AF-Net: Appearance-based Gaze Estimation via Adversarial Network and Attention

Bokun Wang, Yuan Yuan

Abstract—**Traditional appearance-based gaze estimation methods suffer from insufficient generalization in complex natural environments. Addressing the limitations of standard convolutional networks, which are unable to selectively adjust feature layer parameters and do not fully leverage multi-channel input information, this paper proposes a Feature Fusion and Adversarial Network (AF-Net), consisting of an extraction network and a fusion network. The extraction network incorporates an adversarial module that employs adversarial optimization to more accurately extract features relevant to gaze estimation. The fusion network, on the other hand, merges features from multiple channels and adaptively assigns weights to each channel through an attention mechanism, thereby obtaining a more precise gaze direction. Experimental results on three public datasets demonstrate that our method outperforms mainstream CNN approaches in terms of accuracy in unconstrained natural environment gaze estimation.**

Index Terms—**Gaze estimation, Deep learning, Attention mechanisms, Feature fusion**

I. Introduction

THE definition of gaze estimation tasks is to predict the three-dimensional directional vector or two-dimensional fixation point position based on image or video information. As one of the important branches of computer vision, gaze estimation integrates the application of machine learning and image processing technologies, which holds significant research importance. Gaze information can be used to infer various potential psychological or physiological information of the subject, such as attention distribution and cognitive behavioral processes, thereby enabling practical commercial applications like human-computer interaction [1], detecting driver fatigue, and assisting in disease diagnosis.

Based on the implementation principle, gaze estimation can be divided into appearance-based methods and model-based methods, among which appearance-based methods directly use RGB or depth images to obtain the gaze direction through regression functions. In recent years, with the development of deep learning and neural networks, appearance-based methods have rapidly improved in accuracy and have shown potential for use in unconstrained environments with significant variations in lighting and head pose, attracting increasing attention [2]。

The advent of deep learning networks, represented by Convolutional Neural Networks (CNNs), has made it possible to perform gaze estimation solely through input images. As the depth of the network layers increases, CNNs are capable of extracting more abstract and high-level features, which further enhances the accuracy of gaze estimation. However, in gaze estimation based on deep learning, models trained on large datasets often exhibit poor generalization performance, and the phenomenon of overfitting is a significant challenge that limits their applicability. When faced with new environments that have not been trained on, or new individuals not seen before, the predictive accuracy of the network tends to decline significantly compared to the training environment. Currently, researchers typically attempt to improve adaptability through methods such as rapid calibration, but the results are still not satisfactory.

Upon analysis, it has been determined that there are two primary factors that constrain the generalizability of models. The first is the failure to eliminate all irrelevant features related to specific environments from the feature set. An ideal network should more accurately capture factors within the input image that are related to gaze direction. However, environment-related features such as lighting conditions and background settings, as well as subject-related features such as the appearance of the subject, can lead to a decrease in accuracy when the model is applied to different scenarios. Secondly, most current networks only utilize a single input, such as an eye image or a facial image, or they simply concatenate the two without properly integrating the extracted features. Research has shown that the fusion of multiple features can be beneficial for enhancing the robustness of the network [3].Therefore, this paper proposes a Feature Fusion and Adversarial Network (AF-Net), designed to more effectively extract and remove environmental factors from the features, and to organically integrate multiple features from both the face and eyes.

The primary contributions of this paper are twofold:

(1) We propose an extraction network for the precise extraction of features. First, an adversarial module is introduced to eliminate redundant features. This module is not directly involved in feature extraction but is trained in an

Manuscript received on September 29, 2024; revised on February 25, 2025. This work was supported in part by the Science and Technology Research Program of Chongqing Municipal Education Commission (KJ1704072), the Chongqing Basic Science and Frontier Technology Research Program (Grant No. Cstc2017jcyjAX0212), and the Youth Fund Program of the National Natural Science Foundation of China (Grant No. 61703067).

Bokun Wang is a graduate student of Optoelectronic Information Sensing and Technology Laboratory at the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (corresponding author phone: 198-2381-3095; email: bkwang2018@163.com)

Yuan Yuan is a graduate student of the Optoelectronic Device Systems and Innovation Laboratory at the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (email: 2250599628@qq.com)

adversarial manner to optimize the feature extractor, thereby enhancing the precision with which gaze estimation-relevant features are obtained. Second, to preserve more detailed information from the input images, certain convolutional operations within the CNN network are replaced with dilated convolutions.

(2) We introduce a fusion network for the integration of multi-channel features, incorporating a cross-attention module to adaptively reallocate weights for multiple input features from the eyes and face.

The remainder of this paper is organized as follows: The second section provides a brief overview of related work. The third section presents a detailed explanation of the implementation principles of our method. The fourth section verifies the actual performance of the network through experiments. The fifth section offers a comprehensive summary.

II. Related works

A. Methods Based on Appearance

In comparison with model-based methods, appearance-based methods require no specialized equipment, such as infrared cameras, for feature extraction, and their applications extend beyond constrained environments with limited variations in lighting and head angles. This approach initially extracts effective features from the input facial image and then regresses the final gaze direction based on these features.

To date, numerous methods utilizing neural networks for feature extraction have been proposed and have demonstrated promising results. MnistNet [4] was the first to employ a neural network for gaze estimation. It achieved an average error of 6.3° using a 5-layer network with the LeNet architecture. ARE-Net [5] enhances accuracy in more complex lighting environments by leveraging the asymmetry of the eyes, selecting the more reliable eye for gaze estimation. Deep-Pictorial [6] regresses the appearance of the input eye to an image representation and uses this representation to estimate the gaze direction. iTracker [7] constructs a four-stream network that inputs both eyes, the face, and a facial mesh, achieving real-time 2D gaze estimation on smartphones. Wang et al. [8] addressed the overfitting issue in point estimation by integrating Bayesian inference and adversarial learning into a single framework, proposing a Bayesian Convolutional Neural Network to model the posterior distribution of parameters for more robust gaze estimation. Spatial weights CNN [9] introduces a facial weighting mechanism for the network to understand the importance of different regions of the face, achieving an average error of 4.8° based on the AlexNet architecture. PureGaze [10] designed a self-adversarial framework to purify gaze features and eliminate factors unrelated to gaze, with experimental results indicating a significant improvement in cross-dataset performance.

In recent years, the introduction of some new network architectures has also been proven to be beneficial for the task of gaze estimation. GazeOnce [11] proposed a one-stage end-to-end method that can process multi-person gaze estimation in real time. Dilated-CNN [12] uses dilated convolutions for downsampling, which, compared to common convolutional and pooling layer structures, can preserve more pixel-level fine details while significantly increasing the network's receptive field. Its architecture is suitable for a variety of other image processing tasks. The Transformer [13], initially used in natural language processing, has been applied in many fields of computer vision and has achieved good results. Data indicates that it can capture more long-range dependencies compared to CNNs. DVGaze [14] obtains more complete facial information by performing dual-view gaze estimation from two cameras.

III. Method

A. AF-Net

Based on the visual characteristics of the human eye, when focusing on an object, the general direction of both eyes is similar, providing complementary visual cues. Additionally, in most cases, the face determines the basic gaze direction, while the eyes fine-tune the angle. To explicitly leverage this characteristic, this paper proposes the Feature Fusion and Adversarial Network (AF-Net), which integrates features from both eyes alongside facial features, thereby capturing gaze-related information more comprehensively through three channels. Furthermore, while existing feature fusion methods predominantly rely on shallow networks, this paper enhances the fusion effect by introducing a cross-attention module.

As illustrated in Figure 1(a), the Feature Fusion and Adversarial Network comprises two main components: an adversarial extraction network designed to extract multi-channel features and a fusion network that fuses these features.

In the extraction network, FaceCNN and EyeCNN extract preliminary features of the face and eyes, respectively. To achieve a larger receptive field, dilated convolution [15] is incorporated into the VGG-Net-based backbone network in both CNNs. Common methods to increase the receptive field, such as increasing the stride or using pooling layers, often reduce the resolution of the feature maps, which can significantly impair the performance of gaze estimation regression tasks [16]. In contrast, dilated convolution preserves spatial resolution and maintains the number of parameters without significantly increasing computational complexity. In this study, to satisfy the accuracy requirements of the gaze estimation task, we configure the convolutional layers to seven, with the first four layers employing dilated convolutions. Both FaceCNN and EyeCNN utilize dilated convolutions to enhance the multi-channel architecture, with FaceCNN adopting a higher dilation rate. The template image sizes for the two networks are set to 64×64 and 64×32, respectively. To address the differences in feature map sizes across channels, Max pooling and a 2×2 convolution module are introduced to standardize the image dimensions while preserving spatial information. The backbone network is depicted in Figure 1(b).

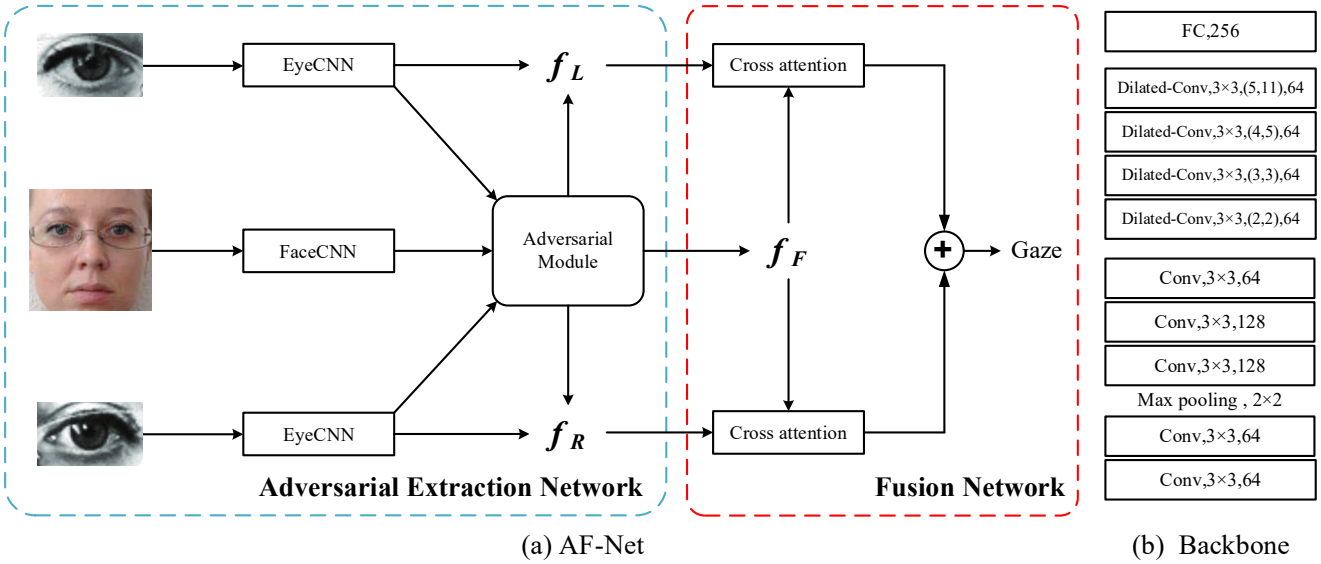

Fig. 1. Overall architecture of the Feature Fusion and Adversarial Network. (a) Architecture of AF-Net. (b) Backbone networks of FaceCNN and EyeCNN.

The output features from the dilated convolution, denoted by v, can be represented by the following equation:

$$v(x,y)=\sum_{k=1}^{K}\sum_{m=0}^{M-1}\sum_{n=1}^{N-1}u(x+nr_1,y+mr_2,k)w_{nmk}+b \quad (1)$$

where N×M×K is the size of the convolutional kernel, w and b represent the weight and bias values, respectively, and (r_1,r_2) denote the dilation rates of the network.

The preliminary features of the face and both eyes are transformed into secondary features by the adversarial module, which are then fed into the fusion network. The fusion network integrates the three features through a cross-attention module to derive the final gaze direction. The following sections will elaborate on the adversarial extraction network and the cross-attention module.

B. Adversarial Extraction Network

The traditional appearance-based deep learning approach for gaze estimation involves a series of steps: first, the raw image is processed by a feature extraction layer, where abstract deep feature maps are generated through multiple convolutional layers. Subsequently, the gaze regression layer, typically implemented as a multilayer perceptron (MLP) or fully connected layers, applies nonlinear functions to predict the final gaze direction. Since raw images captured in natural environments often contain redundant information unrelated to gaze, the excessive presence of irrelevant data can impair the feature extraction layer's ability to represent key features, thereby reducing estimation accuracy. As illustrated in Figure 2, to address this issue without increasing the network's depth or width, this paper proposes an adversarial feature extraction network built upon the traditional feature extraction layer. By introducing an adversarial module, the feature extractor is trained to capture deep features while effectively eliminating redundant information.

Generative Adversarial Networks (GAN) [17] is a network architecture that implements unsupervised learning based on game theory, consisting of two neural networks: a generator network that creates data and a discriminator network that judges the authenticity of the generated data. Through the adversarial optimization between the generator and the discriminator, domain adaptation can be achieved with fewer labeled examples, and it is widely used in various fields of computer vision [18].

Based on the aforementioned research, we construct a network using the idea of adversarial optimization. Since there is no need for a generation operation in the results, this paper discards the generator and adds an additional discriminator, forming an adversarial module with two discriminators. The following sections will detail the adversarial module.

In traditional supervised learning, features x are extracted from the input image p_{data}, and the goal of supervised learning is to optimize the features based on a loss function to minimize the error. In this paper, we employ two mutually antagonistic loss functions to indirectly achieve feature optimization.

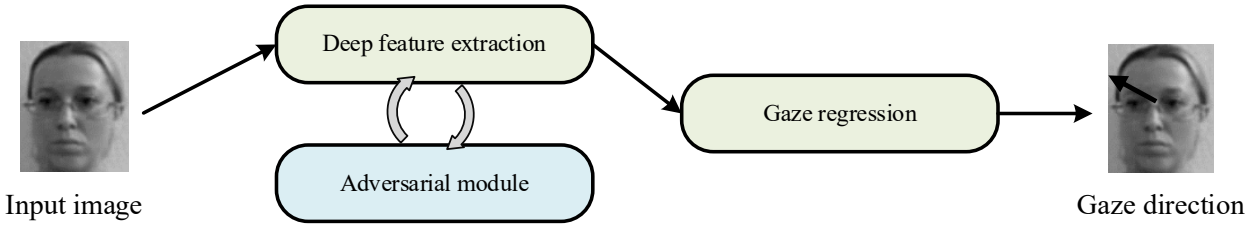

Fig. 2. Adversarial Module

As depicted in Figure 3, the adversarial module is composed of a gaze estimator $Dg(x\ ;\ \theta_g)$ and an appearance estimator $Da(x\ ;\ \theta_a)$. Instead of directly using the features x for gaze direction estimation, the adversarial module optimizes the feature extractor's parameters to obtain more robust features x', which are then used for the final gaze estimation. Specifically, the gaze estimator and the appearance classifier each have their own optimization objectives. While controlling their own parameters, they must also contend with the adverse effects of parameter changes induced by the other. The optimization is implicitly realized through the antagonism of the loss functions.

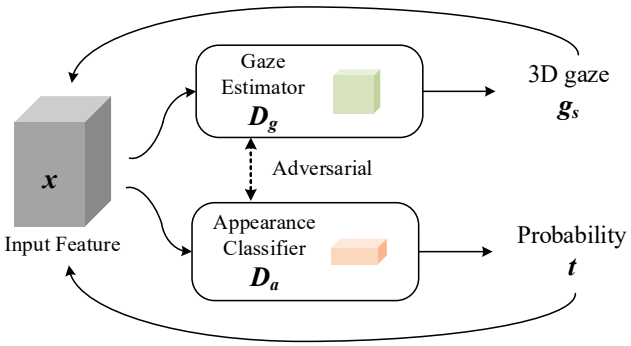

Fig. 3. Adversarial Module Architecture

The gaze estimator $Dg(\cdot)$ directly regresses the direction of gaze and constructs a loss function based on the deviation from the true gaze direction, with the objective of achieving a more precise estimation of the gaze direction. The loss function is defined as follows:

$$\mathrm{L}_g\left(\theta_f,\theta_g\right)=\frac{1}{n_t^{'}}\sum_{i=1}^{n_t^{'}}\| D_{\mathrm{g}}\left(E_f\left(\boldsymbol{x}_i;\theta_f\right);\theta_g\right)-\boldsymbol{y}_i\|^2 \quad (2)$$

where $\boldsymbol{y}$ represents the true values of the azimuth and elevation angles in 3D space. f is the learned feature representation, E_f is the feature extractor with parameters θ_f,, and D_g is the gaze estimator with parameters θ_g.

The appearance classifier $D_a(\cdot)$ produces a probability value t between 0 and 1, which characterizes the likelihood that the input originates from the training domain. The loss function is:

$$\varphi=\frac{1}{n_s}\sum_{i=1}^{n_s}\log\left(D_a\left(E_f\left(\boldsymbol{x}_i;\theta_f\right);\theta_a\right)\right) \quad (3)$$

$$\gamma=-\frac{1}{n_t}\sum_{i=1}^{n_t}\log\left(1-D_a\left(E_f\left(\boldsymbol{x}_i;\theta_f\right);\theta_a\right)\right) \quad (4)$$

$$\mathrm{L}_a\left(\theta_f,\theta_a\right)=\gamma-\varphi \quad (5)$$

defined in a binary cross-entropy manner, where $Da(x\ ;\ \theta_a)$ is the appearance classifier with parameters θ_a, and the output is a scalar probability t, representing the probability that the input comes from the source domain. When the classifier's output probability is close to 0.5, it indicates that it is impossible to distinguish between the feature distribution p_x and the source domain distribution p_{data}, and the appearance classifier reaches a global optimum.

To enable the extracted features to more accurately estimate the gaze direction and to make it difficult to discern the type of environment, the loss functions of the two components are combined to obtain the joint loss function:

$$\theta_f=\arg\min_{\theta_f}\mathrm{L}_g\left(\theta_f,\theta_g\right)-\lambda_a\mathrm{L}_a\left(\theta_f,\theta_a\right) \quad (6)$$

where λ_a is a positive balancing factor.

Since the optimization objectives of Equations (2) and (5) differ, they form an adversarial optimization relationship, together constituting the adversarial module. With the introduction of the adversarial module, for each input image frame, the feature layer dynamically adjusts the weights to filter and select more effective feature parameters. This process enhances the accuracy of gaze estimation, enabling more precise results.

C. *Three-channel Cross Attention Module*

In practical applications, facial angles and other factors often cause significant differences in the features of both eyes. The attention mechanism allows the model to focus on the most prominent features of the target and calculate the salience of each component based on global information [19]. It has been widely applied in many fields [20]. The Three-channel Cross Attention (TCA) module functions to assign different weights to input features based on their reliability. First, the features of both eyes pass through separate self-attention sub-layers, then facial features are integrated through residual connections, and finally, a cross-attention operation is performed between the features of both eyes. It has been verified that the TCA module can allocate weights to the eye and facial features, selectively enhancing or suppressing the input features. The Transformer architecture [21] is a novel network architecture suitable for computer vision tasks. Researchers have demonstrated that integrating pure Transformer models with traditional CNNs can achieve advanced performance [22]. The cross-attention module presented in this paper is based on the Decoder, which has been further improved for the task of gaze estimation. Overall, the module consists of two feature fusion Decoders, each comprising three sub-layers: a self-attention sub-layer, a cross-attention sub-layer, and a fully connected sub-layer. In the self-attention sub-layer, facial features are incorporated, and in the cross-attention sub-layer, the Query values of the two eyes are exchanged. The entire cross-attention module is depicted in Figure 4.

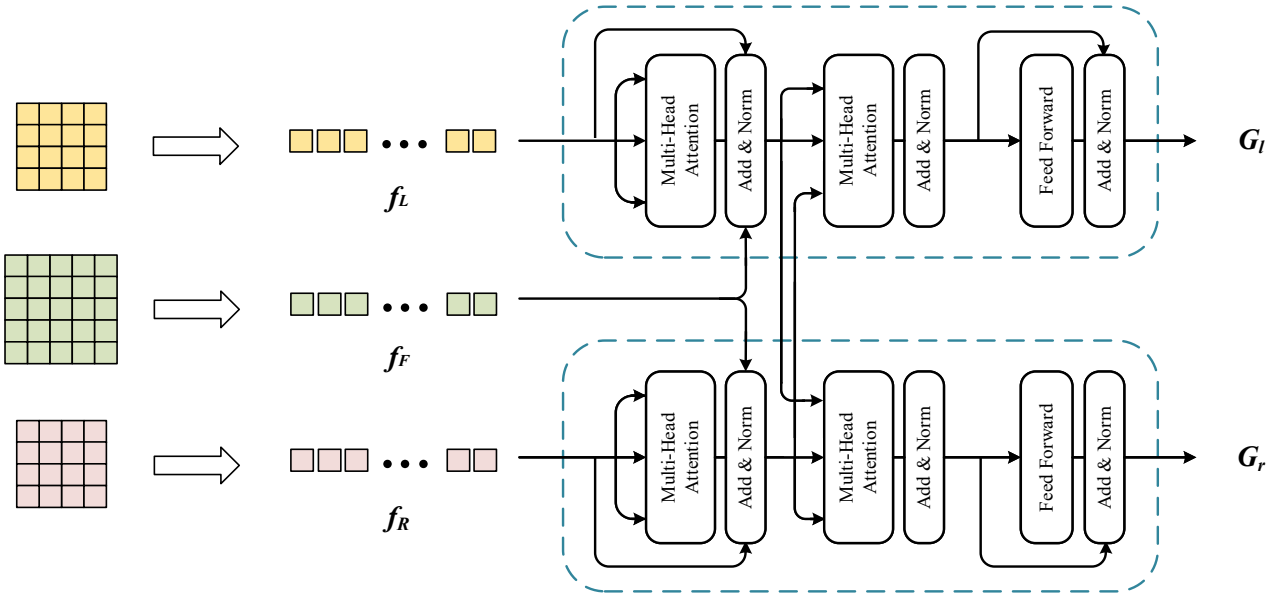

Fig. 4. Three-channel Cross Attention module architecture

The attention function maps the input Query, Key, and Value to the output. The weight of each Value is obtained by processing its Query and Key through dot product, division by the dimensionality parameter, and the softmax function. For the two cross-attention modules, this paper uses the corresponding eye features as Key and Value, and the other eye feature as Query. The formulas are as follows:

$$Output(Q_l, K_r, V_r) = softmax\left(\frac{Q_l K_r^T}{\sqrt{d_k}}\right)V_r \quad (7)$$

$$Output(Q_r, K_l, V_l) = softmax\left(\frac{Q_r K_l^T}{\sqrt{d_k}}\right)V_l \quad (8)$$

where K_l and K_r are the Key vectors of the left and right eye features, respectively; V_l and V_r are the Value vectors of the left and right eye features, respectively; d_k is the dimensionality of the Key vectors K_l and K_r; and $softmax(\cdot)$ denotes processing by the softmax function.

IV. Experimental Result

A. Dataset

The datasets utilized in this experiment are as follows:

MPIIGaze [23]: Comprises over 210,000 images collected over a three-month period in everyday environments by 15 participants. It exhibits significant variations in appearance, lighting, and head pose ranges and is one of the commonly used datasets for unconstrained environments. The standard MPIIGaze dataset includes a standard evaluation subset of 1500 eye images, but it does not contain complete facial images. In this paper, the corresponding facial region images for the evaluation subset are obtained from the MPIIFaceGaze dataset.

ETH-XGaze [24] and Gaze360 [25]: ETH-XGaze includes over 1 million labeled samples from 110 participants across more than 500 gaze directions, while Gaze360 contains 185 outdoor subjects that are closer to unconstrained natural environments. Due to the more varied natural lighting and background environments in both datasets, they are used as training sets in this paper to enhance the robustness of the model.

RT-Gene [26]: Composed of over 120,000 labeled images and over 150,000 unlabeled images, this dataset includes 15 participants. Compared to other datasets, RT-Gene has a larger range of gaze angles. However, due to inconsistent shooting distances, some images exhibit lower pixel quality. This paper uses this dataset to simulate natural environment gaze estimation with larger angle deviations.

EYEDIAP [27]: Composed of 94 short videos recorded with a Kinect sensor and a high-definition camera by 16 participants. It includes two target classifications: discrete locations and continuous trajectories. The dataset covers a head angle range of approximately 90°, but there is less variation in lighting. In this paper, we sample the original videos every 10 frames and construct an evaluation subset.

B. Compared Methods

MnistNet: The foundational network of MnistNet is a 5-layer LeNet architecture, which includes two convolutional layers, two max pooling layers, a fully connected layer, and a linear regression layer. The input is a single eye image.

GazeNet: A method based on a deep convolutional neural network, with the foundational network being a 16-layer VGG architecture, also taking a single eye image as input.

Spatial-Weight: A method that takes a full facial image as input, proposing a spatial weighting mechanism to generate a facial weight map. The network consists of 5 convolutional layers, 2 fully connected layers, and an additional spatial weighting component.

ARE-Net: The foundational network of ARE-Net consists of 6 convolutional layers, 3 max pooling layers, and a fully connected layer. The input includes a binocular image and a head pose vector. This method integrates binocular features, extracts features using AR-Net, and assigns weights through the reliability evaluation of the binoculars by E-Net.

FARE-Net: FARE-Net, building upon ARE-Net, retains the main network design and replaces the head pose vector input with a facial image input.

C. Implementation Details

The experiments in this paper were conducted on a device equipped with an Intel Xeon E5 CPU and 8 NVIDIA GeForce RTX 2080Ti GPUs, with a system environment of Ubuntu 18.04 LTS. The Adam optimizer was used, with a model learning rate of 0.0001 and a batch size of 256. For the adversarial extraction network, the proposed loss function was used, while for the fusion network, the L1 loss function was employed.

To better demonstrate the role of the adversarial module in cross-dataset testing, training was conducted on the ETH-XGaze, Gaze360, and RT-Gene datasets, which exhibit significant angle variations. The best-performing model trained on Gaze360 was then used for cross-dataset testing.

D. Result

1) Within Dataset

First, the performance of the proposed method was evaluated through extensive experiments in terms of average angular error. We compared the proposed method against seven state-of-the-art appearance-based gaze estimation methods on the MPIIGaze, EyeDiap, and RT-Gene datasets. The results, summarized in Table 1, include the backbone networks, input information, and corresponding average angular errors for all tested models.

The results demonstrate that the proposed AF-Net outperforms other methods in terms of angular error reduction. Specifically, compared to FARE-Net, which utilizes a four-stream network input, AF-Net achieves a 9.7% improvement on the MPIIGaze dataset, a 3.5% improvement on the EyeDiap dataset, and a 4.8% improvement on the RT-Gene dataset. Furthermore, compared to Dilated-Net, which employs dilated convolutional networks, AF-Net exhibits a 13.3% performance enhancement on the MPIIGaze dataset. To further analyze the contribution of each module in the proposed model, we conducted ablation studies, as detailed in the following sections.

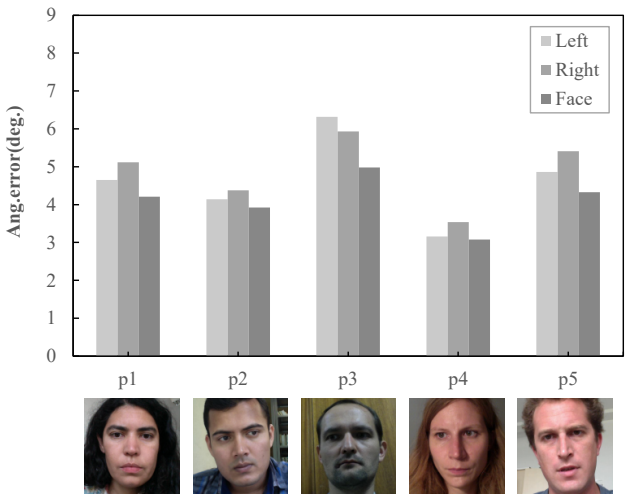

Fig. 5. Angle error of 5 subjects in the MPIIGaze dataset

To highlight the effectiveness of multi-feature fusion, we present the angular errors for the left eye, right eye, and full face of five different subjects from the MPIIGaze dataset. As illustrated in Figure 5, the attention module in the fusion network assigns different weights to multiple channels, assigning higher weights to more reliable channels while suppressing or correcting less reliable ones. By integrating input information from diverse sources, the network can correct low-quality input from a single eye based on data from the other eye and the full facial image, thereby achieving superior overall performance.

To more intuitively demonstrate the method's effectiveness, as shown in Figure 6, a visualization experiment was conducted on the dataset. For 5 testers, the gaze estimation results of EyeCNN and AF-Net for the left and right eyes were visualized. When there is excessive light or insufficient light, the gaze estimation error of a single eye increases significantly. This method effectively suppresses the influence of channels with low attention scores, resulting in a lower overall angular error. The results indicate that the method presented in this paper has good accuracy in various environments.

TABLE I ANGULAR ERROR OF TESTING WITHIN THE DATASET

	Backbone	Dataset	Input	Ang.error(deg.)
iTracker.	VGG-16 AlexNet	MPIIGaze	Face, Eyes	5.64
MnistNet	LeNet	MPIIGaze	Eye, Head pose	6.32
GazeNet.	VGG-16	Ut Multiview MPIIGaze	Eye, Head pose	4.48 5.56
Spatial-Weight	AlexNet	EyeDiap MPIIGaze	Face	6.04 4.87
ARE-Net	AlexNet	EyeDiap MPIIGaze	Eyes	6.17 5.04
Dilated-Net	Dilated-CNN	EyeDiap MPIIGaze	Face, Eyes	5.43 4.52
FARE-Net	AlexNet	RT-Gene EyeDiap MPIIGaze	Face, Eyes	8.42 5.75 4.39
Ours	Dilated-CNN Transformer	MPIIGaze EyeDiap RT-Gene	Face, Eyes	8.06 5.50 3.91

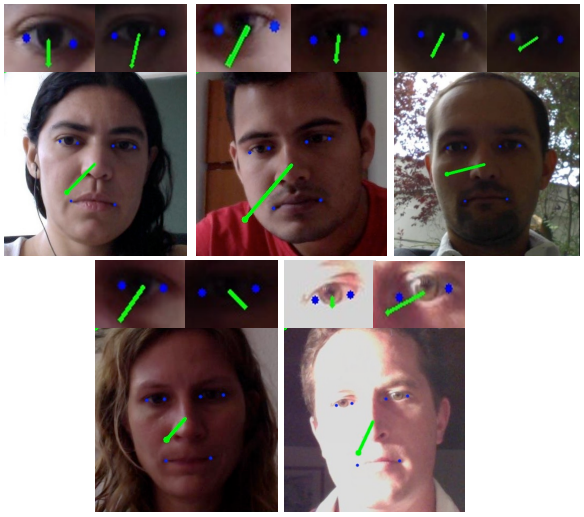

Fig. 6. Visualization experiment results of AF-Net

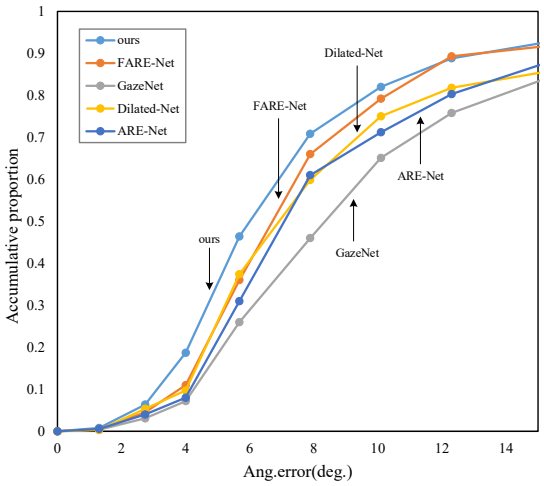

Fig. 7. Cumulative error of the model

To highlight the overall performance of the model, this paper further calculates the cumulative error proportion, which represents the proportion of data with different angular error values within the entire dataset. The calculation formula for the cumulative error proportion is as follows:

$$p = \frac{Num(\{s \mid s \le s_{A.e}, s \in S\})}{Num(s)} \tag{9}$$

where S represents the set of all angular errors, and $S_{A.e}$ represents the specific angular error. As shown in Figure 7, data with errors less than 10° account for 82% of all data, which represents a significant improvement compared to the four other methods. This indicates that the overall error of our network is relatively small and that it has better robustness for input images of varying quality.

2) Cross-dataset Evaluation

Cross-dataset testing reflects the model's adaptability to significant environmental differences, posing a considerable challenge for all appearance-based gaze estimation methods [28]. To simulate common usage scenarios, we train on the Gaze360 dataset, which exhibits significant variations in head pose and environmental lighting, and test on the EyeDiap, MPIIGaze, and RT-Gene datasets, which cover a wide range of angle variations.

TABLE 2 CROSS DATASET EXPERIMENTAL RESULTS OF AF-NET

	MnistNet	GazeNet	Extract Net	AF-Net
MPIIGaze	13.3	15.1	13.7	11.4
EyeDiap	14.6	16.9	14.4	13.2
RT-Gene	16.2	19.5	15.4	14.8

The results are shown in Table 2. First, AF-Net demonstrates a significant performance improvement compared to two classic single-eye-based methods, MnistNet and GazeNet. In the EyeDiap and RT-Gene datasets, AF-Net reduces the angular error by 32.4% and 31.8% compared to GazeNet, respectively. It also shows better performance compared to FARE-Net. Furthermore, to verify the role of the proposed adversarial network in domain adaptation, the performance of the extraction network was tested separately. The Extract Net, as shown in the table, retains all components of the extraction network and replaces the fusion network with two FC networks. The results indicate that Extract Net also achieved significant performance improvements, with accuracy increases of 10.2% and 17.4% compared to GazeNet in the two datasets. This suggests that the performance improvement attributed to the extraction network is the primary factor in the overall network performance improvement, thereby proving the effectiveness of the proposed adversarial module in filtering irrelevant features.

TABLE 3 EXPERIMENTAL RESULTS OF ABLATION STUDY ON AF-NET

Extract Network		Fusion Network		AF-Net		Ang.error(deg.)
-	Adversarial Module	FC	Cross attention	CNN	Dilated convolution	
Y	n	Y	n	Y	n	5.83
Y	n	Y	n	n	Y	5.29
n	Y	Y	n	Y	n	5.27
n	Y	Y	n	n	Y	4.42
Y	n	n	Y	Y	n	5.58
Y	n	n	Y	n	Y	4.65
n	Y	n	Y	Y	n	4.46
n	Y	n	Y	n	Y	3.92

3） Ablation Study

To demonstrate the individual effectiveness of each module, we conducted an ablation study of AF-Net on the MPIIGaze dataset. Specifically, we evaluated the contributions of three key modules: the adversarial module, the cross-attention mechanism, and the dilated network. For the adversarial network, the first group served as a control by completely removing the adversarial module, while the second group retained all components. For the cross-attention mechanism, the control group replaced the cross-attention network with two fully connected layers positioned after the extraction of binocular feature maps, and the facial feature map was entirely discarded. For the dilated convolution, the total number of convolutional layers remained unchanged, but the original dilated convolutions were replaced with standard convolutions.

As shown in Table 3, "Y" indicates the presence of the corresponding module in the network, and " n " indicates that the corresponding module was replaced with the control group. The results of the ablation study show that the network incorporating all three modules achieved the lowest angular error, with each module independently showing a positive impact on improving accuracy. The modules, in order from the greatest to the least impact on accuracy, are the adversarial module, dilated convolution, and the cross-attention module. This confirms that the network can effectively increase the receptive field through dilated convolution, further filter features using the adversarial module, and that the extraction network composed of these two modules effectively improves the quality of the obtained features. Additionally, the extraction network formed by the cross-attention module also actively enhances accuracy.

V. Conclusion

Addressing the domain generalization issue faced by appearance-based gaze estimation under unconstrained conditions, this study initially divides it into two aspects: feature extraction and feature fusion. Compared to other networks, this paper uses dilated convolution instead of regular convolutional layers for preliminary feature extraction, thereby increasing the network's receptive field. Additionally, an adversarial module is introduced to further filter the preliminary features, removing redundant features that are irrelevant to gaze direction. Moreover, the purified features from the parallel inputs of both eyes and the face are fused within a cross-attention module to determine the final gaze direction. Finally, the regression performance of the proposed method is verified through experiments. The experimental results indicate that AF-Net outperforms other mainstream CNN methods in terms of angular estimation accuracy across the three evaluated datasets, demonstrating good robustness and feasibility.

References

[1] M. A. Eid, N. Giakoumidis, and A. El Saddik, “A Novel Eye-Gaze-Controlled Wheelchair System for Navigating Unknown Environments: Case Study With a Person With ALS,” IEEE Access, vol. 4, pp. 558–573, 2016, doi: 10.1109/ACCESS.2016.2520093.

[2] X. Wang, J. Zhang, H. Zhang, S. Zhao, and H. Liu, “Vision-Based Gaze Estimation: A Review,” IEEE Trans. Cogn. Dev. Syst., vol. 14, no. 2, pp. 316–332, Jun. 2022, doi: 10.1109/TCDS.2021.3066465.

[3] D. Cazzato, M. Leo, C. Distante, and H. Voos, “When I Look into Your Eyes: A Survey on Computer Vision Contributions for Human Gaze Estimation and Tracking,” Sensors, vol. 20, no. 13, p. 3739, Jul. 2020, doi: 10.3390/s20133739.

[4] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “Appearance-based gaze estimation in the wild,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, Jun. 2015, pp. 4511–4520. doi: 10.1109/CVPR.2015.7299081.

[5] Y. Cheng, F. Lu, and X. Zhang, “Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression,” in Computer Vision – ECCV 2018, vol. 11218, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11218. , Cham: Springer International Publishing, 2018, pp. 105–121. doi: 10.1007/978-3-030-01264-9_7.

[6] S. Park, A. Spurr, and O. Hilliges, “Deep Pictorial Gaze Estimation,” in Computer Vision – ECCV 2018, vol. 11217, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11217. , Cham: Springer International Publishing, 2018, pp. 741–757. doi: 10.1007/978-3-030-01261-8_44.

[7] K. Krafka et al., “Eye Tracking for Everyone,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2176–2184.

[8] K. Wang, R. Zhao, H. Su, and Q. Ji, “Generalizing Eye Tracking With Bayesian Adversarial Learning,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 11899–11908. doi: 10.1109/CVPR.2019.01218.

[9] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “It’s Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation,” in 2017 IEEE CVPRW, Honolulu, HI, USA: IEEE, Jul. 2017, pp. 2299–2308. doi: 10.1109/CVPRW.2017.284.

[10] Y. Cheng, Y. Bao, and F. Lu, “PureGaze: Purifying Gaze Feature for Generalizable Gaze Estimation,” AAAI, vol. 36, no. 1, pp. 436–443, Jun. 2022, doi: 10.1609/aaai.v36i1.19921.

[11] M. Zhang, Y. Liu, and F. Lu, “GazeOnce: Real-Time Multi-Person Gaze Estimation,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: IEEE, Jun. 2022, pp. 4187–4196. doi: 10.1109/CVPR52688.2022.00416.

[12] Z. Chen and B. E. Shi, “Appearance-Based Gaze Estimation Using Dilated-Convolutions,” in Computer Vision – ACCV 2018, vol. 11366, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds., in Lecture Notes in Computer Science, vol. 11366. , Cham: Springer International Publishing, 2019, pp. 309–324. doi: 10.1007/978-3-030-20876-9_20.

[13] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems, 2017.

[14] Y. Cheng and F. Lu, “DVGaze: Dual-View Gaze Estimation,” Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[15] X. Lei, H. Pan, and X. Huang, “A Dilated CNN Model for Image Classification,” IEEE Access, vol. 7, pp. 124087–124095, 2019, doi: 10.1109/ACCESS.2019.2927169.

[16] Y. Luo, J. Chen, and J. Chen, “CI-Net: Appearance-Based Gaze Estimation via Cooperative Network,” IEEE Access, vol. 10, pp. 78739–78746, 2022, doi: 10.1109/ACCESS.2022.3194123.

[17] I. Goodfellow et al., “Generative adversarial networks,” Commun. ACM, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[18] Hao Zuo, Haoran Wang, and Yun Zhou, "Improving Extended Tree Augmented Naive Classifier with GAN," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2023, 5-7 July, 2023, Hong Kong, pp25-30

[19] Shuaibo Li, Zhengpeng Li, Jiansheng Wu, Jiawei Miao, Yuhang Bai, Xinmiao Yu, and Kejin Li, "Attention Feature Fusion Graph Convolutional Network for Target-Oriented Opinion Words Extraction," Engineering Letters, vol. 31, no.3, pp1273-1280, 2023

[20] Shuo Wang, and Yang Xu, "MI-YOLO: An Improved Traffic Sign Detection Algorithm Based on YOLOv8," Engineering Letters, vol. 32, no. 12, pp2336-2345, 2024

[21] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 03, 2021, arXiv: arXiv:2010.11929. Accessed: Dec. 18, 2023. [Online]. Available: http://arxiv.org/abs/2010.11929

[22] Y. Cheng and F. Lu, “Gaze Estimation using Transformer,” May 30, 2021, arXiv: arXiv:2105.14424. Accessed: Dec. 18, 2023. [Online]. Available: http://arxiv.org/abs/2105.14424

[23] 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, Nov. 2017, doi: 10.1109/TPAMI.2017.2778103.

[24] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation," in Computer Vision – ECCV 2020, vol. 12350, pp. 365–381. doi: 10.1007/978-3-030-58558-7_22.

[25] P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically Unconstrained Gaze Estimation in the Wild," in 2019 IEEE/CVF ICCV, Seoul, Korea (South): IEEE, Oct. 2019, pp. 6911–6920. doi: 10.1109/ICCV.2019.00701.

[26] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments," in Computer Vision – ECCV 2018, vol. 11214, pp. 339–357. doi: 10.1007/978-3-030-01249-6_21.

[27] 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, Safety Harbor Florida: ACM, Mar. 2014, pp. 255–258. doi: 10.1145/2578153.2578190.

[28] A. A. Akinyelu and P. Blignaut, "Convolutional Neural Network-Based Methods for Eye Gaze Estimation: A Survey," IEEE Access, vol. 8, pp. 142581–142605, 2020, doi: 10.1109/ACCESS.2020.3013540.