electronics

MDPI

Article

Gaze Estimation Method Combining Facial Feature Extractor with Pyramid Squeeze Attention Mechanism

Jingfang Wei [1], Haibin Wu [1,*], Qing Wu [1], Yuji Iwahori [2], Xiaoyu Yu [3] and Aili Wang [1,*]

[1] Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Technology, Harbin 150080, China; 2220610105@stu.hrbust.edu.cn (J.W.); wuqing@hrbust.edu.cn (Q.W.)
[2] Department of Computer Science, Chubu University, Kasugai 487-8501, Japan; iwahori@isc.chubu.ac.jp
[3] College of Electron and Information, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan 528402, China; yuxy@zsc.edu.cn
* Correspondence: woo@hrbust.edu.cn (H.W.); aili925@hrbust.edu.cn (A.W.)

Abstract: To address the issue of reduced gaze estimation accuracy caused by individual differences in different environments, this study proposes a novel gaze estimation algorithm based on attention mechanisms. Firstly, by constructing a facial feature extractor (FFE), the method obtains facial feature information about the eyes and locates the feature areas of the left and right eyes. Then, the L2CSNet ($\updownarrow_2$ loss + cross-entropy loss + softmax layer network), which integrates the PSA (pyramid squeeze attention), is designed to increase the correlation weights related to gaze estimation in the feature areas, suppress other irrelevant weights, and extract more fine-grained feature information to obtain gaze direction features. Finally, by integrating L2CSNet with FFE and PSA, FPSA_L2CSNet was proposed, which is fully tested on four representative publicly available datasets and a real-world dataset comprising individuals of different backgrounds, lighting conditions, nationalities, skin tones, ages, genders, and partial occlusions. The experimental results indicate that the accuracy of the gaze estimation model proposed in this paper has been improved by 13.88%, 11.43%, and 7.34%, compared with L2CSNet, FSE_L2CSNet, and FCBA_L2CSNet, respectively. This model not only improves the robustness of gaze estimation but also provides more accurate estimation results than the original model.

Keywords: gaze estimation algorithm; face feature extractor; feature information; PSA attention mechanism

Citation: Wei, J.; Wu, H.; Wu, Q.; Iwahori, Y.; Yu, X.; Wang, A. Gaze Estimation Method Combining Facial Feature Extractor with Pyramid Squeeze Attention Mechanism. *Electronics* **2023**, *12*, 3104. https://doi.org/10.3390/electronics12143104

Academic Editor: Byung Cheol Song

Received: 13 June 2023
Revised: 4 July 2023
Accepted: 11 July 2023
Published: 17 July 2023

1. Introduction

With the rapid development of technology and society, human–computer interaction [1,2] has improved people's work efficiency and quality of life, which plays a crucial role in many applications such as virtual reality [3] and medical research [4,5]. Eye detection and tracking is a challenging task in the field of computer vision. It provides useful information for human–computer interaction. The researchers have utilized gaze estimation to study behavior in many domains. It aims to infer people's focus of attention and behavioral intention by analyzing and understanding the fixation point and direction of the human eye. Sight is the nonverbal information in human communication, which can reveal people's interests, emotions, and cognitive process. Therefore, gaze estimation has broad application value in fields such as human–computer interaction, intelligent monitoring, medical diagnosis, and intelligent driving.

In human–computer interaction, gaze estimation can be used as an "eye mouse" to control the operation of the computer interface by capturing the fixation point of the human eye, achieving a more natural and intuitive interaction experience. This can be used to improve user interface design. By analyzing the user's gaze patterns and eye movements, it is possible to understand the user's attention allocation and interaction behavior toward

interface elements. This is very helpful for optimizing interface layout, enhancing the visibility of key features, and providing personalized interactive experiences. For example, in web design, gaze estimation can determine the user's fixation point and reading path when browsing a webpage, thereby adjusting the layout of content and the position of key information, making it easier for users to find the required information.

In terms of intelligent driving, firstly, gaze estimation can be used for driver fatigue and attention monitoring. By analyzing the driver's gaze patterns and eye movements, it is possible to detect whether they experience fatigue, distraction, or lack of concentration. Once the driver's attention drops, the system can issue a warning in a timely manner, reminding the driver to take necessary rest or attention adjustments to avoid traffic accidents. Secondly, gaze estimation can be used for driver behavior analysis. By analyzing drivers' gaze points and gaze directions during driving, we can understand their attention distribution on roads, traffic signs, and other vehicles. This can provide valuable data support for the development of driver behavior models and the optimization of driving decision systems. In addition, gaze estimation can also be used for monitoring the driver's emotional and cognitive state. By observing the driver's gaze patterns and eye movements, one can infer their emotional state, such as anxiety, fatigue, or distraction. This is very important for intelligent driving systems, as it can adjust driving strategies and provide personalized driving experiences based on the driver's emotional and cognitive state.

In terms of intelligent monitoring, gaze estimation can be used for behavior analysis and anomaly detection. By analyzing people's gaze patterns and eye movements, we can understand their points of interest and areas of interest in monitoring scenarios. This is very helpful for detecting suspicious behavior, abnormal activities, or potential threats. For example, in a shopping mall monitoring system, if someone's gaze frequently shifts to a specific area, it may indicate that they are engaging in theft or other improper behavior. Gaze estimation can help intelligent monitoring systems detect and alarm in a timely manner.

In medical diagnosis, the use of eye-tracking technology enables real-time monitoring of patients' fixation points and directions, providing important information about their attention allocation and cognitive processes. Eye line estimation can be used for ophthalmic diagnosis and treatment. By analyzing the patient's gaze patterns and eye movements, their eye coordination, eye movement function, and abnormal eye movements can be evaluated. This is very helpful for early detection and treatment of eye diseases. For example, in the diagnosis of strabismus, by observing the patient's fixation point and eye movements, the type and degree of strabismus can be determined, and an appropriate treatment plan can be developed. In addition, it can be used for neuroscience research and brain function localization. By analyzing patients' gaze patterns and eye movements, we can understand their attention allocation and cognitive processing processes toward stimuli. This is very helpful for studying the cognitive function of the brain and neurological diseases. For example, in the study of cognitive process, eyesight estimation can determine the cognitive load and attention distribution of patients under different tasks and stimuli, thus revealing the neural mechanism of a cognitive process.

2. Related Work

At first, human gaze estimation information is mainly obtained through mechanical, electrical signals, or traditional algorithms. T. Eggert et al. used a recognition algorithm for electrical signals to achieve human gaze estimation technology. This method mainly uses intelligent sensors to sample electrical signals from skin electrodes and represents the subject's eye movement through changes in this electrical signal [6]. This method requires connecting the device to the head position of the subject, which is relatively simple to implement. However, traditional devices are bulky, inconvenient for the subject to move, and have low accuracy. Later, with the continuous development of technology, many researchers began to use image processing algorithms to achieve gaze estimation. Zhang et al. proposed a nonlinear unscented Kalman filter for gaze estimation [7], which

can overcome the difficulties of nonlinear gaze estimation and improve the accuracy of gaze estimation. Jiashu Zhang et al. proposed to use several groups of points to match the posterior probability density function of eye movement [8], which is more accurate than the estimation effect of the traditional Kalman filter. Although nonlinear Kalman filtering algorithms can improve the accuracy of gaze estimation, they are numerically unstable in practical applications and require more computational time. The process of optimizing the structure is complex and has poor real-time performance. Amudha J et al. used deformable template matching to achieve various challenges in gaze estimation, which achieved a tracking accuracy of 98% [9]. L. Yu et al. proposed an eye gaze estimation method based on particle swarm optimization [10], which reduces the restrictive requirements for hardware and improves the practicability of the system. Although the gaze estimation technology based on genetic algorithm and particle swarm optimization algorithm can achieve accurate gaze estimation, the algorithm design is difficult, the parameter adjustment range is wide, and it is time-consuming and laborious.

With the continuous innovation of technology, deep learning methods have gradually attracted more and more attention from researchers. Compared to the Kalman filter algorithm, genetic algorithm, ant colony algorithm, and so on, deep learning advocates direct end-to-end solutions to the problem of gaze estimation, which can be classified into two types:

1. The model-based gaze estimation methods

The model-based gaze estimation is a state-of-the-art computer vision technique that harnesses the power of machine learning algorithms to simulate the intricate movement of human eyes and accurately predict gaze points. B. Yan et al. presented an innovative technique for estimating gaze direction using differential residual networks [11], which can measure the disparity in gaze between two eyes with higher precision. H. Zhang et al. developed a multitask network model to estimate eye gaze and blinks concurrently [12]. S. H. Kim et al. proposed a continuous engagement evaluation framework utilizing estimated gaze direction and facial expression data [13], substantiated by creating a database of gaze estimates. This method effectively expresses fine-grained state information. Liu et al. proposed a 3D gaze estimation model integrating an automatic calibration mechanism [14], which utilizes eye images to construct a 3D eye model, creating a point cloud with an RGBD camera. This method yields an average precision of 3.7° but requires expensive hardware. J. Ma et al. [15] proposed an eye-specific offset predictor for humans that enables predicting gaze estimation for multiple target figures. J. Zhuang et al. proposed adding a squeeze-and-excitation (SE) attention mechanism to ResNet networks to predict gaze points of flight simulator operators on the screen [16], which fills a void in research on multiscreen and multicamera systems. B. Saha et al. developed a real-time gaze estimation interface based on a subjective appearance method to achieve gaze estimation in a free environment using algorithms like decision trees and random forests [17], which has an accuracy of approximately 98%. However, the error varies in distinct environments and lighting conditions, and the system's stability is suboptimal. Chang C C et al. proposed a simple facial tagging function based on the YOLO model to locate two facial markers and six facial orientations [18]. It achieves an average precision of 99% while detecting facial orientations and facial features. However, the accuracy of locational facial marker points is questionable when subjects move significantly. Z. Wan et al. proposed a fractional perceptual gaze estimation method utilizing the actual pupil axis, which regresses the pupil's spherical coordinates normal to the gaze points [19]. This technique effectively solves the issue of pupil cornea refraction. However, the method did not detect the pupil contour, producing a less accurate model. H. Huang et al. presented a new framework: GazeAttentionNet [20], utilizing global and local attention modules to achieve gaze features. The method obtained high accuracy on a dataset with real-time validation. Nevertheless, the method's accuracy is low, with the average error reaching approximately 2 cm for mobile devices.

2. The appearance-based gaze estimation methods

The appearance-based gaze estimation is a cutting-edge computer vision technique employed to predict a person's gaze direction accurately. This advanced method relies on the analysis of appearance-based features surrounding the human eye to achieve gaze estimation. Zhao et al. proposed a monocular gaze estimation network utilizing mixed attention to predict gaze points from monocular features and their location information [21]. Wan et al. introduced a technique for estimating gaze direction utilizing one mapping surface [22], which accomplished eye-center calibration through the mapping surface, simplifying the model by making assumptions. Zhuang et al. developed a simplified network model for gaze estimation based on the LeNet neural network, called SLeNet [23], utilizing depth-separable convolution to decrease the number of convolution parameters and improve the speed of the model. Nonetheless, the model design of this technique is too simple, and the accuracy needs to be enhanced.

Zhang et al. presented a gaze data processing model based on artificial neural networks and face recognition techniques [24], which analyzed gaze estimation from the perspective of data processing, lessened the false recognition rate, and experimentally verified the effectiveness of the method. Murthy et al. designed an end-to-end gaze estimation system to anticipate gaze points employing infrared eyeglass images captured by wearable vision-estimating eyes [25], which realized comprehensive gaze estimation. Liu et al. developed a technique to predict the disparity in gaze between two eye input images of the same subject by directly training a differential convolutional neural network and predicting the direction of eye gaze using the estimated difference [26]. This method achieves high accuracy but involves a more complicated process of adjusting the parameters of the differential network when the subject's eyelids are closed or when the pupils are affected. Suzuki et al. proposed a sight-tracking dataset-based technique for estimating candidate regions for superimposed information in football videos [27]. Nevertheless, the accuracy of line-of-sight tracking decreases with an increase in superimposed information. Chang et al. proposed a gaze estimation technique based on YOLO and deep learning for localizing and detecting facial orientation by integrating appearance and geometric features [28], which achieved a robust accuracy of 88% without calibration. However, it suffered from slow real-time tracking due to its inability to locate facial feature coordinates. Senarath et al. proposed a three-stage, three-attention deep convolutional neural network for retail remote gaze estimation using image data [29]. Luo et al. proposed a collaborative network-based gaze estimation model with an attention mechanism that assigns appropriate weights between eye and facial features and achieved more accurate gaze estimation [30]. Han et al. proposed a pupil shape-based gaze estimation method, which used a deep network to learn gaze points by extracting various features from the image [31]. X. Song et al. proposed an end-to-end network using U-Net with residual blocks to retain eye features in high-resolution feature maps for efficient gaze estimation [32]. Zhao et al. proposed a unified network for simultaneous head detection and gaze estimation, where the two aspects share the same set of features to promote each other and enhance detection accuracy [33]. However, this method still had limitations in terms of accuracy, with an error of 19.62° at 23 fps.

In summary, the aforementioned algorithms faced two primary challenges. (1) The gaze estimation accuracy is greatly reduced when the subject's face is obscured or in a different environment with varying lighting conditions. (2) The designed network models or algorithms are simple in structure, leading to decreased gaze estimation accuracy.

To address the above problems, this paper proposes a new attention-based mechanism for the gaze estimation algorithm FPSA_L2CSNet, which integrates L2CSNet with facial feature extractor (FFE) and pyramid squeeze attention (PSA), which has the following contribution:

1. A facial feature extractor is integrated with L2CSNet, enabling retrieval of facial details, location of eye features, extraction of key eye points, and efficient narrowing of eye feature extraction range to enhance gaze estimation accuracy.

2. L2CSNet augmented with PSA incorporates multiscale spatial information and cross-channel attention into the model, selectively highlighting feature regions relevant to

vision estimation and suppressing irrelevant weights. Further, it facilitates granular level feature extraction, gaze direction feature extraction, and accurate vision estimation.

3. The proposed model is tested on a real-life dataset and four representative public datasets: the MPIIGaze dataset, the Gaze360 dataset, the ETH-XGaze dataset, and the GazeCapture dataset. The robustness and accuracy of the algorithm are verified by testing on individuals of varying nationalities, skin colors, ages, partial occlusion, genders and complex backgrounds, and lighting conditions.

3. Materials and Methods

First, this paper constructed a feature extractor to obtain the facial features, thereby narrowing down the range of eye feature extraction and accurately locating the key points of the eyes. Then, a high-precision gaze estimation algorithm [34] is proposed in this paper, which uses ResNet50 as the backbone network, uses a separate loss function to calculate the error of each gaze angle, and regresses the gaze result separately. Furthermore, to further improve the overall performance of the model and reduce the error of gaze results, this paper also introduces the PSA attention module, which combines multiscale spatial information and cross-channel attention into the model and selectively highlights the feature areas related to gaze estimation. Finally, the accuracy and robustness of the algorithm are verified in different environments. The algorithm flowchart of FPSA_L2CSNet is illustrated in Figure 1.

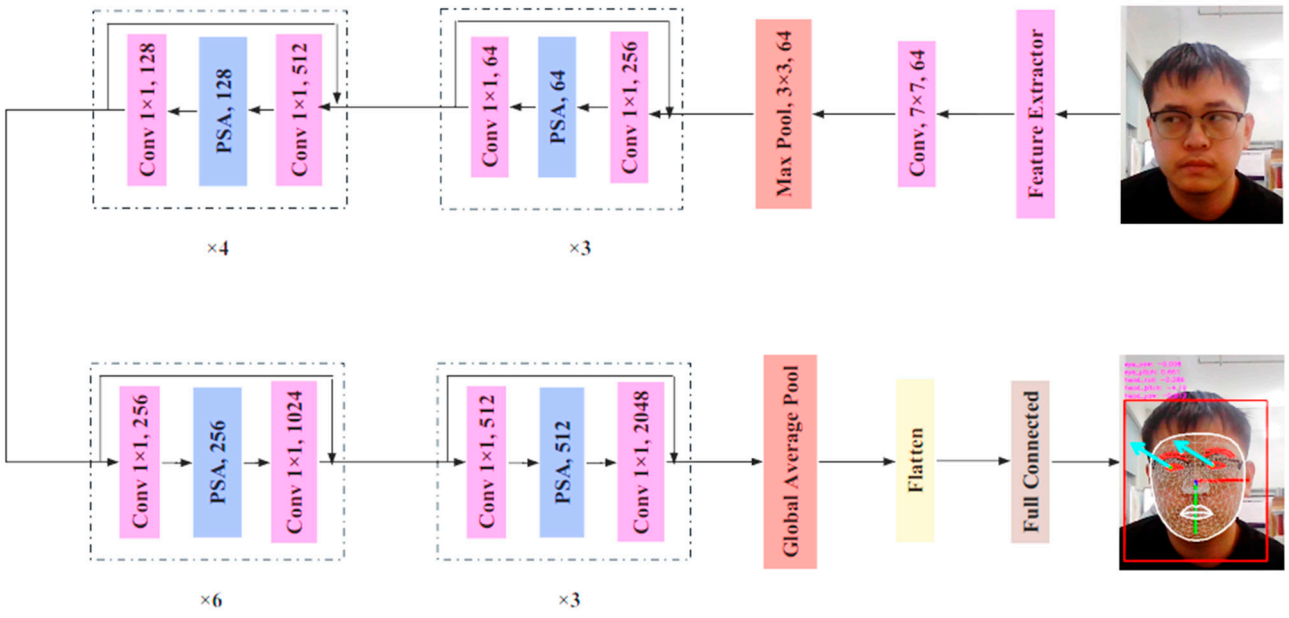


Figure 1. The flowchart of FPSA_L2CSNet model.

Therefore, this paper presents a novel attention-based mechanism, FPSA_L2CSNet, for gaze estimation. This approach leverages attention mechanisms to achieve more precise and efficient gaze estimation. The FPSA_L2CSNet algorithm is trained on different datasets, including MPIIGaze, Gaze360, ETH-XGaze, and GazeCapture, and its performance is thoroughly evaluated via comparison with other models. The results confirm the effectiveness and superiority of FPSA_L2CSNet in achieving more accurate and faster gaze estimation.

3.1. Acquisition of Facial Feature Points

As depicted in Figure 2, the FFE (facial feature extractor) in this study utilizes an RGB input image of 128 × 128 pixels to extract facial features. Due to the inherent variation in pixel composition within the dataset, we shall employ the transformative power of the resize () function to preprocess the images. By skillfully applying bilinear interpolation, we shall artfully scale the image, ultimately attaining a harmonious size of 128 × 128 pixels.

The results are obtained after passing through 5 single BlazeBlocks and 6 double BlazeBlocks. The main route of the single BlazeBlock consists of a 5×5 deep convolution and a 1×1 convolution, with the depth-separable convolution layer being the core component. In contrast to traditional convolution layers that perform convolution operations in both the spatial and channel directions, the depthwise separable convolutional layer focuses on these two directions separately to reduce the parameter amount. Specifically, the depthwise separable convolution uses the same convolution kernel to perform convolution on each input channel, resulting in a set of individual output channels. In this way, each convolution kernel is reused, enabling the reduction in a significant number of parameters. The element-wise convolution is the second part of the depthwise separable convolution, using a 1×1 convolution kernel to transform each individual output channel into the desired shape. In this way, the depthwise separable convolutional layer establishes the connection between depth and width. The side route of the single BlazeBlock consists of max pooling and channel pad, aiming to increase the convolution kernel size and cover the entire receptive field with fewer convolution layers. The skip connections in the single BlazeBlock allow the model to learn higher-level features that match the input better. Typically, cross-layer connections that span multiple layers help capture globally or partially informative features. Additionally, through skip connections, the single BlazeBlock can integrate with deeper grid structures to enhance the model's accuracy.

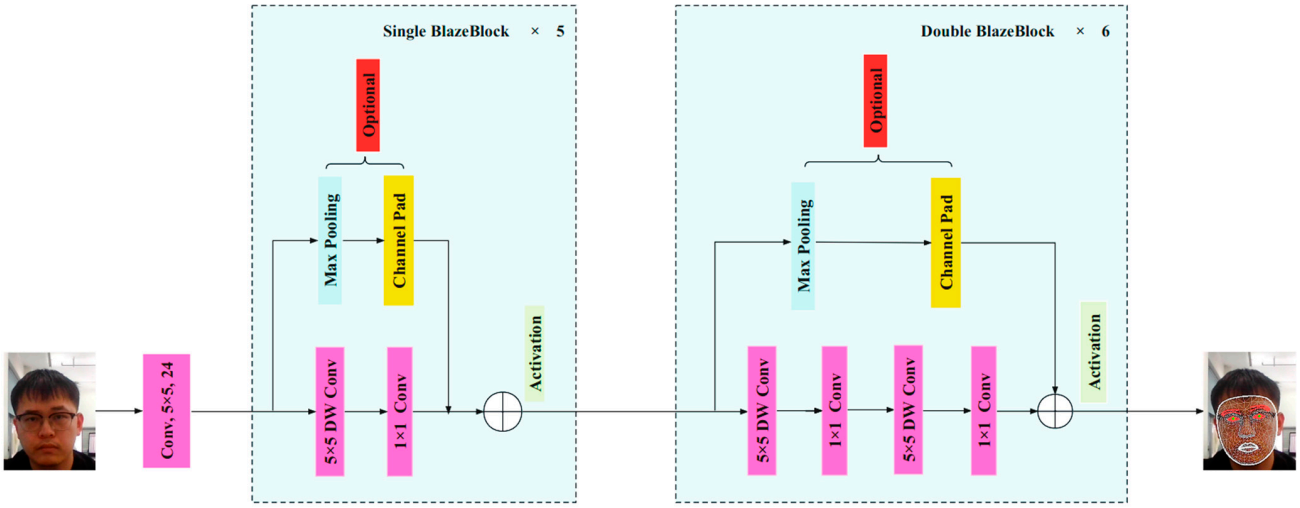


Figure 2. The structure of FFE.

As the network deepens and the features extracted become more advanced, the double BlazeBlock was designed on top of the single BlazeBlock. Compared to the Single BlazeBlock, it simply adds another 5×5 deep convolution and a 1×1 convolution to the main path. In the double BlazeBlock, the two single BlazeBlock modules are similar in composition to single BlazeBlock in that they both consist of a depthwise separable convolution layer and a skip connection. These modules perform convolutional operations on the input feature map to extract high-level features by using a depthwise separable convolutional layer. The skip connection, on the other hand, ensures that the model learns feature information at a shallower level and allows for a fine-grained combination of feature spectra. The two single BlazeBlock modules in double BlazeBlock are similar in composition to single BlazeBlock in that they both consist of a depth-separable convolutional layer and a skip connection. These modules perform convolutional operations on the input feature map to extract high-level features by using a deeply separable convolutional layer. The skip connection, on the other hand, ensures that the model learns feature information at a shallower level and allows for a fine-grained combination of feature spectra. In double BlazeBlock, after processing by two single BlazeBlock modules, the feature map size and computational burden are then further reduced using the downsampling module. This

allows the model to handle large-scale objects better, thus improving the accuracy of the network.

Therefore, the single BlazeBlock can be used for the shallow depth of the network, while the double BlazeBlock can be used for the deeper depth of the network. Through the aforementioned steps, facial feature information, including the feature areas for the left and right eyes, can be obtained.

3.2. ResNet50

Although designing deeper neural networks can lead to better recognition results in the process of deep network design, experiments indicate that as networks become increasingly deep, models will actually perform worse. While overfitting disturbance is eliminated, the real reason behind this problem comes from "vanishing gradients". Vanishing gradients are inherent defects in the backpropagation training algorithm. As the error is returned, gradients from earlier layers of the network will become increasingly smaller.

Equation (1) presents the loss function of the network, where X represents the input of the network, and W represents the weight parameters. The corresponding gradient values from backpropagation are given by Equation (2).

$$Loss = F(X, W) \tag{1}$$

$$\frac{\partial Loss}{\partial X} = \frac{\partial F(X, W)}{\partial X} \tag{2}$$

Furthermore, extending to multilayer networks, the loss function is given by Equation (3), where n denotes the number of layers in the network. According to the chain rule, the gradient of layer i can be derived, as shown in Equation (4). It can be observed that as the error is backpropagated, gradients from earlier layers of the network become progressively smaller.

$$Loss = F_n(X_n, W_n), L_n = F_{n-1}(X_{n-1}, W_{n-1}), \cdots, L_2 = F_1(X_1, W_1) \tag{3}$$

$$\frac{\partial Loss}{\partial X_i} = \frac{\partial F_n(X_n, W_n)}{\partial X_n} * \cdots * \frac{\partial F(X_{i+1}, W_{i+1})}{\partial X_i} \tag{4}$$

To address this issue, ResNet ingeniously introduced residual structures, as shown in Figure 3.

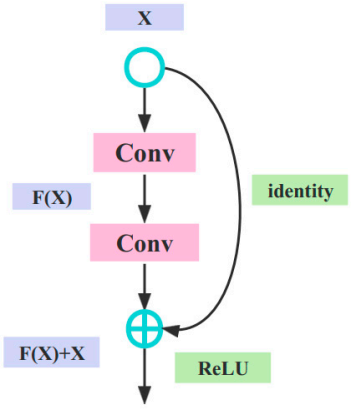


Figure 3. The residual block structure of ResNet.

This means that the output layer is $y = x + F(x)$, which implies that, in the process from Equation (4) to Equation (5), the gradient will not vanish even if the network becomes deeper.

$$\frac{\partial X_{i+1}}{\partial X_i} = \frac{\partial X_i + \partial F(X_i, W_i)}{\partial X_i} = 1 + \frac{\partial F(X_{i+1}, W_{i+1})}{\partial X_i} \tag{5}$$

The introduction of residual structures in ResNet allows for training neural networks that are deeper and more efficient, reducing the risk of overfitting while improving model performance. Specifically, there are several advantages to using residual structures: mitigate the problem of gradient disappearance, solve the problem of degradation, and avoid the problem of overfitting.

ResNet [35] has five different structures, namely ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152, representing different model depths and widths. The main differences between them are the depth and number of parameters of the models. ResNet-18 and ResNet-34 have the same basic structure and are relatively shallow networks. The basic structure of the last three models, ResNet-50, ResNet-101, and ResNet-152, is different from ResNet-18 and ResNet-34, representing deeper networks. Overall, the ResNet network structure has a large amount of computation and is complex, although the use of 1×1 convolution reduces the number of parameters. The depth of the network determines that the number of parameters is large. Therefore, choosing a network with too many layers may lead to overfitting and difficulty in convergence, while choosing a network with too few layers may result in poor training outcomes and low accuracy. Considering the above, in order to improve the accuracy of the network and prevent overfitting, ResNet50 is chosen as the backbone network in this paper. Its residual structure is shown in Figure 4.

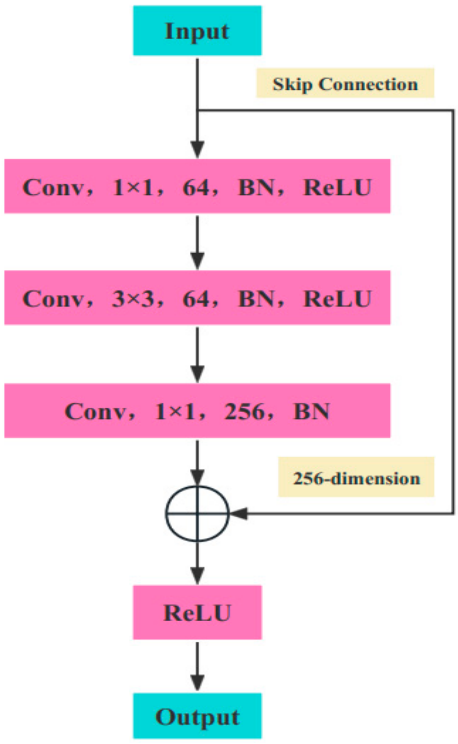


Figure 4. The residual structure of ResNet50.

This model comprises a main branch from top to bottom and skips connections on its right side. Its core idea is to transmit input information directly to the output via skip connections and to learn only the differences between the output and input through the main branch, thereby simplifying the learning process. In the main branch, 1×1 convolution is used to reduce the number of dimensions and generate 64 feature maps. Then, 3×3 convolution is performed to extract the main features. Subsequently, 1×1 convolution is executed to increase dimensionality and generate 256 feature maps. Additionally, batch normalization (BN) is utilized in ResNet to improve the model's generalization ability by subtracting the mean and dividing by the variance of the same batch of data. Furthermore, rectified linear units (ReLU), a nonlinear activation function, are employed to enhance the model's nonlinear adaptability.

Figure 5 shows the overall architecture of ResNet50, which adds branches between every two layers of the network and can sample input images with a convolutional sequence of 2. It is evident from the structure that there are four layers in the ResNet50 structure, and each layer is comprised of 3, 4, 6, and 3 residual blocks, respectively. The "3 4 6 3" in the basic block structure refers to the number of residual units in each layer, and these residual units are the basic subelements with interlayer connections. A large number of convolutional layers in ResNet are composed of these basic subelements arranged according to certain rules. Therefore, these numbers actually reflect the overall network depth of ResNet. The advantage of this design is that it enables ResNet to use the depth of the network more effectively and avoids the performance degradation problem that occurs when the number of layers increases in traditional deep networks.

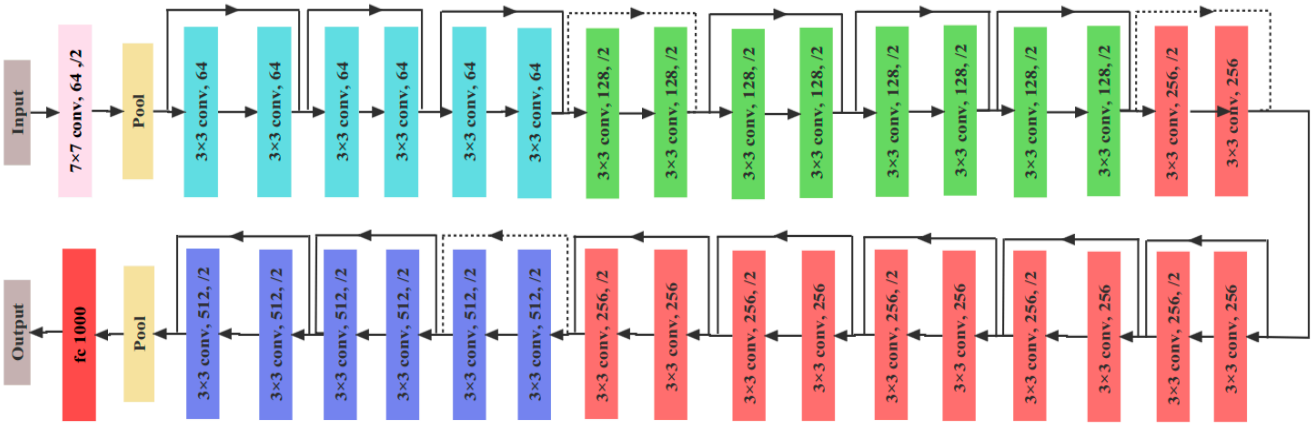


Figure 5. The overall architecture of ResNet50.

3.3. Attention Mechanism

As a data processing method in machine learning, the attention mechanism mainly focuses the network's learning on important areas. Studies have shown that embedding attention modules into convolutional neural networks can greatly improve network performance. In recent years, the method has been widely applied in computer vision, such as image classification, object detection, and semantic segmentation. Overall, attention mechanisms are mainly divided into channel attention and spatial attention; the most commonly used methods are the SE attention mechanism module [36] and the convolutional block attention module (CBAM) [37], respectively. The former has a simpler structure, consisting of two parts: squeeze and excitation. The main purpose of the squeeze part is to represent the importance of each channel feature, while the excitation part multiplies the feature map channels with the weighted channel importance obtained by the squeeze. Thus, it can make the model aware of the importance of each channel. However, the global average pooling of the squeeze module in the SE attention module is too simple to capture complex global information, and the fully connected layer of the excitation module increases model complexity, ignoring spatial information and leading to long calculation time and overfitting. The characteristic of the CBAM is that it focuses on the densest positions of effective information in an image, concatenates the feature maps generated by max pooling and average pooling, and applies convolutional and activation functions to form spatial attention maps. However, the CBAM has not captured spatial information at different scales to enrich the feature space, and its spatial attention only considers local information, failing to establish long-distance dependencies.

The PSA module not only has the ability to process tensors at multiple scales but also extracts spatial information at different scales by compressing the channel dimension of the input tensor. The PSA attention module [38] consists of four parts, as shown in Figure 6. First, the SPC module is used to obtain multiscale feature maps with channel diversity. Then, the SE weight module is used to obtain channel attention across multiple-

scale feature maps. Next, the softmax function is used to adjust channel attention again and obtain the weight of multiple-scale channels. Finally, element-wise multiplication is applied to the recalibrated weights and corresponding feature maps. Based on the above four steps, multiscale feature information can be outputted. In this structure, the spatial information of the input feature maps is obtained by using a multibranch method, which can obtain rich position information and conduct parallel processing for multiple scales. Moreover, the pyramid structure generates different spatial resolutions and depths of multiscale convolution kernels, effectively extracting spatial information at different scales on each channel feature map. Therefore, by integrating multiscale spatial information and cross-channel attention into the original model, multiscale spatial information can be extracted at a more granular level.

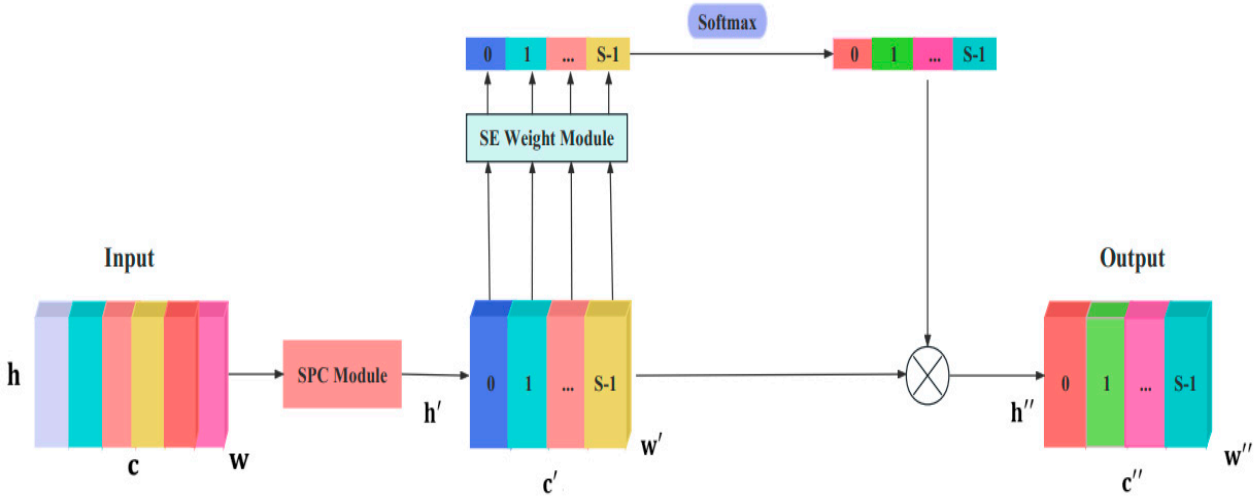


Figure 6. Structure of PSA module.

First, the SPC module is used to obtain feature maps with multiple channel scales. Although each input image in this structure has feature maps with different scales F_i, they have the same channel dimension C^i. Their relationship is expressed as follows:

$$C^i = \frac{C}{S} \quad i = 0, 1, \ldots, m-1 \tag{6}$$

Each branch independently learns multiscale spatial information, allowing them to establish cross-channel interactions in a local manner. The number of parameters changes with the increase in kernel size. In order to handle input tensors with different kernel scales, group convolution is used and applied to the convolution kernel. The quantity n is the kernel size, and Z is the group size. The relationship between them is expressed as follows:

$$Z = 2^{\frac{n-1}{2}} \tag{7}$$

As a result, the generation function for multiscale feature map is expressed as follows:

$$H_i = Conv(n_i \times n_i, Z_i)(X) \quad i = 0, 1, 2 \cdots, m-1 \tag{8}$$

The preprocessed multiscale feature map is expressed as follows:

$$H = Cat([H_0, H_1, \cdots, H_{m-1}]) \tag{9}$$

Next, the SE weight module is used to obtain attention across multiple-scale feature maps in order to obtain channel attention. The weight information and weight vector

for channel attention are obtained from the preprocessed multiscale feature maps. The attention weight vector is expressed as follows:

$$\boldsymbol{Q_i} = \boldsymbol{SEWeight}(\boldsymbol{H_i}),\ \boldsymbol{i} = 1, 2, 3, \cdots \boldsymbol{m} - 1 \tag{10}$$

The SEWeight module is used to obtain weights from input feature maps at different scales and then to fuse feature information of different scales. In addition, to achieve the interaction of feature information, the vectors across dimensions are fused. Therefore, the multiscale channel attention vector is obtained in the following way:

$$\boldsymbol{Q} = Q_0 \oplus Q_1 \oplus \cdots \oplus Q_{m-1} \tag{11}$$

Here, Q_i refers to the attention value of $\boldsymbol{H_i}$, and $\boldsymbol{Q}$ is the multiscale attention weight vector. Through the above steps, the interaction of local and global channel attention is achieved. Next, the channel attention vector can be obtained by fusing and concatenating channel attention. Its expression is written as follows:

$$Attention = Attention_0 \oplus Attention_1 \oplus \cdots \oplus Attention_{m-1} \tag{12}$$

Here, attention refers to the multiscale channel attention weight obtained after the interaction. Next, the weight of multiscale channel attention is multiplied by the feature map, expressed as follows:

$$T_i = F_i \odot Attention_i,\ i = 1, 2, 3, \cdots m - 1 \tag{13}$$

The above equation can effectively preserve the original feature map information. After simplification, the final input process is expressed as follows:

$$Out\ = Cat([T_0, T_1, \cdots, T_{m-1}]) \tag{14}$$

The PSA module is a novel attention mechanism that integrates multiscale spatial information and cross-channel attention into each block of feature groups. Thus, the PSA module can achieve better information interaction between local and global channel attention. Its advantages mainly include the following:

1. PSA has more refined weighting, which uses a more detailed monotonic regression algorithm to obtain the relative relationship between features, allowing it to more finely weight feature maps.

2. The PSA module can obtain the feature dependencies of each position in the feature map at different scales, thereby capturing the global and long-range features of the object better.

3. It can adapt to different resolutions: The PSA module adapts to different image inputs of varying resolutions by pyramid pooling, encoding different resolution feature maps through different sizes of pyramid layers, and then weighting them through corresponding convolution operations. This design improves the robustness of the model to different input resolutions.

The introduction of the PSA module can significantly improve the performance of the model in tasks such as object detection and image classification, even outperforming other powerful attention mechanisms. In Figure 7, the bottleneck module of ResNet consists of two 3×3 convolutions and one 1×1 convolution. The 1×1 convolution is mainly used for dimension reduction or expansion, and the 3×3 convolution is mainly used for feature extraction. The purpose of adding the attention module is to improve the feature extraction ability further, and thus, by simply replacing the 3×3 convolution of the bottleneck module with the PSA attention module, the EPSANet Bottleneck is obtained.

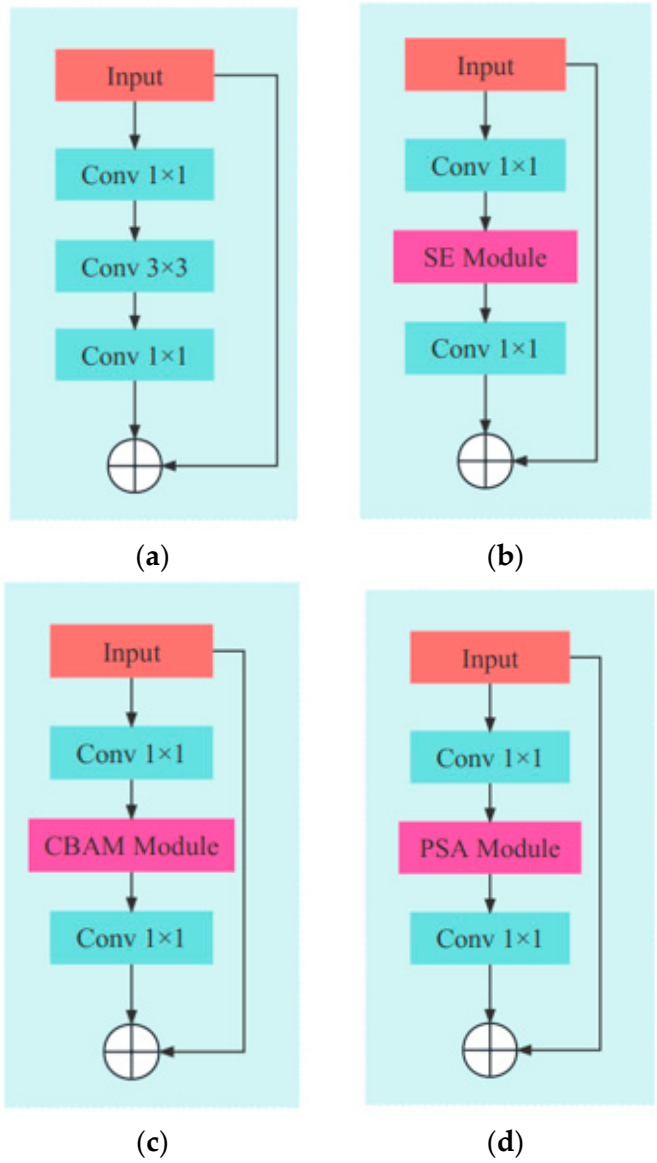


Figure 7. Comparison of bottleneck modules in different networks: (**a**) ResNet Bottleneck, (**b**) SENet Bottleneck, (**c**) CBAMNet Bottleneck, and (**d**) EPSANet Bottleneck.

Next, we adopt two identical loss functions for each gaze angle in this study. Each loss function is a linear combination of cross-entropy loss and mean squared error. The cross-entropy loss is defined as follows:

$$H(y,p) = -\sum_i \ y_i log p_i \tag{15}$$

The cross-entropy loss is defined as follows:

$$\boldsymbol{MSE(y,p)} = \frac{1}{N}\sum_0^N \ (\boldsymbol{y}-\boldsymbol{p})^2 \tag{16}$$

The proposed loss function for each gaze angle is a linear combination of the mean squared error and cross-entropy loss, defined as follows:

$$CLS(y,p) = H(y,p) + \alpha \cdot MSE(y,p) \tag{17}$$

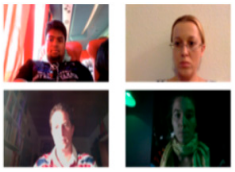
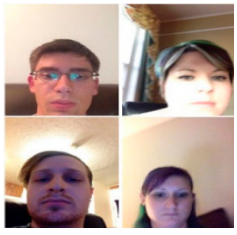
where CLS is the overall loss, p is the predicted value, y is the true value, and α is the regression coefficient.

4. Results

4.1. Selection of Datasets

The MPIIGaze [39] dataset contains 213,695 images of 15 subjects, each subject's facial images were collected at different times of the day with different mean gray intensity percentages of the facial regions. Forty-one dimensions were used to represent the subject gaze information in the dataset. The Gaze360 [40] dataset is the only dataset to obtain 3D gaze symbols of subjects in different indoor and outdoor environments, different lighting, different ages, different races, etc. The ETH-XGaze [41] dataset contains images of 110 subjects, consisting of over one million images of different gaze and extreme head postures. Eight cameras were used to capture the subjects' gaze postures, with an image resolution of up to 6000 × 4000 pixels, which is the highest image resolution dataset to date. The GazeCapture [42] dataset has 2,445,504 color images and contains 1450 2D gaze directions of subjects, which is the largest dataset to date. Table 1 gives the summary of the public datasets for gaze estimation.

Table 1. Summary of the dataset for gaze estimation.

Dataset	Source	Class	Number	Training Data	Test Data	Resolution	Samples
MPII Gaze	Max Planck Institute for Informatics	15	213,695	149,586	64,109	224 × 224	
Gaze360	Massachusetts Institute of Technology	238	172,000	120,400	51,600	3382 × 4096	
ETH-XGaze	ETH Zurich	110	1,532,658	1,072,861	459,797	6000 × 4000	
Gaze Capture	University of Georgia	1450	2,445,504	1,711,853	733,651	224 × 224	

4.2. Experimental Results and Analysis

The FPSA_L2CSNet, our novel network architecture, was implemented and evaluated on a cutting-edge hardware and software setup. The computational environment consisted of an i5-8300H CPU, 4 NVIDIA GeForce RTX2080Ti GPUs, 16 GB RAM, and leveraged CUDA version 10.2 for accelerated processing. The entire development process was conducted using the Python programming language. The network is trained using the Adam optimizer in the PyTorch framework on Ubuntu 20 with a learning rate of 0.00001. We trained our proposed network for 50 epochs using 64 batch sizes. Table 2 shows the comparison results of different methods on four datasets. It is found that the FPSA_L2CSNet proposed in this paper, which incorporates the PSA module, achieves better information interaction between local and global channel attention, and overcomes the lack of spatial information in the SE attention module and the inability of the CBAM to

establish long-range dependencies caused by only considering local information. Therefore, the final accuracy of FPSA_L2CSNet is significantly improved by 13.88%, 11.43%, and 7.34% compared to L2CSNet, FSE_L2CSNet, and FCBAM_L2CSNet, respectively.

Table 2. Comparison of gaze estimation results on four datasets.

Methods	MPIIGaze	Gaze360	ETH-XGaze	GazeCapture	Average Error
ITracker	5.57°	4.89°	10.56°	9.56°	7.65°
FullFace [43]	4.82°	4.32°	9.56°	9.32°	7.01°
DialatedNet [44]	4.68°	5.21°	10.43°	8.47°	7.19°
RTGene [45]	4.78°	5.39°	10.23°	8.98°	7.35°
FARNet [46]	4.29°	4.68°	9.87°	9.54°	7.09°
CANet [47]	4.09°	3.93°	8.97°	9.32°	6.58°
FAZE [48]	3.14°	4.54°	5.26°	3.01°	3.99°
L2CSNet	3.92°	3.89°	8.02°	6.67°	5.63°
FSE_L2CSNet	3.85°	3.52°	6.04°	6.83°	5.06°
FCBAM_L2CSNet	3.68°	3.45°	4.79°	4.79°	4.18°
FPSA_L2CSNet	3.41°	3.35°	3.73°	3.74°	3.56°

In particular, we conducted a comprehensive comparison with the state-of-the-art gaze estimators currently available. Our results revealed that on the MPIIGaze and GazeCapture datasets, the employment of meta-learning trained gaze estimation methods in FAZE led to a slightly lower error compared to the model proposed in this paper. However, upon evaluating FAZE on the Gaze360 and ETH-XGaze datasets, we observed a significantly higher error as opposed to our proposed model.

This disparity in performance can be attributed to the diversity of participants in the Gaze360 and ETH-XGaze datasets, hailing from various regions worldwide, including the United States, the Middle East, Europe, Africa, and Asia. In contrast, the MPIIGaze and GazeCapture datasets predominantly consist of participants from the United States and Europe. Consequently, the FAZE model's limited generalization ability and reliance on a relatively singular dataset structure hinder its adaptability in diverse regions.

The resulting error in the Gaze360 dataset was measured at 4.54°, while our model's error, as proposed in this article, stood at 3.35°. Similarly, the ETH-XGaze dataset yielded an error of 5.26° for FAZE, whereas our model demonstrated an error of 3.73°. Overall, across all four datasets, FAZE exhibited an average error of 3.99°, while our model showcased an average error of 3.56°. Evidently, the method presented in this paper exhibits superior adaptability to public test datasets, showcasing robustness, universality, and stability.

To obtain satisfactory results, this study compares the FPSA_L2CSNet model with L2CSNet, FSE_L2CSNet, and FCBAM_L2CSNet models. The results obtained by testing on the MPIIGaze, Gaze360, ETH-Xgaze, and GazeCapture datasets are shown in Figure 8. In Figure 8a, it can be seen that on the MPIIGaze dataset, due to the addition of the PSA module in the FPSA_L2CSNet network, the input image receives better information interaction between local and global channel attention and overcomes the lack of spatial information of the SE attention module and the inability of the CBAM to establish long-range dependencies caused by only considering local information. Therefore, the network is in a convergent state at the 10th epoch with the lowest average error.

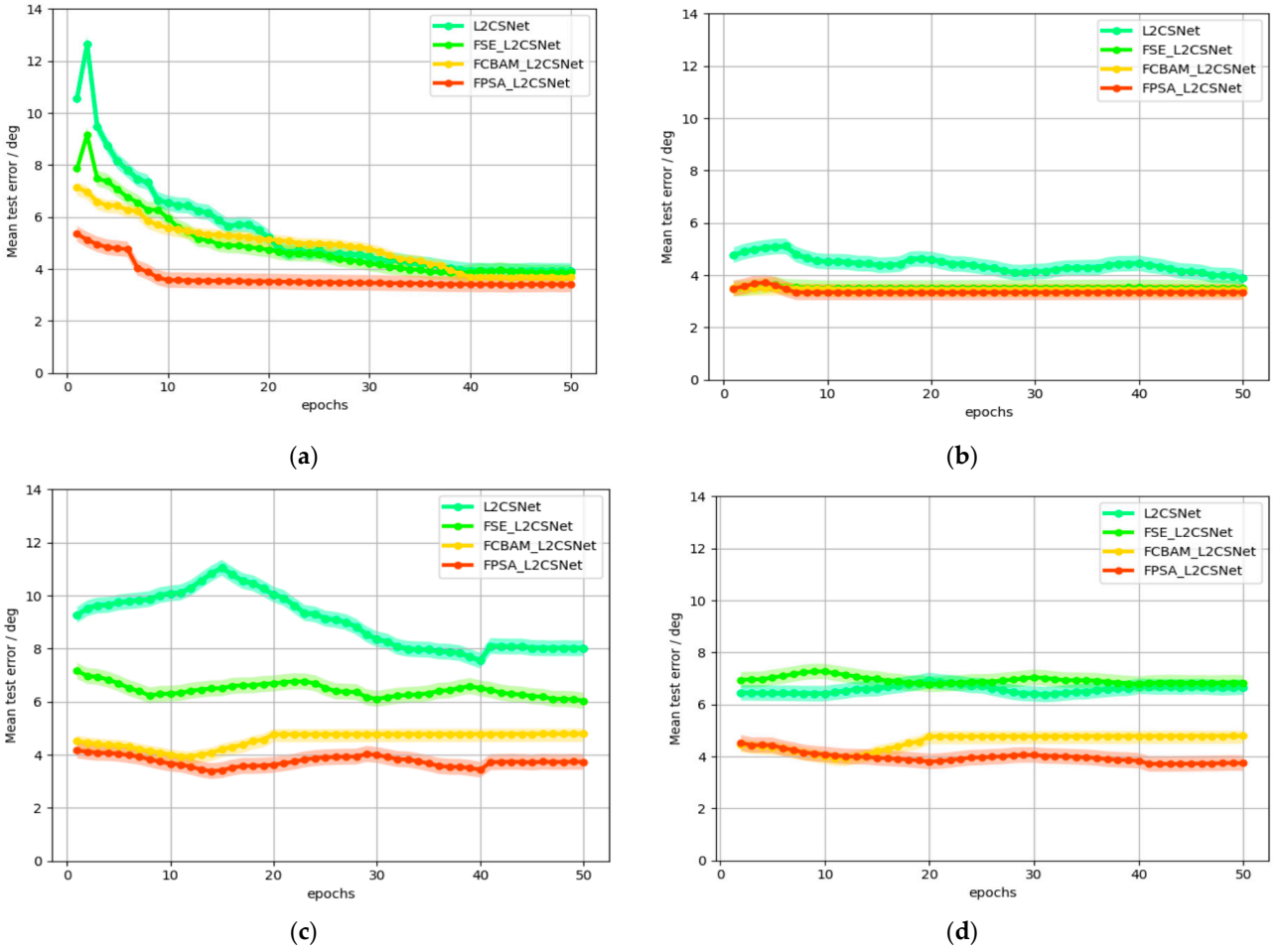


Figure 8. The comparison of gaze estimation accuracy on four datasets: (**a**) MPIIGaze, (**b**) Gaze360, (**c**) ETH-XGaze, and (**d**) GazeCapture.

In Figure 8b, on the Gaze360 dataset, since the FPSA_L2CSNet, FSE_L2CSNet, and FCBAM_L2CSNet added attention modules to the original network, the convergence of the three models is achieved at the 8th epoch, while L2CSNet gradually converges at the 45th epoch. The convergence speed and error of the former three are better than those of L2CSNet. In Figure 8c, on the ETH-XGaze dataset, because there is no added attention module, the oscillations and fluctuations of L2CSNet are the largest, and its error is also the largest. The FSE_L2CSNet ignores spatial information, leading to a large error, while the error of FCBAM_L2CSNet increases after decreasing because the accuracy decreases after saturation as the network depth increases. The architecture of FPSA_L2CSNet is better than the other three, so it has the lowest error. In Figure 8d, on the GazeCapture dataset, the error of FSE_L2CSNet is basically the same as that of L2CSNet because the feature information extraction is insufficient after spatial information is ignored. FCBAMNet has the problem of having too deep a network, leading to a rise in error after decreasing, while FPSA_L2CSNet has the lowest error.

In the testing phase, this article will test the above four methods on the four public datasets. The performance of the four methods varies on different datasets. In this paper, a thermodynamic diagram is used to analyze the error, and the results of the MPIIGaze dataset are shown in Figure 9. The horizontal axis represents the yaw angle, while the vertical axis represents the pitch angle. As the color changes from cyan to red, the error gradually increases. Therefore, the error size of this method is measured by calculating the proportion of cyan areas in the total area. The more cyan areas, the smaller the error, and vice versa. The specific results are shown in Table 3. From the table, it can be seen

that in the four datasets, FPSA_ L2CSNet has the highest proportion of cyan areas and the smallest error.

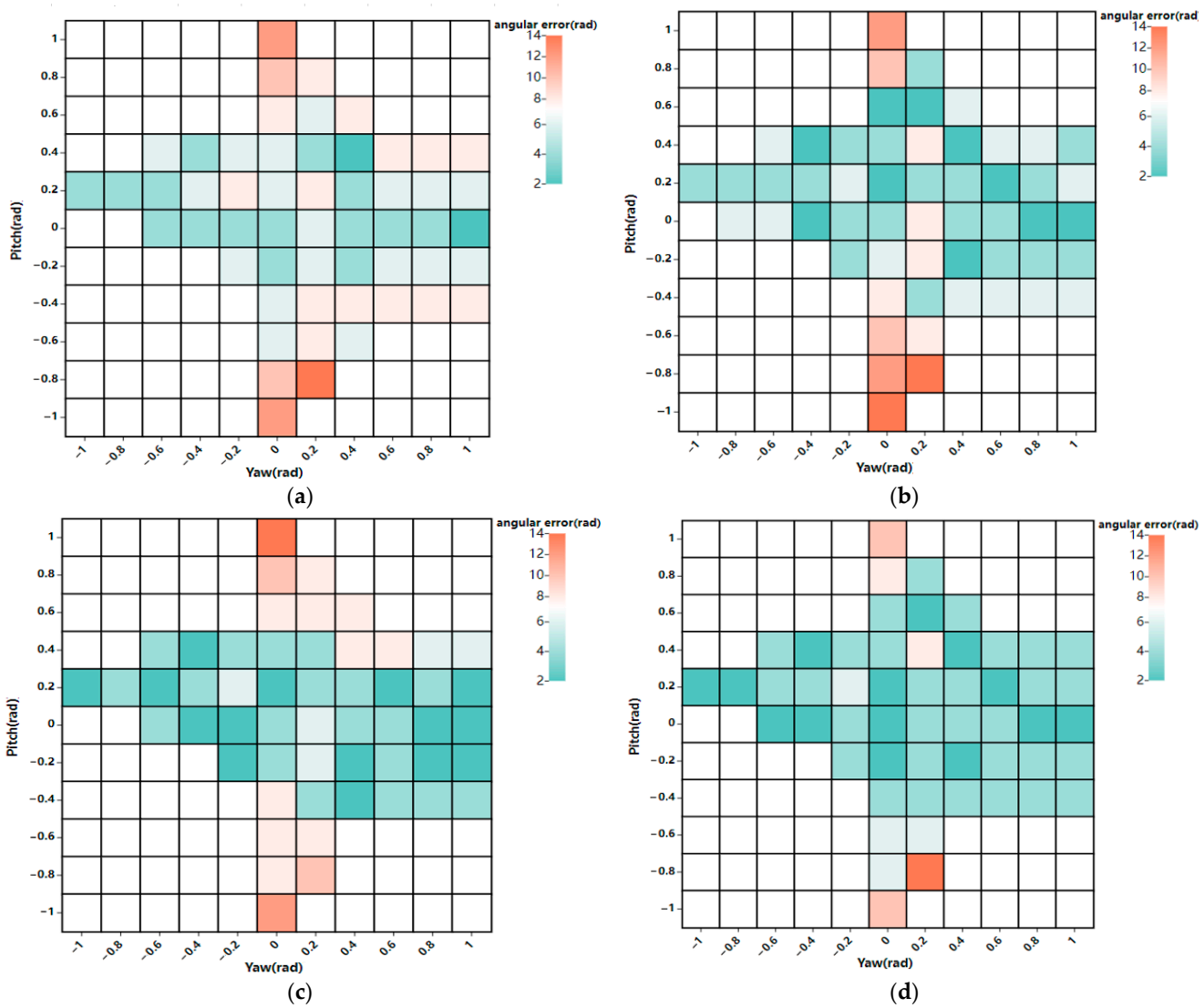


Figure 9. The comparison of gaze estimation error on four datasets: (**a**) L2CSNet, (**b**) FSE_L2CSNet, (**c**) FCBAM_L2CSNet, and (**d**) FPSA_L2CSNet.

Table 3. The accuracy comparison results of different gaze estimation methods.

Methods	MPIIGaze	Gaze360	ETH-XGaze	GazeCapture
L2CSNet	59.65%	63.98%	60.49%	53.62%
FSE_L2CSNet	65.35%	68.26%	66.25%	65.12%
FCBAM_L2CSNet	76.56%	73.45%	69.23%	70.25%
FPSA_L2CSNet	87.28%	83.64%	74.36%	75.95%

Table 3 gives the accuracy comparison results of different gaze estimation methods. It can be seen from the results that the SE attention mechanism may excessively pursue the weighting of useful channels and ignore the information of low-level channels, resulting in overfitting of the model, loss of generalization ability, and decline in accuracy. The CBAM mechanism requires adjustment factors for learning channel attention and spatial attention, leading to an increase in the number of parameters and a decrease in the learning rate, which invisibly affects the accuracy of the model. The method proposed in this article comprehensively measures local and global attention mechanisms, which can help the

model better understand input data while paying attention to features at different scales and synthesizing their information to generate total attention, reducing unnecessary errors caused by fluctuations in information at a single scale, improving the robustness of the model, and using features at different scales to help the model better understand input data. In summary, FPSA_L2CSNet has the highest gaze accuracy and the strongest performance on all four datasets, with the smallest error on the four public datasets.

Finally, this proposed gaze estimation model selects real individuals of different backgrounds, lighting, nationalities, skin tones, ages, partial occlusions, and genders to gaze in different directions in the lab to prove progressiveness and effectiveness. Among them, in complex background conditions, the model can accurately obtain the gaze direction when the testers with obvious differences in skin tone and from different countries in images. For example, the girl in Figure 10a comes from Mongolia with yellow skin, and Figure 10f comes from Madagascar with black skin. For testers under bright lighting conditions and gazing in different directions, the model can accurately estimate the gaze direction as seen in Figure 10b,d. Even when the tester in image Figure 10c is under the dimmest lighting conditions compared to other testers, the model still accurately estimates the gaze direction. Moreover, even when the facial features of the tester in image Figure 10e are partially occluded and the gaze is to the lower left, the accuracy of the estimated gaze direction does not exhibit significant deviation due to the precise extraction of facial feature points. In summary, the robustness, universality, and accuracy of this model have been fully validated.

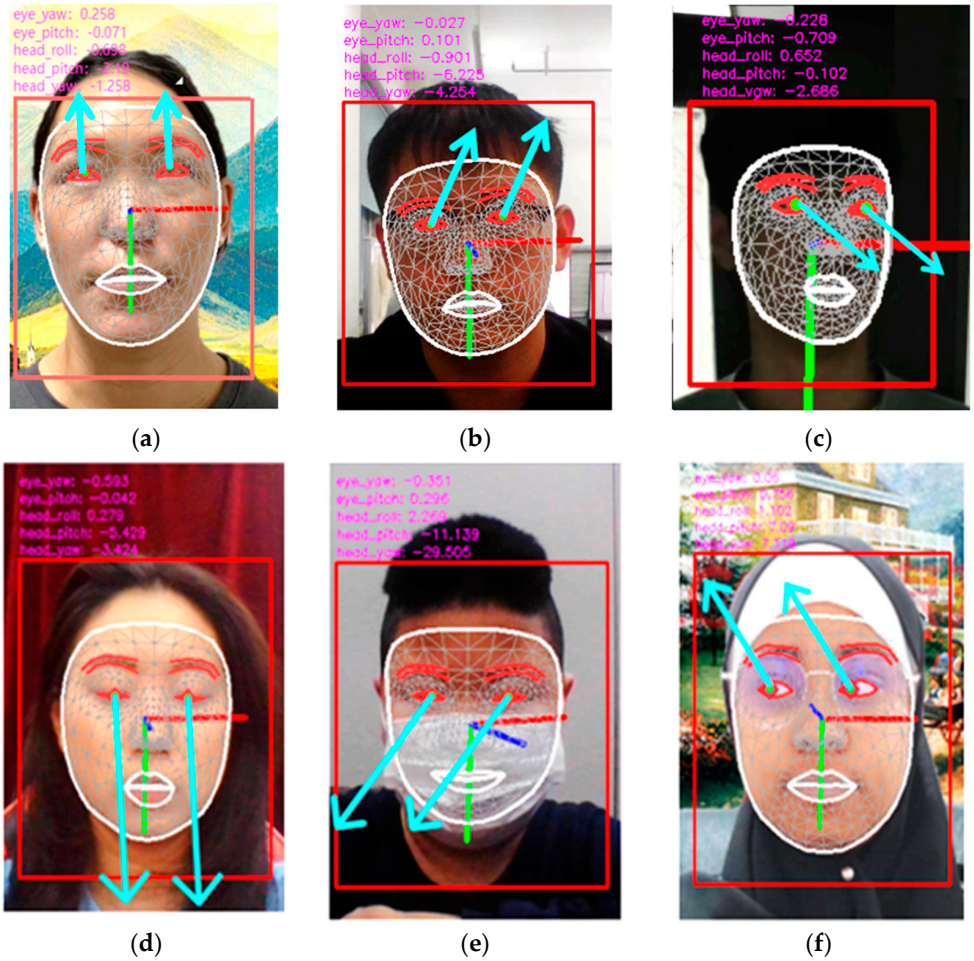


Figure 10. The results of gaze estimation on real datasets: (**a**) upper; (**b**) upper right; (**c**) lower right; (**d**) lower; (**e**) lower left; (**f**) upper left.

5. Conclusions

To address the challenge of low gaze estimation accuracy among individuals in different environments, this study proposes a gaze estimation model based on attention mechanisms: FPSA_L2CSNet. The results obtained from testing on the public dataset show that this model exhibits the lowest error and highest accuracy on the MPIIGaze dataset, with errors of 3.41°, 3.35°, 3.73°, and 3.74°, respectively. Compared with L2CSNet, FSE_L2CSNet, and FCBAM_L2CSNet, the accuracy is improved by 13.88%, 11.43%, and 7.34%, respectively. The results obtained from testing on real-world data show that this proposed model achieves precise gaze estimation under different lighting, background, and partial occlusion conditions for individuals of different nationalities, skin tones, ages, and genders, thus verifying the robustness and universality of the model.

In the future, PSA's attention mechanism could be further enhanced to amalgamate diverse levels of gaze estimation features, which would enable network models to concentrate on line-of-sight information, thereby improving the resilience and precision of the model.

Author Contributions: Conceptualization, J.W., X.Y., Q.W., Y.I., H.W. and A.W.; methodology, J.W. and X.Y.; software, J.W.; validation J.W.; writing—review and editing X.Y., Q.W., Y.I., H.W. and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the high-end foreign experts' introduction program (G2022012010L) and Major Science and Technology Projects of Zhongshan City in 2022 (2022A1020).

Data Availability Statement: MPII Gaze: http://datasets.d2.mpi-inf.mpg.de/MPIIGaze/MPIIGaze.tar.gz; Gaze360: http://gaze360.csail.mit.edu/download.php; ETH-XGaze: https://ait.ethz.ch/xgaze?query=eth; GazeCapture: https://gazecapture.csail.mit.edu/download.php.

Acknowledgments: Qing Wu acknowledges the National Natural Science Foundation of China (Grant No. 62205091), the China Postdoctoral Science Foundation Funded Project (Grant No. 2022M710983), and HeiLongJiang Postdoctoral Foundation (Grant No. LBH-Z22201). The study was supported by the Fundamental Research Foundation for Universities of Heilongjiang Province (2022-KYYWF-0121).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bulling, A.; Gellersen, H. Toward Mobile Eye-Based Human-Computer Interaction. *IEEE Pervasive Comput.* **2010**, *9*, 8–12. [CrossRef]
2. Drewes, H. Eye gaze Tracking for Human Computer Interaction. Ph.D. Thesis, LMU, Munich, Germany, 2010; pp. 9–16.
3. Patney, A.; Kim, J.; Salvi, M. Perceptually-based foveated virtual reality. In Proceedings of the ACM SIGGRAPH 2016 Emerging Technologies, Anaheim, CA, USA, 24–28 July 2016; pp. 1–2.
4. Novak, D.; Riener, R. Enhancing patient freedom in rehabilitation robotics using gaze-based intention detection. In Proceedings of the IEEE International Conference on Rehabilitation Robotics, Seattle, WA, USA, 24–26 June 2013; pp. 123–126.
5. Atkins, M.S.; Tien, G.; Khan, R.S.A.; Meneghetti, A.; Zheng, B. What do surgeons see: Capturing and synchronizing eye gaze for surgery applications. *Surg. Innov.* **2013**, *6*, 14–21. [CrossRef] [PubMed]
6. Eggert, T. Eye movement recordings: Methods. *Neuro-Ophthalmology* **2007**, *40*, 15–34.
7. Zu-Tao, Z.; Jia-Shu, Z. Sampling strong tracking nonlinear unscented Kalman filter and its application in eye tracking. *Chin. Phys. B* **2021**, *6*, 89–95. [CrossRef]
8. Zhang, J.; Zhang, Z. Application of a strong tracking finite-difference extended kalman filter to eye tracking. In Proceedings of the Intelligent Computing: International Conference on Intelligent Computing, Kunming, China, 16–19 August 2006; pp. 65–68.
9. Amudha, J.; Chandrika, K.R. Suitability of Genetic Algorithm and Particle Swarm Optimization for Eye Tracking System. 6th International Conference on Advanced Computing (IACC). *IEEE Access* **2016**, *23*, 165–168.
10. Yu, L.; Xu, J.; Huang, S. Eye-gaze tracking system based on particle swarm optimization and BP neural network. In Proceedings of the 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China, 12–15 June 2016; pp. 1269–1273.
11. Yan, B.; Tang, X. Gaze Estimation Based on Difference Residual Network. In Proceedings of the2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI), Kunming, China, 17–19 September 2021; pp. 651–655.
12. Zhang, H.; Wang, X.; Ren, W.; Noack, B.R.; Liu, H. Improving the Reliability of Gaze Estimation through Cross-dataset Multi-task Learning. In Proceedings of the 2022 International Conference on High Performance Big Data and Intelligent Systems, Tianjin, China, 10–11 December 2022; pp. 202–206.

13. Kim, S.H.; Lee, D.J.; Kim, D.H.; Song, B.C. Continuous Engagement Estimation based on Gaze Estimation and Facial Expression Recognition. In Proceedings of the 2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Phuket, Thailand, 5–8 July 2022; pp. 937–939.
14. Liu, M.; Li, Y.; Liu, H. 3D Gaze Estimation for Head-Mounted Eye Tracking System with Auto-Calibration Method. *IEEE Access* **2020**, *8*, 104207–104215. [CrossRef]
15. Ma, J.; Zhang, X.; Wu, Y.; Hedau, V.; Chang, S.-F. Few-Shot Gaze Estimation with Model Offset Predictors. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 4893–4897.
16. Zhuang, J.; Wang, C. Attention Mechanism Based Full-face Gaze Estimation for Human-computer Interaction. In Proceedings of the 2022 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 23–25 September 2022; pp. 6–10.
17. Saha, B.; Islam, M.J.; Dipto, A.S.; Mostaque, S.K. An Efficient Approach for Appearance Based Eye Gaze Estimation with 13 Directional Points. In Proceedings of the 2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, Rajshahi, Bangladesh, 26–27 December 2021; pp. 1–5.
18. Chang, C.C.; Ou, W.L.; Chen, H.L. Detection of Facial Directions and Features With YOLO-Based Deep-Learning Technology for Pre-Processing of Gaze Estimation. In Proceedings of the 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 18–21 October 2022; pp. 284–285.
19. Wan, Z.; Xiong, C.; Chen, W.; Zhang, H.; Wu, S. Pupil-Contour-Based Gaze Estimation With Real Pupil Axes for Head-Mounted Eye Tracking. *IEEE Trans. Ind. Inform.* **2022**, *18*, 3640–3650. [CrossRef]
20. Huang, H.; Ren, L.; Yang, Z.; Zhan, Y.; Zhang, Q.; Lv, J. GAZEATTENTIONNET: Gaze Estimation with Attentions. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23–27 May 2022; pp. 2435–2439.
21. Zichen, Z.; Lai, W.; Xiaofeng, L.; Zhi, L. Monocular Gaze Estimation Network based on Mixed Attention. In Proceedings of the 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA), Changchun, China, 20–22 May 2022; pp. 377–380.
22. Wan, Z.; Xiong, C.; Li, Q. Accurate Regression-Based 3D Gaze Estimation Using Multiple Mapping Surfaces. *IEEE Access* **2020**, *8*, 166460–166471. [CrossRef]
23. Zhuang, Y.; Zhang, Y.; Zhao, H. Appearance-based gaze estimation using separable convolution neural networks. In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 March 2021; pp. 609–612.
24. Zhang, Y.; Sun, J. Research and Application of Gaze Data Processing Method. In Proceedings of the 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 17–19 June 2022; pp. 885–890.
25. Murthy, L.R.; Biswas, P. Deep Learning-based Eye Gaze Estimation for Military Aviation. In Proceedings of the 2022 IEEE Aerospace Conference (AERO), Big Sky, MT, USA, 5–12 March 2022; pp. 1–8.
26. Liu, G.; Yu, Y.; Mora, K.A.F.; Odobez, J.-M. A Differential Approach for Gaze Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *43*, 1092–1099. [CrossRef] [PubMed]
27. Suzuki, G.; Takahashi, S.; Ogawa, T.; Haseyama, M. An Estimation Method of Candidate Region for Superimposing Information Based on Gaze Tracking Data in Soccer Videos. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics—Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 28–30 September 2020; pp. 1–2.
28. Chang, C.C.; Ou, W.L.; Chen, H.L. YOLO-Based Deep-Learning Gaze Estimation Technology by Combining Geometric Feature and Appearance Based Technologies for Smart Advertising Displays. In Proceedings of the 2022 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE), Tainan, Taiwan, 7–10 November 2022; pp. 1–3.
29. Senarath, S.; Pathirana, P.; Meedeniya, D.; Jayarathna, S. Customer Gaze Estimation in Retail Using Deep Learning. *IEEE Access* **2022**, *10*, 64904–64919. [CrossRef]
30. Luo, Y.; Chen, J. CI-Net: Appearance-Based Gaze Estimation via Cooperative Network. *IEEE Access* **2022**, *10*, 78739–78746. [CrossRef]
31. Han, S.Y.; Cho, N.I. User-Independent Gaze Estimation by Extracting Pupil Parameter and Its Mapping to the Gaze Angle. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1993–2000.
32. Song, X.; Guo, S.; Yu, Z.; Dong, J. An Encoder-Decoder Network with Residual and Attention Blocks for Full-Face 3D Gaze Estimation. In Proceedings of the 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China, 26–28 July 2022; pp. 713–717.
33. Zhao, K.; Hu, Z.; Zhang, Q.; Liu, J. RTHG: Towards Real-Time Head Detection And Gaze Estimation. In Proceedings of the 2022 IEEE International Conference on Robotics and Biomimetics, Jinghong, China, 5–9 December 2022; pp. 735–740.
34. Abdelrahman, A.A.; Hempel, T.; Khalifa, A. L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environment. *arXiv* **2022**, arXiv:2203.03339.
35. He, K.; Zhang, X.; Ren, S. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 56–65.
37. Woo, S.; Park, J.; Lee, J.Y. CBAM: Convolutional block attention module. In Proceedings of the Computer Vision-ECCV, Munich, Germany, 8–14 September 2018; pp. 3–19.
38. Zhang, H.; Zu, K.; Lu, J. EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network. In Proceedings of the Asian Conference on Computer Vision, Macau, China, 4–8 December 2022; pp. 1161–1177.
39. 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 (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4511–4520.
40. Kellnhofer, P.; Recasens, A.; Stent, S. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2020; pp. 195–198.
41. Zhang, X.; Park, S.; Beeler, T. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In Proceedings of the Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 156–159.
42. Krafka, K. Eye Tracking for Everyone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2176–2184.
43. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. It's written all over your face: Fullface appearance-based gaze estimation. In Proceedings of the Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 2299–2308.
44. Chen, Z.; Shi, B.E. Appearance-based gaze estimation using dilated-convolutions. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 309–324.
45. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 334–352.
46. Cheng, Y.; Zhang, X.; Lu, F.; Sato, Y. Gaze estimation by exploring two-eye asymmetry. *IEEE Trans. Image Process.* **2020**, *29*, 5259–5272. [CrossRef] [PubMed]
47. 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, New York, NY, USA, 7–12 February 2020; pp. 10623–10630.
48. Park, S.; Molchanov, P.; Kautz, J. Few-Shot Adaptive Gaze Estimation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9367–9376.