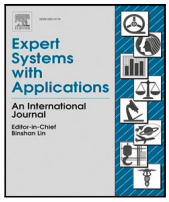
Expert Systems With Applications 256 (2024) 124935

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



SimpleCNN-UNet: An optic disc image segmentation network based on efficient small-kernel convolutions

Yichen Xiao [a,b,c,d], Jing Zhao [a,b,c,d], Yanze Yu [a,b,c,d], Xuan Ding [a,b,c,d], Shengtao Liu [a,b,c,d], Wuzhida Bao [e], Shiping Wen [e], Xingtao Zhou [a,b,c,d,*]

[a] Eye Institute and Department of Ophthalmology Eye & ENT Hospital, Fudan University, Shanghai, 200031, China
[b] NHC Key Laboratory of Myopia (Fudan University), Key Laboratory of Myopia, Chinese Academy of Medical Sciences, Shanghai, 200031, China
[c] Shanghai Research Center of Ophthalmology and Optometry, Shanghai, 200031, China
[d] Shanghai Key Laboratory of Visual Impairment and Restoration, Shanghai, 200031, China
[e] Australian AI Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, 2007, Australia

ARTICLE INFO

MSC:
0000
1111

Keywords:
U-Net
Optic disc image segmentation
Multi-layer feature fusion
Space-channel attention

ABSTRACT

Pathological myopia can lead to a series of eye diseases, including glaucoma and retinal pathologies. One of its most significant changes is the alteration in the size of the optic disc area in fundus images. Therefore, precise segmentation of the optic disc area is particularly important in ocular medical diagnosis. Although many well-established methods in medical image segmentation rely on Fully Convolutional Networks (FCNs), they often struggle to capture global context compared to Transformer models. However, incorporating Transformers generally necessitates larger training datasets, which can pose a significant challenge. To address these issues, Convolutional Neural Networks (CNNs) with large convolutional kernels have been proposed as an alternative for capturing contextual information, but they come with increased parameter counts and higher computational costs during training. In this paper, we introduce SimpleCNN-UNet, a lightweight image segmentation network based on small-kernel convolutions. By strategically stacking these small convolutions, we emulate the receptive field of large-kernel convolutions while substantially reducing the number of parameters. Another novel feature of SimpleCNN-UNet is the Multi-Layer Cross-Attention Gate, designed for efficient feature fusion across different levels. To overcome the limited availability of fundus image data, we employed extensive data augmentation techniques on our existing dataset. Our experimental results on the iChallenge-PM, iChallenge-AMD, iChallenge-GON, and IDRiD datasets demonstrate that SimpleCNN-UNet outperforms other image segmentation networks in terms of performance while also offering faster inference speeds and lower training costs.

1. Introduction

Myopia, a common ophthalmic condition, not only affects visual acuity but can also lead to various eye health issues. In some individuals, myopia may progress into a more severe form known as Pathological myopia which is not only characterized by high degrees of myopia but can also lead to fundus diseases, including retinal detachment, macular degeneration, and pathological changes in the optic disc. Changes in the optic disc region are crucial for assessing and monitoring the progression of pathological myopia. Due to the continuous elongation of the eyeball, the optic disc may undergo morphological changes such as tilting, deformation, or deepening of the cupping (the central concave part of the disc), which may cause or exacerbate vision problems. Therefore, detailed observation and analysis of the optic disc are crucial for patients with pathological myopia, which leads to the development of optic disc image segmentation (Minaee et al., 2021) technology. In traditional ophthalmic examinations, assessment of the optic disc largely depends on the physician's prior knowledge, while the application of image segmentation technology enhances the objectivity and accuracy of this process. Therefore, optic disc image segmentation technology is playing an increasingly important role in modern ophthalmological treatment.

Traditional image segmentation techniques include methods based on thresholding (Sujji, Lakshmi, & Jiji, 2013; Zhu, Xia, Zhang, & Belloulata, 2007), region-based (Lalaoui & Mohamadi, 2013; Wang, Wu, & Pan, 2013), clustering (Jurio, Pagola, Galar, Lopez-Molina, &

* Corresponding author.
E-mail addresses: yichenxiao@fudan.edu.cn (Y. Xiao), zhaojing_med@fudan.edu.cn (J. Zhao), yzyu18@fudan.edu.cn (Y. Yu), 17211260004@fudan.edu.cn (X. Ding), 505560283@qq.com (S. Liu), wuzhida.bao@student.uts.edu.au (W. Bao), shiping.wen@uts.edu.au (S. Wen), xingtaozhou@fudan.edu.cn (X. Zhou).

https://doi.org/10.1016/j.eswa.2024.124935
Received 24 January 2024; Received in revised form 18 June 2024; Accepted 28 July 2024
Available online 6 August 2024
0957-4174/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Paternain, 2010; Manoharan, 2020), and edge detection (Aslam, Khan, & Beg, 2015). Thresholding depends on the appropriate selection of threshold T and is easily affected by noise and uneven illumination; classical threshold selection techniques include the maximum entropy method (Wang et al., 2017) and Otsu's method (Li, Lv, Xu, Li, & Gu, 2020), among others. Region-based methods divide image areas through similar features, with primary methods including region growing (Preetha, Suresh, & Bosco, 2012), region merging (Wang, Jensen, & Im, 2010), and watershed algorithms (Monteiro & Campilho, 2008); they are sensitive to noise and computationally intensive. Clustering divides pixels into clusters, being simple and fast but susceptible to the influence of initial centers. Edge detection identifies boundaries in high-frequency areas, with commonly used edge detection operators (Jing, Liu, Wang, Zhang, & Sun, 2022) including Sobel, Canny, Prewitt, Laplacian, and others.

However, due to real-world images' distinct characteristics from medical images, such as intense colors and clear edges, traditional image segmentation techniques reached a bottleneck. With the advancement of computer technology, image segmentation techniques based on neural network models were introduced to address the issues faced by traditional methods. Scholars have utilized deep learning methods for image segmentation. Jonathan Long et al. first introduced Fully Convolutional Networks for Semantic Segmentation (Long, Shelhamer, & Darrell, 2015), applying deep learning methods to the field of image segmentation and guiding subsequent methods. Based on FCN concepts, Chen and others proposed a milestone model DeepLab (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014) that incorporated atrous convolution to expand the receptive field and significantly improved the accuracy of boundary segmentation using conditional random fields. Building on this, the team subsequently introduced DeepLab v2 (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017), DeepLab v3 (Chen, Papandreou, Schroff & Adam, 2017), and DeepLab v3+ (Chen, Zhu, Papandreou, Schroff, & Adam, 2018). They introduced Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale contextual information. They continuously optimized ASPP, incorporating a global average pooling layer to enhance contextual understanding. Additionally, they added a decoder model to refine segmentation results. Considering the scarcity of medical image samples, Olaf and others proposed U-Net (Ronneberger, Fischer, & Brox, 2015), based on an encoder–decoder architecture and skip-connection design, which has been widely recognized and remains influential. Networks improved upon U-Net, such as U-Net++ (Zhou, Siddiquee, Tajbakhsh, & Liang, 2018), Attention U-Net (Oktay et al., 2018), UNeXt (Valanarasu & Patel, 2022), have achieved better segmentation results. To address the limitations of U-Net in explicitly modeling long-range dependencies, Chen and others combined Transformer with U-Net to propose TransUNet (Chen et al., 2021). After comparing fully convolutional or Transformer and convolution combined U-shaped networks, Cao and others proposed Swin U-Net (Cao et al., 2022), based on a pure Transformer design.

The "Diabetic Retinopathy: Segmentation and Grading Challenge" (Porwal et al., 2020) held at ISBI-2018 provided an essential platform for the optic disc image segmentation field to showcase and evaluate the latest segmentation technologies, introducing a new publicly available diabetic retinopathy image dataset (IDRiD) (Porwal et al., 2018). Building on the achievements in medical image segmentation, numerous scholars are dedicated to developing networks specifically for retinal segmentation. Hasan, Alam, Elahi, Roy, and Martí (2021) developed an end-to-end encoder–decoder network, DRNet, capable of accurately segmenting and locating the OD and the orbital center area, achieving an mIOU of 0.845 on the IDRiD dataset. Maysanjaya, Kesiman, and Indradewi (2022) utilized Mask R-CNN with ResNet50 as the backbone network to segment OD and fundus exudates, achieving IoU values of 0.8431 and 0.9933 on the IDRiD dataset, respectively. Tang et al. (2024) introduced a boundary-aware cascading network, W-Net, employing a multi-level strategy for training and demonstrating excellent OD segmentation performance across various datasets. To reduce reliance on manual annotations by medical experts, Xiong, Liu, Sharan, Coiera, and Berkovsky (2022) proposed a weakly labeled Bayesian U-Net that does not depend on manual annotation of optic disc masks, significantly reducing the time required for optic disc mask annotation and simplifying the segmentation process. Additionally, images captured by different devices exhibit significant differences in brightness, shape, color, and orientation between the optic cup and disc areas, significantly affecting the segmentation performance of models. To address the performance degradation caused by cross-domain segmentation issues, Hua, Fang, Tang, Cheng, and Yu (2023) proposed a fundus domain-generalization segmentation framework called DCAM-NET, significantly enhancing the segmentation model's generalization ability for target domain data and improving the capture of details from the source domain data. Hasan et al. (2021) introduced an end-to-end encoder–decoder network, named DRNet, designed for the segmentation of the Optic Disc (OD) and the localization of OD and Fovea centers. DRNet employs a residual skip connection to mitigate spatial information loss, demonstrating superior performance across various metrics compared to state-of-the-art methods. Fu et al. (2021) developed an automated OD segmentation approach that integrates U-net with a model-driven probability bubble technique. This method incorporates the positional relationship between retinal vessels and the OD into the output layer of U-net, significantly enhancing segmentation accuracy in the presence of bright lesions.

As an alternative to transformers, CNNs with large convolutional kernels have been utilized to extract global contextual information (Liu, Chen et al., 2022; Liu, Mao et al., 2022), which have a larger receptive field. However, large-kernel convolutions also come with higher computational costs and increased parameter counts, and some of them involve additional post-processing, such as reparameterization or sparsity. Additionally, the skip connections in U-Net simply concatenate upsampled deep features with shallow skip features, which does not fully consider feature alignment issues, such as ensuring precise matching of these feature maps in scale and dimension, leading to problems like boundary effects and alignment inaccuracies, thereby affecting the model's segmentation accuracy. Inspired by Simple Convolutional Neural Network (SCNN) (Lai et al., 2023), we propose an efficient fully convolutional segmentation network based on small-kernel convolutions, named SimpleCNN-UNet, which primarily consists of the SCNN block and a Multi-Layer Cross-Attention Gate. The SCNN block is used to fully extract contextual information, and the Multi-Layer Cross-Attention Gate is for merging features from different layers.

Overall, this paper makes the following contributions: (1) We propose a fully convolutional medical image segmentation network with efficient small-kernel convolutions; (2) We introduce the Multi-Layer Cross-Attention Gate, which effectively merges features from different levels; (3) To further optimize the performance of the model, we expanded the optic disc image dataset through data augmentation techniques. This step will not only increase the diversity of the data but also enhance the training efficiency and segmentation effectiveness of the network; (4) Comparative experiments demonstrate that our proposed network has the excellent performance in Optic Disc Image Segmentation.

2. Datasets and preprocessing

2.1. Datasets

This study utilized two sets of datasets, with the first group originating from Baidu Brain's fundus image datasets: iChallenge-PM, iChallenge-GON, and iChallenge-AMD (Brain, 2023); the second group consisted of fundus image datasets IDRiD (Porwal et al., 2018), taken by retinal specialists at an ophthalmology clinic in Nanded, Maharashtra, India. The iChallenge-PM dataset focuses on Pathologic Myopia (PM), containing numerous fundus images specifically for the

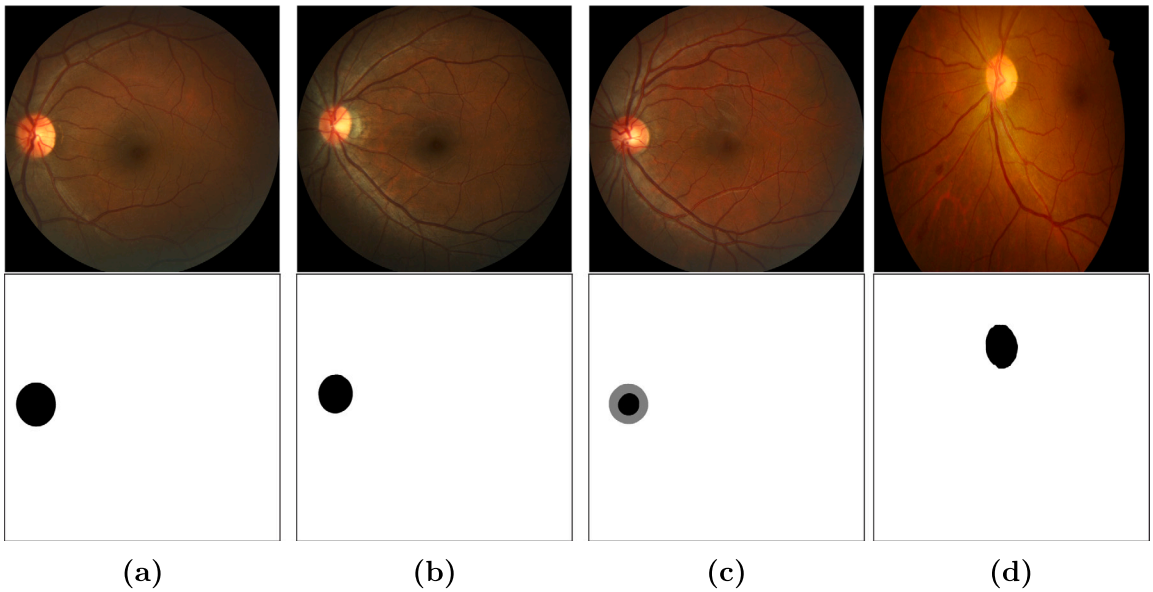


Fig. 1. (a) The sample of iChallenge-PM (b) The sample of iChallenge-AMD (c) The sample of iChallenge-GON (d) The sample of IDRiD.

Table 1
Distribution and details of the four different datasets used for evaluating the proposed SimpleCNN-UNet.

Dataset	Resolution	Bit depth	Size	Dots per Inch
iChallenge-PM	1444 × 1444	24-bits	~850 KB	96
iChallenge-AMD	2124 × 2056	24-bits	~1640 KB	96
iChallenge-GON	2124 × 2056	24-bits	~1640 KB	96
IDRiD	4288 × 2848	24-bits	~800 KB	300

training and evaluation of automatic optic disc and cup segmentation algorithms, including available data of 400 training images and 400 validation images, with 381 of the training images and 379 of the validation images annotated by experts. The iChallenge-AMD dataset is dedicated to the study of Age-related Macular Degeneration (AMD), comprising a vast number of detailed annotated fundus images. Only 400 training images are publicly available for the AMD dataset, with 270 of them annotated by experts. The iChallenge-GON dataset, designed for glaucoma (Glaucoma) diagnostic research, also offers 400 training and 400 validation images, with precise annotations by experts in the optic disc and cup regions of each image. The IDRiD dataset was acquired using a Kowa VX-10 alpha digital fundus camera with a 50-degree field of view (FOV), centering all images near the macula. It includes 81 color fundus images exhibiting signs of Diabetic Retinopathy (DR), with mask annotations provided for the optic disc region. Table 1 summarizes the detailed parameters of images within the aforementioned datasets.

We consolidated and filtered the iChallenge dataset, resulting in 1830 valid images. These were then divided into training/validation and test sets using an 8:2 ratio. The training/validation set was divided into training and validation sets at a 9:1 ratio. Consequently, we obtained 1318 images for the training set, 366 images for the test set, and 146 images for the validation set. To verify the generalizability and robustness of the method proposed in this paper, we divided the IDRiD dataset according to the guidelines of the Diabetic Retinopathy: Segmentation and Grading Challenge—allocating 54 images for the training set and 27 images for the test set. Data examples are presented in Fig. 1. The aspect ratio of all images and their labels in both datasets was standardized to 1:1 for ease of display.

2.2. Data preprocessing and augmentation

In the iChallenge-GON dataset, experts annotated the optic disc and cup separately, as shown in Fig. 2(b). However, since the optic cup is a part of the optic disc and is located at its center, it is necessary to merge the labeled areas of the optic disc and cup before training. We used threshold segmentation to set the grayscale of all pixel values other than 255 to 0, and the segmented label images are shown in Fig. 2.

To address the challenge of insufficient fundus image datasets, we performed data augmentation on the training data to increase its volume and enhance the model's robustness. The augmented image data is shown in Fig. 3. Fig. 3(a) shows the original training image, to which we applied various augmentation techniques: random horizontal and vertical flips, random rotations (−90° to 90°), random grayscale conversion, random addition of grid masks, random affine transformations, random addition of masks sized 5%–10% of the image with aspect ratios between 1 and 2, and random cropping of 1/4th of the image area.

3. Method

3.1. Network structure

This section introduces our proposed SimpleCNN-UNet, depicted in Fig. 4, which features a typical U-shaped structure. It can be divided into two stages: an encoder and a decoder, interconnected with skip connections. During the encoder stage, SCNN Blocks (Lai et al., 2023) extract rich spatial information at different levels, followed by general convolutional blocks to increase the number of channels. In the decoder stage with skip connections, upsampled deep features are fused with shallow features from skip connections through multi-scale cross-attention blocks. Subsequently, general convolutional blocks are used to reduce the number of channels.

3.2. Encoder stage

The encoder consists of five layers, arranged from top to bottom, as depicted in Fig. 4. Each layer comprises an SCNN Block, a general convolutional block, and a downsampling operation. The general convolutional block consists of a convolutional layer with a kernel size of 3 × 3, stride of 1, and padding of 1, followed by batch normalization and a ReLU activation layer, effectively increasing the channel count of

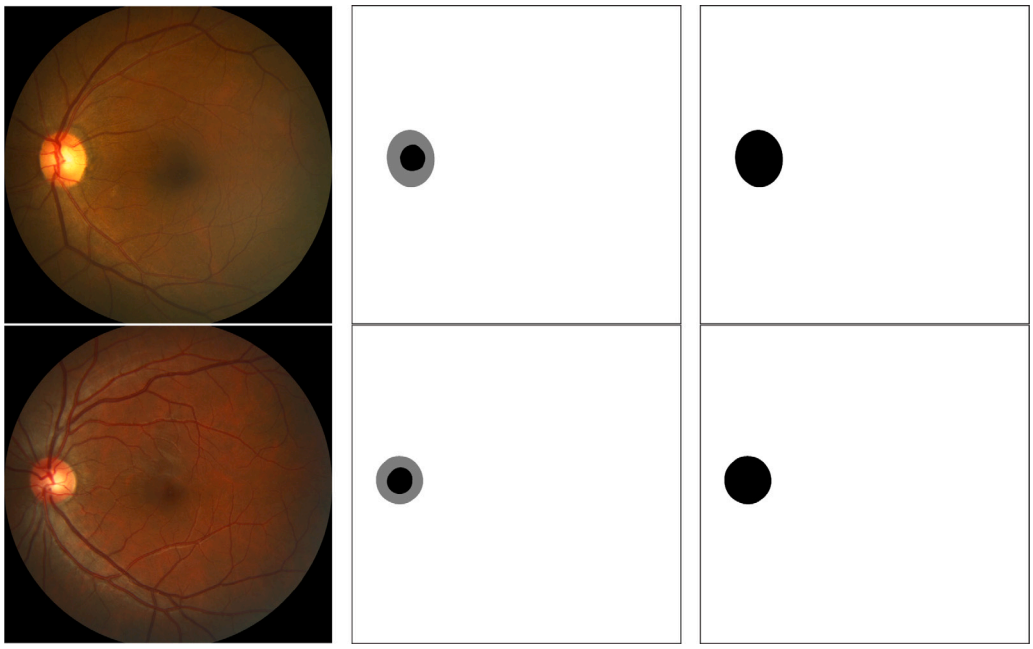


Fig. 2. (a) original GON image data (b) original GON image label (c) the label after threshold segmentation.

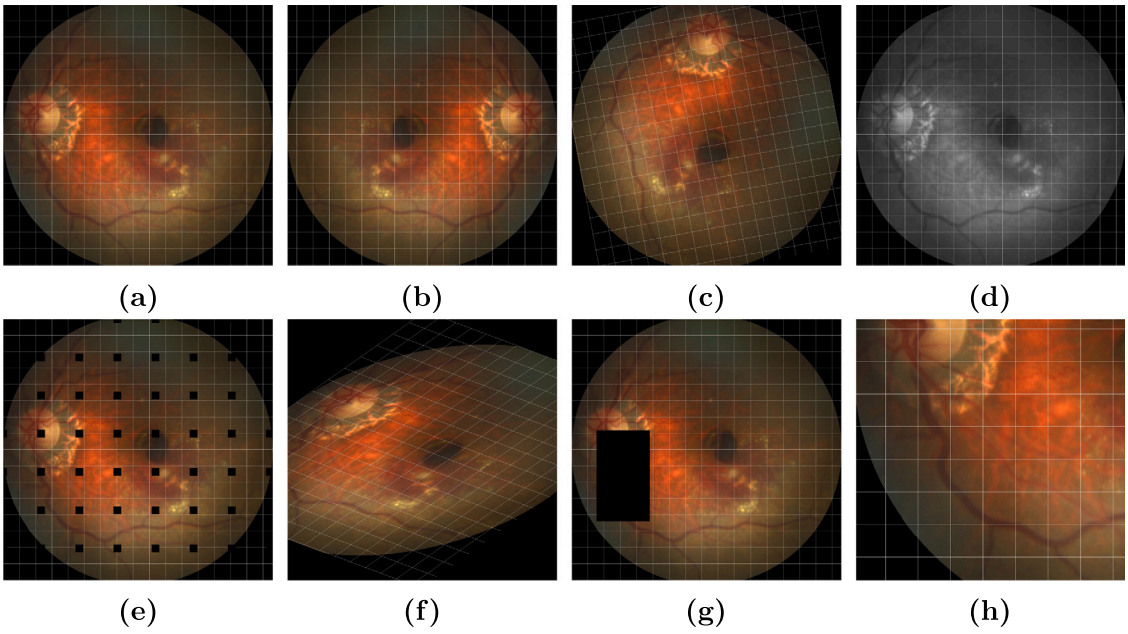


Fig. 3. Data augmentations graphs (a) original image (b–h) flipped image, rotated image, gray image, gridmasked image, affined image, erased image and cropped image.

the feature images. Consistent with U-Net and most of its variants (Oktay et al., 2018; Valanarasu & Patel, 2022; Zhou et al., 2018), we use max pooling for the downsampling step, with a filter window of 2×2 and a stride of 2. Unlike traditional convolution or average pooling, max pooling can better highlight salient features within the feature maps and exhibit stronger noise resistance, which is particularly beneficial when processing medical images where high-frequency features are not prominent.

The primary feature extraction module of the encoder is the SCNN Block, as depicted in Fig. 5. It is mainly composed of two sets of depthwise convolution followed by pointwise convolution, a configuration also known as depthwise separable convolution, which has been proven effective as early as in MobileNet (Howard et al., 2017). Depthwise convolution, with the number of groups equal to the number of channels, is used to extract spatial information of images, followed by pointwise convolution, with a kernel size of 1×1, which merges cross-channel information while maintaining spatial dimensions. The SCNN Block does not employ traditional large kernel size depthwise convolutions (Liu, Mao et al., 2022), but rather utilizes two depthwise convolutions with a kernel size of 3×3 and a stride of 1. This thin and deep architecture is capable of obtaining a receptive field comparable to that of large kernel convolutions, while simultaneously reducing the number of parameters and the difficulty of training. Additionally, the SCNN Block employs an inverted bottleneck design (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018) between the two depthwise convolution —

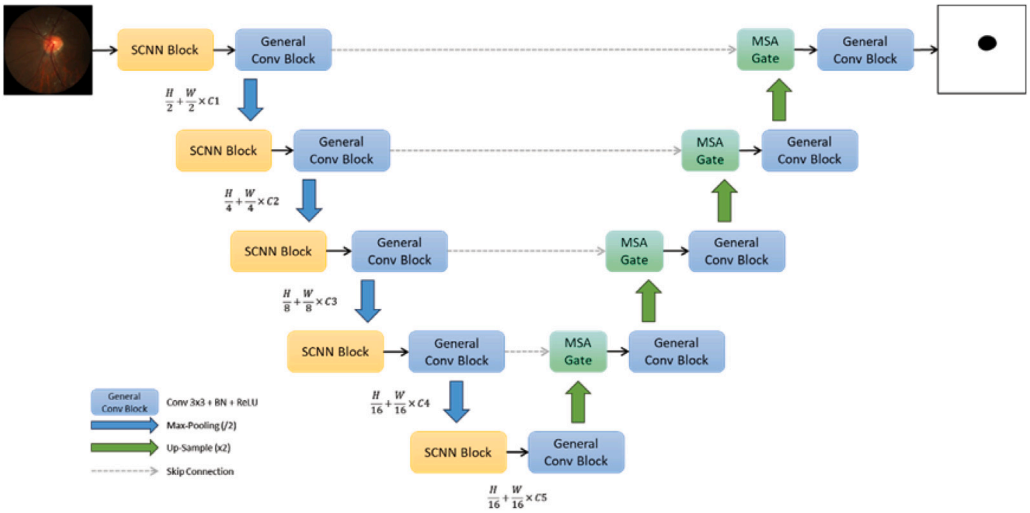


Fig. 4. SimpleCNN-UNet.

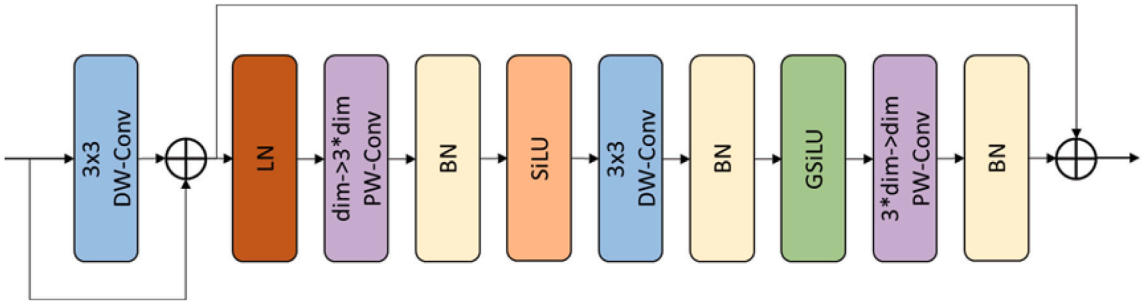


Fig. 5. SCNN Block.

pointwise convolution pairs, expanding the hidden dimensions between the convolutional layers to three times the input dimensions and then restoring, to capture information across different spatial dimensions. The first depthwise convolution and the final residual connection enhance the efficiency of feature propagation. Finally, to optimize the network structure, layer normalization (LN), batch normalization (BN), Sigmoid linear unit (SiLU), and global sigmoid linear unit (GSiLU) are employed between the convolutional blocks. The specific definition of the SCNN Block is as follows:

$$f_l' = DepthwiseConv\left(f_{l-1}\right) + f_{l-1} \tag{1}$$

$$f_l'' = \sigma_2\left(BN\left\{PointwiseConv\left(LN(f_l')\right)\right\}\right) \tag{2}$$

$$f_l''' = \sigma_1\left(BN\left\{DepthwiseConv\left(f_l''\right)\right\}\right) \tag{3}$$

$$f_l = BN\left(PointwiseConv\left(f_l'''\right)\right) + f_l' \tag{4}$$

where f_l represents the output feature map of the lth layer SCNN block, σ_1 represents GSiLU, σ_2 represents SiLU, BN stands for Batch Normalization and LN for Layer Normalization. As the SCNN block does not alter the resolution and channel count of the feature map, we subsequently use a general convolutional block to double the number of channels.

3.3. Decoder stage with skip connection

The decoder has the same number of layers as the encoder, consisting of five layers from bottom to top. Among them, the top four layers each incorporate a Multi-Layer Cross-Attention Gate (MLCAGate), the primary function of which is to maintain scale consistency between features while effectively fusing features of different scales. Similarly, we use general convolutional blocks comprising a convolutional layer with a 3 × 3 kernel, BN layer, and ReLU for channel reduction, and employ nearest-neighbor interpolation for feature map upsampling.

The inherent disparities in dimensionality and scale that exist between deep and shallow features, render the effective alignment of their information unattainable through mere upsampling. To better fuse features across different network layers, we propose the Multi-Layer Cross-Attention Gate (MLCAGate), as illustrated in Fig. 6. This gate incorporates a novel approach for feature fusion across disparate layers, where the dashed line represents the input of shallow features and the solid line represents the input of deep features. Initially, we perform convolution, GELU activation, and batch normalization (BN) on both the upsampled deep features and the skip-connected shallow features. Then, the processed shallow features are element-wise added to the deep features. Since shallow features contain more spatial information while deep features possess richer channel information, we execute an element-wise multiplication of the deep features' channel attention with the shallow features, followed by an element-wise multiplication of the shallow features' spatial attention with the deep features. Finally, we add the products of both to obtain the fused features. The strategy of extracting spatial and channel attention separately has been proven effective in Convolutional Block Attention Module (CBAM) (Woo, Park, Lee, & Kweon, 2018). By combining spatial-channel attention from different levels, the Cross Attention Gate can focus on both deep channel and shallow spatial information, thus enhancing the network's ability to extract multi-level feature information. The specific implementation of the Multi-Layer Cross-Attention Gate is as follows:

$$f_s' = \sigma_1(BN(Conv(f_s))) \tag{5}$$

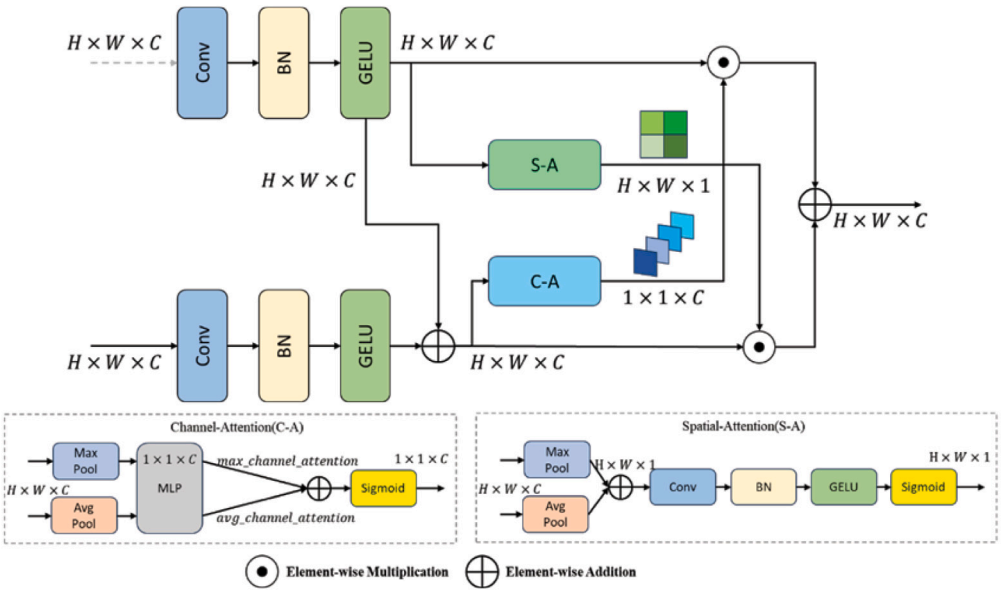


Fig. 6. Multi-Layer Cross-Attention Gate.

$$f'_d = f'_s + \sigma_1(BN(Conv(f_d))) \tag{6}$$

$$C - Map(d) = \sigma_2(MLP(MaxPool(f'_d)) + MLP(AvgPool(f'_d))) \tag{7}$$

$$S - Map(s) = \sigma_2(\sigma_1(BN(Conv(Concat[MaxPool(f'_s); AvgPool(f'_s)])))) \tag{8}$$

$$f_{out} = f'_s \odot C - Map(d) + f'_d \odot S - Map(s) \tag{9}$$

where f_d, f_s and f_{out} represent the upsampled deep features, skip-connected shallow features, and the output features of the MCAGate, respectively, σ_1 stands for RELU, σ_2 for Sigmoid, BN for Batch Normalization, and MLP for a multi-layer perceptron (MLP) with one hidden layer, where the hidden layer's feature map size is C/r×1 × 1, with r being the reduction ratio. It is noteworthy that MLCAGate does not change the resolution and number of channels of the feature map during processing. Consequently, we follow this by using a general convolutional block to reduce the number of channels to half their original sizes, effectively compressing and optimizing the features.

4. Experiments and results

4.1. Evaluation metric

In this study, we employ Intersection over Union (IoU), Accuracy (Acc), and F1 Score (F1) as evaluation metrics. IoU quantifies the overlap between predicted and ground truth segmentation regions. Accuracy measures the overall pixel-wise classification correctness. The F1 Score, balancing precision and recall, assesses model performance on imbalanced datasets with minority classes. Researchers can comprehensively assess and compare different image segmentation models by combining these metrics.

Intersection over Union (IoU). In the field of deep learning-driven image segmentation, IoU is widely used as the primary evaluation metric, which provides a quantifiable standard for measuring the degree of overlap between the predicted segmentation area and the actual segmentation area. Specifically, IoU is defined as the ratio of the intersection area to the union area between the predicted and the actual annotations. Its mathematical expression is:

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

where TP (True Positives) denotes the pixels correctly classified as the target class. FP (False Positives) represents pixels incorrectly classified as the target class. FN (False Negatives) refers to pixels from the target class incorrectly classified as background. TN (True Negatives) denotes the pixels correctly classified as background. The IoU score ranges from 0 to 1, with higher values indicating better segmentation performance, as it signifies a greater overlap between the predicted and ground truth regions.

Table 2
Comparative experiments based on iChallenge-PM, iChallenge-GON and iChallenge-AMD.

Types	Method	PixAcc (%)	IoU (%)	F1 (%)
U shape	U-Net	93.20	87.66	92.30
	U-Net++	95.19	89.51	94.42
	U-Net3+	95.65	90.23	94.63
	Attention U-Net	95.17	89.31	94.29
Same Datasets	Pf-PSP-Net	94.94	89.93	/
	2C-128N-0D	95.00	/	/
	U-Net PB	95.43	89.87	93.97
	Ours	95.98	90.45	94.77

Accuracy (ACC). Accuracy is another essential metric for assessing a model's overall performance across the entire dataset. It is defined as the ratio of correctly classified pixels to the total number of pixels, calculated as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

where Acc is an intuitive metric to evaluate a model's overall segmentation performance, with higher values indicating better performance.

F1-score. The F1 Score (F1) is the harmonic mean of the model's precision and recall. It effectively measures a classification model's predictive performance, especially for minority class samples in imbalanced datasets. The F1 Score is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{14}$$

The F1 Score ranges from 0 to 1, with higher values indicating better model performance.

Table 3
Comparative experiments based on IDRiD Dataset.

Team	Method	IoU (%)	Input Size (Pixels)	External Dataset
ZJU-BH-SGEX	Mask R-CNN	93.38	1024 × 1024	RIGA
VRT	U-Net	93.05	640 × 640	DRIVE, DRIONS-DB
CBER	Handcrafted Features	89.12	536 × 356	Drishti-GS
IITKgpKLIV	SegNet	85.72	536 × 356	/
SDNU	Mask R-CNN	78.92	1984 × 1318	/
/	DRNet	84.50	/	/
/	Mask R-CNN -ResNet50	84.31	512 × 512	/
/	W-Net	90.78	/	/
/	U-Net PB	91.07	512 × 512	/
/	Ours	91.44	512 × 512	/
/	Ours pre-training	93.86	512 × 512	i-Challenge

4.2. Implement details

The experimental implementation details of this study are as follows: The computer hardware configuration includes a high-performance CPU with an Intel 12th Gen Core i7 and 16 GB of RAM, as well as an NVIDIA RTX 4080 GPU with 16 GB of VRAM. These hardware components ensure efficiency and stability throughout the experimental process. The software configuration chose the Ubuntu 16.04 Linux operating system as the base environment and utilized the Pytorch GPU framework, CUDA, and the Python programming language for implementing model training and testing. During the model training process, we used the Adam optimizer to update and optimize model parameters, ensuring efficiency and precision in model training. The initial learning rate was set to 0.01 to balance the training speed and model convergence. To further enhance the model's performance, we set 100 training epochs and reduced the learning rate by a factor of 0.2 at the 20th, 40th, and 80th epochs. This helps the model achieve stable convergence in the later stages of training. To maximize the use of hardware resources while considering computational efficiency, training speed, and the stability and convergence of the model, we set the batch size to 8. Through these meticulous settings and optimizations, our experimental platform provided strong support for model training, ensuring the reliability and validity of the experimental results.

4.3. Quantitative evaluation

In the quantitative evaluation phase of this study, Table 2 comprehensively summarizes the performance comparison between our proposed fundus image segmentation network SimpleCNN-UNet and various U-shaped fully convolutional image segmentation networks (including U-Net, U-Net++, U-Net3+, AttentionU-Net), and other advanced methods using the same dataset, such as Phase-fusion PSPNet(Pf-PSPNet) (Rauf, Gilani, & Waris, 2021), 2C-128N-0D (Fang, Shen, Zheng, Zhu, & Wu, 2021) and U-net with Probability Bubble (U-Net PB) (Fu et al., 2021). It is evident from the table that our SimpleCNN-UNet achieved the best results in all performance metrics. Compared to other methods, SimpleCNN-UNet showed improvements in key metrics such as Pixel Accuracy (PixAcc), Intersection over Union (IoU), and F1 Score, with increases ranging from 0.33%–2.99%, 0.22%–3.18%, and 0.14%–2.68% respectively. Additionally, we individually compared the performance of the algorithms on different datasets, with specific results shown in Tables 4–6. The experimental results demonstrate that the SimpleCNN-UNet proposed in this paper exhibited good performance in optic disc segmentation tasks and achieved the best results on all datasets.

We compared our method with several SOTA methods on the IDRiD dataset. The results are shown in Table 3. Notably, these methods include the recently outstanding DRNet (Hasan et al., 2021), Mask R-CNN-ResNet50 (Maysanjaya et al., 2022) and W-Net (Tang et al., 2024) on the IDRiD dataset, methods proposed by the top five teams in the optic disc segmentation competition (Porwal et al., 2020) and a U-Net Probability Bubble (U-Net PB) approach (Fu et al., 2021) which combines traditional algorithms with deep learning. Since most methods have not disclosed their code, we obtained their experimental data for comparison from respective research papers and official optic disc segmentation competition results. According to the results in Table 2, SimpleCNN-UNet, pre-trained on the i-challenge dataset, achieved the best performance among all methods. Even without training on additional datasets, SimpleCNN-UNet exhibited an accuracy of 91.44%, ranking third. It is worth emphasizing that the image input size processed by our model is only 512 × 512, indicating that our model can learn more effective knowledge from lower-resolution images, and its performance surpasses models that process higher-resolution, finer-grained images, which explicitly demonstrates our model's efficiency and applicability, especially in resource-constrained application scenarios.

Table 7 shows the comparison of our model with U-Net, U-Net++, U-Net3+, and Attention U-Net on three key performance metrics: Parameters, GFLOPs, and FPS. The results indicate that our model is more lightweight, with only 11.21M parameters and a reduction in GFLOPs

Table 4
Comparative experiments based on iChallenge-PM.

Types	Method	PixAcc (%)	IoU (%)	F1 (%)
U shape	U-Net	94.79	86.97	92.83
	U-Net++	92.91	83.87	90.20
	U-Net3+	95.65	88.81	93.91
	Attention U-Net	95.64	86.50	92.56
	Ours	96.45	90.23	94.93

Table 5
Comparative experiments based on iChallenge-AMD.

Types	Method	PixAcc (%)	IoU (%)	F1 (%)
U shape	U-Net	94.63	84.16	91.06
	U-Net++	94.18	84.22	91.04
	U-Net3+	93.85	83.98	88.85
	Attention U-Net	94.51	86.51	92.39
	Ours	94.77	86.68	91.77

Table 6
Comparative experiments based on iChallenge-GON.

Types	Method	PixAcc (%)	IoU (%)	F1 (%)
U shape	U-Net	94.47	91.44	95.53
	U-Net++	95.74	92.25	95.97
	U-Net3+	95.60	92.57	95.13
	Attention U-Net	95.50	91.75	95.69
	Ours	96.06	92.96	96.14

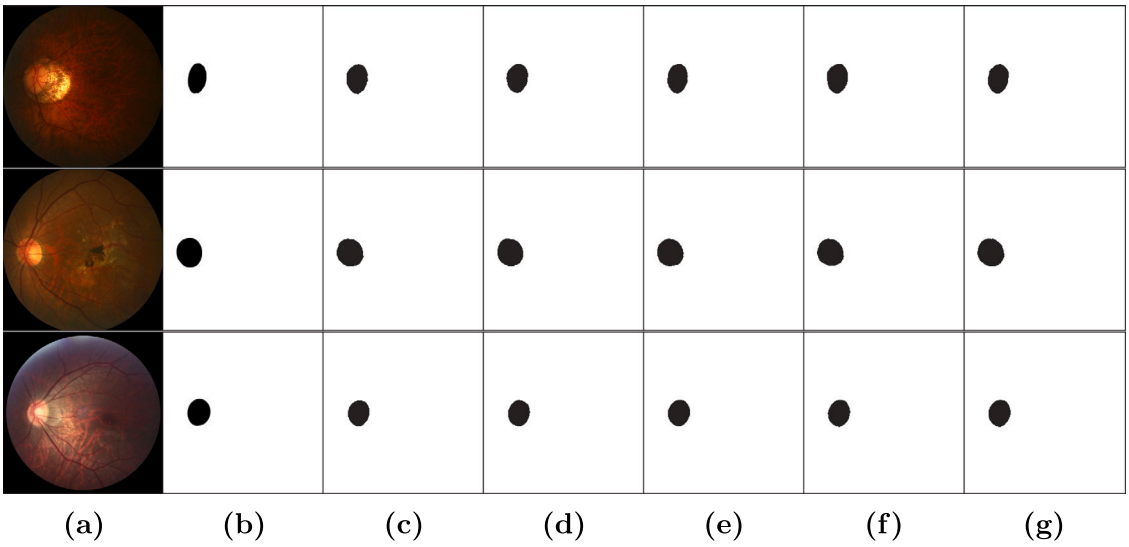


Fig. 7. Qualitative evaluation (a) original image (b) image mask (c) U-Net (d) U-Net++ (e) U-Net3+ (f) Attention U-Net (g) SimpleCNN-UNet.

Table 7
Comparison of model parameters.

Types	Method	Params(M)↓	GFLOPs↓	FPS↑
U shape	U-Net	34.52	65.52	139.32
	U-Net++	26.90	37.62	125.50
	U-Net3+	26.97	199.74	50.60
	Attention U-Net	34.87	66.63	129.92
	Ours	11.21	18.87	278.63

Table 8
Results of ablation study with different strategy combinations and model parameters.

Method	PixAcc (%)	IoU (%)	F1 (%)
Original U-Net	93.20	87.66	92.30
U-Net+SCNN	94.13	89.39	93.89
U-Net+SCNN+AttGate	95.44	90.11	94.13
U-Net+BCNN+AttGate	94.04	89.24	94.08
Ours	95.98	90.45	94.77
Method	**Params(M)↓**	**GFLOPs↓**	**FPS↑**
Original U-Net	34.52	65.52	139.32
U-Net+SCNN	12.46	15.78	245.17
U-Net+SCNN+AttGate	11.21	18.87	278.63
U-Net+BCNN+AttGate	12.13	25.18	199.74
Ours	11.21	18.87	278.63

Ours = U-Net+SCNN+AttGate+Data Augmentation.

to 18.87, implying faster inference speeds with the same computational power. Our model achieves 278.63 FPS regarding processing speed, significantly surpassing other methods. Based on the hardware environment and hyperparameter configuration described in section 4.2, our model's average training time per epoch on the i-challenge integrated dataset is about 400 s, compared to approximately 570 s for U-Net, fully demonstrating our model's significant advantages in efficient and accurate segmentation of the fundus optic disc region.

4.4. Qualitative evaluation

In the qualitative assessment, we compared the image segmentation results of our proposed method with those of the U-shaped networks mentioned in the quantitative evaluation. As shown in Fig. 7, the model performance corresponds to Table 2 in the quantitative experiments. The comparison results clearly show that our method performs better in terms of pixel-level segmentation accuracy. Particularly in handling optic disc edges, SimpleCNN-UNet presented more precise segmentation effects and was closer to the Ground Truth in all aspects.

4.5. Ablation study

Furthermore, we conducted a series of ablation experiments to evaluate the specific contributions of each module we proposed to the model's performance. As shown in Table 8, after adding the SCNN Block to the original U-Net, a significant performance improvement has been made with a reduction in the number of parameters and GFLOPs, which increases the operational speed. Additionally, after incorporating the Multi-Layer Cross-Attention Gate, despite a slight increase in the number of GFLOPs, the accuracy has made improved significantly. The segmentation accuracy was further improved when the model was trained using an expanded dataset. To demonstrate the effectiveness of our efficient small-kernel depth convolution stacking strategy, we included a control group using large-kernel depth convolutions in our ablation study. Specifically, we replaced the two 3×3 small-kernel depth convolutions in the SCNN block with one 7×7 large-kernel depth convolution, and named it the BCNN block. The experimental results showed that the new structure model had increased parameters and GFLOPs, and a significant reduction in FPS. These results fully demonstrate that our stacked small-kernel convolution method can achieve higher training efficiency and reduce resource consumption. These ablation results fully demonstrate the positive impact of each module we proposed in enhancing optic disc image segmentation.

5. Conclusion

It is important to develop the optic disc image segmentation technology in modern ophthalmology, however there exist many issues in current segmentation models such as high computational resource consumption, inadequate feature fusion, and data scarcity. Thus we proposed SimpleCNN-UNet, which uses an efficient small-kernel convolutional SCNN Block, and introduces the Multi-Layer Cross-Attention Gate to fuse features from different levels. For the scarce optic disc image data, we utilized various image enhancement techniques to expand the dataset. Experimental results show that our proposed method outperforms other existing methods on multiple datasets, fully validating its effectiveness. Furthermore, ablation studies further confirm the importance of each module we proposed for optic disc image segmentation. Overall, our research provides new technical tools for ophthalmic diagnosis and treatment. In the future, we plan to apply

the model to the segmentation of other important areas like the optic cup and retinal vasculature.

CRediT authorship contribution statement

Yichen Xiao: Writing, Conceptualization, Methodology. **Jing Zhao:** Writing – original draft. **Yanze Yu:** Visualization, Investigation. **Xuan Ding:** Writing, Validation. **Shengtao Liu:** Software, Validation. **Wuzhida Bao:** Writing, Software. **Shiping Wen:** Editing, Supervision. **Xingtao Zhou:** Review, Supervision.

Declaration of competing 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.

Data availability

No data was used for the research described in the article.

References

Aslam, A., Khan, E., & Beg, M. S. (2015). Improved edge detection algorithm for brain tumor segmentation. *Procedia Computer Science*, *58*, 430–437.

Brain, B. (2023). ichallenge data. https://ai.baidu.com/broad/introduction.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision* (pp. 205–218). Springer.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision* (pp. 801–818).

Fang, X., Shen, Y., Zheng, B., Zhu, S., & Wu, M. (2021). Optic disc segmentation based on phase-fusion PSPNet. In *Proceedings of the 2nd international symposium on artificial intelligence for medicine sciences* (pp. 152–156).

Fu, Y., Chen, J., Li, J., Pan, D., Yue, X., & Zhu, Y. (2021). Optic disc segmentation by U-net and probability bubble in abnormal fundus images. *Pattern Recognition*, *117*, Article 107971.

Hasan, M. K., Alam, M. A., Elahi, M. T. E., Roy, S., & Martí, R. (2021). DRNet: Segmentation and localization of optic disc and Fovea from diabetic retinopathy image. *Artificial Intelligence in Medicine*, *111*, 102001.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hua, K., Fang, X., Tang, Z., Cheng, Y., & Yu, Z. (2023). DCAM-NET: A novel domain generalization optic cup and optic disc segmentation pipeline with multi-region and multi-scale convolution attention mechanism. *Computers in Biology and Medicine*, *163*, Article 107076.

Jing, J., Liu, S., Wang, G., Zhang, W., & Sun, C. (2022). Recent advances on image edge detection: A comprehensive review. *Neurocomputing*.

Jurio, A., Pagola, M., Galar, M., Lopez-Molina, C., & Paternain, D. (2010). A comparison study of different color spaces in clustering based image segmentation. In *Information processing and management of uncertainty in knowledge-based systems. applications: 13th international conference, iPMU 2010, dortmund, Germany, June 28–July 2, 2010. proceedings, part II 13* (pp. 532–541). Springer.

Lai, S., Zhang, H., Shao, W., Liu, H., Cai, D., Wang, W., et al. (2023). Simple CNN for vision. URL https://openreview.net/forum?id=FDve8qGH3M.

Lalaoui, L., & Mohamadi, T. (2013). A comparative study of image region-based segmentation algorithms. *International Journal of Advanced Computer Science and Applications*, *4*(6).

Li, N., Lv, X., Xu, S., Li, B., & Gu, Y. (2020). An improved water surface images segmentation algorithm based on the Otsu method. *Journal of Circuits, Systems and Computers*, *29*(15), Article 2050251.

Liu, S., Chen, T., Chen, X., Chen, X., Xiao, Q., Wu, B., et al. (2022). More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. arXiv preprint arXiv:2207.03620.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976–11986).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Manoharan, D. S. (2020). Performance analysis of clustering based image segmentation techniques. *Journal of Innovative Image Processing*, *2*(1), 14–24.

Maysanjaya, I. M. D., Kesiman, M. W. A., & Indradewi, I. G. D. (2022). Optic disc and exudates segmentation on retinal fundus images using mask R-CNN. In *2022 international conference on information technology research and innovation* (pp. 168–172). IEEE.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3523–3542.

Monteiro, F. C., & Campilho, A. (2008). Watershed framework to region-based image segmentation. In *2008 19th international conference on pattern recognition* (pp. 1–4). IEEE.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., et al. (2018). Indian diabetic retinopathy image dataset (IDRiD). http://dx.doi.org/10.21227/H25W98.

Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., et al. (2020). Idrid: Diabetic retinopathy–segmentation and grading challenge. *Medical Image Analysis*, *59*, 101561.

Preetha, M. M. S. J., Suresh, L. P., & Bosco, M. J. (2012). Image segmentation using seeded region growing. In *2012 international conference on computing, electronics and electrical technologies* (pp. 576–583). IEEE.

Rauf, N., Gilani, S. O., & Waris, A. (2021). Automatic detection of pathological myopia using machine learning. *Scientific Reports*, *11*(1), 16570.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, munich, Germany, October 5–9, 2015, proceedings, part III 18* (pp. 234–241). Springer.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520).

Sujji, G. E., Lakshmi, Y., & Jiji, G. W. (2013). MRI brain image segmentation based on thresholding. *International Journal of Advanced Computer Research*, *3*(1), 97.

Tang, S., Song, C., Wang, D., Gao, Y., Liu, Y., & Lv, W. (2024). W-Net: A boundary-aware cascade network for robust and accurate optic disc segmentation. *Iscience*, *27*(1).

Valanarasu, J. M. J., & Patel, V. M. (2022). Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention* (pp. 23–33). Springer.

Wang, Y., Dai, Y., Xue, J., Liu, B., Ma, C., & Gao, Y. (2017). Research of segmentation method on color image of Lingwu long jujubes based on the maximum entropy. *EURASIP Journal on Image and Video Processing*, *2017*, 1–9.

Wang, Z., Jensen, J. R., & Im, J. (2010). An automatic region-based image segmentation algorithm for remote sensing applications. *Environmental Modelling & Software*, *25*(10), 1149–1165.

Wang, L., Wu, H., & Pan, C. (2013). Region-based image segmentation with local signed difference energy. *Pattern Recognition Letters*, *34*(6), 637–645.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Xiong, H., Liu, S., Sharan, R. V., Coiera, E., & Berkovsky, S. (2022). Weak label based Bayesian U-Net for optic disc segmentation in fundus images. *Artificial Intelligence in Medicine*, *126*, Article 102261.

Zhou, Z., Siddiquee, M., Tajbakhsh, N., & Liang, J. U. (2018). A nested U-Net architecture for medical image segmentation. arXiv preprint arXiv:1807.10165.

Zhu, S., Xia, X., Zhang, Q., & Belloulata, K. (2007). An image segmentation algorithm in image processing based on threshold segmentation. In *2007 third international IEEE conference on signal-image technologies and internet-based system* (pp. 673–678). IEEE.