This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2022.3210219

A Complementary Dual-branch Network for Appearance-based Gaze Estimation from Low-resolution Facial Image

Zhesi Zhu, Dong Zhang, Cailong Chi, Ming Li, *Senior Member, IEEE*, and Dah-Jye Lee, *Senior Member, IEEE*

***Abstract*—Estimating gaze from a low-resolution facial image is a challenging task. Most current networks for gaze estimation focus on using face images of adequate resolution. Their performance degrades when the image resolution decreases due to information loss. This work aims to explore more helpful face and gaze information in a novel way to alleviate the problem of information loss in the low-resolution gaze estimation task. Considering that all faces have a relatively fixed structure, it is feasible to reconstruct the residual information of face and gaze based on the solid constraint of the prior knowledge of face structure through learning an end-to-end mapping from pairs of low- and high-resolution images. This paper proposes a complementary dual-branch network (CDBN) to achieve this task. A fundamental branch is designed to extract features of the major structural information from low-resolution input. A residual branch is employed to reconstruct features containing the residual information as a supplement under the supervision of both the high-resolution image and gaze direction. These two features are then fused and processed for gaze estimation. Experimental results on three widely used datasets, MPIIFaceGaze, EYEDIAP, and RT-GENE, show that the proposed CDBN achieves more accurate gaze estimation from the low-resolution input image compared with the state-of-the-art methods.**

***Index Terms*—Gaze estimation, Super-resolution, Low resolution, Complementary dual-branch network, Residual information.**

I. Introduction

EYE gaze is an important non-verbal cue for human behavior analysis. It is involved in many cognitive processes and reflects the visual attentiveness, internal thoughts, and mental states of a person during social interaction and other human behaviors. Over the past two decades, gaze estimation has attracted research interest in the fields of cognitive science and computer vision. It helps the computer to better understand human intention and plays a prominent role in various intelligent applications such as virtual reality [1], [2], driving assistance systems [3], security surveillance [4], and human-robot interaction (HRI) [5]–[7].

Zhesi Zhu, Dong Zhang, and Cailong Chi are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China, 510006. E-mail: zhuzhs@mail2.sysu.edu.cn, zhangd@mail.sysu.edu.cn, and chiclong@mail2.sysu.edu.cn.

Ming Li is with the Data Science Research Center, Duke Kunshan University, Kunshan, China, 215316. E-mail: ming.li369@dukekunshan.edu.cn

Dah-Jye Lee is with the Department of Electrical and Computer Engineering, Brigham Young University, Provo, Utah, USA, 84602. E-mail: djlee@byu.edu.

Corresponding author: Dong Zhang. Email: zhangd@mail.sysu.edu.cn.

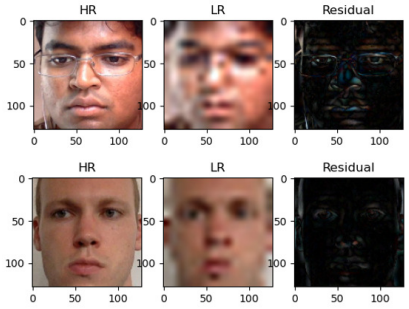


Fig. 1. Comparison of high-resolution images (left column), low-resolution images (center column), and visualized residual images (right column). The high-resolution face images are selected from the MPIIFaceGaze dataset. The low-resolution images are generated by applying bicubic downsampling (8x) to the corresponding high-resolution images. The visualized residual images are the absolute difference between the high- and low-resolution images. All images are scaled to the same size for easy comparison.

Based on the equipment requirements and implementation processes, gaze estimation can be typically grouped into model-based and appearance-based methods [8], [9]. Model-based methods obtain highly accurate gaze by extracting eye features and building geometric eye models with the assistance of dedicated devices. They generally work in a controlled environment due to limitations such as short working distances and high facility costs. Unlike model-based methods, appearance-based methods learn a mapping directly from the input image to the gaze direction. They only require a consumer monocular camera to perform, making them possible to adapt to outdoor application scenarios. Although this simple setup greatly extends their applicability, appearance-based methods are still challenging because they need to handle the appearance diversity affected by various factors such as head pose and individual facial differences.

Advances in deep learning in recent years have made convolution neural networks (CNNs) a sort of efficient method for gaze estimation. With sufficient learning data, CNN-based methods have shown powerful representation capability to handle diversified appearances and achieved remarkable performance. Despite the recent developments, most state-of-the-art CNN-based methods merely consider face images of ideal sizes, e.g., 224×224 pixels. Their estimation accuracy degrades greatly as the image resolution decreases. In scenarios where the resolution of the input image is low, e.g., the human face is captured from a long distance in the wild or the region of the human face is cropped from an image of

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2022.3210219

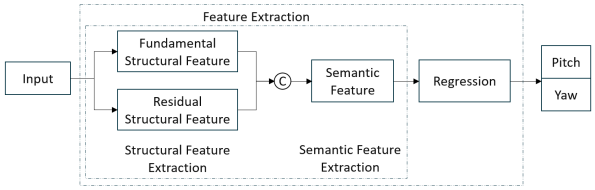


Fig. 2. Framework of our proposed Complementary Dual-branch Network (CDBN). The c in circle represents concatenation of feature maps.

a crowd, accuracy obtained by current methods is far from satisfactory.

Most existing methods strive to extract discriminative features with techniques like attention mechanism [10]–[12] or dilated convolution [12]–[14] to represent the input image. These methods work well when the input is of high resolution (HR) that includes rich facial information. When the quality of the input image degrades, these methods have trouble characterizing the intrinsic features of a low-resolution (LR) face image due to information loss. Fig. 1 shows the comparison of high- and low-resolution images and their corresponding visualized residual images. Detailed facial information that appears in HR images, such as sharp edges and eye details, becomes very fuzzy in LR images. The information difference between the HR and LR image, defined as the residual information in this paper, is reflected in the residual images in Fig. 1. As shown in Fig. 1, the residual information is significant, and recovering it could help improve gaze estimation accuracy at low resolutions.

The main idea of this work is to recover the residual information from the LR face image and supplement the gaze estimation network with this information for better gaze estimation accuracy. Since the structure and components of human faces are relatively fixed, there exists prior knowledge which can be used as a solid constraint to recover the residual information. It has been demonstrated in the research field of face super-resolution [15], [16] that neural networks can grasp this facial prior knowledge and recover the lost information by training with pairs of HR and LR images. Therefore, in addition to the basic features extracted directly from the LR input, features that contain the residual information, defined as residual features in this paper, will also be recovered and introduced into our network through network learning. The recovered residual features act as the supplementary information and lead to improved estimation accuracy.

This paper proposes a new model named Complementary Dual-Branch Network (CDBN) to address the challenge of low-resolution gaze estimation. The framework of the proposed model is shown in Fig. 2. Feature extraction process is explicitly divided into the structural feature extraction stage and the semantic feature extraction stage. A dual-branch module is used to capture structural features in the former stage. One branch extracts the fundamental structural features from the LR image, while the other restores the residual structural features based on the prior knowledge learned from pairs of HR and LR images. The restoration of residual structural features is also supervised under the ground truth of gaze to ensure that the restored information tends to be beneficial for gaze estimation.

Features generated by the dual-branch are then concatenated in the channel dimension and are fed into the subsequent module for semantic features extraction. Gaze estimation regression is finally performed on the semantic features. To evaluate the performance of our proposed CDBN, extensive experiments were carried out on three widely used gaze datasets, i.e., MPIIFaceGaze [17], EYEDIAP [18], and RT-GENE [19]. Experimental results demonstrate that our proposed CDBN achieves smaller angular errors with images of various resolutions compared to the state-of-the-art methods.

For this paper, the main contributions are as follows:

1) We evaluate the performance of the existing state-of-the-art gaze estimation methods with low-resolution input images. Their performance degrades rapidly when the image resolution decreases because the amount of information contained in the LR image is limited.
2) A dual-branch model named CDBN is proposed to alleviate the problem caused by limited information and improve the accuracy of low-resolution gaze estimation. In CDBN, one branch extracts fundamental features directly from the LR input, while the other restores residual features as a supplement. The residual features are constructed based on the prior knowledge learned from pairs of HR and LR images. The learning process is supervised by the ground truth of gaze meanwhile. The recovered residual features introduce prior facial knowledge into the network and help CDBN better understand and analyze the low-resolution image.
3) Extensive experiments have been conducted to demonstrate the effectiveness of the proposed CDBN. Compared to the state-of-the-art models, CDBN achieves smaller angular errors with various resolutions.

II. Related Work

Generally, gaze estimation algorithms can be categorized as either model-based or appearance-based methods. Model-based methods build a subject-specific geometric eye model to estimate gaze. The eye model is fitted by geometric features such as the infrared corneal reflections [20] and pupil center [21]. Although model-based methods achieve superior accuracy, most of them require short working distances and dedicated devices, which greatly limits their applications.

Appearance-based methods learn a mapping directly from an eye or face image to the gaze estimation target. The target is defined as either a gaze point in 2D estimation or a gaze vector in 3D estimation. Appearance-based methods usually use the image captured by a consumer camera as input and therefore have the potential of estimating gaze from the low-resolution image. Various conventional approaches such as Random Forests [22], K-Nearest Neighbors [23], and adaptive linear regression [24] were applied to address the gaze estimation task in the early days. While most recently, the state-of-the-art results were achieved by CNN-based methods.

Zhang et al. were the first to adopt a convolution network to estimate gaze [25]. Krafka et al. implemented a multi-region

2D gaze estimation method to estimate gaze points on the screen of a smartphone or tablet [26]. Zhang et al. presented a spatial weighting CNN that took full-face into account and encoded the weights for the different regions of the face [11]. Chen et al. applied the technique of dilated convolutions to catch the subtle distinctions of eye appearance when gaze angle varied [13]. They subsequently proposed an improved version by adding subject-dependent bias into their original network [13] to handle inter-subject variations in appearance [14]. Dai et al. built a residual network with an attention mechanism to extract contributive features for gaze from the face image and the original input image with the background [27]. Some work, e.g., I2D-Net [12] and FARE-Net [28], achieved good results by considering the difference between left and right eyes. And some work, e.g., CA-Net [10], AGE-Net [12], and the model with LBSAM and GBSAM [29], integrated attention mechanism into the eye branches of CNN models in pursuit of an improved representation capability. Although current CNN-based methods have obtained promising results in terms of accuracy in gaze estimation, their performance still has room for improvement, especially with the low-resolution input image.

As the increasing real-world applications may not guarantee high-quality images, vision tasks in complex scenes, such as occlusion [30]–[32] and low resolution [33]–[36], are becoming crucial and challenging. This paper focuses on the low-resolution problem. An intuitive idea for the challenge of low-resolution is to employ the Single Image Super-Resolution (SISR) technique as an auxiliary task. SISR aims to reconstruct a high-resolution image from a single low-resolution image. It is straightforward to think of a two-stage approach that first applies an off-the-shelf SISR method to reconstruct a high-resolution image and then carries out the actual vision task on the reconstructed image. However, if the optimization of the super-resolution network is not designed especially for the main vision task, the fine details of the reconstructed image may not contribute to improving the performance of the main task. Joint training in a single-stage manner between SISR and the main task is essential to addressing this challenge.

The existing joint training networks can be divided into two categories: cascade networks [37]–[40] and parallel networks [33]–[36]. In cascade networks, the super-resolution network is directly connected to the network of the main task in series. The losses of two networks are combined and optimized together. However, it may not be an optimal choice because the information of the two networks may not be fully fused or shared. Parallel networks can overcome this shortcoming and embed the information more mutually and thoroughly.

Yin et al. proposed a joint facial alignment and super-resolution network to simultaneously detect facial landmarks and super-resolve low-resolution faces [33]. It is a parallel multi-task network that allows two tasks to benefit from each other, which improves the performance of both tasks. Wang et al. presented a dual super-resolution learning paradigm to keep a high-resolution representation for semantic segmentation [34]. This paradigm significantly improves the performance of semantic segmentation for low-resolution input and demonstrates the effectiveness of super-resolution as an auxiliary means. Bai et al. proposed a face detection algorithm that generated clear faces from tiny ones in the picture by adopting a generative adversarial network (GAN) and detected the positions of the tiny faces meanwhile [35]. The favorable results show a positive effect of super-resolution on face detection when dealing with low-resolution faces as small as 10×10 pixels. A similar method has been used for small object detection and has drawn the same conclusion [36]. The success of parallel joint training indicates the potential of the super-resolution technique in handling low-resolution vision tasks. This paper incorporates this idea to help with gaze estimation from the low-resolution image.

III. Proposed Method

A. Overview

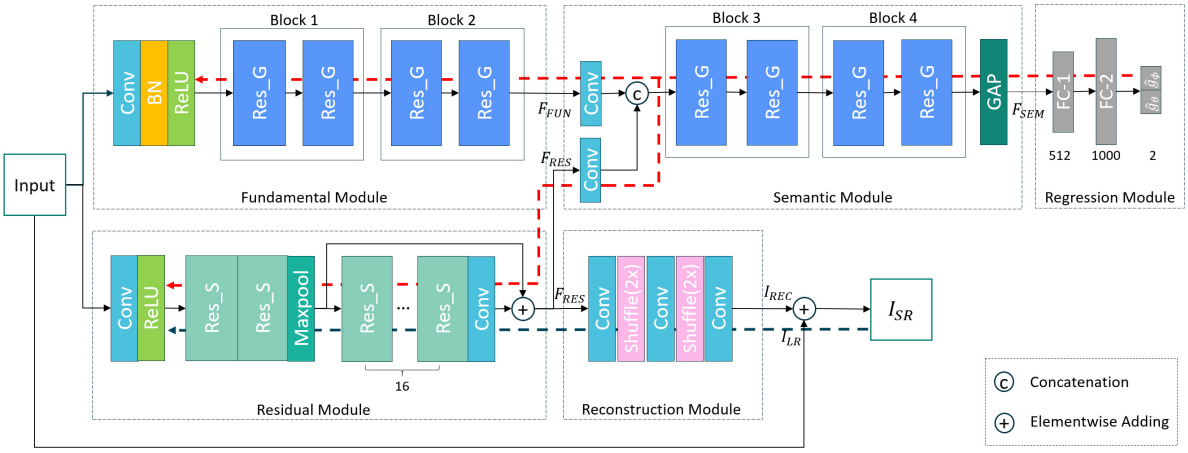
The core idea of the proposed method is to introduce residual information into low-resolution gaze estimation. On the one hand, estimating gaze direction from a low-resolution facial image is a hard problem owing to information loss. Meanwhile, recovering the lost information is an underdetermined inverse problem since multiple possible solutions exist for any given low-resolution image. Such a problem is likely to be solved by constraining the solution space with strong prior information [41], [42], e.g., the prior knowledge of a human face. Advances in the research field of super-resolution show that learning an end-to-end mapping between low-/high-resolution images with a deep convolutional neural network is an efficient way to grasp the prior information and recover the lost information [15], [16]. Capturing prior information through deep CNN reduces ambiguity and uncertainty in the recovering procedure, making the residual features reliable. On the other hand, using only pairs of low- and high-resolution images to supervise the training of residual features is insufficient to ensure that the recovered residual information is beneficial for gaze estimation. It is essential to use gaze-related supervision simultaneously during the recovery process.

As an implementation of the idea, this paper proposes a unified optimization framework to estimate gaze from a low-resolution facial image. As shown in Fig. 3, the proposed network employs a Fundamental Module to capture basic structural features directly from low-resolution input and a Residual Module to restore residual features. The training of the Residual Module is supervised by pairs of high- and low-resolution images and the mapping relation between images and the ground truth of gaze. This design pushes the Residual Module to recover useful residual features that are biased towards gaze estimation. The two parts of features are then concatenated and mapped to a high-level semantic feature space. Gaze regression is finally performed on the semantic features.

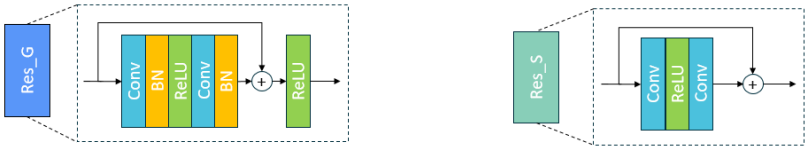
B. Complementary Dual-Branch Network

As shown in Fig. 3, the proposed network consists of five parts: (a) Fundamental Module for extracting basic structural features from the LR input; (b) Residual Module for restoring residual structural features as the supplementary information for gaze estimation; (c) Reconstruction Module for recovering

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2022.3210219



(a) The architecture of CDBN



(b) Res_G block (c) Res_S block

Fig. 3. An overview of the proposed approach. (a) The architecture of the proposed network CDBN. Fundamental Module and Residual Module form a dual branch. Two features, F_{FUN} and F_{RES}, are extracted by this dual-branch. They are fused in channel dimension and go through Semantic Module for high-level features F_{SEM}. F_{SEM} finally goes through two fully connected (FC) layers in Regression Module for gaze prediction. Reconstruction Module serves as an auxiliary module for the extraction of F_{RES}; (b) The structure of the Res_G block; (c) The structure of the Res_S block. Compared to the Res_G block, the Res_S block omits two Batch Normalization (BN) layers [43] and the second ReLU [44] function.

the super-resolution image; (d) Semantic Module for obtaining high-level semantic features, and (e) Regression Module for regressing the gaze direction. Among them, the Fundamental Module and the Residual Module form a complementary dual-branch. The input of the whole network is an RGB image, and its size is set to 128×128.

Fundamental Module. This module is used to extract fundamental features directly from the LR input. It is a part of the ResNet-18 [45] network that consists of a stem followed by four residual blocks (Res_G in Fig. 3(b)). The max-pooling layer is removed from the stem to maintain an appropriate size of feature maps at the output of this module. Since this module is shallow, the extracted features can be regarded as structural features. The size of the output feature maps (F_{FUN} in Fig. 3(a)) from this module is 32×32×128.

Residual Module. This module is used to extract residual features. It is constructed by an improved version of the residual block, namely the Res_S block in Fig. 3(c). The structure of the Res_S block is the same as the original one (Res_G) in ResNet, except that two Batch Normalization (BN) layers [43] and the second ReLU [44] function are omitted because they slow down the convergence speed of super-resolution [33], [46]. A convolution layer with stride 2, along with two Res_S blocks followed by a max-pooling layer, is first applied as a stem, which reduces the size of the feature maps by a factor of 4. Afterward, 16 Res_S blocks and 1 convolution layer are stacked to generate deeper and richer features, while a short skip connection is arranged to keep gradients for the previous layers. The size of the output feature maps (F_{RES} in Fig. 3(a)) from this module is 32×32×64.

Reconstruction Module. This module alternately connects three convolution layers and two Pixel-Shuffle layers [47] to upsample F_{RES} twice, resulting in an image ($\boldsymbol{I}_{REC}$ in Fig. 3(a)) of the same size as the LR input, i.e., 128×128. A long skip connection from the input image to the output of the Reconstruction Module is used to provide rough overall information directly. It forces the Residual Module and the Reconstruction Module to focus on finer details and guarantee these two modules only learn the residual information of the input. The reconstructed image ($\boldsymbol{I}_{SR}$ in Fig. 3(a)) is finally obtained after the skip connection. All convolution layers in the Residual Module and the Reconstruction Module use kernels of size 3×3, and the number of channels for all is set to 64, except the last convolution layer in the Reconstruction Module is set to 3. Note that the proposed network uses the high-resolution image corresponding to the LR input to supervise the learning process of $\boldsymbol{I}_{SR}$, and the Reconstruction Module can be removed after training to reduce storage and computational overhead since only the output of the Residual Module (F_{RES}) is needed for inference. Because super-resolution is a pixel-level task, features that flow in these two modules are regarded as low-level structural features.

Semantic Module. High-level semantic features are explored in this module. First, structural features from the dual branches (F_{FUN} and F_{RES}) are fed into a 1×1 convolution layer with 64 channels separately. Next, the feature fusion operation is performed in the channel dimension and generates feature maps of 128 channels. The combined features then go through four Res_G blocks to extract high-level semantic features. The Res_G blocks are set by the configuration of the last four residual blocks in ResNet-18. Finally, a global

average pooling layer (GAP) is applied to obtain a compact but representative 1D feature vector.

Regression Module. Fully connected (FC) layers are employed in this Module to regress the gaze vector. There are two fully connected layers, and the numbers of units are 512 and 1000, respectively. The output of this module is a 2D angular vector $\hat{\boldsymbol{g}}_{\boldsymbol{s}} = [\hat{g}_\phi, \hat{g}_\theta]$, where $\hat{g}_\phi$ represents the angle of yaw (horizontal gaze angle) and $\hat{g}_\theta$ represents the angle of pitch (vertical gaze angle) in the spherical coordinate system. Then they can be converted to the Cartesian coordinate system of the OpenGL standard via Eq. (1).

$$\hat{x} = cos(\hat{g}_\theta)sin(\hat{g}_\phi), \tag{1a}$$

$$\hat{y} = sin(\hat{g}_\theta), \tag{1b}$$

$$\hat{z} = cos(\hat{g}_\phi)cos(\hat{g}_\theta), \tag{1c}$$

where $\hat{\boldsymbol{g}}_{\boldsymbol{c}} = [\hat{x}, \hat{y}, \hat{z}] \in \mathbb{R}^3$ is the unit vector representation of predicted gaze. Assuming the actual gaze direction is $\boldsymbol{g} \in \mathbb{R}^3$, the evaluation metric of the angular error can be computed by Eq. (2).

$$e_{angle} = arccos\left(\frac{\boldsymbol{g} \cdot \hat{\boldsymbol{g}}_c}{\|\boldsymbol{g}\|\|\hat{\boldsymbol{g}}_c\|}\right) \tag{2}$$

We give a qualitative analysis of how residual information is restored in CDBN. Assuming that the HR image, the LR image, and the residual image (the absolute difference between the HR/LR images) are denoted as $\boldsymbol{I}_{HR}$, $\boldsymbol{I}_{LR}$, and $\boldsymbol{I}_{RES}$, respectively, we can describe the relationship among them by Eq. (3).

$$\boldsymbol{I}_{HR} = \boldsymbol{I}_{LR} + \boldsymbol{I}_{RES}. \tag{3}$$

Then in CDBN, we denote the output features of the Residual Module as F_{RES}, and the output image of the Reconstruction Module before skip connection as $\boldsymbol{I}_{REC}$. When provided LR image $\boldsymbol{I}_{LR}$ as input, the forward propagation in the Residual Module and the Reconstruction Module can be expressed by Eq. (4) and Eq. (5).

$$F_{RES} = H_{RES}(\boldsymbol{I}_{LR}), \tag{4}$$

$$\boldsymbol{I}_{REC} = H_{REC}(F_{RES}), \tag{5}$$

where $H_{RES}(\cdot)$ and $H_{REC}(\cdot)$ refer to the operations of the Residual Module and the Reconstruction Module, respectively. After skip connection, the final SR reconstructed image $\boldsymbol{I}_{SR}$ can be obtained by Eq. (6).

$$\boldsymbol{I}_{SR} = \boldsymbol{I}_{LR} + \boldsymbol{I}_{REC}. \tag{6}$$

Comparing Eq. (3) and Eq. (6), and considering that $\boldsymbol{I}_{SR}$ is reconstructed under the supervision of $\boldsymbol{I}_{HR}$, we can draw a conclusion that

$$\boldsymbol{I}_{REC} \approx \boldsymbol{I}_{RES}. \tag{7}$$

Eq. (7) illustrates that the output image before skip connection ($\boldsymbol{I}_{REC}$) is an estimation of the residual image ($\boldsymbol{I}_{RES}$). In other words, the information passing through the Residual Module and the Reconstruction Module is the estimated residual information as expected. The feature F_{RES}, as a mid-embedded representation of the two modules, characterizes the estimated residual information and participates in the subsequent gaze estimation task as a supplement.

Algorithm 1 CDBN Training Process in One Iteration

Input: High-/low-resolution image pair ($\boldsymbol{I}_{HR}$, $\boldsymbol{I}_{LR}$); gaze label ($\boldsymbol{g}$); learning rate (η); operations and parameters: Fundamental Module (H_{FUN}, θ_{FUN}), Residual Module (H_{RES}, θ_{RES}), Reconstruction Module (H_{REC}, θ_{REC}), Semantic Module (H_{SEM}, θ_{SEM}), Regression Module (H_{REG}, θ_{REG}).

Output: Updated parameters θ^*_{FUN},θ^*_{RES},θ^*_{REC},θ^*_{SEM} and θ^*_{REG};

1: Extract features from dual branches,
$F_{FUN} = H_{FUN}(\boldsymbol{I}_{LR})$, $F_{RES} = H_{RES}(\boldsymbol{I}_{LR})$;
2: Predict gaze direction,
$\hat{\boldsymbol{g}} = H_{REG}(H_{SEM}(F_{FUN}, F_{RES}))$;
3: Generate reconstructed SR image,
$\boldsymbol{I}_{SR} = H_{REC}(F_{RES}) + \boldsymbol{I}_{LR}$;
4: Calculate the loss L_{sr} and L_{gaze} via Eq. (8) and Eq. (9), respectively;
5: Update parameters of Regression Module, Semantic Module, and Reconstruction Module,
$\theta^*_{REG} \leftarrow \theta_{REG} - \eta\frac{\partial}{\partial\theta_{REG}}L_{gaze}$;
$\theta^*_{SEM} \leftarrow \theta_{SEM} - \eta\frac{\partial}{\partial\theta_{SEM}}L_{gaze}$;
$\theta^*_{REC} \leftarrow \theta_{REC} - \eta\frac{\partial}{\partial\theta_{REC}}L_{sr}$;
6: Update parameters of dual branches,
$\theta^*_{FUN} \leftarrow \theta_{FUN} - \eta\frac{\partial}{\partial\theta_{FUN}}L_{gaze}$;
$\theta^*_{RES} \leftarrow \theta_{RES} - \eta\frac{\partial}{\partial\theta_{RES}}(L_{gaze} + L_{sr})$.

C. Optimization

As mentioned in Section III-A, this work aims to exploit the maximum amount of information for low-resolution gaze estimation, including fundamental and residual information. It is done by simultaneously optimizing the gaze estimation and SR reconstruction tasks. For both tasks, the Mean Absolute Error (MAE) loss is minimized via Eq. (8) and Eq. (9).

$$L_{sr} = \frac{1}{N}\sum_{i=1}^{N}||\boldsymbol{I}^{(i)}_{SR} - \boldsymbol{I}^{(i)}_{HR}||_1, \tag{8}$$

$$L_{gaze} = \frac{1}{N}\sum_{i=1}^{N}||\hat{\boldsymbol{g}}^{(i)} - \boldsymbol{g}^{(i)}||_1, \tag{9}$$

where superscript i represents the i^{th} image; $\boldsymbol{I}^{(i)}_{SR}$ represents the restored super-resolution image and $\boldsymbol{I}^{(i)}_{HR}$ represents its HR counterpart; $\hat{\boldsymbol{g}}^{(i)}$ represents the estimated 2D gaze vector, and $\boldsymbol{g}^{(i)}$ represents the ground truth of gaze. The whole objective function L is a weighted average of L_{sr} and L_{gaze}, as shown in Eq. (10).

$$L = \lambda_1 L_{sr} + \lambda_2 L_{gaze} \tag{10}$$

where λ_1 and λ_2 are weight parameters that are set to 1 and 2 in our experiments.

The backpropagation process is represented by two dotted lines in Fig. 3(a). The parameters of the Fundamental Module are updated under the supervision of L_{gaze} to obtain the fundamental information from the LR input. At the same time, the training of the Residual Module is supervised by both losses, i.e., L_{sr} and L_{gaze}. The supervision of L_{sr} guarantees

the information extracted from this module is that for image restoration (i.e., residual information), and the supervision of L_{gaze} ensures the restored residual features tend to be beneficial for gaze estimation. The training process of one iteration is summarized in Algorithm 1 for better understanding.

IV. Experiments

A. Datasets and preprocessing

We evaluated the proposed approach on three popular gaze datasets: MPIIFaceGaze [17], EYEDIAP [18], and RT-Gene [19].

MPIIFaceGaze: This dataset is a full-face version of the MPIIGaze dataset. It was collected in real-world conditions with illumination and head pose variations. It contains 213,659 images in the size of 1280×720 that were captured from 15 subjects by webcams. It provides an evaluation subset, which contains 3000 samples randomly selected from each subject, for a total of 45000 samples. Same as other works [11, 13, 17, 25], experiments were carried out on this evaluation subset and performed a leave-one-person-out cross-validation.

EYEDIAP: This dataset contains a set of video clips of 16 subjects. The videos were collected under two visual target sessions, i.e., screen target and 3D floating ball. Same as the routine mentioned in [11, 14], only the screen target sessions were used for evaluation and were sampled per 15 frames from videos (640×480), starting to count from the first frame. Note that data from only 14 subjects can be used in the experiment due to the lack of videos in the screen target sessions for two subjects. We randomly divided the 14 participants into five groups and performed a 5-fold cross-validation.

RT-Gene: This dataset contains 122,531 facial images of 15 subjects with a resolution of 224×224. It was collected with wearable eye-tracking glasses to acquire accurate gaze annotations. In order to remove the eye-tracking glasses from captured images, the authors of the dataset used semantic inpainting to fill the masked regions of the eye-tracking glasses with skin texture [19]. As a result, the RT-Gene dataset provides another inpainted version of the images. We only used the original dataset for the experiment since there is too much noise in the inpainted set [12, 14]. We divided the original dataset into three subsets for 3-fold cross-validation according to the evaluation protocol provided by the dataset.

For the MPIIFaceGaze and EYEDIAP datasets, we followed the data processing method proposed in [17] to segment the facial region from the background and resized the cropped facial images to 128×128. These were designated as high-resolution (HR) images. To simulate low-resolution (LR) scenes in real-world applications, bicubic downsampling with factors of 2, 4, and 8 was applied to the HR images, and three series of LR images with resolutions of 64×64, 32×32, and 16×16 were obtained. As for the method which requires eye image input, eye images were cropped from HR images in fixed-size rectangles (35×21) centered around the landmark and were downsampled with the same scale factors. Finally, the LR images were reversed to match the size of the networks' input using bicubic interpolation. For the RT-Gene dataset, since the authors have already cropped eyes and faces, we did not do any further processing except for down-/up-samplings.

B. Implementation Details

The experiments in this work were conducted with the PyTorch platform. Cuda 10.1 was also employed to speed up model training.

In CDBN, besides the eight Res_G blocks initialized with ResNet-18 pre-trained on ImageNet, the parameters of the other parts were randomly initialized using the Kaiming initialization [49]. Optimization was done with the Adam algorithm [50]. The model was trained for 30 epochs with a batch size of 100. The learning rate was initialized as 0.0005 and reduced by a factor of 0.1 every ten epochs.

The following state-of-the-art methods were evaluated for comparison. Experimental settings in the corresponding publications, including model architectures and hyper-parameters, were used to implement and test their networks.

- **iTracker [26]:** A multi-region method that takes two eye images, the face image, and the face grid, as input to implement 2D gaze estimation. We simply changed it to 3D gaze estimation by training with 3D gaze labels.
- **RT-Gaze [19]:** RT-Gaze uses a two-stream VGG-like CNN to process two eye images and predict a gaze. The head pose vector is also appended to the FC layers to introduce head pose information.
- **FullFace [11]:** A deep neural network with a spatial weighting mechanism that takes the full facial image as input.
- **Dilated-Net [13]:** A three-stream CNN that takes both eye images and a face image as input. By replacing traditional convolutions with dilated convolutions in eye streams, small appearance changes can be captured effectively.
- **Gaze360 [48]:** Since our experiments were conducted with static images, we chose the static version of Gaze360 as the baseline model, which takes ResNet-18 as the backbone model while using Pinball loss.
- **I2D-Net [12]:** A three-stream CNN that is constructed with dilated convolution layers. In the eye stream, it employs a difference layer to eliminate common features from the left and right eyes of a participant that are not pertinent to gaze estimation.
- **GEDDNet [14]:** This is an improved version of Dilated-Net. It modifies the network structure on the basis of Dilated-Net and adds a subject-dependent bias to handle the appearance variation between subjects.

C. Experiment on the MPIIFaceGaze dataset

The performance comparisons of our proposed CDBN and other state-of-the-art methods on the MPIIFaceGaze dataset are shown in Table I. The gaze performance was evaluated by the mean angular error between the predicted gaze vector and the ground-truth gaze vector (see Eq.(2)). Each row shows the errors of different methods with the same resolution, and each column shows the errors of the same method with different resolutions.

As shown in Table I, the performance of all methods deteriorated as the input image size decreased. Our CDBN model had the slowest degradation trend among all methods. More

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2022.3210219

TABLE I
MEAN ANGULAR ERROR AND MODEL SIZE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE MPIIFACEGAZE DATASET.

Model / Image size	iTracker [26]	RT-Gaze [19]	Full Face [11]	Dilated-Net [13]	Gaze360 [48]	I2D-Net [12]	GEDDNet [14]	CDBN(ours)
128×128	6.13°	4.79°	4.86°	4.65°	**4.47°**	4.38°	4.59°	4.67°
64×64	6.25°	5.30°	4.89°	4.93°	4.71°	4.74°	4.78°	**4.68°**
32×32	6.70°	6.59°	5.48°	5.57°	5.34°	5.55°	5.44°	**5.13°**
16×16	8.84°	9.62°	7.25°	7.27°	7.13°	7.35°	7.18°	**6.89°**
#Params	6.29M	31.67M	196.6M	4.71M	11.95M	87.73M	4.05M	13.08M

TABLE II
MEAN ANGULAR ERROR COMPARISON WITH STATE-OF-THE-ART METHODS ON THE EYEDIAP DATASET.

Model / Image size	iTracker [26]	RT-Gaze [19]	Full Face [11]	Dilated-Net [13]	Gaze360 [48]	I2D-Net [12]	GEDDNet [14]	CDBN(ours)
128×128	6.53°	5.28°	5.97°	5.46°	5.59°	5.30°	5.32°	**5.13°**
64×64	6.65°	5.34°	5.99°	5.50°	5.68°	5.35°	5.42°	**5.21°**
32×32	7.26°	5.53°	6.13°	5.71°	5.84°	5.60°	5.66°	**5.43°**
16×16	8.20°	7.18°	7.42°	6.69°	6.74°	6.64°	6.65°	**6.56°**

TABLE III
MEAN ANGULAR ERROR COMPARISON WITH STATE-OF-THE-ART METHODS ON THE RT-GENE DATASET.

Model / Image size	RT-Gaze [19]	Full Face [11]	Dilated-Net [13]	Gaze360 [48]	I2D-Net [12]	GEDDNet [14]	CDBN(ours)
128×128	7.50°	8.42°	8.40°	6.93°	8.03°	8.17°	**6.67°**
64×64	7.86°	8.51°	8.47°	7.03°	8.13°	8.19°	**6.69°**
32×32	8.81°	8.83°	8.76°	7.49°	8.29°	8.55°	**7.08°**
16×16	12.49°	11.52°	10.53°	10.66°	10.45°	10.39°	**10.16°**

specifically, when the image size was 128×128, CDBN had a comparable performance compared to other state-of-the-art methods. When the image size was 64×64, the angular error of CDBN was slightly better than that of competitive methods. When the image size was 32×32, the performance of other methods worsened rapidly due to extensive loss of information of the input image. CDBN caught facial prior knowledge by employing the residual branch, used it to reconstruct part of the gaze features, and thus maintained superior performance. Compared to other methods, CDBN obtained an angular error of 5.13° , which was 3.9%, 5.7%, 6.4%, 7.6%, and 7.9% lower than Gaze360, GEDDNet, FullFace, I2D-Net, and Dilated-Net, respectively. When the image size was reduced to 16×16, CDBN was still in the lead and obtained an angular error of 6.89°, which was 3.4%, 4.0%, 5.0%, 6.3%, and 5.2% lower than those five methods in the same order.

We also list the size of each model in the bottom row of Table I. The model size of CDBN is smaller than FullFace, I2D-Net, and RT-Gaze, and is close to the rest of models. CDBN achieved excellent performance with a reasonable model size.

D. Experiment on the EYEDIAP and the RT-GENE datasets

To further evaluate the performance of CDBN and other state-of-the-art methods, we conducted experiments on two other popular datasets, i.e., EYEDIAP and RT-GENE. The results are shown in Table II and Table III.

On the EYEDIAP dataset, the proposed CDBN obtained competitive results. It obtained the smallest angular error for all four image resolutions, and the performance improvement was the most significant for the image size of 64×64 and 32×32. When the image size was 64×64, CDBN obtained the smallest angular error of 5.21°, which was 2.6%, 3.9%, 5.3%, 8.3%, and 13.0% lower than I2D-Net, GEDDNet, Dilated-Net, Gaze360, and FullFace, respectively. When the image size was reduced to 32×32, CDBN obtained the smallest angular error of 5.43°, which was 3.0%, 4.1%, 4.9%, 7.0%, and 11.4% lower than those five methods in the same order. The experimental result shows that our proposed method effectively alleviated performance degradation as image size decreased.

The RT-GENE dataset is more challenging than the other two because it has larger head-pose and gaze-angle variation. Furthermore, the occlusion of eye-tracking glasses makes the network harder to learn valuable features for gaze estimation. As reported in Table III, CDBN had superior performance on RT-GENE compared to other models for the image size of 128×128, 64×64, and 32×32. In particular, the performance improvement of CDBN was most prominent for 32×32. CDBN obtained an angular error of 7.08°, which was 5.5%, 14.6%, 17.2%, 19.2%, and 19.8% lower than Gaze360, I2D-Net, GEDDNet, Dilated-Net, and FullFace, respectively. When

TABLE IV
MEAN ANGULAR ERROR COMPARISON WITH DIFFERENT CONFIGURATIONS OF CDBN ON THE MPIIFACEGAZE DATASET.

Model \ Image size	64×64	32×32	16×16
Fundamental branch	4.75°	5.29°	7.08°
Residual branch	4.92°	5.35°	7.21°
Dual branch	4.68°	5.13°	6.89°

TABLE V
MEAN ANGULAR ERROR COMPARISON WITH DIFFERENT CONFIGURATIONS OF CDBN ON THE EYEDIAP DATASET.

Model \ Image size	64×64	32×32	16×16
Fundamental branch	5.19°	5.44°	6.69°
Residual branch	5.87°	5.95°	6.75°
Dual branch	5.21°	5.43°	6.56°

TABLE VI
MEAN ANGULAR ERROR COMPARISON WITH DIFFERENT CONFIGURATIONS OF CDBN ON THE RT-GENE DATASET.

Model \ Image size	64×64	32×32	16×16
Fundamental branch	6.83°	7.29°	10.42°
Residual branch	8.06°	8.32°	11.20°
Dual branch	6.69°	7.08°	10.16°

the image size was reduced to 16×16, the performance of CDBN took a hit but was still the most accurate. It obtained an angular error of 10.16°, with a performance improvement ranging from 2.2% (lower than GEDDNet) to 18.7% (lower than RT-Gaze).

E. Ablation Study

In this section, ablation experiments were performed by removing one of the two branches to investigate the contribution of each branch in CDBN. The results are shown in Table IV to Table VI. In the fundamental branch model, we only used the Fundamental Module to extract features. In the residual branch model, we only used the Residual Module to extract features. We employed both modules in the dual-branch model.

Table IV shows that the dual-branch model had the best performance on MPIIFaceGaze, while the fundamental-branch model was the second and the residual-branch model was the last. It demonstrates that the deletion of either branch would negatively impact the performance of CDBN at low resolutions. The better performance of the fundamental-branch model than the residual-branch model indicates that features extracted by the fundamental branch contained more information and played a dominant role in gaze estimation.

A similar conclusion can be made on the other two datasets. As shown in Table V, When the image size was 64×64, the fundamental-branch model obtained comparable performance to the dual-branch model on EYEDIAP, while the residual-branch model performed worse. When the image size was reduced to 32×32 or 16×16, the dual-branch model began to show its efficiency and achieved better results compared to the single-branch model. Table VI also shows that the dual-branch model performed better than the single-branch model at three low resolutions on RT-GENE. It is worth mentioning that the performance improvement of the dual-branch model on EYEDIAP was less than that on the other two datasets, proving that the dual-branch model performed better with more data support.

TABLE VII
THE NUMBER OF PARAMETERS AND RUNTIMES OF FOUR POPULAR GAZE ESTIMATION NETWORKS AND CDBN.

Model	#Params	Images / sec.	
		GPU	CPU
Gaze360 [48]	11.95M	214.8	21.0
iTracker [26]	6.29M	**317.2**	**31.2**
Dilated-Net [13]	4.71M	147.4	16.4
GEDDNet [14]	4.05M	156.3	18.5
CDBN(ours)	13.08M	117.8	13.9

F. Visualization

In order to intuitively understand the effect of CDBN, some visualization results are shown in Fig. 4 to Fig. 6 in this section.

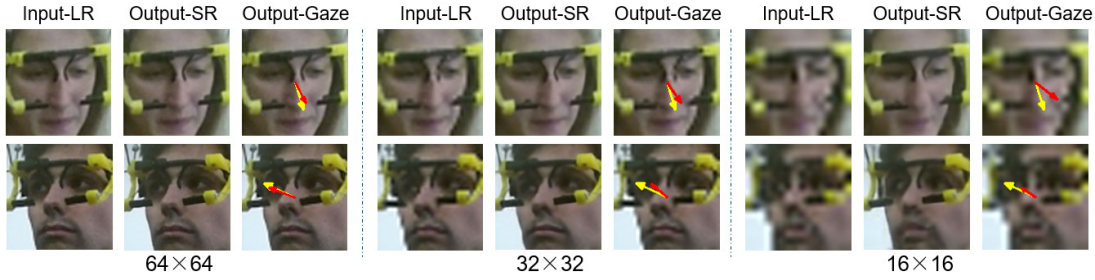
Fig. 4 visualizes the input LR images, the reconstructed SR images, and the images with the predicted gaze directions on the three employed datasets. Comparison between the input LR images and the reconstructed SR images shows that facial details, such as accurate facial component shapes and textures and especially eyes, were recovered successfully by the residual branch. These facial details, or residual information, serve as complementary information to the LR input and significantly contribute to low-resolution gaze estimation. As shown in the images with the predicted gaze directions, with the assistance of residual information, our gaze estimation method obtained promising results even if the input image is blurred. Fig. 5 illustrates the visualized comparison of predicted gaze directions when the input size was 32×32. It shows that CDBN obtained more accurate predicted gaze directions compared to other competing methods.

We further visualized the output feature maps of the dual-branch in Fig. 6. Fig. 6(a) depicts the selected sample of an original HR image from the MPIIFaceGaze dataset, and Fig. 6(b) depicts the corresponding LR image at the resolution of 16×16. The LR image was fed into CDBN to extract the fundamental and residual features. As shown in Fig. 6(c), the Fundamental Module paid more attention to global representation. Features from this module described the appearance and contour of the face from a global perspective, containing information about the flat regions in a facial image. In contrast, Fig. 6(d) shows that the Residual Module concentrated more on edge contents. Features from this module described the edges of facial components, containing information about the sudden changes in a facial image, such as the edges of the eyeballs in (2) and (12) of Fig. 6(d). The two types of features

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2022.3210219

(a) Visualization of the MPIIFaceGaze dataset

(b) Visualization of the EYEDIAP dataset



(c) Visualization of the RT-GENE dataset

Fig. 4. Visualization of the input and output of CDBN on the three employed datasets. For each specific input resolution, the leftmost image is the LR input, the middle image is the reconstructed SR image, and the rightmost image shows the predicted gaze drawn on the LR input. The yellow arrow indicates the ground truth gaze and the red arrow indicates the estimated gaze. All images are resized to 128×128 for display.

Fig. 5. Comparison of predicted gaze directions for the input size of 32×32. Three examples, one from each dataset, are included. The yellow arrow indicates the ground truth gaze and the red arrow indicates the estimated gaze.

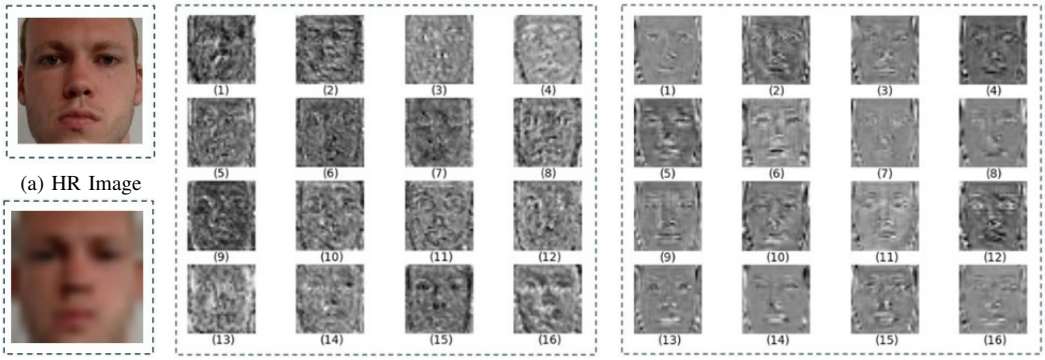
complement each other and positively impact low-resolution gaze estimation.

G. Processing Speed

To further investigate the feasibility of the proposed method in real-world intelligent systems, we conducted an experiment to test the processing speed of the proposed network. Four state-of-the-art networks with a similar number of parameters were compared with ours on both GPU and CPU platforms. The GPU platform employed a single NVIDIA TITAN Xp. The version of CUDA library was 10.1. The CPU platform used an Intel Core i7-6700 @ 3.40 GHz processor.

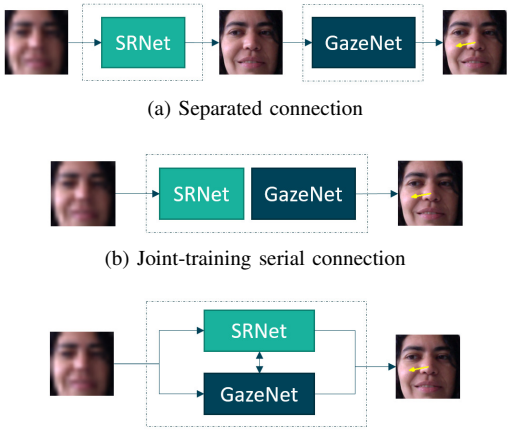
Each network was evaluated on the same target platform ten times to get reliable experimental results. The average runtime of each network is shown in Table VII. It shows that the actual inference speed of other competitive networks was faster than CDBN since CDBN has two branches to perform. However, this dual-branch design captures more helpful information

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2022.3210219



(a) HR Image (b) LR Image (c) Fundamental Structural Feature Maps (d) Residual Structural Feature Maps

Fig. 6. Visualization of dual-branch feature maps. Only the first 16 feature maps are shown. (a) A sample of original HR image; (b) The corresponding input LR image with the resolution of 16×16; (c) Feature maps extracted from fundamental branch; (d) Feature maps extracted from residual branch.



(a) Separated connection

(b) Joint-training serial connection

(c) Joint-training parallel connection

Fig. 7. Three topologies of super-resolution network (SRNet) and gaze network (GazeNet).(a) The topology of the separated connection model; (b) The topology of the joint-training serial connection model; (c) The topology of the joint-training parallel connection model.

and provides superior performance. Compared to iTracker, the performance improvement of CDBN was up to 20% for the image size of 32×32 on MPIIFaceGaze (5.13° vs. 6.70° or a 23.4% improvement) and EYEDIAP (5.43° vs. 7.26° or a 25.2% improvement). Compared to Dilated-Net and GEDDNet, the performance improvement of CDBN for the image size of 32×32 on MPIIFaceGaze was 7.9% (5.13° vs. 5.57°) and 5.7% (5.13° vs. 5.44°), respectively.

H. Network Topology

Since there are many ways to combine the SR task with the gaze estimation task, this section aims to study the impact of different network topologies. For the convenience of description, the combination of the Residual Module and the Reconstruction Module is defined as SRNet, and the combination of other modules is defined as GazeNet. Fig. 7 shows three potential topologies of these two nets.

In the topology shown in Fig. 7(a), the super-resolution process is performed with SRNet first, then the gaze is estimated from the reconstructed SR image with GazeNet. SRNet and GazeNet are optimized separately. The topology of Fig. 7(b) connects SRNet and GazeNet in series, and both nets

TABLE VIII
MEAN ANGULAR ERROR COMPARISON OF THREE NETWORK TOPOLOGIES ON THE MPIIFACEGAZE DATASET.

Model \ Image size	64×64	32×32	16×16
Separated connection	4.78°	5.60°	7.99°
Joint-training serial connection	4.64°	5.30°	7.03°
Joint-training parallel connection	4.68°	5.13°	6.89°

are trained together in an end-to-end manner. Our proposed method is depicted in Fig. 7(c), where two networks are connected in parallel. The HR image and the gaze label are used to supervise the training phase in all three topologies.

We conducted an experiment on the MPIIFaceGaze dataset with the same experimental configuration mentioned in Section IV-B. The experimental results are listed in Table VIII. When the image size was 64×64, the three topologies had comparable accuracy. When the image size was reduced to 32×32 or 16×16, there were apparent differences in the performance of the three topologies. More specifically, the separated connection model (Fig. 7(a)) obtained the worst performance among the three models. The reason was that the separately optimized strategy made the optimization direction inconsistent for both tasks. The joint-training serial connection model (Fig. 7(b)) performed better than the separate connection model, demonstrating the importance of joint tuning. Our proposed method (Fig. 7(c)) obtained the best performance as it employed a parallel connection of the SRNet and GazeNet, where the gaze label supervised the super-resolution process more directly than the other two models. It helped reconstruct more efficient details biased towards the gaze estimation task.

V. CONCLUSION

In this paper, we evaluate existing gaze estimation methods and find their disadvantages in dealing with low-resolution images. We then propose a novel network named CDBN to address this low-resolution challenge. The proposed CDBN employs a dual-branch design to alleviate the lack of information in the LR image and improve the estimation accuracy of gaze direction. In CDBN, the fundamental branch directly extracts

features from the LR image. The residual branch recovers the residual features as a supplement for gaze estimation based on the facial prior knowledge. The optimization of the residual branch is supervised together by pairs of LR/HR images and gaze labels, guaranteeing that the recovered residual features are beneficial to the gaze estimation. Two features are then concatenated and mapped into a high-level semantic feature space in the subsequent module. Gaze regression is finally performed on the semantic features.

We evaluated our method on three widely used gaze datasets, MPIIFaceGaze, EYEDIAP, and RT-GENE, at three different low resolutions. The experimental results demonstrated that our method achieves a lower angular error with a low-resolution input compared to other state-of-the-art approaches. The ablation study and network topology results further prove the effectiveness of the dual-branch design. We also tested our network in terms of inference speed. We believe the inference speed of our proposed method can be further improved with some optimization techniques, which is part of our future work.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (62173353, 62171207), Guangzhou Municipal People's Livelihood Science and Technology Plan (201903010040), Science and Technology Program of Guangzhou, China (202007030011).

REFERENCES

[1] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn, "Towards foveated rendering for gaze-tracked virtual reality," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–12, Nov. 2016.

[2] G. A. Koulieris, K. Akşit, M. Stengel, R. K. Mantiuk, K. Mania, and C. Richardt, "Near-eye display and tracking technologies for virtual and augmented reality," *Computer Graphics Forum*, vol. 38, no. 2, pp. 493–519, May 2019.

[3] Z. Yu, X. Huang, X. Zhang, H. Shen, Q. Li, W. Deng, J. Tang, Y. Yang, and J. Ye, "A multi-modal approach for driver gaze prediction to remove identity bias," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 768–776.

[4] Q. Cheng, D. Agrafiotis, A. M. Achim, and D. R. Bull, "Gaze location prediction for broadcast football video," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4918–4929, Dec. 2013.

[5] P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction," in *Advances in Physiological Computing*. Springer, London, 2014, pp. 39–65.

[6] J. Guo, Y. Liu, Q. Qiu, J. Huang, C. Liu, Z. Cao, and Y. Chen, "A novel robotic guidance system with eye-gaze tracking control for needle-based interventions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 1, pp. 179–188, Mar. 2021.

[7] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 83–90.

[8] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[9] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," *arXiv preprint arXiv:2104.12668*, 2021.

[10] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 623–10 630.

[11] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 2299–2308.

[12] M. L R D and P. Biswas, "Appearance-based gaze estimation using attention and difference mechanism," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 3137–3146.

[13] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Asian Conference on Computer Vision(ACCV)*. Springer, Cham, 2018, pp. 309–324.

[14] Z. Chen and B. Shi, "Towards high performance low complexity calibration in appearance based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access)*, pp. 1–1, 2022.

[15] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y. K. Wong, "Learning spatial attention for face super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 1219–1231, 2021.

[16] Y. Chen, V. Phonevilay, J. Tao, X. Chen, R. Xia, Q. Zhang, K. Yang, J. Xiong, and J. Xie, "The face image super-resolution algorithm based on combined representation learning," *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30 839–30 861, Nov. 2021.

[17] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 162–175, Jan. 2019.

[18] 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 *Proceedings of the Symposium on Eye Tracking Research and Applications(ETRA)*, 2014, pp. 255–258.

[19] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 334–352.

[20] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1124–1133, Jun. 2006.

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

[22] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1821–1828.

[23] T. Schneider, B. Schauerte, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation," in *2014 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 1167–1172.

[24] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2033–2046, Oct. 2014.

[25] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4511–4520.

[26] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2176–2184.

[27] L. Dai, J. Liu, Z. Ju, and Y. Gao, "Attention mechanism based real time gaze tracking in natural scenes with residual blocks," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 696–707, 2022.

[28] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Transactions on Image Processing*, vol. 29, pp. 5259–5272, Mar. 2020.

[29] L. Dai, J. Liu, and Z. Ju, "Binocular feature fusion and spatial attention mechanism based gaze tracking," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 2, pp. 302–311, 2022.

[30] J. Zhang, J. Sun, J. Wang, Z. Li, and X. Chen, "An object tracking framework with recapture based on correlation filters and siamese networks," *Computers & Electrical Engineering*, vol. 98, p. 107730, Mar. 2022.

[31] R. Xia, Y. Chen, and B. Ren, "Improved anti-occlusion object tracking algorithm using unscented rauch-tung-striebel smoother and kernel correlation filter," *Journal of King Saud University-Computer and Information Sciences*, 2022.

[32] J. Zhang, W. Feng, T. Yuan, J. Wang, and A. K. Sangaiah, "Scstcf: spatial-channel selection and temporal regularized correlation filters for visual tracking," *Applied Soft Computing*, vol. 118, p. 108485, Mar. 2022.

[33] Y. Yin, J. Robinson, Y. Zhang, and Y. Fu, "Joint super-resolution and alignment of tiny faces," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 693–12 700.

[34] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3773–3782.

[35] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 21–30.

[36] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 206–221.

[37] J. Shao and Q. Cheng, "E-fcnn for tiny facial expression recognition," *Applied Intelligence*, vol. 51, no. 1, pp. 549–559, Jan. 2021.

[38] Z. Zhang, L. Wan, W. Xu, and S. Wang, "Estimating a 2d pose from a tiny person image with super-resolution reconstruction," *Computers & Electrical Engineering*, vol. 93, p. 107192, Jul. 2021.

[39] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," in *International Conference on Neural Information Processing*. Springer, cham, 2021, pp. 387–395.

[40] D. Cai, K. Chen, Y. Qian, and J.-K. Kämäräinen, "Convolutional low-resolution fine-grained classification," *Pattern Recognition Letters*, vol. 119, no. 1, pp. 166–171, Mar. 2019.

[41] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[42] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[44] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[46] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.

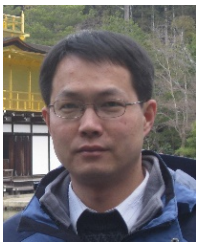
[47] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1874–1883.

[48] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6911–6920.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

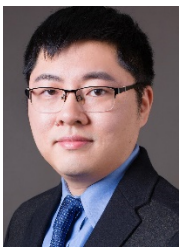
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015, pp. 1–13.

Zhesi Zhu received his B.S. degree from Sun Yat-sen University, China, in 2020. He is currently a postgraduate student in the school of Electronics and Information Technology, Sun Yat-sen University. His research interests include gaze estimation and computer vision.



Dong Zhang received his B.S.E.E. and M. S. degrees from Nanjing University, China, in 1999 and 2003, respectively, and Ph.D. degree from Sun Yat-sen University, China, in 2009. He is currently an associate professor in the school of Electronics and Information Technology, Sun Yat-sen University. His research interests include image processing, pattern recognition, and information hiding.

Cailong Chi received his B.S. degree from Shenzhen University, China, in 2020. He is currently a postgraduate student in the school of Electronics and Information Technology, Sun Yat-sen University. His research interests include human pose estimation and computer vision.



Ming Li received his Ph.D. in Electrical Engineering from University of Southern California in May 2013. He is currently an associate professor of the Data Science Research Center at Duke Kunshan University, a research scholar at the ECE department of Duke University, and the adjunct professor at Wuhan University. His research interests are in the areas of speech processing and multimodal behavior signal analysis with applications to human centered behavioral informatics notably in health, education and security.

Dah-Jye Lee received his B.S. degree from National Taiwan University of Science and Technology in 1984, M.S. and Ph.D. degrees in electrical engineering from Texas Tech University in 1987 and 1990, respectively. He also received his MBA degree from Shenandoah University, Winchester, Virginia in 1999. He worked in the machine vision industry for eleven years prior to joining BYU in 2001. He is currently a Professor in the Department of Electrical and Computer Engineering at Brigham Young University. His research work focuses on artificial intelligence, robotic vision, high-performance visual computing, and visual inspection automation.