Computer Science and Information Systems 21(3):831–850 https://doi.org/10.2298/CSIS231116020X

CTA-Net: A Gaze Estimation network based on Dual Feature Aggregation and Attention Cross Fusion

Chenxing Xia[1,2,3], Zhanpeng Tao[1,⋆], Wei Wang[4], Wenjun Zhao[1], Bin Ge[1], Xiuju Gao[4], Kuan-Ching Li[6], and Yan Zhang[7]

[1] College of Computer Science and Engineering, Anhui University of Science and Technology
232001 Huainan, China
847990008@qq.com
[2] Institute of Energy, Hefei Comprehensive National Science Center
230031 Hefei, China
cxxia@aust.edu.cn
[3] Anhui Purvar Bigdata Technology Co. Ltd
232001 Huainan, China
[4] Anyang Cigarette Factory, China Tobacco Henan Industrial Co
Anyang, CHINA
[5] College of Electrical and Information Engineering, Anhui University of Science and Technology
232001 Huainan, China
[6] Department of Computer Science and Information Engineering, Providence University
43301 Taichung City, Taiwan
[7] The School of Electronics and Information Engineering, Anhui University
Hefei, China

Abstract. Recent work has demonstrated the Transformer model is effective for computer vision tasks. However, the global self-attention mechanism utilized in Transformer models does not adequately consider the local structure and details of images, which may result in the loss of information and local details, causing decreased estimation accuracy in gaze estimation tasks when compared to convolution or sequential stacking methods. To address this issue, we propose a parallel CNNs-Transformer aggregation network (CTA-Net) for gaze estimation, which fully leverages the advantages of the Transformer model in modeling global context while the convolutional neural networks (CNNs) model in retaining local details. Specifically, Transformer and ResNet are deployed to extract facial and eye information, respectively. Additionally, an attention cross fusion (ACFusion) Block is embedded with CNN branch, which decomposes features in space and channels to supplement lost features, suppress noise, and help extract eye features more effectively. Finally, a dual-feature aggregation (DFA) module is proposed to effectively fuse the output features of both branches with the help feature a selection mechanism and a residual structure. Experimental results on the MPIIGaze and Gaze360 datasets demonstrate that our CTA-Net achieves state-of-the-art results.

Keywords: Appearance-based gaze estimation, Deep neural networks, Dilated convolution, Fusion, Transformer

⋆ Corresponding author

1. Introduction

Estimating gaze from a single low-cost RGB sensor is an important research topic in computer vision, where eye or facial images are typically used as inputs to estimate the real gaze direction and locate gaze points. Gaze estimation has important applications in fields such as human-computer interaction [32], education [10], and medical diagnosis [36] [30].

Existing gaze estimation methods can be roughly categorized into two categories: model-based methods [9] [23] [25] and appearance-based methods [13] [26] [22]. Model-based methods focus on learning the geometric model of the entire eye and perform gaze estimation through manual calibration of features. However, this approach heavily relies on complex experimental equipment, which limits its effectiveness in harsh and unconstrained environments. In recent years, appearance-based methods have attracted much attention due to the development and application of deep learning, which only requires a regular RGB camera to capture images and directly learns the mapping function from facial appearance to human gaze [1]. CNNs have the ability to learn highly complex mapping functions, making them suitable for gaze regression. For example, Dilated-Net [4] , RT-Gene-Net [12], CA-Net [5]. Although these CNNs-based gaze estimation methods have achieved good performance, they always rely on using dilation convolution operations with different padding rates to extract contextual information, which may cause the loss of local information related to gaze and make the contextual information unrelated. Therefore, developing more effective gaze estimation models is critical for achieving better performance, robustness, and generalization in gaze estimation tasks.

Along with the Transformer model [28], thanks to its self-attention mechanism that selectively captures long-term dependencies between all tokens, has demonstrated outstanding performance in natural language processing tasks. In recent years, researchers have explored the use of Transformers in visual tasks, including gaze estimation. For instance, Cheng *et al.* [6] designed a Hybrid Transformer, which combines CNN and Transformer to extract low-level features and model global interactions, respectively. Cai *et al.* [2] used a linear combination of different Transformer frameworks for prediction and achieved a high ranking in the ETH-XGaze competition leaderboard. These methods effectively address the limitations of the Transformer architecture in modeling fine-grained details by using convolutional networks to introduce spatial biases in modeling local information. However, given that the operation of flattening image patches in the Transformer architecture may negatively impact the internal structural information of low-resolution images . Additionally, Transformer-based approaches often suffer from a quadratic growth of computational complexity with spatial size, which can result in cumbersome network architectures. Therefore, there is a need for a new network framework that can effectively address these issues while maintaining the integrity of structural information in low-resolution images.

In this paper, a parallel CNNs-Transformer aggregation network is proposed for gaze estimation (CTA-Net), which explicitly embeds global context and local information. Different from most existing Gaze estimation methods using a single feature encoder to extract feature, our CTA-Net adopts a parallel CNNs-Transformer structure to extract local and global cues via CNNs and Transformer networks from facial and eyes images, respectively. Moreover, since existing methods either ignore the correlation between the two eyes or handle the eye images separately for final output, we propose an eye image

Attention Cross Fusion (ACFusion) Block by using different combinations and interaction mechanisms. Specifically, we generate multiple attention feature maps through an attention mechanism to effectively filter out redundant or noisy information in feature channels, and further enhance the internal information of features and improve the performance of gaze estimation and target detection by interacting with binocular attention features. Finally, a dual-feature aggregation (DFA) module is introduced to merge the output features of different encoders for joint gaze regression.

Our contributions are as follows:

• We propose a parallel CNNs-Transformer aggregation network for gaze estimation (CTA-Net), which simultaneously considers the local details of features and global high-level semantic information.

• We design a dual-feature aggregation (DFA) module to fully integrate feature information from different encoders, promoting global information from the Transformer branch and reducing the noise that CNN features may contain.

• We present an eye image Attention Cross Fusion Block(ACFusion), which generates multiple attention feature maps to aggregate attention feature information for binocular image interaction.

2. RELATEDWORK

In the realm of gaze estimation, our exploration of related work focuses on two pivotal aspects: Gaze Estimation and Transformer-based approaches. These choices are motivated by the historical evolution of gaze estimation methodologies and the recent transformative impact of Transformer architectures in computer vision.

2.1. Gaze Estimation

Before the advent of CNN-based approaches to gaze estimation, regression functions were commonly used to create specific gaze mapping functions. These methods, such as neural networks [31], adaptive linear regression [21], Gaussian process regression [33], and dimension reduction [20], showed reasonable accuracy in constrained settings, but they were significantly less accurate in unconstrained settings due to the highly non-linear nature of the mapping function. However, with the rapid development of deep learning in recent years, Zhang *et al.* [38] proposed the first CNN-based gaze estimation method that far exceeded the performance of function regression methods using only a simple CNN. Since then, many improved and extended CNN-based gaze estimation methods [7] have emerged. For instance, Yu *et al.* [19] designed a multi-task gaze estimation model with landmark constraints, and Fischer *et al.* [12] used VGG-16 to extract features from binocular images to estimate gaze. Zhang *et al.* [38] proposed GazeNet, which inputed eye images into a 16-layer architecture where the head pose information was connected to the first fully connected layer after the convolutional layer. Cheng et al. [4] introduced dilated convolution in their gaze estimation method, which was combined with joint inference for head and eye avatars. Moreover, Cheng *et al.* [8] developed the FAR-Net to estimate the 3D gaze points of both eyes by combining the asymmetric properties of both eyes, which has the best performance in several public data sets. Krafka *et al.* [18] presented a multi-channel architecture to takes as input a left-eye image, a right-eye image, and a

cropped image of the face and face grid information. Cheng *et al.* [5] later proposed a coarse-to-fine network to integrate face and eye images by estimating a basic gaze from the face image and then refining the basic gaze with the eye image. To better investigate CNN-based gaze estimation, many large-scale gaze dataset tasks have been mentioned [16] [40] [37]. Despite the great success of these methods in gaze estimation, more accurate models are always needed and gaze estimation still faces challenges.

2.2. Transformer

The Transformer was initially introduced by Vaswani *et al.* [28] in the realm of natural language processing (NLP), as a unique self-attention mechanism to replace convolutional and recurrent networks. Compared with recurrent networks, the global computation of self-attention layers can selectively capture long-term dependencies between all tokens, and efficiently capturing continuity between semantic information and exhibiting outstanding performance. The success of the Transformer in NLP has inspired research in the field of computer vision, and it has been applied to various visual tasks. For instance, Carion *et al.* [3] used the encoder-decoder structure of the Transformer as the detection head (DETR) by extracting image features through a CNN and inputting the features into the detection head for prediction. Dosovitskiy *et al.* [11] proposed a purely self-attention-based visual Transformer (ViT), which solved the image classification problem by directly predicting possible categories using the Transformer encoder after dividing the image into non-overlapping 16×16 patches. Zheng *et al.* [41] applied the Transformer to replace the encoder in the natural image segmentation task to achieve state-of-the-art results. Cheng *et al.* [6] first introduced the Transformer into the gaze estimation task and proposed a Hybrid Transformer, which extracts bottom-level features using a CNN and models global interaction through the Transformer. Cai *et al.* [2] combined various Transformer frameworks to make predictions, and Huang *et al.* [15] proposed a lightweight transformer network for gaze estimation by using self-attention mechanisms. However, when replacing convolution with the Transformer or stacking them sequentially, it may lead to the loss of semantic information and features, affecting gaze estimation accuracy. Therefore, new methods are needed to address this problem.

3. Methods

3.1. Overall architecture

The architecture of CTA-NET is shown in Fig. 1, which mainly consists of a Gaze Transformer-ResNet encoder, an attention cross fusion (ACFusion), and a dual-feature aggregation (DFA) module. We first adopt a parallel Transformer-ResNet network to extract facial and eye information, respectively. Specifically, the Transformer branch is utilized to process facial images and starts with global self-attention, which is then followed by the restoration of detailed local features. Meanwhile, the CNN branch processes eye images, and the ACFusion module is employed to enhance the expression ability of local features by increasing the receptive field through two layers. The features extracted from both branches are fed into our proposed DFA module, which selectively fuses the information using attention and Hadamard addition. The multi-level fused feature maps are combined using residual connections, and then passed through a fully connected layer for gaze estimation.

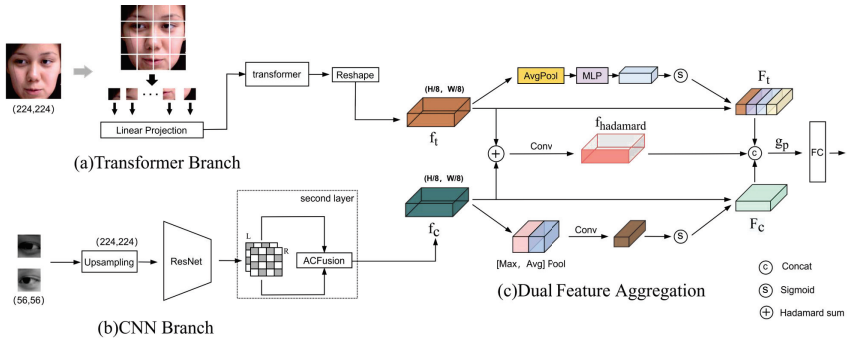

Fig. 1. An overview of the CNNs and Transformer aggregation network framework(CTA), which includes two parallel branches(Transformer and CNN), and our proposed Dual Feature Aggregation module

3.2. Gaze Transformer-ResNet Encoder

Unlike previous methods[6][2], we do not send the facial and eye images together to the Transformer for gaze regression after processing with the CNN. This decision stems from our consideration that the Transformer relies on the concept of position encoding, and the amalgamation of features from distinct facial regions may compromise the inherent positional information of the subject. Such compromise has the potential to diminish the accuracy of gaze estimation. In lieu of this approach, our paper introduces a two-stream framework, meticulously designed to concurrently extract features from diverse image types. The merit of this parallel processing framework lies in its ability to not only capture comprehensive global information but also retain sensitivity to local details. This dual focus proves advantageous in achieving more precise and accurate gaze estimation.

Transformer Branch The Transformer branch adopts a traditional encoder-decoder architecture, starting with the input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ being divided into $N = \frac{H}{S} \times \frac{W}{S}$ blocks with S typically set to 16. These blocks are flattened and linearly projected into a D_0 -dimensional output, generating the original embedding sequence $\mathbf{e} \in \mathbb{R}^{N\times D_0}$. To incorporate prior knowledge, a learnable position embedding is added to $\mathbf{e}$. Then, the generated embedding $\mathbf{z}^0 \in \mathbb{R}^{N\times D_0}$ is used as the input to the Transformer encoder, which consists of L layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) [27]. The self-attention module is the core of the Transformer encoder, and updates the state of each embedded patch by globally aggregating information across all layers. By applying the softmax function, the inner product between the query(q) vector and the key(k) vector is normalized, resulting in a relative weight assigned to each elemenAt of the global information. This facilitates the aggregation and updating of the global information, as expressed through the following formula:

$$\mathrm{SA}\left(\mathbf{z}_i\right) = \mathrm{softmax}\left(\frac{\mathbf{q_i}\mathbf{k}^T}{\sqrt{D_h}}\right)\mathbf{v}, \tag{1}$$

where $[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{W}_{qkv}, \mathbf{W}_{qkv} \in \mathbb{R}^{D_0 \times 3D_h}$is the projection matrix, $\mathbf{z}_i \in \mathbb{R}^{1 \times D_0}$ and $\mathbf{q}_i \in \mathbb{R}^{1 \times D_h}$ are the i-th rows of $\mathbf{z}$ and $\mathbf{q}$, respectively. MSA is extended by concatenating multiple self-attention modules SA and projecting the latent dimension to $\mathbb{R}^{D_0}$. The MLP is a dense layer stack consisting of fully connected layers with GELU activation and Dropout, where the first fully connected layer increases the number of input nodes by four times, and the second fully connected layer recovers the original number of nodes. The output of the last transformer layer is layer-normalized to obtain the encoded sequence $\mathbf{z}^L \in \mathbb{R}^{N \times D_0}$. The encoded sequence $\mathbf{z}^L$ is reshaped to produce the final output $\mathbf{f}_{\mathrm{t}} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times D_0}$, which is then jointly inputted into the DFA module with the feature map outputted by the CNN branch for sufficient fusion.

CNN Branch Our CNN branch is designed based on the ResNet network architecture. The traditional ResNet contains 4 layers to obtain enough low-level information while avoiding excessive consumption of model resources from overly deep networks, which can increase the model's sensitivity to noise, interference, and adversarial attacks. In our approach, we specifically chose the output of the second layer $CONV \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$ as the feature information for subsequent fusion. This choice aims to enhance the overall generalization ability. To better preserve the details of the image and reduce information loss, we introduce an upsampling operation before feature fusion. Specifically, we first use bicubic interpolation to upsample the left and right eye images (56×56→224×224). Compared with traditional bilinear interpolation, bicubic interpolation has better edge-preserving ability and performs better in visual tasks that require accuracy. Next, the ResNet network is utilized to extract features to generate the original feature images $\mathbf{f}_L \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$ and $\mathbf{f}_R \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$. These paired features are then input into our proposed feature-enhanced attention module ACFusion to fuse the left and right eye features. Furthermore, The feature $\mathbf{f}_{\mathrm{c}} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_1}$ obtained by the CNN branch is then combined with the feature map output by the Transformer branch and input into the DFA module for comprehensive fusion.

3.3. Attention Cross Fusion

This article proposes a module named ACFusion Fig. 2 to address various issues in the task of gaze estimation caused by the fusion of information from the left and right eyes, such as feature noise, information loss, and region imbalance. The ACFusion module combines residual connections with attention mechanisms to effectively fuse the feature information from the left and right eyes. Specifically, the ACFusion module employs a channel attention mechanism to assign different weights to feature maps, emphasizing important features while suppressing noise. For the left eye feature $\mathbf{t}_{\mathrm{L}}$, the ACFusion module first calculates the weight of each channel and adjusts the feature map based on these weights to retain valuable features. The channel attention mechanism is then used to further enhance the left eye feature map. The attention-enhanced left eye feature map is multiplied by the right eye feature $\mathbf{t}_{\mathrm{R}}$, and their spatial information is encoded and mapped using a 3x3 convolutional operation to highlight or suppress the necessary regions. The resulting convolutional features are calculated using the ReLU function to retain useful feature information. To further improve the quality of the fused feature representation, the ACFusion module employs a spatial attention mechanism to enhance the

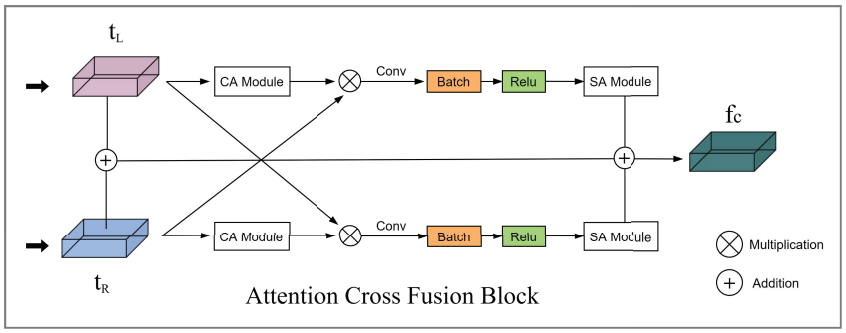

Fig. 2. Schematic diagram of the attention cross-fusion module. As shown in the figure, it includes a channel attention module(CA Module), convolution module, and spatial attention module(SA Module). The feature map is converged by each module in turn and cross-fused with the original feature map to get the final output. The fusion operation of the eye image is carried out simultaneously

global contextual information of the feature map. In this process, the refined left and right eye feature maps are progressively combined with the original feature map to obtain an information-enhanced feature map. Finally, these feature maps are concatenated and output as the fused feature $\mathbf{f}_c$.

The ACFusion module effectively utilizes the left and right eye feature information. Compared with previous single extraction methods, the ACFusion module can more effectively capture the left and right eye feature information while reducing feature noise and information loss. By introducing attention mechanisms, the ACFusion module can also better adapt to the imbalance of image regions, further improving the quality of the fused feature representation.

3.4. Dual Feature Aggregation

Due to the issues of dimension mismatch, context integration, and imbalanced feature representation in the fusion of encoding features from CNN and Transformer, in order to balance the feature expression capabilities of both and ensure comprehensive representation and integration of global and local contextual information, we propose a novel DFA Block (see Fig. 1(c)) that combines attention and multi-layer feature fusion mechanisms. Specifically, we leverage the SE-Block introduced in [14] to incorporate channel attention with residual connections to process the global information $\mathbf{f}_t$ from the Transformer branch, resulting in the featured representation $\mathbf{F}_t$:

$$\mathbf{F}_t = \text{Residual}\left([\mathbf{f}_t,\ \text{ChannelAttn}\ (\mathbf{f}_t)]\right). \tag{2}$$

Here, $Residual()$ denotes the residual connection operation, while $ChannelAttn()$ denotes the complete channel attention module, and $SpatialAttn()$ denotes the complete spatial attention module. Considering that high-level features in CNNs typically possess a larger receptive field and stronger semantic information, we adopt spatial attention from CBAM [34] as a spatial filter combined with residual connections to process the feature

information $\mathbf{f}_{\mathrm{c}}$ from the CNN branch, enhancing local details and suppressing irrelevant regions to obtain the feature $\mathbf{F}_{\mathrm{c}}$:

$$\mathbf{F}_{\mathrm{c}} = \mathrm{Residual}\left([\mathbf{f}_{\mathrm{c}}, \mathrm{SpatialAttn}\left(\mathbf{f}_{\mathrm{c}}\right)]\right). \tag{3}$$

Next, Hadamard addition is applied to element-wise sum the outputs from both branches to alleviate the issue of gradient vanishing, preserve the positional information of the original feature maps, and enhance the expressive capacity of the model:

$$\mathbf{F}_{\text{hadamard}} = \mathrm{Conv}\left(\mathbf{f}_{\mathrm{t}} \oplus \mathbf{f}_{\mathrm{c}}\right).$$

Here, $\mid \oplus \mid$ denotes Hadamard addition, and Conv refers to a 3x3 convolutional layer. Finally, the interactive feature $\mathbf{F}_{\text{hadamard}}$ is concatenated with the attention features $\mathbf{F}_{\mathrm{t}}$ and $\mathbf{F}_{\mathrm{c}}$, followed by a final fully connected layer. This step allows for effective integration of global and local contextual information, resulting in the feature representation $\mathbf{g}_p$ that effectively captures the global and local context at the current spatial resolution:

$$\mathbf{g}_{\mathrm{p}} = \left(\mathbf{F}_{\mathrm{t}}\right) \mathrm{Concat}\left(\mathbf{F}_{\mathrm{c}}\right) \mathrm{Concat}\left(\mathbf{F}_{\text{hadamard}}\right). \tag{4}$$

Here, the $Concat$ operation entails merging the outputs of multiple independent self-attention modules along the feature dimension. The proposed block combines two attention mechanisms to enhance the learning of global and local contextual information.Firstly, the SE block introduced in [14] is employed to fuse channel attention with residual connections, aiding the flow of global information in the Transformer branch. Secondly, spatial attention is utilized as a spatial filter combined with residual connections to selectively emphasize local details and suppress irrelevant regions. This approach leverages the fact that high-level features in neural networks typically possess larger receptive fields and stronger semantic information. By combining these attention mechanisms, our goal is to effectively capture both global and local contextual information. To further enhance the model's expressive power, we utilize the Hadamard addition operator to element-wise combine the outputs of the two branches, effectively integrating both global and local contextual information. Finally, the interactive feature $\mathbf{F}_{\text{hadamard}}$ and the attention features $\mathbf{F}_{\mathrm{t}}$ and $\mathbf{F}_{\mathrm{c}}$ are concatenated and passed through a final fully connected layer. This step enables the model to effectively integrate global and local contextual information, resulting in the feature representation $\mathbf{g}_{\mathrm{p}}$ that captures contextual information at the current spatial resolution.

3.5. Loss Function

Deep supervision is a technique that introduces intermediate supervision signals at multiple stages of the network. In our approach, we employ deep supervision to enhance the training process[24]. Specifically, we incorporate the Transformer branch and fusion branch into our architecture, serving as additional sources of supervision. These branches aim to capture different aspects and features of the input data.By adding these intermediate supervision signals, our goal is to improve gradients and facilitate convergence during network training. Deep supervision signals provide valuable information at different levels of abstraction, enabling the network to learn more effectively and efficiently. This

approach not only helps optimize the overall loss function but also enhances the interpretability and generalization capability of the network. We formulate a comprehensive objective function as follows:

$$\mathcal{L} = \alpha \mathrm{L}\left(\mathrm{G}, \mathbf{f}_{\mathrm{t}}\right) + \beta \mathrm{L}\left(\mathrm{G}, \mathbf{f}_{\mathrm{c}}\right) + \gamma \mathrm{L}\left(\mathrm{G}, \mathbf{g}_{\mathrm{p}}\right), \tag{5}$$

where α, β, and γ are adjustable hyperparameters that control the weights of each component. $\mathrm{L}\left(\mathrm{G}, \mathbf{f}_{\mathrm{t}}\right)$ represents the weighted IoU loss between the predicted facial gaze values $\mathbf{f}_{\mathrm{t}}$ and the ground truth (G). This term encourages accurate localization of facial pixels and precise depiction of facial regions. $\mathrm{L}\left(\mathrm{G}, \mathbf{f}_{\mathrm{c}}\right)$ represents the binary cross-entropy loss between the predicted facial gaze values and the ground truth, promoting pixel-wise classification accuracy of facial regions. Finally, represents a specific loss term that captures additional information related to facial geometry. By combining these different loss terms, the network is guided to learn robust and accurate facial gaze values. Through experimentation and iterative optimization of hyperparameters, we can find the optimal balance between the different components of the loss function, thereby improving the overall performance of the network. In summary, adopting weighted IoU loss, binary cross-entropy loss, and deep supervision techniques contributes to enhancing the effectiveness and robustness of our gaze prediction network. These strategies enable the network to leverage both pixel-level and high-level cues to accurately depict and localize facial regions, making it suitable for various practical applications such as facial analysis, recognition, and virtual reality.

4. Experiments

4.1. Setups

Dataset Our experiments are conducted on two main-stream datasets: MPIIGaze [40] and Gaze360 [17]. MPIIGaze is a kind of data set of gaze estimation based on appearance. It consists of MPIIGaze and MPIIFaceGaze, with MPIIGaze containing 15 subjects and 3,000 eye images per subject. The 3000 eye images are made up of 1500 left-eye images and 1500 right-eye images. MPIIFaceGaze, as an extension of MPIIGaze, contains the face image corresponding to each eye image in MPIIGaze. Note that MPIIGaze provides a standard evaluation method that selects 3000 images for each subject to make up the evaluation set. With leave-one-out evaluation, we do the experiment in the evaluation set, not in the whole evaluation set.

Gaze360 contains part of the back image without facial features, so we need to delete the image without face detection result according to the face detection annotation provided. Gaze360 contains a training set with 84K images for 54 subjects and a test set with 16K images for 15 subjects.

Data preprocessing We normalized both datasets in the same way. Specifically, we used virtual camera rotation and translation to eliminate the roll angle of the head and maintain the same distance between the virtual camera and the reference point (the center of the face). Additionally, we cropped eye images of size 56×56 from the normalized facial images, which were automatically detected using a face detection algorithm. The eye images were then histogram equalized and converted to grayscale to eliminate lighting effects.

Implementation details CTA-Net was built using the PyTorch framework and trained using an NVIDIA-A40 GPU. We used the Adam optimizer to train the entire model and used a linear learning rate warm-up with a warm-up phase of 5. For training on MPIIGaze, we use a batch size of 512 and 80 iterations, with a learning rate set to 0.0005 and a decay rate of 0.5. On Gaze360, we use a 256 batch size and 80 iterations, with a learning rate set to 0.0005 and a decay rate of 0.5. It is worth noting that we trained MPIIGaze using the leave-one-out method, taking an average of 15 sessions.

Evaluation metrics We use gaze Angle error as the evaluation measure of most gaze estimation methods. Assuming that the ground truth gaze direction is $\mathrm{G} \in \mathbb{R}^3$ and the predicted gaze vector is $\mathrm{g}' \in \mathbb{R}^3$, the gaze Angle error can be calculated as follows:

$$\mathcal{L}_{\text{angular}} = \frac{\mathrm{G} \cdot \mathrm{g}'}{\|\mathrm{G}\| \, \|\mathrm{g}'\|}. \tag{6}$$

The smaller the Angle error value is, the closer the model result is to the real value.

4.2. Comparison with appearance-based methods

To demonstrate the performance of our proposed method compared to other appearance-based gaze estimation methods on two datasets, MPIIFaceGaze and Gaze 360, we conducted a comprehensive evaluation. Considering the unique characteristics of the CTA-Net network architecture, we compared it with a total of nine convolutional network-based gaze estimation methods, including Full-face [39], Dilated-Net [4], RT-Gene [12], and GazeTR-Hybrid [6](the transformer-based gaze estimation method) to showcase its design advantages. Due to the unavailability of source code for some recent appearance-based methods, we referenced Cheng's survey [7] for some of our data.

Table 1. Performance in MPIIFaceGaze dataset. Due to the improved accuracy of RT-Gene through ensemble of four models, we also provided the results of model ensemble and referred to it as RT-Gene (4 models) to distinguish it from RT-Gene

Methods	MPIIFaceGaze	Gaze360 (Front 180°)	Realtime
iTracker (AlexNet)[18]	5.6°	None	$37ms$
MeNets[35]	4.9°	None	$36ms$
FullFace[39]	4.8°	14.99°	$37ms$
Dilated-Net[4]	4.8°	13.73°	$34ms$
RT-Gene[12]	4.8°	None	$34ms$
RT-Gene(4 ensemble)	4.3°	12.26°	$36ms$
Bayesian Approach[29]	4.3°	None	$32ms$
FAR-Net[8]	4.3°	None	$33ms$
CA-Net[5]	4.27°	12.26°	$34ms$
Gaze360[17]	None	11.40°	$33ms$
GazeTR-Hybrid[6]	4.00°	10.62°	$28ms$
CTA-Net(ours)	3.91°	10.44°	$30ms$

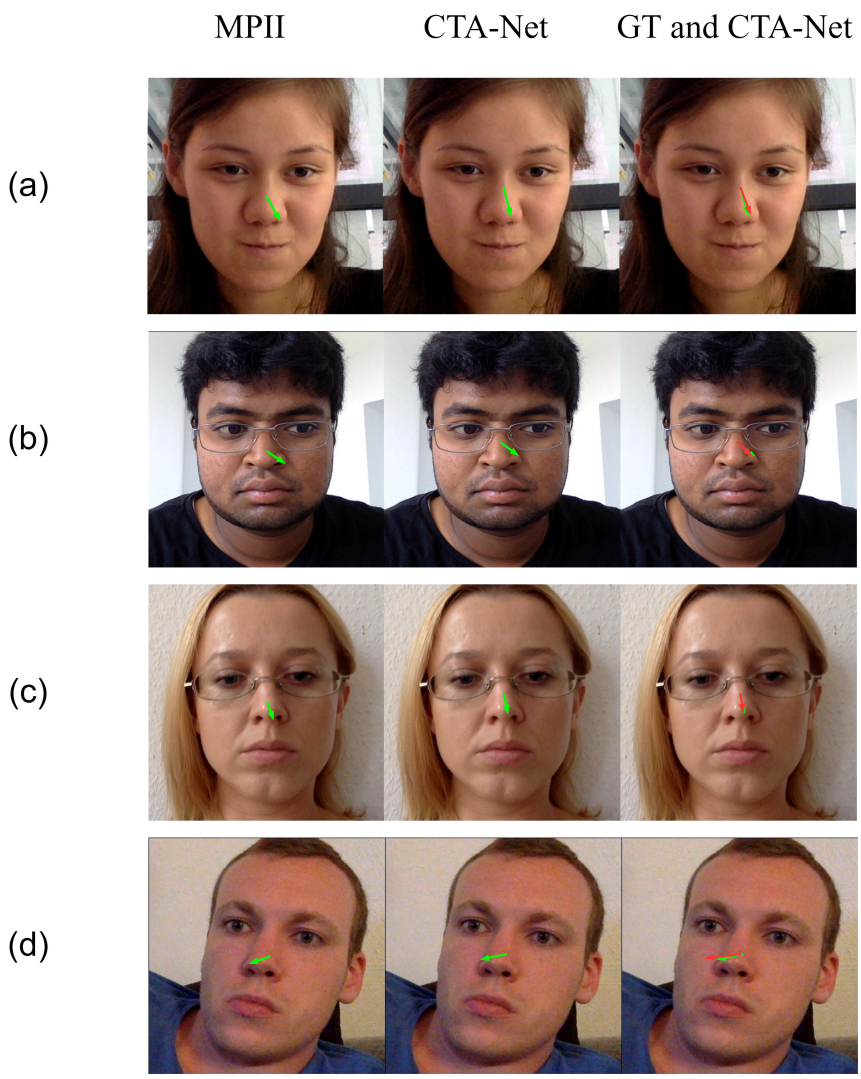

Fig. 3. The dataset is presented in a visualized format, where the ground truth is represented by a red arrow

Table 1 presents the results on the MPIIFaceGaze dataset. Our CTA-Net achieves lower angular error compared to the other methods. Specifically, our CTA-Net outperforms the highly accurate GazeTR-Hybrid with an angular error of 3.91°, representing an improvement of nearly 0.1° on the MPIIFaceGaze dataset. Table 1 displays the results on the Gaze 360 dataset. Following the division of the Gaze 360 dataset into train-val-test sets and evaluation ranges by Kellnhofer *et al.* [17], we adopt the same evaluation criteria, focusing on the frontal 180° range. This allows for a fair comparison with all relevant methods trained and evaluated on datasets within the 180° range. The proposed CTA-Net

achieves state-of-the-art gaze estimation performance with an average angular error of 10.44° in the frontal 180° range.

Fig. 3 illustrates the visualization results of the proposed method on the MPIIGaze dataset. Panels (a)-(d) show visualizations of eye directions for different subjects. Each group presents the visualizations for the same subject under different methods. We also compared the visualizations from the MPII model and the ground truth. The experimental results demonstrate that our method can adapt to different eye appearances while maintaining high accuracy and robustness.In summary, our research extends beyond theoretical contributions to underscore the practical efficacy of the proposed model. Through rigorous visual analyses in real-world application scenarios, depicted in Fig. 4, we confirm the model's robust generalizability. The positive outcomes observed in diverse contexts not only enhance theoretical understanding but also position the model as a versatile and effective solution for addressing real-world challenges. This dual validation, both in theory and practical application, establishes our model as a promising and impactful advancement in the field.

Fig. 4. The Citation Representation of Models in the Real World

4.3. Ablation study

We propose ablation experiments for the network structure and the proposed module respectively, and evaluate the effectiveness of the parallel branch design and the fusion module by changing the serial-parallel structure of the backbone network and the design choice of the fusion module.

Ablation experiments of structures. In order to better reflect the effectiveness of the parallel structure of CNN and Transformer, we design two simple network frameworks respectively, as shown in the figure, and recorde our experimental results on MPIIFaceGaze and Gaze360.

In this study, we present two architectures for gaze estimation: (a) a unified approach where both facial and eye images are processed by the same convolutional network before being fed into a transformer decoder and a fully connected layer for the final output, and

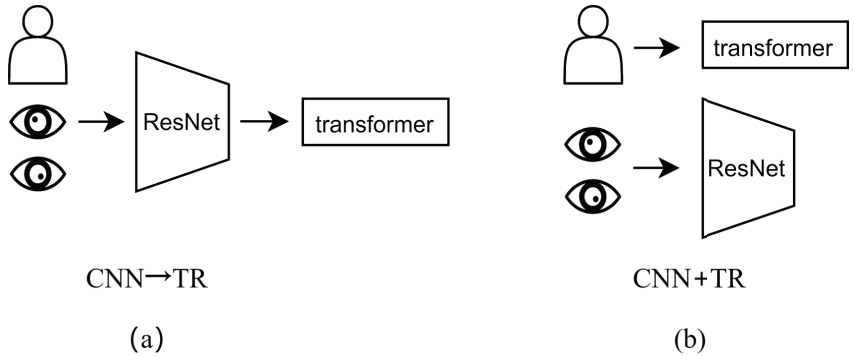

Fig. 5. The structure of CNN is mainly a pre-trained backbone network, including a pooling layer and fully connected layer, while the structure of the transformer is consistent with that in the text. Figure (a) is a serial structure, and Figure (b) is a parallel structure

(b) a parallel architecture inspired by CTA-Net, which utilizes separate convolutional and transformer networks that are connected via a fully connected layer, shown in Fig. 5. To ensure experimental accuracy, we do not incorporate any additional modules and relied solely on ImageNet pre-trained networks for our CNN backbone. Our results (Table 2) demonstrate that the parallel architecture outperforms the unified approach by a margin of 0.13° and 0.25° on MPIIFaceGaze and Gaze360 datasets, respectively.

Table 2. Ablation experiments of structures. 'CNN→TR'and 'CNN+TR'indicate a serial structure and a parallel structure, respectively

Methods	Backbones	Pre-train	MPIIFaceGaze	Gaze360
CNN→TR	ResNet18	ImageNet	4.47°	12.31°
CNN→TR	ResNet50	ImageNet	4.50°	13.16°
CNN+TR	ResNet18	ImageNet	4.44°	12.57°
CNN+TR	ResNet50	ImageNet	4.37°	12.21°

In addition, we conduct a thorough analysis of the rationality of the CNN branch design. Specifically, we perform comparative experiments to evaluate different upsampling methods and output layers of the CNN branch network. The tested upsampling methods include linear and bicubic interpolation, while the four different levels of the ResNet50 network are used as output layers. The overall architecture of the model remain unchanged, with the transformer branch generating an encoding sequence, which is reshaped to match the output channels of the CNN branch. The results, as shown in Table 3, indicates that the bicubic interpolation method exhibited higher accuracy compared to the linear interpolation method. In addition, our performance analysis of the output layer selection of CTA-Net demonstrates that it represented the optimal choice. These findings provide compelling evidence for the benefits of our proposed CNN branch design strategy.

Table 3. Ablation experiments of CNN branch

CNN-ResNet50		MPIIFaceGaze		Gaze360	
		bilinear	bicubic	bilinear	bicubic
Layer L	*L=1*	3.94°	3.94°	10.60°	10.59°
	L =2	3.93°	3.91°	10.51°	10.44°
	L =3	3.97°	3.95°	10.77°	10.71°
	L =4	4.02°	4.01°	10.61°	10.61°

Ablation experiments of modules.

In order to evaluate the impact of ACFusion and DFA modules on gaze estimation accuracy, we design CT-Net for module comparison, which replaces the module details in CTA-Net with fully connected layers for direct output. Results on both datasets are shown in Table 4(‘CTF-Net’stands for CT-Net+ACFusion network.), where our proposed method demonstrates a significant improvement in accuracy when incorporating the AC-Fusion module, with an increase of 0.16° and 1.92° on MPIIFaceGaze and Gaze360, respectively. At the same time, both datasets exhibit worse performance compared to CTA-Net, which proves the advantages of the proposed attention components. In addition, we present the gaze estimation accuracy of each participant in the MPIIFaceGaze dataset and compare different methods. Out of 15 participants, our proposed method achieves better gaze estimation accuracy for 9 participants, as shown in Fig. 6.

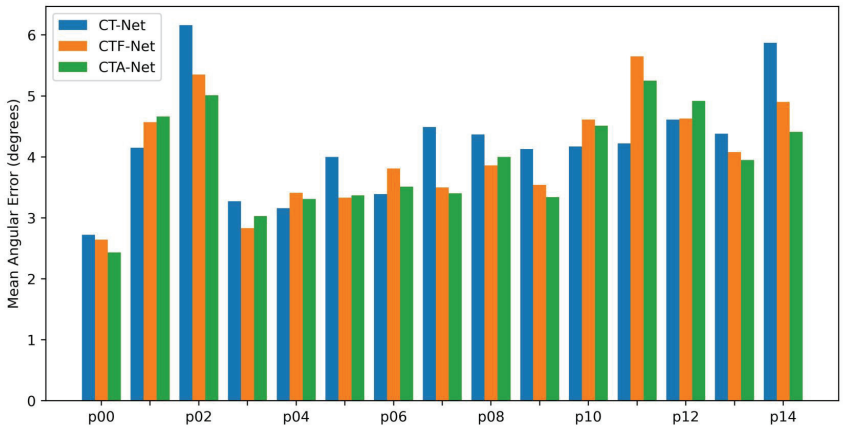

Fig. 6. The structure of CNN is mainly a pre-trained backbone network, including a pooling layer and fully connected layer, while the structure of the transformer is consistent with that in the text. Figure (a) is a serial structure, and Figure (b) is a parallel structure

In addition, we conduct ablation experiments to investigate the impact of the internal design of the modules.

Attention ablation:

Table 4. Ablation experiments of modules

Methods	MPIIFaceGaze	Gaze360
CT-Net	4.18°	12.51°
CTF-Net	4.02°	10.59°
CTA-Net(Single)	4.41°	14.01°
CTA-Net	3.91°	10.44°

We specifically focus on the ACFusion module and evaluate its influence on gaze estimation performance. The ACFusion module incorporates an attention mechanism to assign different weights to the left and right eyes, enabling the estimation of gaze residuals. To examine the role of the attention module, we propose an experimental configuration called "Without-Att", where the weights for the left and right eyes were fixed at 0.5, effectively removing the attention module for generating eye features. Furthermore, to explore the effect of the stacking order within the ACFusion module, we introduce a different stacking order called "Reverse-Att", which interchanges the spatial attention and channel attention. The results of these ablation experiments are presented in Table 5.

Compared to the complete ACFusion model, the attention ablation experiment shows a decrease in performance by 1.58° on the Gaze360 dataset. Similarly, the performance of the reverse attention order experiment is reduced by 2.27° compared to the CTA-Net. These results demonstrate the advantages of the attention component and the importance of the stacking order. By analyzing the findings from these ablation experiments, we can draw conclusions about the internal design of the ACFusion module and further validate its significance in gaze estimation tasks. This deeper understanding will contribute to unraveling the functionality and benefits of the ACFusion module.Furthermore, attention ablation experiments were conducted on the extraction of monocular feature information (CTA Net Single). As depicted in Figure 4, the accuracy of independently processing eye images decreased by nearly 0.5, rendering it ineffective in guiding gaze estimation. The fundamental reason for this phenomenon is that monocular feature information may lack sufficient context and comprehensiveness, posing a challenge in capturing the complexity of the visual system. Although eyes play a crucial role in visual perception, processing eye images independently may not fully capture the interaction and synergy between the eyes and the entire visual process. Therefore, our experimental results underscore the importance of considering binocular or more extensive visual information in gaze estimation tasks to achieve more accurate and comprehensive gaze estimation results.

Hadamard ablation:

To evaluate the impact of the DFA module on gaze estimation performance, we conduct an ablation experiment called "Without-Had". The DFA module is a critical component in our approach, utilizing Hadamard addition to generate feature maps. We design an ablation experiment specifically targeting the Hadamard addition operation within the DFA module. In the Hadamard ablation experiment, we retain the DFA module but remove the Hadamard addition operation, implying that we no longer employ Hadamard addition for generating feature maps. This allows us to assess the importance of the Hadamard addition operation within the DFA module. By comparing the performance

of these two ablation experiments with the complete model, we can evaluate the influence of the DFA module and the Hadamard addition operation on gaze estimation.

Based on the results from Table 5 and the ablation experiments, we observe a decrease in accuracy of 0.97° in the DFA module when the Hadamard addition operation is removed. This highlights the significance of the Hadamard addition operation in the DFA module for accurate gaze estimation

Table 5. Attention ablation and Hadamard ablation

Methods		Gaze360
Attention ablation	Without- Att	12.02°
Attention ablation	Reverse- Att	12.71°
Hadamard ablation	Without-Had	11.41°
CTA-Net		10.44°

4.4. Limitation and Future Work

Our model currently faces challenges due to a large number of overall parameters, resulting in prolonged training times. Furthermore, the complexity of the model structure introduces the risk of overfitting, especially when dealing with limited training data. This complexity may compromise the model's ability to generalize to new data. In the future, we are committed to thoroughly exploring these identified limitations and proposing practical solutions. Additionally, we will carefully consider the challenges posed by extensive training times and the risk of overfitting in our work. This is aimed at enhancing the efficiency and robustness of the model.

5. Conclusion

Our proposed dual-stream framework for accurate gaze estimation demonstrates exceptional functionality and meticulous internal design techniques. Through image fusion and the utilization of attention mechanisms, our approach preserves crucial positional information and effectively combines features from the left and right eyes. The attention cross-fusion (ACFusion) module addresses challenges such as feature noise and information loss, while the Dual Feature Aggregation (DFA) block integrates global and local contextual information. These designs yield superior accuracy in gaze estimation.

Although extensive experimental results on two mainstream datasets demonstrate the effectiveness of our proposed approach, it is important to note that the design based on this framework requires high-quality datasets and is inevitably influenced by environmental factors. Future research can focus on refining feature extraction methods to handle more complex scenarios and further improve accuracy and robustness.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62102003), Anhui Postdoctoral Science Foundation (2022B623),Medical Special Cultivation Project

of Anhui University of Science and Technology (YZ2023H2B003),Huainan City Science and Technology Plan (2023A316), Natural Science Foundation of Anhui Province (2108085QF258), the University Synergy Innovation Program of Anhui Province (GXXT-2022-038), Central guiding local technology development special funds (202107d06020001), University-level general projects of Anhui University of science and technology (xjyb2020-04).

References

1. Biswas, P., et al.: Appearance-based gaze estimation using attention and difference mechanism. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3143–3152 (2021)
2. Cai, X., Chen, B., Zeng, J., Zhang, J., Sun, Y., Wang, X., Ji, Z., Liu, X., Chen, X., Shan, S.: Gaze estimation with an ensemble of four architectures. arXiv preprint arXiv:2107.01980 (2021)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision. pp. 213–229. Springer (2020)
4. Chen, Z., Shi, B.E.: Appearance-based gaze estimation using dilated-convolutions. In: Proceedings of the Asian Conference on Computer Vision. pp. 309–324. Springer (2018)
5. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10623–10630 (2020)
6. Cheng, Y., Lu, F.: Gaze estimation using transformer. arXiv preprint arXiv:2105.14424 (2021)
7. Cheng, Y., Wang, H., Bao, Y., Lu, F.: Appearance-based gaze estimation with deep learning: A review and benchmark. arXiv preprint arXiv:2104.12668 (2021)
8. Cheng, Y., Zhang, X., Lu, F., Sato, Y.: Gaze estimation by exploring two-eye asymmetry. IEEE Transactions on Image Processing 29, 5259–5272 (2020)
9. Cristina, S., Camilleri, K.P.: Model-based head pose-free gaze estimation for assistive communication. Computer Vision and Image Understanding 149, 157–170 (2016)
10. Dari, S., Kadrileev, N., Hüllermeier, E.: A neural network-based driver gaze classification system with vehicle signals. In: Proceedings of the International Joint Conference on Neural Networks. pp. 1–7. IEEE (2020)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Fischer, T., Chang, H.J., Demiris, Y.: Rt-gene: Real-time eye gaze estimation in natural environments. In: Proceedings of the European Conference on Computer Vision. pp. 334–352 (2018)
13. Hansen, D.W., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3), 478–500 (2009)
14. Huang, C.H., Wu, H.Y., Lin, Y.L.: Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172 (2021)
15. Huang, H., Ren, L., Yang, Z., Zhan, Y., Zhang, Q., Lv, J.: Gazeattentionnet: Gaze estimation with attentions. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2435–2439. IEEE (2022)
16. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications 28, 445–461 (2017)
17. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6912–6921 (2019)

18. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2176–2184 (2016)
19. Liu, G., Yu, Y., Mora, K.A.F., Odobez, J.M.: A differential approach for gaze estimation with calibration. In: Proceedings of the British Machine Vision Conference. vol. 2, p. 6 (2018)
20. Lu, F., Chen, X., Sato, Y.: Appearance-based gaze estimation via uncalibrated gaze pattern recovery. IEEE Transactions on Image Processing 26(4), 1543–1553 (2017)
21. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(10), 2033–2046 (2014)
22. Mora, K.A.F., Odobez, J.M.: Person independent 3d gaze estimation from remote rgb-d cameras. In: Proceedings of the IEEE International Conference on Image Processing. pp. 2787–2791. IEEE (2013)
23. Morimoto, C.H., Mimica, M.R.: Eye gaze tracking techniques for interactive applications. Computer Vision and Image Understanding 98(1), 4–24 (2005)
24. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7479–7489 (2019)
25. Stampe, D.M.: Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. Behavior Research Methods, Instruments, & Computers 25, 137–142 (1993)
26. Sugano, Y., Matsushita, Y., Sato, Y.: Appearance-based gaze estimation using visual saliency. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(2), 329–341 (2012)
27. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al.: Resmlp: Feedforward networks for image classification with data-efficient training. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Proceedings of the Advances in Neural Information Processing Systems 30 (2017)
29. Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with bayesian adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11907–11916 (2019)
30. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging 41(7), 1688–1698 (2022)
31. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
32. Wang, Z., Wang, H., Yu, H., Lu, F.: Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system. IEEE Transactions on Human-Machine Systems 51(5), 524–534 (2021)
33. Williams, O., Blake, A., Cipolla, R.: Sparse and semi-supervised visual mapping with the sˆ 3gp. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 230–237. IEEE (2006)
34. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3–19 (2018)
35. Xiong, Y., Kim, H.J., Singh, V.: Mixed effects neural networks (menets) with applications to gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7743–7752 (2019)
36. Yang, B., Cui, J., Tong, Y., Wang, L., Zha, H.: Recognition of infants' gaze behaviors and emotions. In: Proceedings of the International Conference on Pattern Recognition. pp. 3204–3209. IEEE (2018)

37. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: Proceedings of the European Conference on Computer Vision. pp. 365–381. Springer (2020)
38. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4511–4520 (2015)
39. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–60 (2017)
40. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(1), 162–175 (2017)
41. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6881–6890 (2021)

Chenxing Xia received the Ph.D. degree from Hunan University in 2019. He is currently an Associate Professor at the College of Computer Science and Engineering, Anhui University of Science and Technology. His research interests include saliency detection, computer vision, image processing, and depth prediction.

Zhanpeng Tao received the Anhui University of Science and Technology. He is currently pursuing the master degree with Anhui University of Science and Technology. His research interests include computer vision, machine learning, monocular 3D object detection,gaze estimation and depth estimation.

Wenjun Zhao received the bachelor degree from Wanjiang University of Technology. He is currently pursuing the master degree with Anhui University of Science and Technology. His research interests include computer vision, machine learning, monocular 3D object detection, and depth estimation.

Wei Wang received his bachelor's degree from the City College of Wuhan University of Science and Technology, He is currently a technician at Anyang Cigarette Factory, China Tobacco Henan Industrial Co., Ltd.,and his job title is an engineer. his research interests include automatic control and tobacco machinery.

Bin Ge received the Ph.D. degree in computer application technology from Hefei university of technology, Hefei China, in 2016. He began to teach at Anhui University of Science and Technology in 1999. Currently, he is a professor of information security at Anhui University of Science and Technology. His research interests include information security, Internet of things technology and artificial intelligence.

Xiuju Gao received the M.S. degree in Computer Science and Electronic Engineering from Hunan University, Changsha China, in 2016. She is currently an assistant at the College of Electrical and Information Engineering, Anhui University of Science and Technology. Her research interests include image/video processing and computer vision.

Kuan-Ching Li is a Distinguished Professor in the Dept of Computer Science and Information Engineering (CSIE) at Providence University, Taiwan, where he also serves as the Director of the High-Performance Computing and Networking Center. He received the Licenciatura in Mathematics, and MS and Ph.D. degrees in electrical engineering from the University of Sao Paulo (USP), Brazil, in 1994, 1996, and 2001, respectively. He published more than 250 scientific papers and articles and is author, co-author or editor of more than 25 books published by Taylor & Francis, Springer, and McGraw-Hill. Professor Li is the Editor in Chief of the Connection Science and serves as an associate editor for several leading journals. Also, he has been actively involved in many major conferences and workshops in program/general/steering conference chairman positions and has organized numerous conferences related to computational science and engineering. He is a Fellow of IET and a Senior Member of the IEEE. His research interests include parallel and distributed computing, Big Data, and emerging technologies.

Yan Zhang received the Ph.D. degree in computer application technology from Anhui University in 2010. She is a professor and Ph.D. supervisor at Anhui University of Science and Technology. Her research interests include information security, data mining, etc.

Received: November 16, 2023; Accepted: February 05, 2024.