\mathcal{D}_E$) AND GAZE360 ($\mathcal{D}_G$). THE BEST RESULT IN EACH COLUMN IS **BOLD**

Models	$\mathcal{D}_{GC}$ [29] $Ang\ error(\circ)\downarrow$	$\mathcal{D}_{GC}\rightarrow\mathcal{D}_M$ [26] $Ang\ error(\circ)\downarrow$	$\mathcal{D}_{GC}\rightarrow\mathcal{D}_E$ [23] $Ang\ error(\circ)\downarrow$	$\mathcal{D}_{GC}\rightarrow\mathcal{D}_G$ [28] $Ang\ error(\circ)\downarrow$	FLOPs (G)↓	Params (M)↓
FAZE (DenseNet) [18]	3.5°	5.2°	-	-	7.8	31.6
LRSNet (ResNet-18) [32]	**3.3**°	4.9°	6.0°	-	-	-
ETH-XGaze (ResNet-50) [56]	**3.3**°	4.5°	13.7°	30.2°	4.1	**23.5**
Ours (ResNet-18)	3.5°(3.51°)	**4.4**°(4.37°)	**5.9**°(5.92°)	**25.3**°(25.33°)	**2.7**	29.5

named *single branch w/ Block4* and *single branch w/ Block5*, respectively. As shown in Tab. I, the results of these variants confirm that our full model can achieve more favorable performance than them.

C. Main Comparisons

We compare our method with 12 state-of-the-art appearance-based gaze estimation methods, including GazeNet [26], AR-Net [57], ARE-Net [57], CA-Net [31], MeNets [13], FAR-Net [30], FARE-Net [30], FAZE [18], LRSNet [32], FullFace [27], DPGaze [47] and ETH-XGaze [56]. For a fair comparison, the results of all compared methods are either from the original papers or obtained by running their released codes.

1) Quantitative Results: In Tab. III, we show the comparisons on the GazeCapture dataset and in the cross-dataset setting. Overall, our proposed method outperforms other state-of-the-art approaches. Concretely, our result (4.37°) outperforms the second best method ETH-XGaze [56] (4.5°) by over 3% on the cross-dataset from GazeCapture to MPIIGaze evaluation. For the cross-dataset from GazeCapture to EyeDiap and Gaze360 evaluations, our method consistently outperforms the previous state-of-the-art methods.

In Tab. IV, we show the comparisons using within evaluation protocol on MPIIGaze, EyeDiap and Gaze360 datasets. We compare our model against 11 state-of-the-art methods, including LRSNet [32], GazeNet [26], AR-Net [57], ARE-Net [57], CA-Net [25], MeNet [13], FAR-Net [30], FARE-Net [30], FullFaze [27], DPGaze [47] and ETH-XGaze [56]. Within MPIIGaze evaluation, CA-Net [31] achieves the best result (4.1°). Our proposed method achieves the second best result (4.3°) in angular error. Moreover, our method achieves the best result on Gaze360.

2) Computational Complexity Comparison: We report parameter amount (Params) and FLOPs of our various variants in Tab. I and most compared methods in Tab. III. Unfortunately,

TABLE IV

COMPARISON OF THE PROPOSED MODEL WITH CURRENT STATE-OF-THE-ARTS WITHIN DATASET EVALUATIONS ON MPIIGAZE ($\mathcal{D}_M$), EYEDIAP ($\mathcal{D}_E$) AND GAZE360 ($\mathcal{D}_G$). THE TOP THREE RESULTS ARE SHOWN IN RED, BLUE AND GREEN, RESPECTIVELY

Models	$\mathcal{D}_M$ [26]	$\mathcal{D}_E$ [23]	$\mathcal{D}_G$ [28]
	Ang error(∘) ↓	*Ang error*(∘) ↓	*Ang error*(∘) ↓
LRSNet [32]	4.5°	6.6°	N/A
GazeNet [26]	5.8°	6.8°	N/A
FullFace [27]	4.9°	6.6°	15.0°
AR-Net [57]	5.7°	6.4°	N/A
ARE-Net [57]	5.0°	6.1°	N/A
MeNets [13]	4.9°	N/A	N/A
DP-Gaze [47]	4.5°	10.3°	N/A
FAR-Net [30]	4.5°	6.1°	N/A
FARE-Net [30]	4.4°	5.9°	N/A
ETH-XGaze [56]	4.8°	6.5°	N/A
CA-Net [31]	4.1°	5.3°	N/A
Ours	4.3°	6.3°	13.2°

most methods don't release the codes, so their Params and FLOPs are missing in Tab. IV. As shown in Tab. I, our method puts a little more computational complexity compared to various variants. Notably, the FLOPs of our method is 2.7G, accounting for 66% of the second-place method ETH-XGaze in Tab. III. The parameter amount of our method (*i.e.*, 29.5M) is slightly higher than ETH-XGaze (*i.e.*, 23.5M). In general, the proposed method is an efficient and effective method.

3) Qualitative Results: In Fig. 7, we show the qualitative examples of our method in the cross-dataset evaluation from GazCapture to MPIIGaze. Our proposed approach consistently generates gaze directions close to the ground-truths on different samples.

There are several challenging and complicated scenes for gaze estimation: 1) low-quality images (No. 11, 13, 14, 15, 17 and 18); 2) abnormal illumination (No. 1, 7, 8 and 13); and 3) wearing glasses (No. 3, 4, 9 and 10). Besides, there are some good images (No. 2, 5, 6 and 12). Corrupted face images might fail to provide valuable cues, and even provide incorrect information. Thanks to the strong power of each component in the proposed model, our model overcomes various extreme conditions.

V. CONCLUSION AND DISCUSSION

In this paper, we introduce auxiliary self-learning to gaze estimation, which can deal with corrupted images (*e.g.*, blurred images) at test time. Moreover, we propose a Modulation-based Adaptive Network (MANet), which is the first method utilizing high-level cues of the eye-specific regions to modulate face features in the gaze estimation task. MANet contains feature extraction, eye-guiding network and gaze prediction network. The proposed eye-guiding network captures meaningful information from two eyes to generate adaptive features, which adjust the gaze prediction network by the modulation-feedback mechanism to filter noise information for accurate gaze directions. Further, we combine self-learning and MANet to learn to adapt to challenging cases by joint training. In addition, our model runs at 900 *fps* for practical real-time eye tracking applications.

Notwithstanding the above advantages, a limitation of our method is that it requires the coordinates of eye pairs for the eye-guiding network, which sets some constraints for practical applications. We will design a network to automatically locate the eye regions or more important regions to extract high-level gaze cues in the future work. Moreover, we will use more augmentations for self-learning training which will make our model more robust.

TABLE V

DESCRIPTION OF MLP

Operation	Configuration	Channel
MLP	Linear	1024 → 512
	BatchNorm1d	
	Linear	512 → 512
	ReLU	

TABLE VI

DESCRIPTION OF Conv OF $\boldsymbol{F}^3$ AND $\boldsymbol{F}^4$ IN THE EYE-GUIDING NETWORK. HERE, k DENOTE THE KERNEL SIZE, p IS THE PADDING, AND s IS THE FILTER STRIDE

Operation	Layer	Configuration	Channel
$\texttt{Conv}(\boldsymbol{F}^3)$	Conv2d	k(3,3), $p=1$, $s=1$	128 → 256
		BatchNorm2d	
		ReLU	
$\texttt{Conv}(\boldsymbol{F}^4)$	Conv2d	k(3,3), $p=1$, $s=1$	256 → 256
		BatchNorm2d	
		ReLU	

TABLE VII

DESCRIPTION OF $h(\cdot)$ OF $\boldsymbol{F}^4$ AND $\boldsymbol{F}^5$, RL_4, AND RL_5 IN THE GAZE PREDICTION NETWORK. HERE, k INDICATES THE KERNEL SIZE, p IS THE PADDING, AND s IS THE FILTER STRIDE

Operation	Layer	Configuration	Channel
$h(\boldsymbol{F}^4)$	Conv2d	k(3,3), $p=1$, $s=1$	256 → 256
		BatchNorm2d	
		ReLU	
	Conv2d	k(3,3), $p=1$, $s=1$	256 → 512
		BatchNorm1d	
		ReLU	
$h(\boldsymbol{F}^5)$	Conv2d	k(3,3), $p=1$, $s=1$	512 → 512
		BatchNorm2d	
		ReLU	
	Conv2d	k(3,3), $p=1$, $s=1$	512 → 512
		BatchNorm2d	
		ReLU	
RL_4	MaxPool2d	k(3,3), $p=0$, $s=1$	512 → 512
		BatchNorm2d	
		ReLU	
	Conv2d	k(3,3), $p=0$, $s=1$	512 → 512
		BatchNorm2d	
		ReLU	
	Conv2d	k(3,3), $p=0$, $s=1$	512 → 512
		BatchNorm2d	
		ReLU	
RL_5	Conv2d	k(3,3), $p=0$, $s=1$	512 → 512
		BatchNorm2d	
		ReLU	

REFERENCES

[1] T. Alshawi, Z. Long, and G. AlRegib, "Unsupervised uncertainty estimation using spatiotemporal cues in video saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2818–2827, Jun. 2018, doi: 10.1109/TIP.2018.2813159.

[2] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1913–1927, Aug. 2020, doi: 10.1109/TPAMI.2019.2905607.

[3] W. Chen *et al.*, "Gaze estimation via the joint modeling of multiple cues," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Apr. 7, 2021, doi: 10.1109/TCSVT.2021.3071621.

[4] L. Sun, Z. Chen, Q. M. J. Wu, H. Zhao, W. He, and X. Yan, "AMPNet: Average- and max-pool networks for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4321–4333, Nov. 2021, doi: 10.1109/TCSVT.2021.3054471.

[5] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017, doi: 10.1109/TCSVT.2016.2595324.

[6] X. Zhang, Y. Sugano, and A. Bulling, "Everyday eye contact detection using unsupervised gaze target discovery," in *Proc. 30th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2017, pp. 193–203.

[7] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, "Conversational gaze aversion for humanlike robots," in *Proc. ACM/IEEE HRI*, Mar. 2014, pp. 25–32.

[8] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for EEG-based human intention recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3033–3044, Jul. 2020, doi: 10.1109/TCYB.2019.2905157.

[9] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. C. Bovik, "Study of 3D virtual reality picture quality," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 89–102, Jan. 2020.

[10] J. He *et al.*, "On-device few-shot personalization for real-time gaze estimation," in *Proc. IEEE ICCVW*, Oct. 2019, pp. 1149–1158.

[11] Y. Yu, G. Liu, and J.-M. Odobez, "Improving few-shot user-specific gaze adaptation via gaze redirection synthesis," in *Proc. IEEE CVPR*, Jun. 2019, pp. 11929–11938.

[12] Y. Yu and J.-M. Odobez, "Unsupervised representation learning for gaze estimation," in *Proc. IEEE CVPR*, Jun. 2020, pp. 7312–7322.

[13] Y. Xiong, H. J. Kim, and V. Singh, "Mixed effects neural networks (MeNets) with applications to gaze estimation," in *Proc. IEEE CVPR*, Jun. 2019, pp. 7735–7744.

[14] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "TabletGaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets," *Mach. Vis. Appl.*, vol. 28, nos. 5–6, pp. 445–461, Aug. 2017.

[15] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *J. Mach. Learn. Res.*, vol. 20, no. 184, pp. 1–25, 2019.

[16] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019, *arXiv:1903.12261*.

[17] K. Wang, R. Zhao, H. Su, and Q. Ji, "Generalizing eye tracking with Bayesian adversarial learning," in *Proc. IEEE CVPR*, Jun. 2019, pp. 11899–11908.

[18] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," in *Proc. IEEE ICCV*, Oct. 2019, pp. 9367–9376.

[19] J.-B. Grill *et al.*, "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. NeurIPS*, vol. 33, 2020, pp. 21271–21284.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.

[21] Z. Fu, Z. Mao, C. Yan, A.-A. Liu, H. Xie, and Y. Zhang, "Self-supervised synthesis ranking for deep metric learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 13, 2021, doi: 10.1109/TCSVT.2021.3124908.

[22] S. Guo, E. Rigall, L. Qi, X. Dong, H. Li, and J. Dong, "Graph-based CNNs with self-supervised module for 3D hand pose estimation from monocular RGB," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1514–1525, Apr. 2021, doi: 10.1109/TCSVT.2020.3004453.

[23] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," in *Proc. Symp. Eye Tracking Res. Appl.*, Mar. 2014, pp. 255–258.

[24] G. Liu, Y. Yu, K. A. F. Mora, and J. Odobez, "A differential approach for gaze estimation with calibration," in *Proc. BMVC*, 2018, p. 6.

[25] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. ECCV*, 2018, pp. 339–357.

[26] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019, doi: 10.1109/TPAMI.2017.2778103.

[27] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE CVPRW*, Jul. 2017, pp. 2299–2308.

[28] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE ICCV*, Oct. 2019, pp. 6911–6920.

[29] K. Krafka *et al.*, "Eye tracking for everyone," in *Proc. IEEE CVPR*, Jun. 2016, pp. 2176–2184.

[30] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Trans. Image Process.*, vol. 29, pp. 5259–5272, 2020, doi: 10.1109/TIP.2020.2982828.

[31] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proc. AAAI*, 2020, pp. 10623–10630.

[32] X. Zhang, Y. Sugano, A. Bulling, and O. Hilliges, "Learning-based region selection for end-to-end gaze estimation," in *Proc. BMVC*, 2020, pp. 1–13.

[33] C. Palmero, J. Selva, M. A. Bagheri, M. B. Ca, and S. Escalera, "Recurrent CNN for 3D gaze estimation using appearance and shape cues," in *Proc. BMVC*, 2018, pp. 1–13.

[34] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE CVPR*, Jun. 2020, pp. 9729–9738.

[36] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semi-supervised recurrent convolutional attention model for human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1747–1756, May 2020, doi: 10.1109/TNNLS.2019.2927224.

[37] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE CVPR*, Jun. 2018, pp. 4320–4328.

[38] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. ICLR*, 2016, pp. 1–13.

[39] Z. Zheng and Y. Yang, "Unsupervised scene adaptation with memory regularization *in vivo*," in *Proc. IJCAI*, Jul. 2020, pp. 1–7.

[40] K. Wang and Q. Ji, "Real time eye gaze tracking with 3D deformable eye-face model," in *Proc. IEEE ICCV*, Oct. 2017, pp. 1003–1011.

[41] C. H. Morimoto, A. Amir, and M. Flickner, "Detecting eye position and gaze from a single camera and 2 light sources," in *Proc. Int. Conf. Pattern Recognit.*, vol. 4, 2002, pp. 314–317.

[42] E. Wood and A. Bulling, "EyeTab: Model-based gaze estimation on unmodified tablet computers," in *Proc. ETRA*, Mar. 2014, pp. 207–210.

[43] K. A. F. Mora and J.-M. Odobez, "Geometric generative gaze estimation (G3E) for remote RGB-D cameras," in *Proc. IEEE CVPR*, Jun. 2014, pp. 1773–1780.

[44] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 802–815, Feb. 2012.

[45] L. Sun, M. Song, Z. Liu, and M. T. Sun, "Real-time gaze estimation with online calibration," *IEEE Multimedia*, vol. 21, no. 4, pp. 28–37, Oct./Dec. 2014, doi: 10.1109/MMUL.2014.54.

[46] N. M. Arar, H. Gao, and J.-P. Thiran, "A regression-based user calibration framework for real-time gaze estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2623–2638, Dec. 2017, doi: 10.1109/TCSVT.2016.2595322.

[47] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in *Proc. ECCV*, 2018, pp. 741–757.

[48] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," in *Proc. ETRA*, Jun. 2018, pp. 1–10.

[49] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE CVPR*, Jul. 2017, pp. 2242–2251.

[50] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-Synthesis for appearance-based 3D gaze estimation," in *Proc. IEEE CVPR*, Jun. 2014, pp. 1821–1828.

[51] K. Wang, R. Zhao, and Q. Ji, "A hierarchical generative model for eye image synthesis and eye gaze estimation," in *Proc. IEEE CVPR*, Jun. 2018, pp. 440–448.
[52] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," in *Proc. IEEE ICCV*, Dec. 2015, pp. 3756–3764.
[53] H. Deng and W. Zhu, "Monocular free-head 3D gaze tracking with deep learning and geometry constraints," in *Proc. IEEE ICCV*, Oct. 2017, pp. 3162–3171.
[54] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," in *Proc. NeurIPS*, 1993, pp. 753–760.
[55] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE CVPR*, Jun. 2015, pp. 4511–4520.
[56] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *Proc. ECCV*, 2020, pp. 365–381.
[57] Y. Cheng, L. Feng, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. ECCV*, 2018, pp. 105–121.
[58] M. K. K. Reddy, M. Rochan, Y. Lu, and Y. Wang, "AdaCrowd: Unlabeled scene adaptation for crowd counting," *IEEE Trans. Multimedia*, vol. 24, pp. 1008–1019, 2022, doi: 10.1109/TMM.2021.3062481.
[59] M. Rochan, M. K. K. Reddy, L. Ye, and Y. Wang, "Adaptive video highlight detection by learning from user history," in *Proc. ECCV*, 2020, pp. 261–278.
[60] Y. Lu, F. Yu, M. K. K. Reddy, and Y. Wang, "Few-shot scene-adaptive anomaly detection," in *Proc. ECCV*, 2020, pp. 125–141.
[61] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. NeurIPS*, 2016, pp. 1–9.
[62] M. K. K. Reddy, M. A. Hossain, M. Rochan, and Y. Wang, "Few-shot scene adaptive crowd counting using meta-learning," in *Proc. IEEE WACV*, Mar. 2020, pp. 2803–2812.
[63] L. Zhang, S. Wang, X. Chang, J. Liu, Z. Ge, and Q. Zheng, "Auto-FSL: Searching the attribute consistent network for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Apr. 29, 2021, doi: 10.1109/TCSVT.2021.3076523.
[64] C. Zhang, C. Li, and J. Cheng, "Few-shot visual classification using image pairs with binary transformation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2867–2871, Sep. 2020, doi: 10.1109/TCSVT.2019.2920783.
[65] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.
[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Jun. 2016, pp. 770–778.
[67] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. ECCV*, 2018, pp. 816–832.
[68] X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proc. ETRA*, Jun. 2018, pp. 1–9.
[69] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2014, pp. 1–15.

Yong Wu received the B.E. degree from Anhui Science and Technology University, Bengbu, China, in 2015, and the M.S. degree from Shantou University, Shantou, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include domain adaptation, gaze estimation, and visual tracking.

Gongyang Li received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai. His research interests include image/video object segmentation and saliency detection.

Zhi Liu (Senior Member, IEEE) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2005. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He was a TPC Member/Session Chair of ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, and WIAMIS 2013. He has co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an Area Editor of *Signal Processing: Image Communication*. He served as a Guest Editor for the Special Issue on "Recent Advances in Saliency Models, Applications and Evaluations" in *Signal Processing: Image Communication*.

Mengke Huang received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2014. He is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai. His research interests include saliency detection and 360° visual saliency.

Yang Wang received the B.Sc. degree in computer science from the Harbin Institute of Technology, Harbin, China, the M.Sc. degree in computer science from the University of Alberta, Edmonton, AB, Canada, and the Ph.D. degree in computer science from Simon Fraser University, Burnaby, BC, Canada. He was previously a NSERC Post-Doctoral Fellow with the University of Illinois at Urbana–Champaign, Champaign, IL, USA. He is currently an Associate Professor of computer science with the University of Manitoba, Winnipeg, MB, Canada. His research interests include computer vision and machine learning.