mathematics

MDPI

Article

An Irregular Pupil Localization Network Driven by ResNet Architecture

Genjian Yang [1], Wenbai Chen [1,*], Peiliang Wu [2], Jianping Gou [3] and Xintong Meng [1]

[1] School of Automation, Beijing Information Science and Technology University, Beijing 100192, China; 2022020397@bistu.edu.cn (G.Y.); 2023020368@bistu.edu.cn (X.M.)
[2] School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China; peiliangwu@ysu.edu.cn
[3] College of Computer and Information Science, College of Software, Southwest University, Chongqing 400715, China; cherish.gjp@gmail.com
* Correspondence: chenwb@bistu.edu.cn

Abstract: The precise and robust localization of pupils is crucial for advancing medical diagnostics and enhancing user experience. Currently, the predominant method for determining the center of the pupil relies on the principles of multi-view geometry, necessitating the simultaneous operation of multiple sensors at different angles. This study introduces a single-stage pupil localization network named ResDenseDilateNet, which is aimed at utilizing a single sensor for pupil localization and ensuring accuracy and stability across various application environments. Our network utilizes near-infrared (NIR) imaging to ensure high-quality image output, meeting the demands of most current applications. A unique technical highlight is the seamless integration of the efficient characteristics of the Deep Residual Network (ResNet) with the Dense Dilated Convolutions Merging Module (DDCM), which substantially enhances the network's performance in precisely capturing pupil features, providing a deep and accurate understanding and extraction of pupil details. This innovative combination strategy greatly improves the system's ability to handle the complexity and subtleties of pupil detection, as well as its adaptability to dynamic pupil changes and environmental factors. Furthermore, we have proposed an innovative loss function, the Contour Centering Loss, which is specifically designed for irregular or partially occluded pupil scenarios. This method innovatively calculates the pupil center point, significantly enhancing the accuracy of pupil localization and robustness of the model in dealing with varied pupil morphologies and partial occlusions. The technology presented in this study not only significantly improves the precision of pupil localization but also exhibits exceptional adaptability and robustness in dealing with complex scenarios, diverse pupil shapes, and occlusions, laying a solid foundation for the future development and application of pupil localization technology.

Keywords: pupil localization; center determination; ResNet; dense dilated convolutions; NIR

MSC: 68T45

check for updates

Citation: Yang, G.; Chen, W.; Wu, P.; Gou, J.; Meng, X. An Irregular Pupil Localization Network Driven by ResNet Architecture. *Mathematics* **2024**, *12*, 2703. https://doi.org/10.3390/math12172703

Academic Editor: Konstantin Kozlov

Received: 29 July 2024
Revised: 22 August 2024
Accepted: 29 August 2024
Published: 30 August 2024

1. Introduction

Pupil localization technology is a method for precisely measuring the position of the pupil and its movement relative to the head and is aimed at efficiently tracking eye movements. This technology reveals the distribution of human visual attention, cognitive processing, and emotional states. As a result, it has shown significant application value in multiple fields such as psychology, human–computer interaction, driving safety assessment, medical diagnosis, market research, and virtual reality. For instance, Cao utilized pupil localization technology in the field of human–computer interaction, greatly enhancing the interactive experience between users and computers as well as related devices [1]. Similarly, Ahmad and colleagues have applied this technology in the realm of intelligent driving, assessing the safety of the driving process by analyzing changes in the driver's visual state [2]. In the context of gaze estimation tasks for VR devices, the precision of pupil tracking plays a pivotal role in determining the effectiveness and fluidity of the human–computer interaction. When pupil tracking is highly accurate, it ensures that the user's gaze is correctly interpreted by the system, allowing for intuitive and responsive interactions. This precision is especially critical in VR environments, where the user's sense of immersion and control is directly tied to how well the system can follow and respond to their eye movements. Any discrepancy in tracking can disrupt the experience, leading to misaligned actions, user frustration, and a diminished sense of presence within the virtual environment.

In the field of pupil detection, existing localization methods primarily fall into two categories: traditional methods and those based on deep learning, each with its unique advantages and limitations. Regarding traditional methods, Song et al. have emphasized a multi-template matching algorithm based on XLD contours [3]. This method locates the pupil edge using XLD contours and ellipse-fitting techniques, achieving precise pupil localization. It relies on classic image processing technology and is suitable for scenarios where real-time performance is not a critical requirement. With the development of deep neural networks and advancements in computing power, numerous innovative detection algorithms have emerged. A typical example of a deep learning-based approach is the pupil localization method using a hybrid vision transformer network proposed by Wang et al. [4]. This method initially uses a Convolutional Neural Network (CNN) to extract local feature maps from eye images and then feeds these features into the encoder of a vision transformer to capture global relationships, ultimately predicting the pupil center's position accurately. This approach combines the local-feature-learning capability of CNNs with the global information processing of vision transformers, enhancing the accuracy and efficiency of pupil localization. On the other hand, Jia and others have proposed a coarse-to-fine neural network architecture, which includes stages of rough classification and fine regression, further improving the accuracy of pupil detection [5]. This phased approach enhances the precision and robustness of localization by fine-tuning after the preliminary positioning. Common detection algorithms in pupil detection network workflows can be categorized into single-stage and two-stage object detection algorithms. Single-stage object detection algorithms, like the YOLO series [6] and SSD algorithms [7], perform object detection by dividing the image into multiple small areas and setting anchor boxes of different sizes on feature maps of different scales. Chen and others proposed a single-stage object detection algorithm using a lightweight backbone network for feature extraction, aiming to reduce the network size and to improve the inference speed, making it suitable for eye detection tasks on small low-power computers [8]. In contrast, two-stage object detection algorithms, like the R-CNN series [9], first extract regions of detected objects and then classify and regress these object areas. Zhang and others proposed an improved MTCNN infrared human eye detection algorithm, which offers good detection performance while ensuring real-time capabilities, although it has limitations in target detection applications involving complex scenes [10]. To better meet the real-time requirements of eye tracking and to better adapt to mobile devices with limited computing resources, we simplified the training and inference process, avoiding the steps of candidate region generation and selection. This significantly reduces the computation requirements, resulting in an end-to-end single-stage method that improves processing speed and better adapts to mobile devices with limited computing power.

The development of pupil localization technology has gone through multiple stages, with modern localization methods broadly categorized into four types: mechanical recording, electrical recording, image recording, and near-infrared recording (NIR) [11]. Mechanical recording measures changes in pupil position through electromagnetic induction signals. Although this method excels in accuracy, it may cause eye discomfort and has relatively high manufacturing costs. In contrast, electrical recording uses electrode devices to monitor changes in pupil position, offering lower costs and greater operability but with larger measurement errors, which impact the data reliability to some extent. Image recording captures eye images through cameras, followed by subsequent analysis and processing. This method demonstrates excellent measurement precision; however, its performance is somewhat affected by variations in light intensity. NIR employs near-infrared photosensitive devices, measuring the state of the pupil's position by analyzing the light reflected from different structures of the eye. As a non-invasive technique, NIR significantly enhances user comfort and performs well in terms of measurement accuracy. Furthermore, with the advancement of virtual reality devices, NIR shows substantial market potential in the consumer sector [12]. Consequently, this study has collected and organized a large volume of NIR data and, based on the characteristics of this data, proposes an advanced pupil localization network.

This study introduces an end-to-end, single-stage pupil localization system that is specifically designed for NIR ocular images. The system achieves precise pupil localization with good real-time performance, showing an excellent ability to meet the computational power requirements of mobile devices. The core of this system lies in the adoption of the ResDenseDilateNet network in combination with our original Contour Centering Loss. ResNet, as a deep residual network, is widely recognized for its efficiency and stability in the field of image recognition. Its skip connections effectively mitigate the vanishing gradient problem, allowing the network to extract more complex features at deeper levels. On the other hand, DDCM, through the combination of dense connections and dilated convolutions, is capable of capturing multi-scale contextual information and handling high-resolution features effectively. The ResDenseDilateNet network, integrating dilated convolution [13] with a densely connected architecture [14], enhances the accuracy and computational efficiency of pupil center localization through its efficient parameter configuration and superior feature extraction capabilities. The Contour Centering Loss, a novel loss function, further optimizes the accuracy of pupil localization by minimizing the discrepancy between the pupil contour and the predicted center. This approach, merging modern deep learning techniques with precise geometric analysis, not only boosts the performance of the pupil localization system but also provides new perspectives and tools for future research and applications in related fields. The main contributions of this inquiry are summarized as follows:

(1) An end-to-end, single-stage pupil localization network, ResDenseDilateNet, is proposed that significantly enhances the accuracy and computational efficiency of pupil localization by integrating dilated convolution with a densely connected architecture.
(2) A loss function, named Contour Centering Loss, has been designed specifically for irregular circles, which effectively optimizes localization precision by minimizing the discrepancy between the pupil contour and its predicted center.
(3) By merging deep learning techniques with geometric analysis, the performance of pupil localization has been substantially improved, offering vital technical support for the development of interactive technologies in the electronics domain and applications in medical diagnostics.

2. Methods

This research introduces an innovative end-to-end, single-stage pupil localization network that has been meticulously developed to address the challenges of pupil segmentation and localization in NIR images. The network, christened ResDenseDilateNet, is illustrated in Figure 1. ResDenseDilateNet adeptly merges the principles of a Residual Network (ResNet) [15] with a Dense Dilated Convolutions Merging Module. Furthermore, the network employs a specially designed loss function, termed the Contour Centering Loss, that has been exclusively tailored for accurate pupil localization. This unique loss function focuses on minimizing the disparity between the predicted pupil contour and its actual center, significantly enhancing the precision of localization. The integration of ResNet effectively counters the prevalent issue of gradient vanishing during deep neural network training. With the incorporation of skip connections [16], ResNet allows gradients to directly traverse through a specific layer [17], aiding the network in mastering identity mapping and ensuring consistent model performance even with an increased network depth. The Dense Dilated Convolutions Merging Module, by expanding the spacing within the convolutional kernel, enlarges the network's receptive field. This strategic augmentation enables the network to retain computational efficiency while substantially improving its ability to capture long-range information, thus markedly enhancing the model's comprehension of spatial relationships in images.

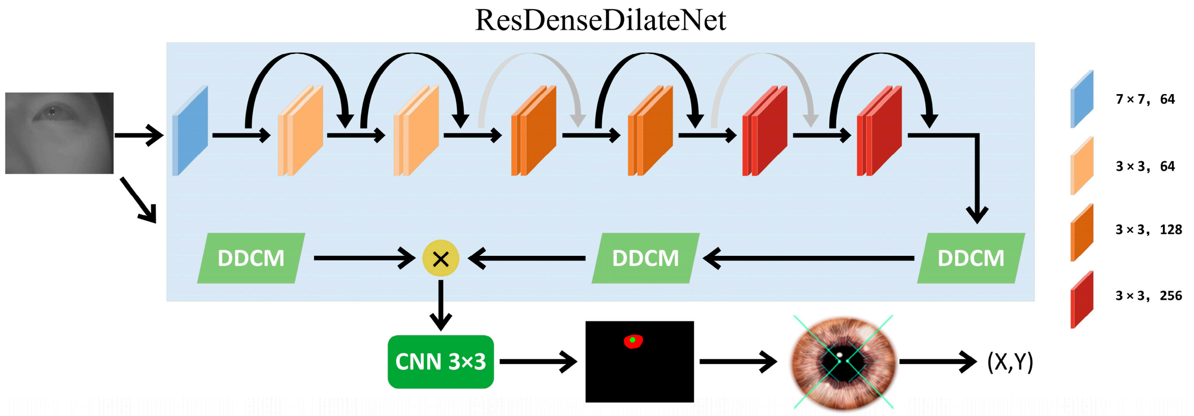

Figure 1. Diagram of the ResDenseDilateNet architecture.

ResDenseDilateNet is capable of efficiently handling the complexities involved in pupil localization while maintaining a high sensitivity to details. The implementation of this approach not only improves the accuracy of pupil localization but also enhances the model's adaptability to various pupil textures and sizes.

2.1. Dilated Convolutions

Dilated convolution, also known as convolution with dilation, represents an advanced design in convolutional network modules that is aimed at enhancing and broadening the capabilities of traditional convolutional neural networks [18]. By incorporating techniques based on dilated convolution, these modules effectively aggregate contextual information from multiple scales while maintaining the resolution and coverage of images or feature maps unchanged. The uniqueness of this module lies in its support for the exponential expansion of the receptive field without relying on pooling layers or subsampling layers to increase the size of the receptive field. Within dilated convolution modules, a series of filters with exponentially growing strides are applied, facilitating in-depth analysis of the input data [19]. This approach allows the network to capture a broader range of contextual information without adding an extra computational burden or losing vital spatial information. This characteristic makes dilated convolution particularly suitable for applications requiring fine spatial resolution, such as image segmentation, object detection, and video analysis. Dilated convolution operates by inserting a fixed number of spaces between each element of the standard convolutional kernel [20], thereby increasing the effective size of the convolutional kernel without actually increasing the number of its weights.

The operational principle of dilated convolution is intuitively elucidated in Figure 2. Specifically, panel (a) illustrates the outcome of a 1-dilated convolution, which is akin to a standard convolution, where the receptive field for each element spans a 3×3 area. Panel (b) unveils the effect of a 2-dilated convolution, under which circumstances the receptive field for each element is expanded to a 7×7 matrix. Panel (c) showcases the impact of a 4-dilated convolution, whereby the receptive field for each element is further enlarged to 15×15. The precise calculation formula for the size of the receptive field is presented in Equation (1).

$$F_{i+1} = (2^{i+2} - 1) \times (2^{i+2} - 1) \tag{1}$$

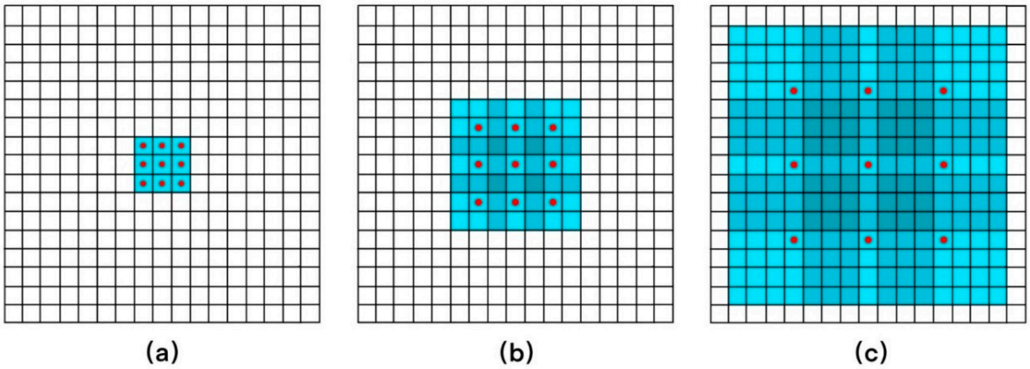

Figure 2. Principle Diagram of Dilated Convolution. (**a**) The outcome of a 1-dilated convolution; (**b**) The effect of a 2-dilated convolution; (**c**) The impact of a 4-dilated convolution.

The actual size of the dilated convolution kernel can be calculated using Equation (2), where k denotes the size of the original convolution kernel, a represents the dilation rate, and K signifies the size of the convolution kernel after dilation [21]. Dilated convolution achieves an expansion of its size by inserting spatial units between adjacent elements of the original convolution kernel, thereby broadening the data region processed without directly increasing the number of parameters within the convolution kernel. This method allows the network to effectively enlarge its receptive field without significantly increasing its computational burden, catering to the extraction of features across varying scales. By adjusting the dilation rate, researchers can flexibly control the spatial resolution of the model while maintaining relative complexity constant, which is particularly valuable for processing data with complex spatial structures [22].

$$K = k + (k-1)(a-1) \tag{2}$$

2.2. Dense Dilated Convolutions Merging Module (DDCM)

The DDCM is a cutting-edge component in neural network architecture that is designed specifically to boost the performance of Convolutional Neural Networks (CNNs) when dealing with image and video analysis tasks [23]. Central to the DDCM is its integration of dilated convolutions with dense connectivity, as illustrated in Figure 3. The module layers multiple dilated convolutional layers, each utilizing a unique dilation rate. This setup is intended to enable the network to learn features across various scales simultaneously, thereby enhancing its ability to comprehend the complex structures present in images. Moreover, the incorporation of dense connectivity means that the output from each layer is not just transferred to the subsequent layer but also combined with the outputs from all previous layers. This approach of dense information flow ensures that the network effectively leverages the features learned from earlier layers throughout its layers.

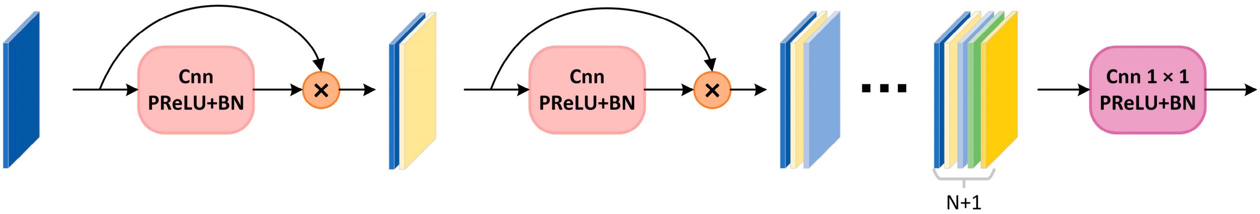

Figure 3. Schematic diagram illustrating the principle of the DDCM.

This architecture, which merges dilated convolutions with dense connectivity, significantly enhances the model's efficiency and precision in processing advanced image features. In the specific task of pupil localization, the model is required to accurately identify the subtle details of the pupil while also understanding the overall context of the image. This challenge demands that the model concurrently processes fine local information and extensive background knowledge. The DDCM, through its innovative structural design, provides the model with the multi-scale information necessary for capturing complex visual patterns. This structure allows each layer to not only utilize the detailed features from the previous layer but also to integrate information from all the preceding layers, thereby achieving higher accuracy and robustness in identifying pupil positions. This approach effectively addresses challenges posed by variations in lighting and other environmental factors, especially when dealing with NIR images.

2.3. Contour Centering Loss

Existing pupil center localization techniques primarily rely on directly determining the center point position from the outputs of detection or segmentation algorithms. However, given that the pupil often presents as an irregular circle, coupled with significant variations in pupil shape across different individuals, these methods frequently result in substantial localization errors and may adversely affect pupil tracking performance. To address this challenge, this study introduces an innovative loss function design specifically tailored to accurately determine the center positions of irregular circular objects. This loss function aims to optimize the precision of pupil center localization while enhancing the system's adaptability to the diversity of pupil shapes, thereby improving overall robustness. Through the meticulous design of this loss function, we can significantly reduce localization errors due to the irregularity of pupil shapes while maintaining a high degree of accuracy. Figure 4 illustrates a conceptual diagram of the principles behind this loss function, providing a visual explanation of its mechanism for the precise center localization in irregular shapes.

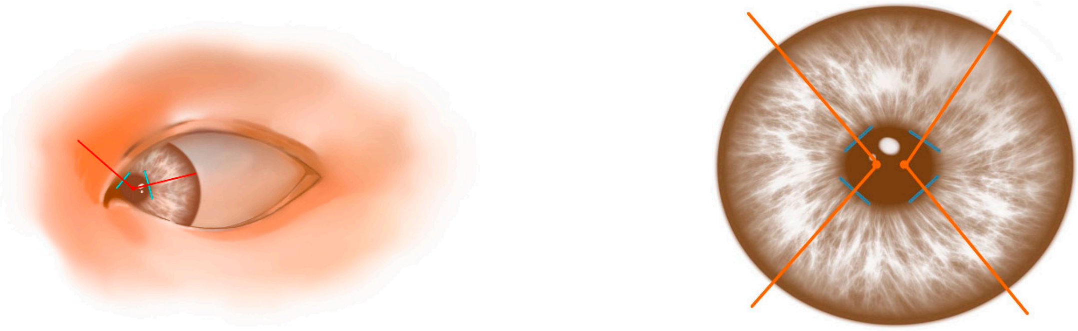

Figure 4. Diagram illustrating the principles of Contour Centering Loss. The blue lines represent tangents, and the red lines represent normals.

The loss function is contingent upon the computation of the curvature at points on the pupil boundary, as defined by Equation (3). Here, K_i represents the curvature, and the points (x_{i-1}, y_{i-1}) and (x_{i+1}, y_{i+1}) are the coordinates of two points adjacent to (x_i, y_i) along the segmented boundary. The formula calculates the measure of the curvature by determining the difference in the tangent values of the angle formed by these three points. During pupil movement, it is common for the eyelids to partially obscure the boundary, resulting in segments that exhibit a straight-line appearance with curvature values approaching zero. Such occurrences can substantially impede the accurate localization of the pupil's center. To address this, the method introduces a curvature threshold to effectively eliminate boundary points that yield aberrant curvature information due to occlusion, thereby enhancing the precision of pupil center localization.

$$K_i = |\text{arctan2}(y_{i+1} - y_i, x_{i+1} - x_i) - \text{arctan2}(y_i - y_{i-1}, x_i - x_{i-1})| \tag{3}$$

The pupil typically exhibits an elliptical characteristic, and uniform sampling along its boundary may lead to a displacement in the estimated center of the pupil. To address this issue and to select a higher density of points in areas of lower curvature, thus reducing the positional offset of the pupil's center, this study employs a sampling strategy based on the reciprocal of the curvature, as outlined in Equation (4). Within this equation, a small fixed constant ϵ is incorporated to ensure the stability of the sampling approach and to prevent division by zero in regions where the curvature is minimal.

$$R_i = \frac{1}{K_i + \epsilon} \tag{4}$$

Subsequently, the normal equations at the aforementioned sampled points are meticulously calculated, as demonstrated in Equation (4). Initially, it is necessary to calculate the gradient m_i at the sampling point P_i as well as at its adjacent points P_{i-1} and P_{i+1}. This process enables the precise definition of the normals to the curve at each sampled point. Upon obtaining the normals, the coordinates of the intersection point between two normals are calculated, as demonstrated in Equation (6). This equation outlines the method for determining the precise coordinates where the two geometric lines intersect, providing a crucial step in the analysis process.

$$\begin{aligned} m_i &= \tfrac{y_{i+1}-y_{i-1}}{x_{i+1}-x_{i-1}} \\ n_i &= -\tfrac{1}{m_i} \\ y - y_i &= n_i(x - x_i) \end{aligned} \tag{5}$$

$$\begin{aligned} x_{\text{int}} &= \tfrac{n_1x_1-n_2x_2+y_2-y_1}{n_1-n_2} \\ y_{\text{int}} &= n_1(x_{\text{int}} - x_1) + y_1 \end{aligned} \tag{6}$$

In practical applications, image noise and imperfections in boundary segmentation lead to the formation of a dispersed region of intersection points. The DBSCAN clustering algorithm [24], a density-based clustering method, is adept at identifying areas of high density and considering them as distinct clusters. Therefore, we employ the DBSCAN algorithm to manage the scattered intersection points within the image. This algorithm clusters tightly grouped points by analyzing the density relationship between each point and its neighbors, subsequently calculating the centroid of the cluster. The centroid represents the optimal estimate of the central point within the original dispersed area using the calculation process outlined in Equation (7). This method robustly addresses the challenges posed by image noise and imperfect boundary segmentation, thereby enhancing the accuracy of pupil center determination.

$$\begin{aligned} x_{\text{center}} &= \frac{1}{n}\sum_{i=1}^{n} x_{\text{int}-i} \\ y_{\text{center}} &= \frac{1}{n}\sum_{i=1}^{n} y_{\text{int}-i} \end{aligned} \tag{7}$$

Finally, the deviation loss is calculated using the mean squared error (MSE) function [25], as depicted in Equation (8). Here, c_i represents the coordinates of the actual center point, $\hat{c}_i$ denotes the coordinates of the predicted center point, and N signifies the number of image samples. This function constrains and calibrates the offset of the center coordinates. Thus, we have successfully developed a comprehensive loss function tailored for pupils and other irregular circular shapes, enhancing the model's sensitivity to central point information and achieving greater accuracy and robustness.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(c_i - \hat{c}_i)^2 \tag{8}$$

3. Experiments and Results

3.1. Experimental Environment and Data

The experiment utilized a deep learning architecture based on the ResNet network, which operated on a high-performance computer equipped with a 13th-generation Intel

Core i7 processor and two NVIDIA GeForce RTX 4090 graphics cards, each with 24 GB of VRAM. The system ran on Ubuntu 20, with Pytorch 1.11.0 selected as the deep learning framework due to its suitability for complex image recognition tasks. The training configuration for the ResNet model included a batch size of 64, and to better handle high-resolution images, the input image size was adjusted to 320 × 320 pixels, which aided in capturing more detailed image features. The optimizer chosen was the Stochastic Gradient Descent (SGD), with a learning rate set at 0.01 and a decay momentum of 0.0001. Additionally, the experiment incorporated weight decay and data augmentation techniques to enhance the model's generalization capability.

This study collected and manually annotated a dataset for eye tracking and pupil localization in virtual reality (VR) glasses application scenarios. The dataset comprises 22,330 near-infrared (NIR) images from 120 participants. The purpose of data collection was clearly communicated, and participation was entirely voluntary. The participants included university students as well as employees across various age groups. As shown in Figure 5, capturing a comprehensive range of eye states, including eye opening and closing, as well as pupil positions. The creation of the dataset took into account individual differences, environmental lighting interference, and variables such as eyeglass wear to ensure its wide applicability and representativeness in real-world applications. At the same time, NIR technology meets the current demands for iris recognition applications and is better suited for VR devices in terms of iris-based payment and security verification.

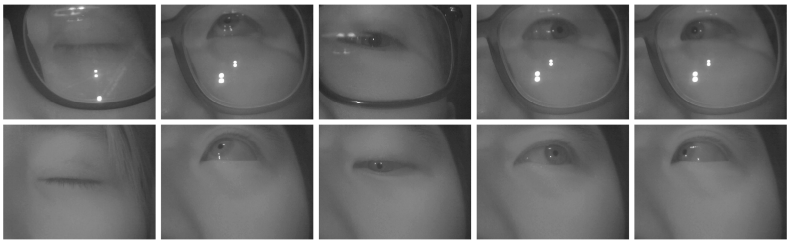

Figure 5. Examples of images from the dataset.

During the data collection process, the left and right eyes of each participant were recorded independently to minimize the impact of individual physiological differences on the accuracy of the data, with a schematic of the collection process illustrated in Figure 6. Participants were asked to open and close their eyes under various gaze directions (up, down, left, right, forward), employing a multi-angle collection strategy aimed at providing a rich sample set that reflects the diversity of eye states. This, in turn, enhances the training effectiveness and generalization capability of eye-tracking algorithms. Particularly noteworthy is that the collected near-infrared (NIR) images vividly reveal the details of the eye structure, which is crucial for precise pupil localization and the accurate determination of eye states. The chosen shooting distance of 7 cm not only complies with the current usage scenarios of VR glasses but also optimizes image quality, avoiding distortion issues that might arise from too close a distance. Furthermore, all images were manually annotated, offering detailed information on the open or closed state of the eyes and the precise position of the pupil.

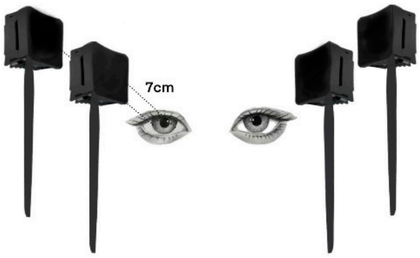

Figure 6. Schematic diagram of the human eye data collection process.

3.2. Evaluation Metrics

In this study, we introduced several evaluation metrics to thoroughly analyze and measure the comprehensive performance of the pupil localization model; these included the mean Intersection over Union (mIoU) [26], mean absolute error (MAE), and relative positional error (RPE) [27], as illustrated in Equation (9). These metrics aim to comprehensively reveal the model's efficacy and precision from multiple dimensions. The mIoU, an amalgamation of the Intersection over Union (IoU) across all classes, offers a balanced and comprehensive measure of performance. It provides deep insights into the model's performance by holistically considering the accuracy of the model in correctly segmenting the pupil's region and its completeness. The mIoU reflects the proportion of overlap between the predicted segmentation and the ground truth among all possible regions, offering a robust evaluation of both the precision and recall of the model's segmentation accuracy. This combined evaluation method is particularly suited for scenarios demanding high accuracy and completeness in segmentation tasks. The MAE provides an intuitive and precise quantification of the average deviation between the model's predicted pupil positions and the actual observations [28]. In the context of pupil localization, the magnitude of the MAE directly impacts the final system's usability and user experience, as any deviation in prediction could lead to a misjudgment of user focus or the incorrect interpretation of visual information. The RPE on the other hand, considers the relative difference between predicted and actual positions, a crucial metric in applications requiring high-precision geolocation or refined visual tracking. It not only measures the model's performance in terms of localization accuracy but also reflects its robustness in dealing with complex or changing environments.

$$\begin{aligned}
\text{mIoU} &= \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i+FN_i} \\
MAE &= \frac{1}{H\times W}\sum_{r=1}^{H}\sum_{c=1}^{W}\left|P(r,c)-G(r,c)\right| \\
\text{RPE} &= \frac{\sqrt{(x_{\text{est}}-x_{\text{true}})^2+(y_{\text{est}}-y_{\text{true}})^2}}{D}
\end{aligned} \tag{9}$$

In Equation (9), TP denotes the number of true positives, or correctly predicted positive instances, while FP represents the number of negative instances incorrectly predicted as positive, and FN refers to the number of positive instances incorrectly predicted as negative. H and W correspond to the height and width of the input image, respectively, and $P(r,c)$ and $G(r,c)$ represent the pixel points of the predicted probability map and the true labels. x_{true} and y_{true} are the coordinates of the actual pupil center, whereas x_{est} and y_{est} are the coordinates of the determined pupil center. D serves as a normalization factor, which, in the context of pupil localization experiments, is the diameter of the pupil. Through comprehensive evaluation, it is possible not only to gain a holistic understanding of the performance of the pupil localization model but also to discern the challenges and limitations the model may encounter in practical applications.

3.3. File Formats for Graphics

In the fields of computer vision and biometric systems research, precise segmentation of the pupil is critical, especially in the process of localizing the pupil center. Accurate identification of the pupil boundary not only directly impacts the accuracy of localization but also significantly affects the overall system performance. In light of this, the current study aims to develop an efficient and innovative deep learning model, namely ResDenseDilateNet, that is focused on enhancing the precision and robustness of pupil segmentation. ResDenseDilateNet combines the benefits of residual learning, densely connected networks, and dilated convolutions, creatively addressing key challenges in pupil segmentation. The core of this network design lies in improving the capture of pupil edge details while reducing the sensitivity to noise, ensuring highly precise and reliable segmentation results. To comprehensively evaluate the performance of ResDenseDilateNet, a series of comparative experiments were conducted, meticulously comparing it against several classical and cutting-edge deep learning networks in the field, as shown in Table 1.

Table 1. Comparative experiments on the pupil segmentation effect.

Parameters	mIoU	MAE	Parameters (M)
SegNet	0.098	0.043	43.87
UNet	0.095	0.048	36.04
PSPNet	0.096	0.044	49.59
SFNet	0.098	0.043	37.76
ResDenseDilateNet	0.098	0.042	19.91

From Table 1, it is evident that this study meticulously analyzed five different deep neural network architectures, including the widely applied SegNet, UNet, PSPNet, SFNet, and our innovatively proposed ResDenseDilateNet. To comprehensively and deeply evaluate the performance of these models on specific tasks, three key quantitative metrics were employed: mIoU, MAE, and the number of model parameters (Parameters, in millions). These metrics together form an all-encompassing evaluation system aimed at deeply mining and comparing the performance and efficiency of these models across various dimensions. In terms of the mIoU, all networks compared exhibited relatively close scores, indicating a certain level of similarity in overall performance among these models. However, it is noteworthy that our proposed ResDenseDilateNet took the lead in this metric. Although the advantage is not overwhelmingly significant, it still reflects the potential and effectiveness of ResDenseDilateNet in terms of comprehensive accuracy. Yet, when we turn to the MAE metric, the performance of ResDenseDilateNet stands out prominently, with a score of 0.042, significantly outperforming all other models, especially when compared with the industry benchmark UNet model, showing an improvement margin of 12.5 percentage points. This significant performance enhancement not only marks a notable advance in the accuracy of pupil boundary determination but also signifies the importance of such precision improvements in pupil localization tasks, particularly for determining the center of the pupil. Additionally, in terms of model parameter efficiency, ResDenseDilateNet, with only 19.91 million parameters, significantly surpasses other models. This result reveals the innovation and optimization in the structural design of ResDenseDilateNet. A lower number of parameters not only means reduced storage requirements for the model but also implies faster inference speeds and lower energy consumption in practical applications, such as real-time processing and deployment on mobile devices. This high-efficiency characteristic makes ResDenseDilateNet an ideal choice for resource-constrained environments.

3.4. Evaluating the Accuracy of Pupil Localization

In this study, we introduce an innovative central loss function designed to enhance the accuracy of pupil center determination, especially in cases involving irregular shapes or partial occlusions of the pupil. This new loss function deeply considers the challenges faced by existing localization techniques when dealing with high variability and complex visual environments and aims to address the insufficient accuracy of traditional algorithms under these conditions. By precisely locating the pupil center, this method significantly improves the accuracy and reliability of eye-tracking technologies and visual attention analysis. To comprehensively evaluate the performance of our proposed central loss function, a series of detailed and systematic experiments were designed. These experiments not only tested the effectiveness of the loss function under various challenging conditions but also conducted an in-depth comparative analysis with several advanced loss functions widely used in the fields of object detection and image segmentation today. These include CIoU (Complete Intersection over Union), SIoU (Scale-Invariant Intersection over Union), and EIoU (Enhanced Intersection over Union), each demonstrating significant effectiveness in improving localization accuracy, adapting to scale changes, and enhancing model generalization capabilities, serving as important tools in the field of computer vision.

During the experimental phase of this research, 12 independent and diverse test datasets were meticulously selected to ensure that the evaluation results are comprehensive, highly objective, and credible. These carefully chosen datasets cover a range of pupil images from simple to complex and from standardized to those presenting unique challenges, ensuring the diversity of experimental conditions and the generalizability of the results. Each dataset was used to thoroughly assess and compare the performance of different loss functions on the critical task of pupil center determination. We paid special attention to the core metric of relative error, as it is a key parameter for measuring the efficacy of pupil localization algorithms. The magnitude of relative error directly reflects the deviation between the model's predicted pupil center position and the actual position, serving as an intuitive and effective metric for evaluating the localization accuracy. This metric is crucial for understanding and improving pupil localization technology, especially when applied in precise eye-tracking and visual attention research fields. In Table 2, we meticulously present the detailed comparison results using these different loss functions on the selected test datasets. Through this data, we can deeply analyze and interpret the performance differences among various loss functions in pupil localization tasks and their specific impact on the final localization accuracy. Such comparisons not only provide a solid evaluation foundation for our central loss function but also offer valuable insights and implications for the research on pupil localization technology in the broader field of computer vision.

Table 2. Comparison of the influence of different loss functions on the RPE.

Trial	Loss			
	CIoU (%)	**SIoU (%)**	**EIoU (%)**	**Center (%)**
1	0.51	0.64	0.88	0.76
2	0.78	0.59	1.09	0.80
3	0.67	1.18	1.16	0.69
4	0.59	1.73	0.16	0.11
5	1.71	1.23	0.75	0.43
6	1.20	0.99	0.77	0.33
7	1.13	0.74	0.89	0.59
8	1.88	0.98	0.54	0.46
9	1.27	0.96	1.28	0.48
10	0.77	1.73	0.58	0.66
11	1.21	1.32	0.70	0.42
12	1.36	0.59	0.45	0.35

Through an in-depth analysis of the data presented in Table 2, it is clearly observed that the central loss function developed in this study demonstrates exceptional accuracy and notable stability in the task of determining the pupil center position compared with several other loss functions widely used in the field of computer vision, such as CIoU, SIoU, and EIoU. The primary advantage of this novel central loss function lies in its precise prediction capability for the pupil center position, particularly in processing pupil images with irregular shapes and complex textures, where it exhibits outstanding recognition and localization accuracy. More critically, the central loss function has shown a high degree of consistency and stability across diverse testing conditions and datasets with varying characteristics. This indicates that the function can provide solid and reliable pupil position determination in a wide range of complex and variable visual environments, significantly reducing the fluctuation and error rate of prediction results. In practical application scenarios, such stability and robustness are crucial as they ensure that the algorithm remains efficient and accurate under various conditions. The characteristics of the central loss function not only enhance its application potential in visual recognition and precise localization but also offer important directions for the future development of pupil localization technology. Its excellent performance suggests broad applications in fields such as eye tracking, facial recognition, augmented reality, and human–computer interaction, greatly improving the performance and user experience of these technologies.

To more comprehensively evaluate the algorithm proposed in this study and highlight its significant advantages over traditional methods in handling complex pupil localization tasks, ablation experiments were conducted, with the results shown in Figure 7. Specifically, our experimental framework mainly covered the following key variants: (1) As a performance benchmark, we first utilized the unmodified ResNet architecture. (2) Building on this, we introduced the DDCM to explore its enhancing effect on the model performance. (3) Subsequently, we integrated our original Contour Centering Loss into ResNet to assess its effect on improving accuracy. (4) Finally, we simultaneously incorporated DDCM and Contour Centering Loss into the ResNet architecture to verify the synergistic effect of these two technologies. To precisely quantify and compare the effects of these configurations, we selected RPE as the key evaluation metric. This metric not only reflects the model's capability in precise localization but also serves as an important tool for measuring algorithm performance. Through these thorough ablation experiments, we can not only intricately dissect the impact of each individual component on the overall algorithm performance but also validate the efficiency and superior performance of our proposed composite algorithm in challenging visual tasks.

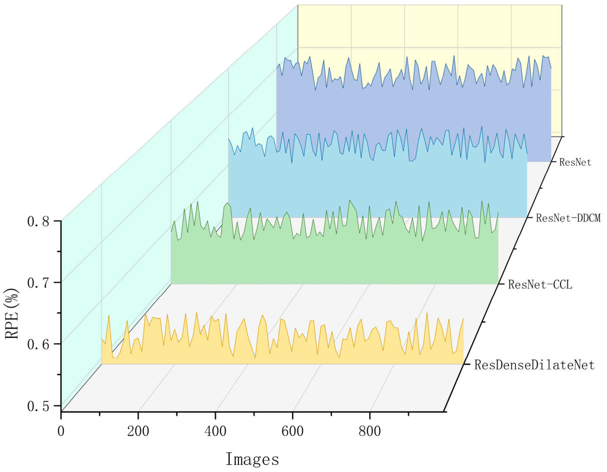

Figure 7. Comparative analysis of ablation experiments on the RPE.

From Figure 7, it is apparent that the baseline performance of the original ResNet model on the RPE reveals its inherent limitations in processing complex visual tasks, highlighting the urgent need for improvements to the existing architecture. By integrating the DDCM into the ResNet architecture, we achieved a significant expansion of the receptive field. The use of dense dilated convolution techniques effectively enhanced the network's spatial resolution and feature-extraction capability, allowing the model to capture multi-scale contextual information in images more intricately. This structural improvement resulted in quantifiable performance enhancements in localization accuracy, enabling the model to parse visual information across a broader spatial scale. Subsequently, we tested the integration of the Contour Centering Loss function into the ResNet architecture separately, with the aim of improving the model's precision in recognizing image edges and contour features. The introduction of Contour Centering Loss significantly boosted the model's responsiveness to contour information, thereby enhancing its localization accuracy. This loss function particularly emphasizes the model's capability in boundary localization and pupil center determination, which is crucial for pupil localization tasks that rely heavily on precise contour detection. The significant performance improvement of the ResNet model with both DDCM and Contour Centering Loss on the RPE metric highlights the complementary effect of the proposed modules in enhancing the performance of deep neural networks. The synergistic action of DDCM's receptive field expansion and Contour Centering Loss's enhancement of contour features notably improved the model's feature representation capability and spatial localization accuracy. This fusion strategy not only enhanced the network's ability to handle complex visual scenarios but also provided empirical evidence for the further optimization of deep learning architectures.

4. Conclusions

In this paper, we introduce ResDenseDilateNet, an efficient single-stage pupil localization network designed to significantly enhance the precision and robustness of pupil localization. The network innovatively combines the efficient processing capabilities of ResNet with the refined feature capture properties of the DDCM, enabling more accurate identification and understanding of the subtle features of the pupil, thus effectively making it applicable to a variety of complex and variable visual environments. Another highlight within this study is the introduction of the Contour Centering Loss function, an innovative loss function specifically designed for dealing with irregularly shaped or partially occluded pupils. It not only significantly improves the accuracy of pupil localization but also enhances the network's adaptability in handling the diversity of pupil shapes and complex occlusion conditions. Through this approach, ResDenseDilateNet demonstrates exceptional flexibility and stability while maintaining high precision. The successful application of this technology promises a widespread impact and application prospects in fields such as human–computer interaction, medical diagnosis, and virtual reality. In future research, we plan to focus on the in-depth optimization of the network architecture with the aim of significantly enhancing its ability to capture multi-scale features and dynamically adjust weights. This will improve both its processing speed and operational efficiency while also ensuring better performance in complex scenarios, making it more aligned with the growing demands of mobile and embedded applications.

Author Contributions: Conceptualization, P.W.; Methodology, G.Y.; Writing—original draft, X.M.; Project administration, J.G.; Funding acquisition, W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part in by the National Natural Science Foundation of China (62276028), the Major Research Plan of the National Natural Science Foundation of China (92267110), and the Beijing Natural Science Foundation–Xiaomi Innovation Joint Fund (L233006).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Cao, X. Eye Tracking in Human-computer Interaction Recognition. In Proceedings of the IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 18–20 August 2023; pp. 203–207.
2. Ahmad, A.; Rosli, S.A.; Chen, A.-H. Eye Tracking System Measurement of Saccadic Eye Movement with Different Illuminance Transmission Exposures during Driving Simulation. In Proceedings of the IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), Kuala Lumpur, Malaysia, 7–9 December 2022; pp. 270–273.
3. Song, B.; Du, W.; Duan, N.; Li, X. Research on pupil location algorithm of non-contact tonometer. *Electron. Meas. Technol.* **2022**, *45*, 112–117.
4. Wang, L.; Wang, C. Pupil Localization Method based on Vision Transformer. *J. Xi'an Technol. Univ.* **2023**, *43*, 561–567.
5. Xiang, Z.; Zhao, X.; Fang, A. Pupil center detection inspired by multi-task auxiliary learning characteristic. *Multimed. Tools Appl.* **2022**, *81*, 40067–40088. [CrossRef]
6. Zhou, Y. A yolo-nl object detector for real-time detection. *Expert Syst. Appl.* **2024**, *238*, 122256. [CrossRef]
7. Wang, L.; Wang, X.; Li, B. Data-driven model SSD-BSP for multi-target coal-gangue detection. *Measurement* **2023**, *219*, 113244. [CrossRef]
8. Chen, L.; Zheng, W. Research on Human Eye Key Point Detection Algorithm Based on Retina Face. *Comput. Simul.* **2023**, *40*, 213–216+354.
9. Xin, F.; Zhang, H.; Pan, H. Hybrid dilated multilayer faster RCNN for object detection. *Vis. Comput.* **2024**, *40*, 393–406. [CrossRef]

10. Zhang, C.; Chen, J. Real-Time Eye Detection Based on Multi-Task Convolutional Neural Networks. *Inf. Comput.* **2022**, *34*, 83–85.
11. da Silva Ferreira, M.V.; de Moraes, I.A.; Passos, R.V.L.; Barbin, D.F.; Barbosa, J.L., Jr. Determination of pitaya quality using portable NIR spectroscopy and innovative low-cost electronic nose. *Sci. Hortic.* **2023**, *310*, 111784. [CrossRef]
12. Qu, J.; Zhang, Y.; Tang, W.; Cheng, W.; Zhang, Y.; Bu, L. Developing a virtual reality healthcare product based on data-driven concepts: A case study. *Adv. Eng. Inform.* **2023**, *57*, 102118. [CrossRef]
13. Khalfaoui-Hassani, I. Dilated convolution with learnable spacings. *arXiv* **2024**, arXiv:2408.06383.
14. Podder, P.; Alam, F.B.; Mondal, M.R.H.; Hasan, M.J.; Rohan, A.; Bharati, S. Rethinking Densely Connected Convolutional Networks for Diagnosing Infectious Diseases. *Computers* **2023**, *12*, 95. [CrossRef]
15. Razavi, M.; Mavaddati, S.; Koohi, H. ResNet deep models and transfer learning technique for classification and quality detection of rice cultivars. *Expert Syst. Appl.* **2024**, *247*, 123276. [CrossRef]
16. Zhang, S.; Zhang, C. Modified U-Net for plant diseased leaf image segmentation. *Comput. Electron. Agric.* **2023**, *204*, 107511. [CrossRef]
17. Jafar, A.; Lee, M. High-speed hyperparameter optimization for deep ResNet models in image recognition. *Clust. Comput.* **2023**, *26*, 2605–2613. [CrossRef]
18. Gao, R. Rethink dilated convolution for real-time semantic segmentation. *arXiv* **2021**, arXiv:2111.09957.
19. Li, Y.; Lu, J.; Chen, H.; Wu, X.; Chen, X. Dilated convolutional transformer for high-quality image deraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 4199–4207.
20. Wang, Z.; Wang, Z.; Zeng, C.; Yu, Y.; Wan, X. High-quality image compressed sensing and reconstruction with multi-scale dilated convolutional neural network. *Circuits Syst. Signal Process.* **2023**, *42*, 1593–1616. [CrossRef]
21. Guo, B.; Wang, Y.; Zhen, S.; Yu, R.; Su, Z. SPEED: Semantic prior and extremely efficient dilated convolution network for real-time metal surface defects detection. *IEEE Trans. Ind. Inform.* **2023**, *19*, 11380–11390. [CrossRef]
22. Chen, J.; Hong, H.; Song, B.; Guo, J.; Chen, C.; Xu, J. MDCT: Multi-kernel dilated convolution and transformer for one-stage object detection of remote sensing images. *Remote Sens.* **2023**, *15*, 371. [CrossRef]
23. Cui, R.; Yang, R.; Liu, F.; Geng, H. HD2A-Net: A novel dual gated attention network using comprehensive hybrid dilated convolutions for medical image segmentation. *Comput. Biol. Med.* **2023**, *152*, 106384. [CrossRef]
24. Civera, M.; Sibille, L.; Fragonara, L.Z.; Ceravolo, R. A DBSCAN-based automated operational modal analysis algorithm for bridge monitoring. *Measurement* **2023**, *208*, 112451. [CrossRef]
25. Jin, H.; Montúfar, G. Implicit bias of gradient descent for mean squared error regression with two-layer wide neural networks. *J. Mach. Learn. Res.* **2023**, *24*, 1–97.
26. Behera, S.K.; Rath, A.K.; Sethy, P.K. Fruits yield estimation using Faster R-CNN with MIoU. *Multimed. Tools Appl.* **2021**, *80*, 19043–19056. [CrossRef]
27. Angelidis, A.; Vosniakos, G.C. Prediction and compensation of relative position error along industrial robot end-effector paths. *Int. J. Precis. Eng. Manuf.* **2014**, *15*, 63–73. [CrossRef]
28. Chen, W.; Yang, G.; Zhang, B.; Li, J.; Wang, Y.; Shi, H. Lightweight and fast visual detection method for 3C assembly. *Displays* **2024**, *82*, 102631. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.