electronics

MDPI

Article

Pupil Localization Algorithm Based on Improved U-Net Network

Gongzheng Chen *, Zhenghong Dong, Jue Wang and Lurui Xia

Graduate School, Space Engineering University, Beijing 101416, China
* Correspondence: chengongzheng64@163.com

Abstract: Accurately localizing the pupil is an essential requirement of some new human–computer interaction methods. In the past, a lot of work has been done to solve the pupil localization problem based on the appearance characteristics of the eye, but these methods are often specific to the scenario. In this paper, we propose an improved U-net network to solve the pupil location problem. This network uses the attention mechanism to automatically select the contribution of coded and uncoded features in the model during the skip connection stage of the U-net network in the channel and spatial axis. It can make full use of the two features of the model in the decoding stage, which is beneficial for improving the performance of the model. By comparing the sequential channel attention module and spatial attention module, average pooling and maximum pooling operations, and different attention mechanisms, the model was finally determined and validated on two public data sets, which proves the validity of the proposed model.

Keywords: eye localization; attentional mechanism; U-net; skip connection

1. Introduction

As computers become increasingly prevalent in modern society, the focus of people's interaction technology is shifting from the computer as the center to the human center, and cross-domain man–machine barrier technology has become a new research hotspot. Eye movement tracking technology is a novel type of human–computer interaction technology, whose principal interaction mode is staring interaction, though it has also developed many interaction modes. Currently, gaze interaction has been successfully applied in various fields, such as human–machine interface, virtual reality, medical care, and so on. Gaze interaction is divided into two phases, pupil localization, and fixation locus description. Accurate pupil localization is one of the most critical and fundamental requirements of eye-tracking technology, and it is an essential component of the human–computer interaction task.

Pupil localization is subject to many factors, such as the shape of the eye and light conditions. Early eye movement detection products use infrared cameras for detection. In these products, the corneal reflex is used to estimate the pupil center, resulting in high accuracy of pupil position. However, these products have several limitations, such as expensive devices and, most importantly, they are cumbersome to wear and cause eye irritation. In recent years, computer vision has gradually entered the public field of vision and has been widely used in various image tasks. Unlike professional equipment, computer vision technology can directly determine the location of the pupil, which has therefore attracted the attention of researchers. Although detecting pupils using a non-wearable camera is easy in ordinary scenes, obstacles such as image brightness, angle of view, and resolution are still obstacles to improving the accuracy of pupil detection. To address this, a Fully Convolution Network (FCN) has been successfully applied to the semantic segmentation task, which is similar to pupil location. During training, a face picture and a heat map of pupil positions are used as the input to the FCN, with the predicted heat map

Citation: Chen, G.; Dong, Z.; Wang, J.; Xia, L. Pupil Localization Algorithm Based on Improved U-Net Network. *Electronics* **2023**, *12*, 2591. https://doi.org/10.3390/electronics12122591

Academic Editor: Gemma Piella

Received: 11 May 2023
Revised: 6 June 2023
Accepted: 6 June 2023
Published: 8 June 2023

converted into pupil position during testing. The main framework used in this paper is the U-net network, which is improved through the attention mechanism. This architecture can make full use of coded features and uncoded features in the skip connection stage of the U-net network by using the attention mechanism, rather than simply connecting the two features as in the original paper [1]. For each stage, the contributions of coded features and uncoded features to the model are different, but the architecture provided in this paper automatically selected them on the channel and spatial axis at each skip connection stage of the U-net network.

Our contributions are as follows:

1. Use of the attention mechanism to connect coded features and uncoded features in the skip connection stage of the U-net network.
2. Our module can not only use the attention mechanism to learn the noteworthy information from both coded features and uncoded features on the channel and spatial axis but also automatically adjust the contribution of coded features and uncoded features to the model by the Softmax function. The model was finally determined by comparing the sequential channel attention module (CAM), spatial attention module (SAM), and average pooling operations and maximum pooling operations, and selecting different attention modules on model performance.
3. The effectiveness of the model has been verified in BioID [2] and GI4E [3] data sets and achieved competitive performance.

The rest of this paper is organized as follows. In Section 2, we review the related work on pupil location, encompassing traditional methods, deep learning, and attention mechanisms. In Section 3, the structure is presented in detail. In Section 4, we discuss the sequential of CAM and SAM, the influence of average pooling and maximum pooling, and different attention mechanisms on the model accuracy. We also compared our proposal on two datasets. Finally, in Section 5, we conclude the manuscript with some final remarks.

2. Related Work

This chapter is divided into three aspects to introduce the related work on pupil localization: traditional methods of pupil localization, the pupil localization method based on deep learning, and the attention mechanism.

2.1. Traditional Methods of Pupil Localization

In the early stage, pupil location was mainly based on eye appearances, such as eye color and geometric features. For instance, Young et al. [4] used Hough transform to detect the iris to track eyes. However, this method was not popularized because it requires head-mounted cameras, which were only suitable for specific scenes. Skodras [5] utilized color information and radial symmetry rows for pupil localization in low-resolution images. Valenti et al. [6] used the method of isophote curvature to locate the pupil position by measuring the relationship between the eye and the face. Timm [7] locates the pupil by calculating the dot product of displacement and the gradient vector. Zhao et al. [8] used the reflective properties of the pupil surface and infrared light to detect the position of the pupil. George et al. [9] used geometric features to detect pupil position. However, the earlier appearance-based methods require prior information from the eyes and can only detect pupils in specific scenes, which is limited by the environment.

Based on this, pupil localization based on machine learning has gradually become a research hotspot. This method mainly uses machine learning and appearance information to solve the problem of pupil location. It extracts key features of the eye using machine learning, trains the model, and then obtains the pupil position through the model. In recent years, many studies have focused on detecting pupil position based on machine learning. For example, Chen [10] used support vector machines and the Haar feature to locate pupil positions. Savakis et al. [11] used HOG feature extraction of the pupils and random forest to detect pupils. Markus et al. [12] located pupils using random regression trees. Compared with the method based on appearance detection of pupil position, the

machine learning-based method can extract pupil features by training a large amount of data, improving the robustness of the model and achieving higher accuracy.

2.2. Pupil Localization Method Based on Deep Learning

The emergence of Alexnet has led to significant breakthroughs in deep learning in image classification, image segmentation, target detection, scene recognition, semantic segmentation, and other directions. Shams et al. [13] used a fast ROUT feature algorithm and a deep belief-based neural network (DBNN) to quickly locate the pupil position. Similarly, other networks such as [14,15] have been applied to detect pupil position using deep networks for learning representational features. Since pupil localization is similar to a semantic segmentation task, the pupil localization task can be transformed into a semantic segmentation task. The FCN is widely used in semantic segmentation tasks. In the field of biomedicine, cell segmentation is a long-standing task with a broad range of applications. Accurately segmenting cells in pathological images is the first step toward precise tumor analysis by computers, and it significantly influences future computer-aided pathological analysis. Xia et al. [16] applied FCNs to pupil localization and achieved high accuracy. Due to the impact of the accuracy of eyewear on the accuracy of pupil localization, Jun et al. [17] first detected whether glasses are worn in the input image. If the glasses are worn, they utilized the GAN to remove the glasses before inputting the processed image to the FCN. Olaf et al. [1] proposed the U-net network, which is similar to the FCN, and employs an encoder-decoder architecture to achieve end-to-end pixel-level prediction. Since 2015, the U-net network has been widely used in the biomedical field. The encoding and decoding structure of the U-net network was initially used in image compression and denoising and later introduced into the field of computer vision for image segmentation. It continues the core idea of FCN and realizes pixel-level classification. Based on this, many variants of the U-net model [18,19] have been proposed.

2.3. Attention Mechanism

The Attention Mechanism is a method that mimics the human visual attention mechanism. In traditional neural network structures, all input features are treated equally, without clear differentiation of which features are more important for the task. The Attention Mechanism can automatically assign different weights to the input based on its content, allowing the model to focus on the important information and features and improve performance. The Attention Mechanism was initially proposed in natural language processing to weigh the importance of different words in text sequences to improve model performance. The recently popular Transformer [20] also uses Attention Mechanism. It comprises three main components: Query, Key, and Value. The basic idea is to calculate the weight of each data point by three vectors and then add them up according to weight to form the final output result. The Attention Mechanism has been widely applied in various deep learning tasks, such as image classification [21,22], object detection [23], and speech recognition [24]. Due to the flexibility of the Attention Mechanism, there are many different variants adapted for specific applications. The Self-Attention Mechanism [20] is a common variant that only focuses on input data without external sources to capture long-range dependencies in sequential data. The Multi-Head Attention Mechanism [25] splits input data into different dimensions, calculates different attention weights, and concatenates them for better performance. Capsule Attention Mechanism and Local Attention Mechanism are variants optimized for specific application scenarios. These variant methods can be selected according to specific application scenarios to improve model performance. The U-Net network core is the skip-connection structure, and our proposed improvement is to introduce an Attention Mechanism during the skip-connection stage to automatically select the proportion of encoded features and uncoded features in channel and spatial axes. This makes full use of the feature parameters in each stage to further improve the pixel-level prediction accuracy.

3. Methods

This chapter introduces the designed network that utilizes the U-net network as its main architecture, as illustrated in Figure 1. The skip connection structure serves as the U-Net network's core, where coded and uncoded features are fused and spliced. In contrast to the direct feature connection approach used in the original paper [1], our proposed architecture can automatically select the optimal proportion of the two features on both the channel and spatial axis during the skip connection stage. This adaptive feature selection enables our network to leverage the learned feature parameters during the encoding process at each decoding stage, leading to superior pixel-level prediction accuracy. As shown in Figure 2, our architecture is divided into a channel fusion module and a space fusion module, which are displayed in Figures 3 and 4, respectively. They are divided into three parts, namely Fuse, Squeeze, and Select, respectively.

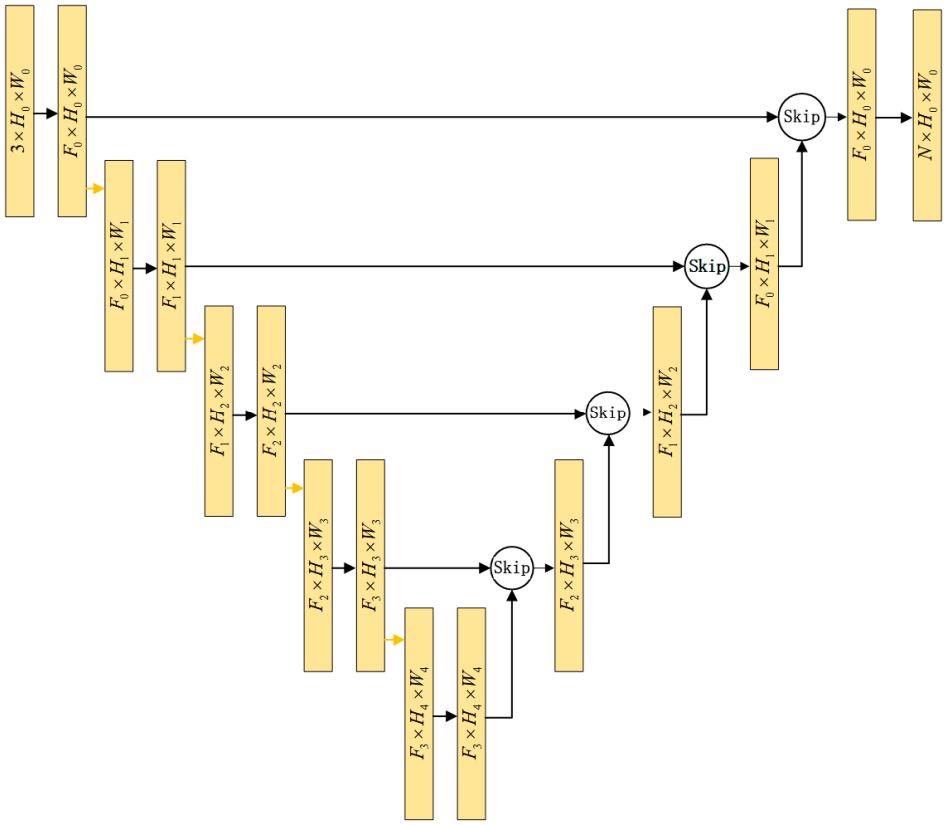

Figure 1. Improved U-net.

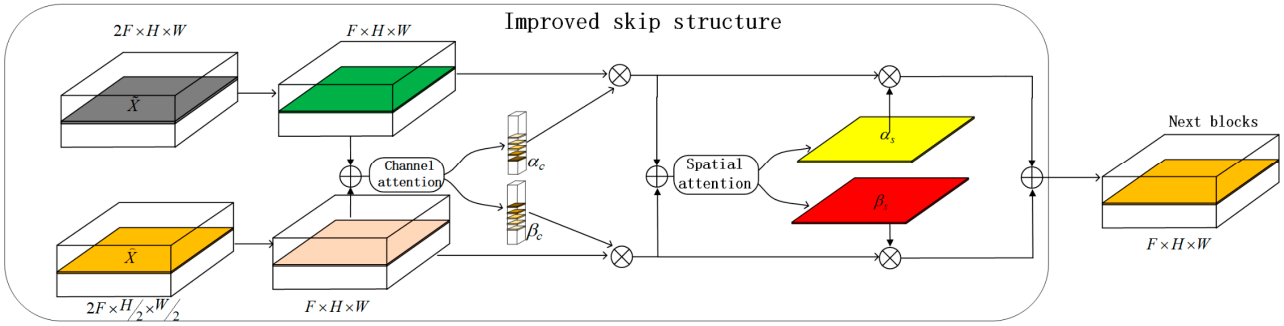

Figure 2. Improved skip structure.

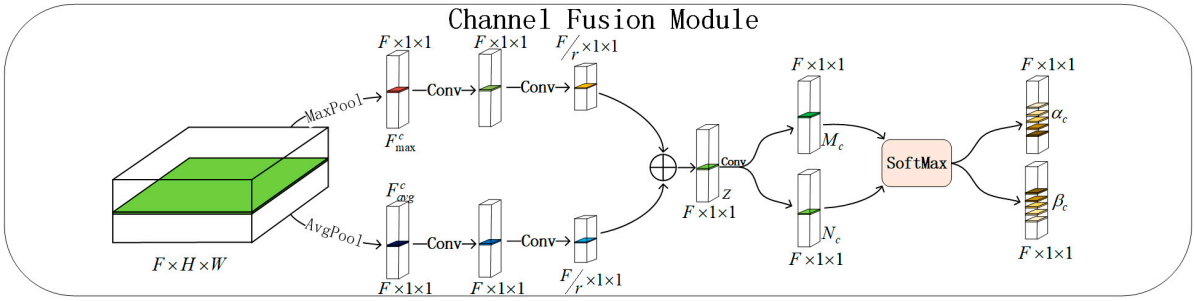

Figure 3. Channel fusion module.

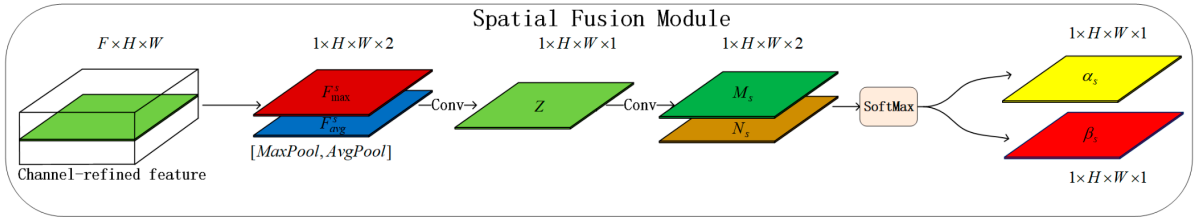

Figure 4. Spatial fusion module.

3.1. Channel Fusion Module

Fuse: The first step is to integrate the encoded feature $\widehat{X}$ and unencoded feature $\widetilde{X}$ by element-wise summation. As their dimensions of height and width are different, it is necessary to upsample the encoded feature to match the size of the unencoded feature. In the skip connection stage, since the number of channels is reduced, the two features are input into a 1×1 convolutional layer for dimension reduction. Finally, the reduced features are added together to generate X with a dimension of $F \times H \times W$.

Squeeze: Subsequently, global average pooling operation and max pooling operation are used to aggregate spatial information and generate two channel descriptors: F^c_{avg} represents the average-pooled feature and $F^c_{\max}$, represents the max-pooled feature. the feature map is downscaled from $F \times H \times W$ to $1 \times 1 \times F$. Next, the two descriptors pass through a simple 1×1 convolutional layer to capture contextual information between channels. Afterward, the feature goes through another 1×1 convolutional layer to decrease the feature map size from $1 \times 1 \times F$ to $1 \times 1 \times \frac{F}{r}$, while also fully capturing channel dependencies by reducing the scale factor r, where r is set to 3. Finally, the two features are added to generate Z with the following formula:

Select: Finally, to fully utilize the information from the squeeze operation, Z is sent through a 1×1 convolutional layer to generate two feature descriptors, M_c and N_c. The Softmax operation automatically selects the weights α_c and β_c of the two features on the channel, resulting in the final output result, Y_c. This process can be described as:

$$X = Conv(f_B(\widehat{X})) + Conv(\widetilde{X}) = W_0((f_B(\widehat{X}))) + W_1(\widetilde{X}) \tag{1}$$

$$\begin{aligned} Z &= Conv(Conv(Avgpool(X))) \\ &+ Conv(Conv(Maxpool(X))) \\ &= \delta(W_4(W_2(F^c_{avg}))) + \delta(W_5(W_3(F^c_{\max}))) \end{aligned} \tag{2}$$

$$M_c = Conv(Z) = W_6(Z) \tag{3}$$

$$N_c = Conv(Z) = W_7(Z) \tag{4}$$

$$\alpha_c = \frac{\exp(M_c)}{\exp(M_c) + \exp(N_c)} \tag{5}$$

$$\beta_c = \frac{\exp(N_c)}{\exp(M_c) + \exp(N_c)} \tag{6}$$

$$Y_c = \alpha_c \cdot Conv(f_B(\widehat{X})) + \beta_c \cdot Conv(\widetilde{X}) \tag{7}$$

where, $\widehat{X}$ and $\widetilde{X}$ represent the encoded feature and the unencoded feature, respectively, and their dimensions are $2F \times \frac{H}{2} \times \frac{W}{2}$ and $2F \times H \times W$, respectively. Where F, H, W represents dimension, height, and width, respectively. Where W_0, W_1, W_2, W_3 is the convolution of 1×1. $W_4 \in R^{\frac{F}{r} \times F}$, $W_5 \in R^{\frac{F}{r} \times F}$, $W_6 \in R^{F \times \frac{F}{r}}$, $W_7 \in R^{F \times \frac{F}{r}}$, δ denotes the ReLU function X'.

3.2. Spatial Fusion Module

Fuse: The first step is to multiply the weights generated by the channel attention module with their corresponding features to obtain the input X' for the spatial attention module, where X' equals Y_c.

Squeeze: X' undergoes global average pooling and maximum pooling along the channel dimension to highlight effective regions and obtain spatial descriptors F^s_{avg} and $F^s_{\max}$, both with a dimension of $1 \times H \times W$. These features then pass through a 7×7 convolution to capture contextual information along the spatial axis.

Select: Finally, another 7×7 convolution is applied for Softmax operation and the Softmax operation automatically selects the weights α_s and β_s for the two features along the spatial axis. This process can be described as:

$$\begin{aligned} Z &= Conv([Avgpool(X^{'}); MaxPool(X^{'}))) \\ &= \delta({f_1}^{7\times7}([F^s_{avg}; F^s_{\max}])) \end{aligned} \tag{8}$$

$$M_s = Conv(Z) = {f_2}^{7\times7}(Z) \tag{9}$$

$$N_s = Conv(Z) = {f_3}^{7\times7}(Z) \tag{10}$$

$$\alpha_s = \frac{\exp(M_s)}{\exp(M_s) + \exp(N_s)} \tag{11}$$

$$\beta_s = \frac{\exp(N_s)}{\exp(M_s) + \exp(N_s)} \tag{12}$$

$$Y_s = \alpha_s \cdot \alpha_c \cdot Conv(f_B(\widehat{X})) + \beta_s \cdot \beta_c \cdot Conv(\widetilde{X}) \tag{13}$$

where $f^{7\times7}$ represents the convolution kernel of 7×7, δ denotes the ReLU function.

4. Experiment

To verify the validity of the proposed model, we used the database from the literature [26] as the experimental training set, which comprised 13,466 images. These images included 5590 images from the LFW dataset and 7876 images downloaded from the Internet. Tests were performed on two public datasets, BioID [2] and GI4E [3].

The BioID dataset consists of 1521 images, which were taken by 23 people under different lighting conditions. The resolution of the image is 286×384, and the left and right eye centers of each image are labeled. This dataset is considered the most challenging dataset.

The GI4E dataset consists of 1236 images, which were taken by 103 people looking in 12 different directions. The resolution of the image is 800×600, and the location of the eye center is also marked in the database.

4.1. Evaluation Criteria

The maximum normalized error is used to evaluate the performance of the model during testing. The process can be formulated as:

$$err = \frac{\max(\sqrt{(x'_l - x_l)^2 + (y'_l - y_l)^2}, \sqrt{(x'_r - x_r)^2 + (y'_r - y_r)^2})}{\sqrt{(x_l - x_r)^2 + (y_l - y_r)^2}} \quad (14)$$

where, (x_l', y_l') and (x_r', y_r') are the estimated position of the left and right eye centers, and (x_l, y_l) and (x_r, y_r) are the ground truth of left and right eye centers. When $err \leq 0.25$, the estimated position of the pupil is between the corner of the eye; when $err \leq 0.1$, the estimated position of the pupil is within the radius of the iris; when $err \leq 0.05$, the estimated position of the pupil is within the radius of the pupil.

4.2. Implementation Details

In this experiment, the training set and test set are cropped through face recognition to obtain the facial region. The face image after cutting was adjusted 96 × 96. Since the real pupil position is only one pixel, the training process may become unstable due to the imbalance between classes, so we conducted dilation on the pupil heat map. We used Morphological dilation. The process is shown in Figure 5. During the training, we randomly selected 10,000 pictures from the literature [26] for training and verified them with another 3466 pictures. We carried out horizontal flip and rotation operations on the training set to expand the size of the data set and improve the performance of the model. In the experiment, we trained the model on the PyTorch platform, its version is 2.0.0. One Quadro GV100 is used for training, the learning rate attenuated from 0.1 polynomial to 0.000001 in 100 epochs, the batch size was 32, and we used the mini-batch stochastic gradient descent (SGD) algorithm to optimize the parameters of the network, momentum was set to 0.9, weight decay was set to 0.0005.

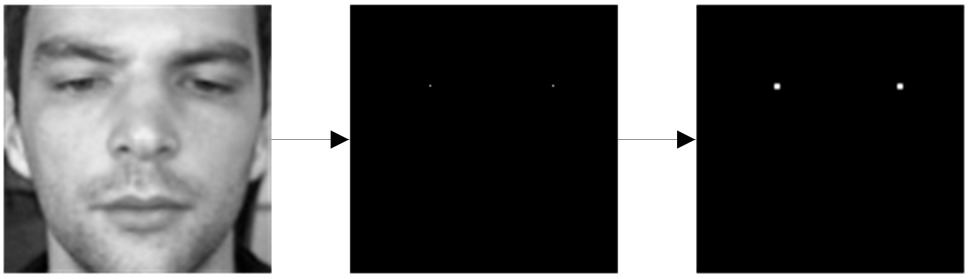

Figure 5. Dilate the pupil position morphologically.

4.3. Explorative Study

This study employed the BioID dataset to conduct comparative experiments and evaluate the effectiveness of the proposed model. The experiment is divided into three parts: the effects of CAM and SAM sequences on model accuracy, the effects of average pooling operation and maximum pooling operation on model accuracy, and the effects of different attention modules are discussed, respectively.

4.3.1. Average Pooling Operation and Maximum Pooling Operation

This experiment compares the effectiveness of three different pooling operations on model performance: average pooling alone, maximum pooling alone, and both pooling operations together. As different pooling operations capture unique information, changing them can significantly impact the performance of the model. Only CAM was used in this experiment.

The experimental results are shown in Table 1. It can be found that the accuracy of both average pooling and maximum pooling is the highest. Therefore, our model is determined to use both pooling operations.

Table 1. Comparison methods of pooling mode on BioID.

Networks	Accuracy%	
	Err $\leq$ 0.025	**Err $\leq$ 0.05**
U-net(avg)	61.01%	94.15%
U-net(max)	60.82%	94.02%
U-net(avg&max)	62.33%	94.54%

4.3.2. Sequences of CAM and SAM

In this experiment, we compared the effects of five different sequences of CAM and SAM on the model accuracy. As these modules focus on distinct information, their varied sequences can impact how the model connects encoded and unencoded features, thereby altering the model's learning ability.

The experimental results are shown in Table 2. It can be found that the accuracy of using CAM first is higher than using SAM first. Therefore, our model is determined to use CAM first and then SAM.

Table 2. Comparison methods of the CAM and SAM on BioID.

Networks	Accuracy%	
	Err $\leq$ 0.025	**Err $\leq$ 0.05**
U-net	59.11%	93.89%
U-net(CAM only)	62.33%	94.54%
U-net(SAM only)	62.39%	94.21%
U-net(CAM + SAM)	64.96%	94.81%
U-net(SAM + CAM)	62.46%	94.28%
U-net(CAM&SAM in parallel)	63.12%	94.35%

4.3.3. Different Attention Modules

This experiment aimed to compare the effects of various attention mechanisms on model performance. Specifically, we focused on CBAM, Cross Dual-Attention Module [27] (CDA), and Efficient Channel Attention (ECA) [28]. Attention modules are powerful tools in deep learning models because they help concentrate the model's attention on crucial features. These modules aim to enhance the representation and discrimination ability of deep learning models, making them more suitable for different tasks and data scenarios.

Table 3 shows the experimental results, which suggest that the proposed method is better suited for the current task. This method mainly utilizes attention mechanisms to automatically select the contributions of unencoded and encoded features to the model, enabling effective information filtering and integration based on the importance of key features. In comparison, methods such as ECA and CDA focus primarily on the learned features themselves and do not properly integrate unencoded and encoded features. Therefore, considering the characteristics of the current task, the proposed method offers greater advantages and applicability.

Table 3. Comparison methods of the different attention modules on BioID.

Networks	Accuracy%	
	Err $\leq$ 0.025	**Err $\leq$ 0.05**
U-net(ECA)	63.16%	94.46%
U-net(CDA)	64.23%	94.62%
U-net(ours)	64.96%	94.81%

4.3.4. Final Experimental Results

The final model was determined based on the results of the three aforementioned experiments: CAM was used first and then SAM was used, and the average pooling operation and the maximum pooling operation were used simultaneously with the CBAM attention mechanism. The experimental results presented in Table 4 demonstrate that our proposed model outperforms state-of-the-art methods on the BioID and GI4E datasets. The efficacy of our method is further illustrated through the experimental results of pupil localization presented in Figure 6. The white dot represents the localized position of the pupil. The above refers to our proposed method, while below are the pupil localization results obtained using FCN. Through visualization of pupil localization, it can be qualitatively demonstrated that our method outperforms FCN. Choi et al. [17] used a fully convolutional network to segment the pupil region. Then the position of the segmented pupil is calculated. G.M. Araujo et al. [29] propose a novel approach for the detection of landmarks on faces. They introduce a new detector, the Inner Product Detector (IPD), based on correlation filters. The main advantages of the proposed method are the tolerance to small variance on the desired patterns, a low computational cost, and the generalization for other features. A. George et al. [9] proposed a two-stage algorithm for IC (Iris Center) localization. The proposed method utilizes the geometrical characteristics of the eye. In the first stage, a fast convolution-based approach is used to obtain the coarse location of the IC. The IC location is further refined and polished in the second stage using boundary tracing and ellipse fitting to obtain the precise IC position. H. Cai et al. [30] mainly utilize the dramatic illumination changes between the iris and sclera. More specifically, novel hierarchical kernels are designed to convolute the eye images and a differential operation is applied to the adjacent convolution results to generate various response maps. The final eye center is localized by searching the maximum response value among the response maps.

Table 4. Comparison results of our models with other state-of-the-art methods on the test set of BIoID and GI4E datasets.

Networks	**Accuracy%**	
	Err $\leq$ 0.025	**Err $\leq$ 0.05**
BioID database		
Timm [7]	37.00%	82.50%
Araujo [29]	31.50%	88.30%
George [9]	48.00%	85.10%
Cai [30]	~	92.80%
Choi [17]	60.00%	93.30%
Xia [16]	~	94.40%
Our	64.96%	94.81%
GI4E database		
Timm [7]	40.00%	92.40%
Villanueva [3]	42.00%	93.90%
George [9]	72.00%	89.30%
Xia [16]	~	99.10%
Cai [30]	85.70%	99.50%
Choi [17]	90.40%	99.60%
Our	90.80%	99.60%

The average processing time per image is another metric for evaluating the performance of pupil localization algorithms. We conducted a comparison of processing times for locating eye centers on the BioID database. Subsequently, we deployed it on a standard laptop with GeForce GTX 1080 Ti. We compared the average processing time of our method with that of other methods, presented in Table 5. Araujo et al. [29] introduce the Inner Product Detector (IPD), a novel detection method that uses correlation filters. The proposed approach has several advantages, including tolerance to small variations in the desired patterns, low computational cost, and the ability to generalize to other features.

Leo et al. [31] propose a novel unsupervised approach for automatically detecting the center of the eye. Our algorithmic core involves representing the shape of the eye through differential analysis of image intensities, combined with the local variability of appearance represented by self-similarity coefficients. Gou et al. [32] propose a coarse-to-fine pupil detection framework based on shape-augmented cascade regression models learning from adversarial synthetic images. Although our method did not achieve the fastest processing time on average, 7.69 ms is still fast enough to meet the real-time application requirements. Moreover, our method outperformed other methods in terms of accuracy, making it an ideal choice for real-time applications that demand high-precision pupil localization. This shows the effectiveness of our method in accurately detecting and locating pupils.

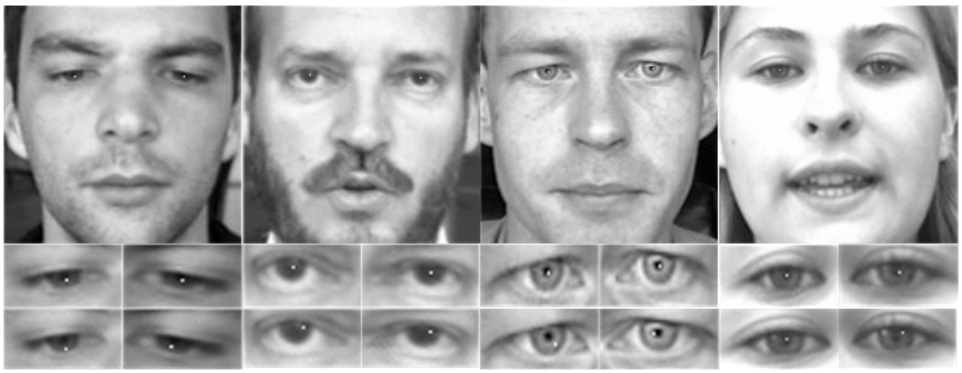

Figure 6. Visualization of the proposed eye pupil localization result.

Table 5. Comparison of our method with other methods in average processing time.

Method	Araujo et al. [29]	Leo et al. [31]	Gou et al. [32]	Xia [16]	Ous
Time (ms)	83	333	63	5	7.69

5. Discussion

For pupil localization, we transform it into semantic segmentation. Image features are learned through an improved U-net network. For the U-net network, how coded features and uncoded features are used is the key factor affecting model performance. The architecture proposed in this paper can utilize the attention mechanism to have a selective connection in both the channel and spatial dimensions. It can not only make full use of the encoded features in the skip connection stage but also automatically select the proportion of the encoded features and unencoded features on the channel and spatial axis so that the contribution of the two features to the model can be fully learned in the decoding stage. The experimental results of multiple error thresholds confirm the superiority of the proposed structure compared with existing methods. In addition, our method can operate in real time.

Author Contributions: Conceptualization, G.C.; Methodology, G.C.; Validation, G.C.; Investigation, Z.D. and J.W.; Writing—original draft, G.C.; Writing—review & editing, G.C.; Supervision, J.W. and L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Boild dataset is from https://www.bioid.com/About/BioID-Face-Database, accessed on 2 March 2023.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
2. BioID Dataset. Available online: https://www.bioid.com/About/BioID-Face-Database. (accessed on 2 March 2023).
3. Villanueva, A.; Ponz, V.; Sesma-Sanchez, L.; Ariz, M.; Porta, S.; Cabeza, R. Hybrid method based on topography for robust detection of iris center and eye corners. *ACM Trans. Multimed. Comput. Commun. Appl.* **2013**, *9*, 25. [CrossRef]
4. Young, D.; Tunley, H.; Samuels, R. *Specialised Hough Transform and Active Contour Methods for Real-Time Eye Tracking*; University of Sussex, Cognitive & Computing Science: Brighton, UK, 1995.
5. Skodras, E.; Fakotakis, N. Precise localization of eye centers in low resolution color images. *Image Vis. Comput.* **2015**, *36*, 51–60. [CrossRef]
6. Valenti, R.; Gevers, T. Accurate eye center location and tracking using isophote curvature. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–8.
7. Timm, F.; Barth, E. Accurate eye centre localisation by means of gradients. In Proceedings of the International Conference on Computer Vision Theory and Applications VISIGRAPP, Algarve, Portugal, 5–7 March 2011; pp. 125–130. [CrossRef]
8. Zhao, S.; Grigat, R.R. Robust eye detection under active infrared illumination. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; IEEE: Piscataway, NJ, USA; pp. 481–484.
9. George, A.; Routray, A. Fast and accurate algorithm for eye localisation for gaze tracking in low-resolution images. *IET Comput. Vision* **2016**, *10*, 660–669. [CrossRef]
10. Chen, S.; Liu, C. Eye detection using discriminatory Haar features and a new efficient SVM. *Image Vis. Comput.* **2015**, *33*, 68–77. [CrossRef]
11. Savakis, A.; Sharma, R.; Kumar, M. Efficient eye detection using HOG-PCA descriptor. In Proceedings of the Imaging and Multimedia Analytics in a Web and Mobile World, San Francisco, CA, USA, 5–6 February 2014; SPIE: Bellingham, WA, USA, 2014; pp. 115–122.
12. Markuš, N.; Frljak, M.; Pandžić, I.S.; Ahlberg, J.; Forchheimer, R. Eye pupil localization with an ensemble of randomized trees. *Pattern Recognit.* **2014**, *47*, 578–587. [CrossRef]
13. Shams, M.Y.; Hassanien, A.E.; Tang, M. Deep belief neural networks for eye localization based speeded up robust features and local binary pattern. In *LISS 2021, Proceedings of the 11th International Conference on Logistics, Informatics and Service Sciences*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 415–430.
14. Park, S.; Zhang, X.; Bulling, A.; Hilliges, O. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 14–17 June 2018; pp. 1–10.
15. Park, S.; Spurr, A.; Hilliges, O. Deep pictorial gaze estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 721–738.
16. Xia, Y.; Yu, H.; Wang, F.-Y. Accurate and robust eye center localization via fully convolutional networks. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 1127–1138. [CrossRef]
17. Choi, J.H.; Lee, K.I.; Song, B.C. Eye pupil localization algorithm using convolutional neural networks. *Multimed. Tools Appl.* **2020**, *79*, 32563–32574. [CrossRef]
18. Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv* **2018**. [CrossRef]
19. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
21. Li, X.; Wu, J.; Lin, Z.; Liu, H.; Zha, H. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 254–269.
22. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, Quebec, Canada, 7–12 December 2015.
24. Chan, W.; Jaitly, N.; Le, Q.V.; Vinyals, O. Listen, attend and spell. *arXiv* **2015**, arXiv:1508.01211.
25. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
26. Sun, Y.; Wang, X.; Tang, X. Deep convolutional network cascade for facial point detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3476–3483.

27. Wu, Z.; Allibert, G.; Meriaudeau, F.; Ma, C.; Demonceaux, C. HiDAnet: RGB-D Salient Object Detection via Hierarchical Depth Awareness. *arXiv* **2023**, arXiv:2301.07405. [CrossRef] [PubMed]
28. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
29. Araujo, G.M.; Ribeiro, F.M.; Silva, E.A.; Goldenstein, S.K. Fast eye localization without a face model using inner product detectors. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1366–1370.
30. Cai, H.; Liu, B.; Ju, Z.; Thill, S.; Belpaeme, T.; Vanderborght, B.; Liu, H. Accurate eye center localization via hierarchical adaptive convolution. In Proceedings of the 29th British Machine Vision Conference, Newcastle, UK, 3–6 September 2018; British Machine Vision Association: Durham, UK, 2018.
31. Leo, M.; Cazzato, D.; De Marco, T.; Distante, C. Unsupervised eye pupil localization through differential geometry and local self-similarity matching. *PloS ONE* **2014**, *9*, e102829. [CrossRef] [PubMed]
32. Gou, C.; Zhang, H.; Wang, K.; Wang, F.-Y.; Ji, Q. Cascade learning from adversarial synthetic images for accurate pupil detection. *Pattern Recognit.* **2019**, *88*, 584–594. [CrossRef]