IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 May 2024, accepted 25 July 2024, date of publication 29 July 2024, date of current version 7 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435370

RESEARCH ARTICLE

GAFUSE-Net: A Low Resolution Gaze Estimation Method Based on the Enhancement of Gaze-Relevant Features

JIN WANG, (Member, IEEE), KE WANG, AND SHUOYU CAO
School of Artificial Intelligence and Computer Science, Nantong University, Nantong 226019, China

Corresponding author: Jin Wang (wj@ntu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China for Young Scholar under Grant 62002179, and in part by the Program for Nantong Basic Science Research Plan under Grant JC22022061.

ABSTRACT Gaze estimation significantly enhances user interfaces, road safety, accessibility for the disabled, and consumer behavior analysis. Traditional methods depend on high-resolution images and ideal conditions, neglecting low-resolution scenarios. We propose GAFUSE-Net, designed for low-resolution gaze estimation. GAFUSE-Net includes a novel Facial Attention-Enhanced feature extraction (FAE) module and a Local image Super-resolution based Eye feature extraction (LSE) module for low resolution gaze estimation. The FAE module extracts gaze-relevant features from facial images. The LSE module uses super-resolution to enhance low-resolution eye images, followed by feature extraction with DeepEyeNet. Combined facial and eye features enable accurate gaze estimation. Experiments on the MPIIFaceGaze dataset show GAFUSE-Net's effectiveness, achieving mean angular errors of 4.37°, 4.62°, 4.84°, and 6.19° at resolutions of 128×128, 64×64, 32×32, and 16×16, respectively. These errors are 0.2%, 1.3%, 5.7%, and 10.2% lower than state-of-the-art methods. The experimental results highlight the efficiency of GAFUSE-Net across various resolutions, significantly advancing the practical implementation of gaze estimation.

INDEX TERMS Gaze estimation, low resolution, feature fusion, super-resolution, gaze-relevant.

I. INTRODUCTION

Gaze estimation is a critical subfield in the realm of computer vision, focusing on determining the direction or target observed by an individual's eyes. Gaze is an integral component of human social interaction and can reveal a substantial amount of implicit information by deciphering the focal point of gaze. For example, shopping malls can analyze the most popular products based on customer gaze data [1], invigilators can detect cheating in students by observing their gaze direction, and mobile applications can deliver targeted advertisements by tracking a user's gaze location. Additionally, gaze estimation technology has been widely applied in various domains, including virtual reality [2], driving assistance systems [3], and human-machine interaction [4], among others.

In recent years, with the continuous development of deep learning, convolutional neural network (CNN)-based methods have gradually become the mainstream approaches for gaze estimation. However, such methods typically require large datasets for training and often consider only scenarios under ideal conditions. The facial images used for training are usually high-resolution (HR) images, which contain abundant information. In real-world scenarios, due to factors such as low camera resolution and the increased distance between the camera and the face, the input facial images often lack clarity. When input images are of low resolution (LR), the accuracy of gaze estimation using these methods declines significantly as the resolution decreases.

What distinguishes our study from prior research is the consideration of distinct image resolutions, in contrast to their assumption that all input images are high resolution. We simulate various resolution conditions using four distinct sizes. Fig 1 illustrates facial and ocular images under different resolutions. As image resolution decreases, a discernible

The associate editor coordinating the review of this manuscript and approving it for publication was Songwen Pei.

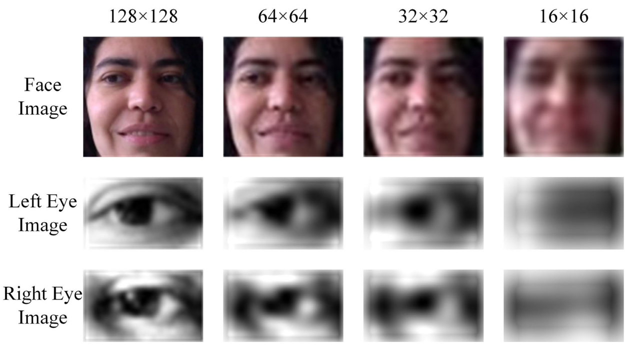

FIGURE 1. Facial and ocular images under different resolutions.

degradation in information occurs. This erosion of detail complicates feature extraction, significantly challenging gaze estimation tasks.

Super-resolution (SR) is an image processing technique aimed at reconstructing a high-resolution image from one or multiple low-resolution images. Super-resolution has also been proven to enhance image quality visually [5]. It is extremely beneficial in numerous applications, such as medical imaging [6], remote sensing [7], and video processing [8]. With the advancement of deep learning technology, particularly the successful application of convolutional neural networks (CNNs) in image processing, modern super-resolution methods often employ deep learning models like SRCNN [9], ESPCN [10], FSRCNN [11], and EDSR [12]. These models learn the intricate mapping relationships between low- and high-resolution images. Super-resolution has demonstrated favorable performance in various scenarios. Some works [13], [30] have combined super-resolution with gaze estimation, validating its effectiveness in enhancing gaze estimation. However, these methods perform super-resolution on complete facial images, which is time-consuming and costly, with limited improvements in gaze estimation accuracy in low-resolution environments. Traditional gaze estimation methods still face the following challenges: (1) existing feature extraction networks experience considerable information loss when extracting features from low-resolution images, making it challenging to obtain valuable feature information; (2) in low-resolution environments, a substantial amount of gaze-related features are lost in current gaze estimation techniques, leading to a marked decrease in estimation accuracy.

This paper presents a novel model, named GAFUSE-Net, to address the challenges of gaze estimation in low resolution scenarios, and the framework of the model is illustrated in the Fig 2. Our approach achieves superior results beyond those of previous works. In our work, A tri-stream network is designed for gaze estimation. The facial branch uses Facial Attention-Enhanced feature extraction (FAE) module to extract the gaze related features of the entire face. The left and right eye branches pass through Local image Super-resolution based Eye feature extraction (LSE) module to restore the low resolution features of the eyes and then extract features of both eyes. To assess the performance of this proposed GAFUSE-Net, this paper conduct experiments in multiple resolution scenarios of the classic dataset MPIIFaceGaze [14]. The results demonstrate that GAFUSE-Net achieve smaller angular errors on images of various resolutions, yielding state-of-the-art results.

The main contributions of our work are summarized as follows:

1. We propose the FAE module, which effectively amplifies the focus on gaze related regions and accomplish the extraction of key features, thereby enhancing gaze estimation performance.
2. We propose the LSE module, designed to rapidly and efficiently recover gaze related features in low resolution eye images, addressing the significant decrease in gaze estimation accuracy at low resolutions.
3. Leveraging the proposed functionalities, We design a tri-stream network named GAFUSE-Net, we conduct extensive experimentation across various resolutions on the MPIIFaceGaze dataset, achieving state-of-the-art performance, The improvements are especially pronounced under extreme LR conditions.

II. RELATED WORK

Research on human gaze is significant for a multitude of reasons and has widespread implications across various research domains and practical applications [15], [16], [17]. This paper will review related work on gaze estimation, with a particular focus on studies concerning low-resolution gaze estimation, which is the primary subject of this research.

A. GAZE ESTIMATION

Currently, gaze estimation techniques can be broadly categorized into model-based methods and appearance-based methods. Model-based gaze estimation primarily involves constructing a geometric model of the eye, integrating facial landmarks, depth information of the pupil center, and other geometric relations to accurately estimate the gaze direction [18], [19]. Unlike model-based methods, appearance-based gaze estimation often does not rely on the detection of eye feature points but directly infers gaze direction from the optical appearance of the eyes [20], [21], [22], [23]. Zhang et al. [14] were among the first to employ convolutional neural networks for gaze estimation. Krafka et al. [24] harnessed convolutional neural networks to estimate gaze points on smartphone screens. Zhang et al. [22] used full-face images as inputs, estimating gaze direction by considering weightings across different facial regions. Fischer et al. [23] proposed a new gaze estimation network to better handle diverse images in a new dataset. Chen and Shi [25] suggested an appearance-based gaze estimation approach using dilated convolution, which allows for the extraction of advanced features without compromising spatial resolution. In a subsequent work, Chen and Shi [26] refined this method. Kellnhofer et al. [27] put forward a gaze

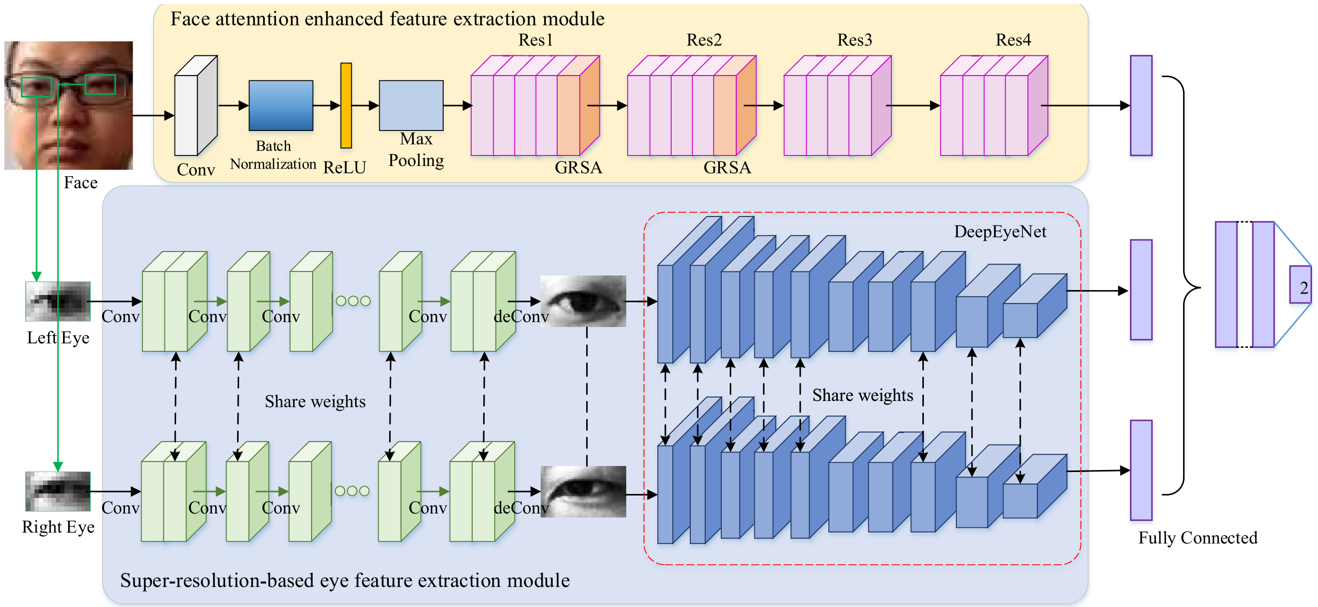

FIGURE 2. **The overall network architecture. A tri-stream network is designed for gaze estimation. The facial branch uses FAE module to capture gaze related features of the entire face. The left and right eye branches navigate through LSE module, they first traverse a local image super-resolution network, primarily aimed at restoring the low resolution features of the eyes. Subsequently, features of both eyes are extracted, using the deep eye feature extraction module DeepEyeNet. Notably, the branches corresponding to the left and right eyes share weight information. The outputs from all three branches are concatenated to produce the final gaze estimation.**

estimation method utilizing spatio-temporal LSTM, adopting a novel loss function for estimating differential quantiles. Murthy and Biswas [28] presented two strategies: the first uses a differential layer to remove left and right eye features unrelated to the gaze estimation task, while the second approach uses an attention mechanism to estimate gaze direction.

B. LOW RESOLUTION GAZE ESTIMATION

While current gaze estimation methods based on CNNs have witnessed substantial advancement, most networks predominantly rely on high-quality image inputs. When presented with low resolution images, the accuracy of gaze estimation tends to degrade substantially. Recent research has started addressing gaze estimation in the context of low resolution scenarios. Nonaka et al. [29] proposed a gaze estimation method based on the coordination of time, eyes, head, and body. They introduced a Bayesian framework that utilizes the learned angular-temporal dependencies between the orientation of the head and body and the direction of gaze to predict the gaze direction. Yun et al. [30] employed a super-resolution module based on high-frequency attention blocks to improve the performance of low resolution gaze estimation. Zhu et al. [13] proposed a complementary dual-branch network that extracts principal structural features from low resolution images and uses a residual branch to reconstruct features containing residual information.

Currently, most gaze estimation research overlooks low-resolution scenarios. Even in the few studies addressing low-resolution gaze estimation, performance significantly declines at extremely low resolutions. Our work introduces a triple-stream network that enhances gaze-related features in facial and eye images under low resolution. This approach noticeably improves gaze estimation accuracy, demonstrating robustness in challenging conditions.

III. A LOW RESOLUTION GAZE ESTIMATION METHOD BASED ON THE ENHANCEMENT OF GAZE-RELEVANT FEATURES

This section delineates the proposed Local image super-resolution Gaze Estimation Network, named GAFUSE-Net, with its overall architecture illustrated in Fig 2. GAFUSE-Net adopts a tri-stream structure, accepting three distinct image inputs: a facial image, a left-eye image, and a right-eye image. Specifically, the facial image undergoes the FAE module to extract gaze related features. This enhances low resolution global features from both spatial and channel dimensions. The branches for the left-eye and right-eye images are analogous, they navigate through LSE module. Initially, the local low resolution eye images are processed through the super-resolution network FSRCNN, which is employed for precise upscaling of these images to high resolution, thereby fortifying the local eye images to augment gaze estimation efficacy under low resolution conditions. This is followed by the innovative DeepEyeNet module, which we present to extract local features from eye images. Finally, the features through a fully connected layer, the final gaze estimation reslut is deduced.

A. FACIAL ATTENTION-ENHANCED FEATURE EXTRACTION MODULE

Within gaze estimation, significant CNN feature extraction models encompass AlexNet [31], VGGNet [32], and

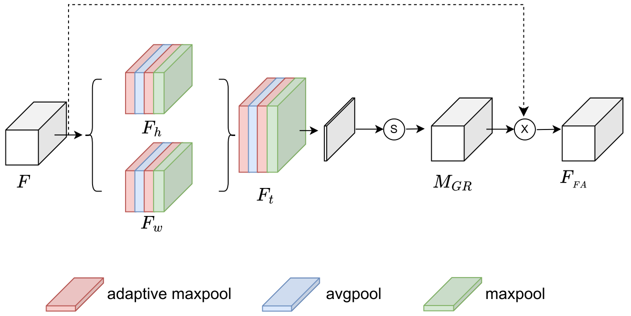

FIGURE 3. Structure of the GRSA module. The F_h and F_w represent operations applied to the height and width dimensions of F, respectively. The F_t is the integrated spatial feature. The spatial attention map M_{GR} is applied to F through a weighted multiplication, resulting in the final feature map F_{FA}, which highlights important spatial regions.

ResNet [33]. ResNet pioneers the residual learning approach, facilitating the training of even deeper networks than its antecedents. The ResNet effectively tackles issues like gradient explosion and vanishing gradient, amplifying its adaptability. For this study's facial feature extraction, we utilize the ResNet18 model, a variant in the ResNet series, deemed adequate for extracting gaze-centric facial characteristics.

Introduced by He et al. [33], Residual Neural Networks center on the novel architecture of residual blocks. These blocks tackle the persistent challenges of gradient vanishing and gradient explosion prevalent in deep neural network training. The core concept behind the residual block posits that discerning the residual between an output and its related input is potentially more direct than understanding the outright output. The standard residual block's mathematical formulation is presented as follow:

$$F_{\text{out}} = \ell\left(f\left(F_{\text{in}}, W_i\right) + F_{\text{in}}\right) \tag{1}$$

where ℓ represents the ReLU activation function, f denotes the weight operation within the residual block, W signifies the weights associated with this operation, and F_{in} is the input to the residual block. Meanwhile, F_{out} corresponds to the output of the residual block.

We improve ResNet18 to enhance its feature extraction capability. Attention mechanisms, grounded in channel and space, are integrated at the conclusion of each residual stage to enhance gaze estimation accuracy. The schematic of the facial feature extraction network is depicted in the upper branch of Fig 2. Here, Res1, Res2, Res3, and Res4 represent the four residual stages, all exhibiting consistent structures.

When the resolution of input facial images is low, useful information can be lost, and traditional ResNet feature extraction networks perform poorly in extracting features from low resolution images. The areas in facial images that play a crucial role in gaze estimation are referred to as gaze related areas. By focusing on the features of gaze related areas, the accuracy of gaze estimation models can be improved. To address this, we design the Gaze Related region Spatial Attention (GRSA) module, designed specifically to amplify the focus on gaze related regions and accomplish the extraction of key features. Fig 3 shows the main structure of this module. We embed the GRSA module following the initial two stages of the ResNet18 architecture, to augment the capability of the model in extracting low resolution facial features. Compared to conventional spatial attention, the GRSA module processes features in both the height and width dimensions separately, enabling a more comprehensive understanding of the spatial distribution of input features. Moreover, it employs adaptive average pooling to capture the global context information of feature maps, allowing the model to accurately capture the spatial structure information of images even when operating on low resolution images devoid of detailed information. Consequently, this further enhances the performance of convolutional neural networks in gaze estimation tasks.

Specifically, the input features $F \in R^{C\times H\times W}$ are initially subjected to adaptive average pooling to reduce their spatial dimensions. Subsequently, separate average pooling and max pooling operations are applied to both the height and width dimensions. The process at this stage can be described by the following formula:

$$\begin{cases} F^{a}_{h,w} = AdaptiveAvgPool_{h,w}(F) \\ F^{a}_{h,w}{}' = AvgPool_{h,w}(F) \end{cases}$$
$$\begin{cases} F^{m}_{h,w} = AdaptiveAvgPool_{h,w}(F) \\ F^{m}_{h,w}{}' = MaxPool_{h,w}(F) \end{cases} \tag{2}$$

where in, *AdaptiveAvgPool* denotes adaptive average pooling, with the subscripts h, w indicating operations along the height and width dimensions, *MaxPool* represents max pooling, and *AvgPool* signifies average pooling.

Subsequently, the pooled features obtained are merged along both the height and width dimensions, as described by the following formula:

$$F_{h,w} = concat(F^{a}_{h,w}, F^{a}_{h,w}{}', F^{m}_{h,w}, F^{m}_{h,w}{}') \tag{3}$$

where *concat* denotes the concatenation operation. Following this, the merged features along the height and width dimensions are concatenated once more to obtain a comprehensive spatial feature representation, the formula is presented as follows:

$$F_t = concat(F_h, F_w) \tag{4}$$

Subsequent to the merging of features, the combined feature undergoes processing through a one-dimensional convolution and batch normalization. This is followed by the application of a Sigmoid activation function to normalize the convolved features, resulting in the generation of a spatial attention map, the formula is presented as follows:

$$M_{GR} = \sigma(conv(BN(F_t))) \tag{5}$$

Herein, σ represents the activation function, *conv* denotes convolution operations, and *BN* signifies batch normalization.

This spatial attention map is then element-wise multiplied with the original feature vector, ultimately yielding a feature map that places enhanced emphasis on spatial locations, the formula is presented as follows:

$$F_{FA} = M_{GR} \otimes F \tag{6}$$

Here, $\otimes$ denotes the operation of element-wise multiplication.

B. EYE FEATURE EXTRACTION MODULE BASED ON LOCAL IMAGE SUPER-RESOLUTION

In our methodology, we propose a novel Local image super-resolution Based Eye Feature Extraction (LSE) module, that targets ocular imagery for feature extraction. Instead of reconstructing the full facial image at super-resolution, we solely amplify the eye region. The rationale behind this is that, within gaze estimation, the eye region embodies the most critical features affecting the estimation. Enhancing only the eye image's resolution allows us to decrease the processing time, also augmenting gaze estimation accuracy efficiently.

For super-resolution reconstruction, LSE module uses the FSRCNN [11] network. FSRCNN is a specialized deep convolutional neural network tailored for single-image super-resolution. Originating as an advancement over SRCNN [9], its main goal was to enhance the speed and efficiency of super-resolution reconstruction.

FSRCNN primarily segments into three phases: feature extraction, shrinking and expanding, and deconvolution. The initial phase prioritizes feature extraction through convolutional layers, extracting low-level features from the input low-resolution image to capture essential details vital for high-resolution reconstruction. Subsequently, the dimensionality of these features is reduced, concentrating on information critical for super-resolution, thereby enhancing computational efficiency. The expanding phase subsequently restores the feature dimensions to levels appropriate for high-resolution reconstruction, setting the stage for the final phase. Here, a deconvolution layer reconstructs the high-resolution image from the processed features, aiming to significantly enhance resolution while maintaining the original image's integrity and details.

Followed by feature extraction via DeepEyeNet module, our proposed deep feature extraction CNN tailored for the eye region. DeepEyeNet is inspired by the VGG16 architecture [32], but it exhibits several advantages. Firstly, The implementation of batch normalization after each convolutional layer. Batch normalization facilitates the acceleration of the training process, enhances the convergence speed of the model, and serves as a regularization technique to reduce the likelihood of model overfitting. Secondly, The adoption of adaptive average pooling following the final convolutional layer to resize the feature map to 1 × 1. This adjustment allows for more flexible handling of input images of varying sizes, decreases the number of model parameters, contributes to reducing the risk of overfitting, and diminishes the overall model size. Thirdly, The reduction in model depth and parameter count is strategically tailored for more effective extraction of features from eye images with lower resolution.

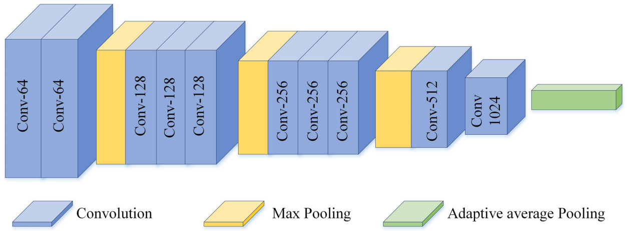

FIGURE 4. The network structure of DeepEyeNet module.

Consequently, DeepEyeNet demonstrates superior adaptability in extracting features from eye images. Specifically, DeepEyeNet module is composed of ten convolutional blocks. Each block integrates a Conv2d convolutional layer, followed by batch normalization and ReLU activation. Max-pooling operations are employed at the end of the 2nd, 5th, and 8th convolutional blocks, effectively reducing the spatial dimensions of feature maps. The output channels from the convolutional layers follow the sequence: 64 × 2, 128 × 3, 256 × 3, 512, and 1024. The final convolutional layer connects to an adaptive average pooling layer, refining the output feature map dimensions to 1 × 1 × 1024. Owing to its depth, this CNN systematically compresses the spatial dimensions of the feature maps as it progresses, allowing intricate feature extraction from super-resolved ocular images crucial for accurate gaze estimation. The network structure of DeepEyeNet module is illustrated in Fig 4.

Due to the inherent symmetry of left and right eyes, they traverse similar structures, enabling shared weight parameters between these pathways. The corresponding mathematical representation is presented below:

$$F_E' = FC_E\left(FLAT\left(\gamma(F_{\mathrm{SR}})\right)\right) \tag{7}$$

where γ denotes the feature extraction operation of DeepEyeNet module, F_E' signifies the output ocular feature, while FC_E represents the fully connected operation for the eye branch, F_{SR} represents the eye features after super-resolution, $FLAT$ represents the flatten operation for feature.

Upon obtaining features for the face and both eyes, the network proceeds through a fully connected layer, integrating features from the face, left eye, and right eye. It subsequently produces a two-dimensional feature set as the gaze estimation result. The corresponding formula is presented as follows:

$$\xi_{pred} = FC_T\left(concat\left(FC_{FA}\left(F_{FA}\right), F_E'\right)\right) \tag{8}$$

where FC_{FA} denotes the fully connected operation for the face branch, FC_T represents the fully connected operation for the total branches, *concat* denotes the concatenation for tensors.

GAFUSE-Net uses MSELoss as the loss function for gaze estimation. The MSELoss, also known as Mean Squared Error, represents the expected value of the squared differences between the predicted and true values. A smaller MSELoss indicates a smaller error. Thus, the loss function for gaze estimation in this module is given by:

$$\mathcal{L}_{gaze} = \frac{1}{n} \sum \left\| \xi_{gt} - \xi_{pred} \right\|_2^2 \tag{9}$$

where the ground true gaze direction value is denoted as ξ_{gt}, and the predicted gaze estimation value is represented by ξ_{pred}.

IV. EXPERIMENT

A. COMPARATIVE EXPERIMENT

1) DATASET

We evaluate the performance of our proposed network model by training it on the widely-recognized MPIIFaceGaze dataset for gaze estimation. Zhang et al. [22] introduced the MPIIFaceGaze dataset, which contains 213,659 images from 15 unique subjects, each with a resolution of 1280 × 720. The dataset encompasses images under diverse head orientations, postures, and lighting scenarios. The evaluation subset comprises 3,000 random samples from each of the 15 subjects, totaling 45,000 samples. Using this dataset enhances the reliability of our experimental findings. In alignment with studies [20], [22], we utilize the dataset's evaluation protocol, employing the leave-one-person-out cross-validation method.

We adapt the data preprocessing method suggested by Zhang et al. [22] We downsampled the dataset to mimic various low resolution conditions. The high resolution images are sized at 128 × 128. We applied scaling factors of 2 ×, 4 ×, and 8 × to downsample the HR images via bicubic interpolation. This produced three sets of LR images: 64 × 64, 32 × 32, and 16 × 16. To validate our findings' robustness, we experiment on these four datasets, assessing results across varied resolutions, using the leave-one-person-out cross-validation method.

2) EXPERIMENTAL ENVIRONMENT AND EVALUATION METRICS

We execute our experiments on a Linux platform with an NVIDIA GeForce RTX 4090 GPU, utilizing the PyTorch framework. Our model underwent training for 30 epochs, employing a batch size of 64. The learning rate is initialized at 0.001, with a decay rate of 0.1 set at an 8,000 step interval.

For evaluating performance, we employ the established angular error metric, commonly used in gaze estimation. The metric measures the angular difference between the predicted gaze direction and the ground truth. A lower metric value indicates better performance.

3) COMPARISON AND EXPERIMENTAL RESULTS

To verify the effectiveness of our proposed GAFUSE-Net architecture, we conduct comparative experiments against state-of-the-art methods using the MPIIFaceGaze dataset. To ensure a fair comparison, the experimental configurations for each method are adopted as per their respective original publications, including both the model architectures and hyperparameters, allowing us to accurately replicate their reported performance. The following provides an introduction to the comparative networks:

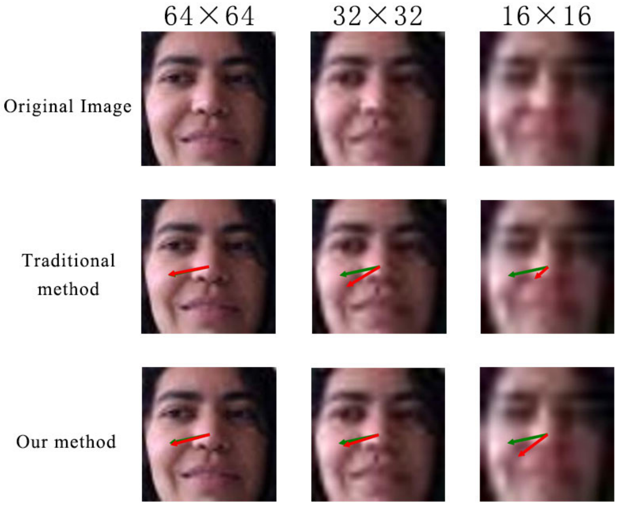

FIGURE 5. Visualization of results. The green arrow indicates the actual direction of gaze, while the red arrow represents the model's predicted direction of gaze.

Mnist [14] is the first to adopt convolutional neural networks in gaze estimation. It is an approach that uses multimodal convolutional neural networks for appearance-based gaze estimation.

GazeNet [20] is a deep gaze estimation method based on appearance, in which 13 convolutional layers are inherited from the 16-layer VGG network.

Full Face [22] is a deep neural network with an integrated spatial weighting mechanism, accepting comprehensive facial images as input.

Gaze360 [27] presents an LSTM-driven spatiotemporal gaze model, incorporating a unique loss function for error percentile estimation, boosting its precision.

RT-Gaze [23] is a gaze estimation network incorporating binocular fusion and head pose, primarily utilizing VGG-16 for eye feature extraction, enhancing network robustness by attaching head pose vector to fully connected layers.

GEDDNet [26] introduces a method decomposing gaze angles into general estimates and subject-specific biases, utilizing image-derived data to improve accuracy in universal gaze models.

CDBN [13] features a dedicated branch for structural information extraction from low-res images and a parallel residual branch for feature reconstruction. Both branches synergize for gaze estimation.

Table 1 below showcases the performance comparison of our proposed method, GAFUSE-Net, with other advanced techniques on the MPIIFaceGaze dataset. In the table, the top row represents the dataset's resolutions, listed as 128 × 128, 64 × 64, 32 × 32, and 16 × 16, illustrating results from the highest to the lowest resolutions. Each column depicts the angular error of different models at various resolutions.

TABLE 1. Comparison of our proposed method with other advanced methods at different resolutions. Results are reported in average angular error(°).

Methods	128×128	64×64	32×32	16×16
Mnist [14]	6.62°	6.72°	6.90°	7.83°
GazeNet [20]	6.25°	6.77°	6.51°	8.22°
Full-Face [22]	4.92°	4.99°	5.68°	8.01°
Gaze360 [27]	4.48°	4.68°	5.19°	7.75°
RT-Gaze [23]	5.34°	5.75°	5.96°	7.71°
GEDDNet [26]	4.60°	4.74°	5.38°	7.46°
CDBN [13]	4.67°	4.68°	5.13°	6.89°
GAFUSE-Net	**4.37°**	**4.62°**	**4.84°**	**6.19°**

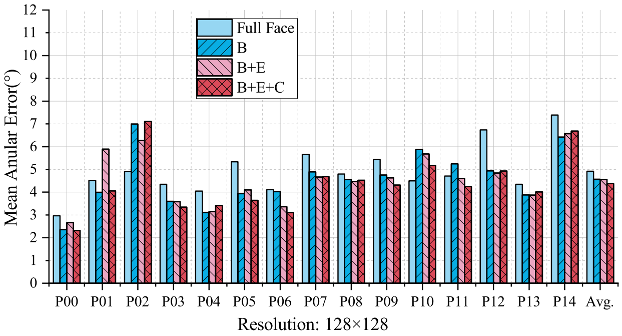

FIGURE 6. Average angular error comparison at a resolution of 128 × 128.

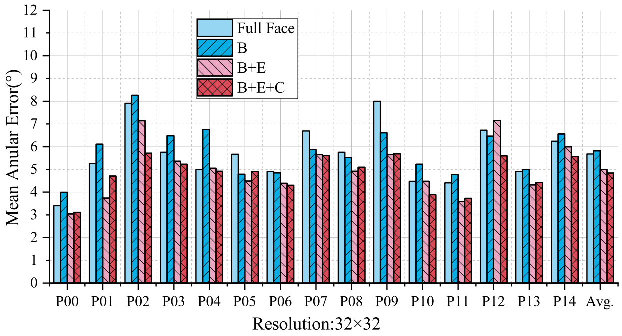

FIGURE 8. Average angular error comparison at a resolution of 32 × 32.

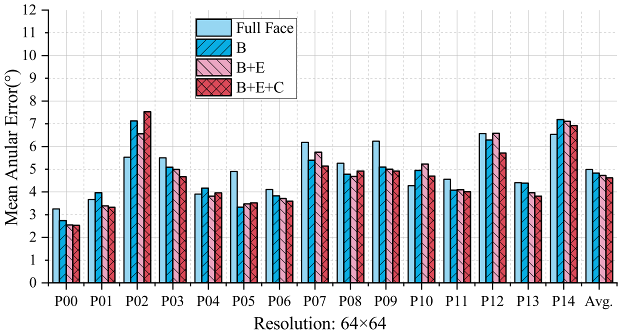

FIGURE 7. Average angular error comparison at a resolution of 64 × 64.

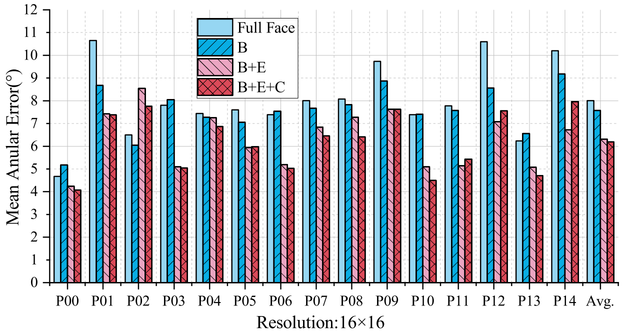

FIGURE 9. Average angular error comparison at a resolution of 16 × 16.

As shown in Table 1, GAFUSE-Net consistently achieves a lower angular error across all resolutions compared to other leading methods, indicating its superior performance. Specifically, at a high resolution of 128 × 128, GAFUSE-Net achieve an error of 4.37°, reducing the error by 2.5% compared to the Gaze360. At a lower resolution of 64 × 64, GAFUSE-Net post an error of 4.62°, representing a 1.3% improvement over the CDBN. At an even lower resolution of 32 × 32, GAFUSE-Net's angular error stood at 4.84°, 5.7% better than second-best. Particularly notable is GAFUSE-Net's performance at the most challenging resolution of 16 × 16, where it achieve an angular error of 6.19°. This marks a substantial 10.2% over the second-best. Our method enhances gaze feature extraction at low resolutions with a face attention-based module and the GRSA module. FSRCNN and DeepEyeNet enhance eye images, enabling high-quality gaze feature recovery at very low resolutions.

These results underline GAFUSE-Net's exceptional capability in low resolution scenarios. While nearly all models witness performance degradation as resolution diminishes, some experience drastic declines. However, GAFUSE-Net's performance deterioration is the most gradual among its peers. Its standout performance at extremely low resolutions testifies to its efficacy in gaze estimation under challenging low resolution conditions.

To visually demonstrate the superiority of out GAFUSE-Net, the gaze estimation results have been visualized. The images used for visualization are from the first picture of topic No. 0, with the results shown in Fig 5. Due to the lack of distinction in results at higher resolutions, three lower resolutions were used for comparison (64 × 64, 32 × 32, 16 × 16). It can be observed from the figure that, at a resolution of 64 × 64, both our method and the traditional method [22] almost reach the ground truth values. However, as the resolution decreases, the angle error of the traditional method gradually increases, while our method continues to achieve smaller angle errors. This demonstrates that out method can effectively estimate the gaze in low-resolution scenarios.

B. ABLATION STUDIES

To validate the efficacy of each component within our proposed GAFUSE-Net, we undertook ablation experiments. For these tests, we employ ResNet18 as our baseline model. Sequentially, we integrate each module proposed in this study into the baseline for ablation testing, and performance is assessed on the MPIIFaceGaze dataset across various resolutions, namely 128 × 128, 64 × 64, 32 × 32, and 16 × 16. These experiments clearly elucidate the contribution of each module in our method to the overall network performance, particularly highlighting the enhancements they offer to gaze estimation in low resolution scenarios.

Model 1: This serves as our baseline, where only ResNet18 is used as the primary network to extract facial features (B). These features are then fed into fully connected layers to yield gaze estimation result.

Model 2: Building on Model 1, we incorporate LSE module(E). This module proposes a dual-eye branch, which emphasizes the reconstruction of the eye region, a critical area for gaze estimation at a super-resolved scale prior to extracting features.

Model 3: Advancing from Model 2, the facial model is revised to integrate our FAE module(C). This paper proposes the GRSA module, which enhances the gaze related regions within facial images and extracts key features beneficial for gaze estimation. This targeted focus also enhances the learning capability for features in a low resolution environment, thereby improving gaze estimation under such conditions. The results of the ablation study are presented in Table 2.

TABLE 2. Ablation study of GAFUSE-Net at different resolutions. Results are reported in average angular error(°).

Models	128×128	64×64	32×32	16×16
1:B	4.58°	4.83°	5.82°	7.57°
2:B+E	4.56°	4.73°	5.00°	6.31°
3:B+E+C	**4.37°**	**4.62°**	**4.84°**	**6.19°**

The experimental results indicate that solely utilizing ResNet18 as the backbone network yields modest performance. As the input image resolution diminishes, there is a substantial increase in the gaze estimation error. However, with the incorporation of Module E, this error witnesses significant rectification. At a resolution of 128 × 128, the error is reduced by 0.4%, at 64 × 64, the error is reduced by 2.0%, at 32 × 32, it decreases by 14.1%, and at 16 × 16, it diminishes by 16.6%. These findings underscore the efficacy of the tri-stream network coupled with the local image super-resolution gaze estimation technique, especially in enhancing performance for low resolution gaze estimation. With the addition of Module C, the errors across these resolutions are further reduced by 4.2%, 2.3%, 3.4%, and 1.9%, respectively. This indicates that the GRSA module also plays a pivotal role in enhancing gaze estimation in low resolution scenarios.

The MPIIFaceGaze dataset encompasses 15 distinct subject themes, and a leave-one-person-out cross-validation approach is used to assess the average performance across these themes. To demonstrate the generalizability of our method in low resolutions, we further compared the average angular error for each theme against the baseline model (Full Face) as well as the models from our ablation study: Model 1, Model 2, and Model 3. The Fig 6, Fig 7, Fig 8 and Fig 9 delineates the average angular errors across subjects for our approach and the baseline model on the MPIIFaceGaze dataset at different resolutions. The x-axis denotes the 15 subject themes, while the y-axis represents the average angular error. For each subject theme, there are four bars, representing different models. "Avg." indicates the average error across all subjects. Comparative analysis reveals that our method outperforms the Full Face approach for the majority of the subjects. This suggests that despite vast disparities in individual appearances among subjects, our method consistently augments gaze estimation accuracy for most participants. The ablation study also further attesting to the effectiveness of each module.

TABLE 3. The results from the comparative evaluation of different attentions. Results are reported in average angular error(°).

Models	Mean Angular Error
Baseline	5.00°
Baseline+SE	4.91°
Baseline+ECA	4.95°
Baseline+GRSA	**4.84°**

C. EFFECTIVENESS OF ATTENTION MECHANISM

To validate the effectiveness of the attention mechanism, we conduct additional experiments, comparing the impact of different attention mechanisms on improving gaze estimation at low resolutions. We chose to contrast with classic attention mechanisms, namely the SE attention mechanism [34] and the ECA attention mechanism [35]. The baseline for this comparison is Model 2 from the ablation study, which incorporates all mechanisms except for the GRSA module. Experiments are conducted at a resolution of 32 × 32. The results from this comparative evaluation of different attention mechanisms are presented in Table 3. From the experiments, it can be observed that in shallow networks like ResNet18, the traditional channel attention mechanism, such as the SE attention mechanism, offers a certain enhancement in performance. In comparison to the SE attention mechanism, the ECA attention mechanism employs a one-dimensional convolution to learn the weights of each channel. This method, while slightly underperforming compared to SE attention, still shows improvement over the baseline. The GRSA module, on the other hand, by processing separately in terms of height and width, offers a more comprehensive understanding of the spatial distribution of feature maps.

This approach significantly enhances the extraction of key features under low resolution conditions, achieving superior gaze estimation performance at low resolutions.

V. CONCLUSION

In this work, we proposed GAFUSE-Net for gaze estimation, designed to excel in low-resolution scenarios. Our model outperforms existing methods by handling both high and low-resolution images effectively. Key innovations include the Facial Attention-Enhanced (FAE) and Local image Super-resolution based Eye (LSE) modules. Experiments on the MPIIFaceGaze dataset demonstrate GAFUSE-Net's versatility, achieving mean angular errors of 4.37°, 4.62°, 4.84°, and 6.19° at resolutions of 128×128, 64×64, 32×32, and 16×16, respectively. These results show error reductions of 0.2%, 1.3%, 5.7%, and 10.2% compared to state-of-the-art methods. These results significantly enhance the applicability of gaze estimation in real-world scenarios.

Despite these strengths, GAFUSE-Net's three-stream architecture increases computational complexity. This may limit its practicality in resource-constrained environments. Future work will focus on creating a lighter network to reduce computational load while maintaining performance. We also aim to enhance multi-scale feature learning to further improve accuracy in low-resolution conditions. These improvements will address current limitations and expand gaze estimation applications.

DATA AVAILABILITY

The study uses a public gaze dataset, can be found in: https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/gaze-based-human-computer-interaction/its-written-all-over-your-face-full-face-appearance-based-gaze-estimation.

CONFILICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] B. Wang, T. Hu, B. Li, X. Chen, and Z. Zhang, "GaTector: A unified framework for gaze object prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19566–19575.

[2] 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 Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Nov. 2016.

[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 *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 768–776.

[4] X. Zhang, Y. Sugano, and A. Bulling, "Evaluation of appearance-based methods and implications for gaze-based applications," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–13.

[5] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin Transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 1833–1844.

[6] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. M. S. M. de Marvao, T. Dawes, D. O'Regan, and D. Rueckert, "Cardiac image super-resolution with global correspondence using multi-atlas PatchMatch," in *Proc. 16th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Nagoya, Japan. Cham, Switzerland: Springer, Sep. 2013, pp. 9–16.

[7] F. Ling and G. M. Foody, "Super-resolution land cover mapping by deep learning," *Remote Sens. Lett.*, vol. 10, no. 6, pp. 598–606, Jun. 2019.

[8] H. Liu, Z. Ruan, P. Zhao, C. Dong, F. Shang, Y. Liu, L. Yang, and R. Timofte, "Video super-resolution based on deep learning: A comprehensive survey," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 5981–6035, Dec. 2022.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[10] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Bishop, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1874–1883.

[11] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 391–407.

[12] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[13] Z. Zhu, D. Zhang, C. Chi, M. Li, and D.-J. Lee, "A complementary dual-branch network for appearance-based gaze estimation from low-resolution facial image," *IEEE Trans. Cogn. Develop. Syst.*, vol. 15, no. 3, pp. 1323–1334, Mar. 2023.

[14] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.

[15] W. Yu, F. Zhao, Z. Ren, D. Jin, X. Yang, and X. Zhang, "Mining attention distribution paradigm: Discover gaze patterns and their association rules behind the visual image," *Comput. Methods Programs Biomed.*, vol. 230, Mar. 2023, Art. no. 107330.

[16] X.-B. Zhang, C.-T. Fan, S.-M. Yuan, and Z.-Y. Peng, "An advertisement video analysis system based on eye-tracking," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity)*, Dec. 2015, pp. 494–499.

[17] B. Noyes, A. Biorac, G. Vazquez, S. Khalid-Khan, D. Munoz, and L. Booij, "Eye-tracking in adult depression: Protocol for a systematic review and meta-analysis," *BMJ Open*, vol. 13, no. 6, Jun. 2023, Art. no. e069256.

[18] X. Zhou, H. Cai, Y. Li, and H. Liu, "Two-eye model-based gaze estimation from a Kinect sensor," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1646–1653.

[19] L. Sun, Z. Liu, and M.-T. Sun, "Real time gaze estimation with a consumer depth camera," *Inf. Sci.*, vol. 320, pp. 346–360, Nov. 2015.

[20] 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.

[21] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 100–115.

[22] 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 Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2299–2308.

[23] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 334–352.

[24] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2176–2184.

[25] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated convolutions," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 309–324.

[26] Z. Chen and B. E. Shi, "Towards high performance low complexity calibration in appearance based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1174–1188, Jan. 2023.

[27] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6911–6920.

[28] L. R. D. Murthy and P. Biswas, "Appearance-based gaze estimation using attention and difference mechanism," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3137–3146.

[29] S. Nonaka, S. Nobuhara, and K. Nishino, "Dynamic 3D gaze from afar: Deep gaze estimation from temporal eye-head-body coordination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2182–2191.

[30] J. Yun, Y. Na, H. Kim, H. Kim, and S. Yoo, ''HAZE-Net: High-frequency attentive super-resolved gaze estimation in low-resolution face images,'' in *Proc. Asian Conf. Comput. Vis.*, Sep. 2022, pp. 3361–3378.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[32] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[33] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] J. Hu, L. Shen, and G. Sun, ''Squeeze-and-excitation networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, ''ECA-Net: Efficient channel attention for deep convolutional neural networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

JIN WANG (Member, IEEE) received the Ph.D. degree from Nanjing University of Science and Technology, China, in 2009. He was an Exchange Scholar with the School of Computing, Informatics, Decision Systems Engineering, Arizona State University, in 2014. He is currently an Associate Professor with the School of Artificial Intelligence and Computer Science, Nantong University. His research interests include pattern recognition and computer vision. He is a member of the IEEE Computer Society and the ACM.

KE WANG received the B.S. degree from Jiangsu University, China, in 2022. He is currently pursuing the M.S. degree in computer technology with Nantong University, Nantong, China. His research interests include gaze estimation and object detection.

SHUOYU CAO received the B.S. degree in Internet of Things engineering from the Applied Technology College, Soochow University, Soochow, China. He is currently pursuing the M.S. degree in computer technology with Nantong University, Nantong, China. His research interests include gaze estimation and super-resolution.