ScleraSegNet: an Improved U-Net Model with Attention for Accurate Sclera Segmentation

Caiyong Wang[1,2], Yong He[2,3], Yunfan Liu[2], Zhaofeng He[4], Ran He[1,2], Zhenan Sun[1,2, ✉]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, P.R. China
[2]CRIPAC, NLPR, CASIA, Beijing, P.R. China
[3]Hunan University of Technology, Zhuzhou, Hunan, P.R. China
[4]Beijing IrisKing Co., Ltd, Beijing, P.R. China

{caiyong.wang, yong.he, yunfan.liu}@cripac.ia.ac.cn
{zfhe, rhe, znsun}@nlpr.ia.ac.cn

Abstract

Accurate sclera segmentation is critical for successful sclera recognition. However, studies on sclera segmentation algorithms are still limited in the literature. In this paper, we propose a novel sclera segmentation method based on the improved U-Net model, named as ScleraSegNet. We perform in-depth analysis regarding the structure of U-Net model, and propose to embed an attention module into the central bottleneck part between the contracting path and the expansive path of U-Net to strengthen the ability of learning discriminative representations. We compare different attention modules and find that channel-wise attention is the most effective in improving the performance of the segmentation network. Besides, we evaluate the effectiveness of data augmentation process in improving the generalization ability of the segmentation network. Experiment results show that the best performing configuration of the proposed method achieves state-of-the-art performance with F-measure values of 91.43%, 89.54% on UBIRIS.v2 and MICHE, respectively.

1. Introduction

Sclera is the white outer layer of the eyeball surrounding the iris. The blood vessel structure of sclera is unique to each person, hence it could be used for identification [23]. Sclera recognition is initially acted as a supportive recognition technology for iris recognition, especially when in off-angle or off-axis eye gaze [8], iris information fusion with sclera can increase the applicability of iris biometrics. Recently, sclera has begun to be widely studied as a stand-alone biometric trait [1, 11]. A complete sclera recognition
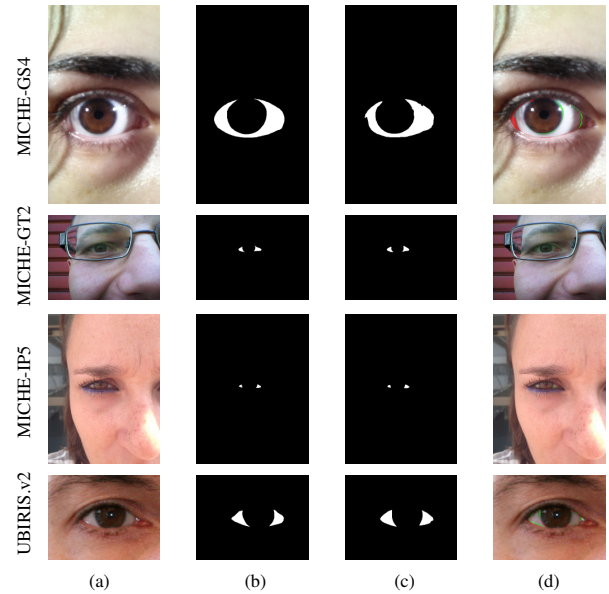

Figure 1. The first column shows the sclera images from different datasets. The second column displays the sclera segmentation ground truths manually labeled by [14]. By applying the improved U-Net with CBAM [20], our approach achieves high accuracy segmentation results across multiple datasets, as illustrated in the third column. The fourth column shows the segmentation errors in comparison with the ground truths where green and red pixels represent the false positives and false negative pixels, respectively.

process often consists of five steps: sclera image acquisition, sclera segmentation, sclera vessel feature extraction, template matching and decision [23]. As in the preprocessing stage, sclera segmentation has a great impact on the accuracy of sclera recognition. Incorrect sclera segmentation could either cause identity-related information contained in blood vessels to be lost or introduce other distractive textures such as eyelids and eyelashes, both damaging the ac-

curacy of sclera recognition [14].

In order to encourage the development of advanced sclera segmentation algorithms, four competitions have been held in main biometric conferences including BTAS, ICB, IJCB until now [3, 4, 5, 6]. In the initial exploration of sclera segmentation, many traditional segmentation algorithms, such as pixel clustering or handcrafted feature descriptors with SVM classifiers, are employed to complete the task. With the development of deep learning, Fully Convolutional Network (FCN) based segmentation algorithms become the mainstream and achieve state-of-the-art performance on sclera segmentation. Most of existing FCN based sclera segmentation methods directly apply off-the-shelf semantic segmentation models, *e.g.*, SegNet [2], RefineNet [12], to sclera image segmentation, by simply changing the number of segmentation classes from N to 2 (sclera area vs. background). In addition, Lucio *et al.* [14] propose two new segmentation methods based on Fully Connected Network and Generative Adversarial Network, respectively. Their methods are divided into two steps: the first is periocular region detection for narrowing the segmentation range, and the second is performing sclera segmentation in the detected patch. Although their best performing method outperforms SegNet, it is not an end-to-end solution and also has high computational complexity.

In this paper, we propose a new method, named ScleraSegNet, for sclera segmentation. The proposed method is based on U-Net [17], a simple yet effective semantic segmentation model. Instead of simply applying the original U-Net to sclera, we made a significant improvement by embedding attention mechanism. Attention mechanism helps U-Net extract more discriminative features for alleviating the interference of noise, hence the improved U-Net achieves high accuracy segmentation results across multiple sclera datasets, as illustrated in Figure 1. Besides, an in-depth analysis of training process and experimental results is provided. The main contributions of this paper are summarized as follows: 1) We improve the original U-Net model with attention mechanism and evaluate the effectiveness of attention mechanism in improving the performance of the segmentation network; 2) We evaluate the effectiveness of data augmentation in improving the generalization ability of the segmentation network; 3) We perform extensive experiments and demonstrate that the proposed method obtains a leading performance on multiple datasets.

The rest of this paper is organized as follows. In Section 2, the proposed method and training/testing process are described in detail. Section 3 presents experiment results and detailed analysis. Finally, we conclude our paper in Section 4.

2. Technical details

The proposed ScleraSegNet is built based on U-Net [17]. However, compared to the original U-Net, we introduce an attention module in the central bottleneck part between the contracting path and the expansive path to learn more discriminative features for separating sclera and non-sclera pixels. We firstly introduce the architecture of the ScleraSegNet in Section 2.1. Then, architectures of bottleneck equipped with different attention modules are described in Section 2.2 in detail. Finally, we present the training and testing process of ScleraSegNet in Section 2.3.
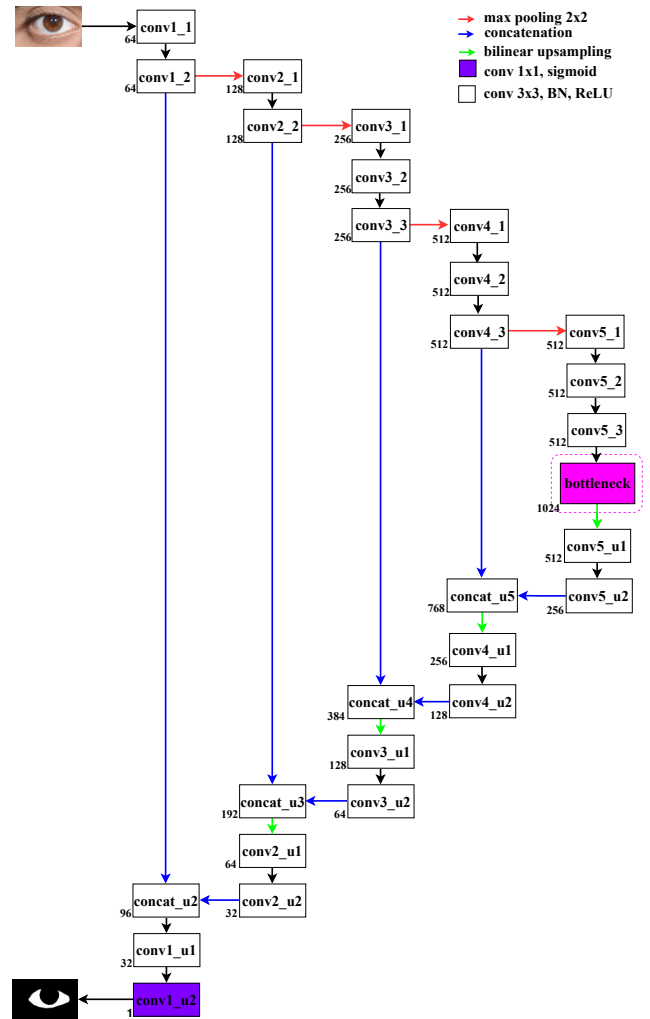
2.1. Structure of ScleraSegNet

Figure 2. Overview of the framework of ScleraSegNet. The number of channels is annotated at the lower left of each box. Best viewed in color.

The network architecture of ScleraSegNet is illustrated in Figure 2, which consists of a contracting path and a symmetric expansive path. The contracting path adopts VGG16

with fully connected layeres discarded as the encoder. The encoder consists of a series of convolutional units, each includes a sequence of one convolutional layer, one batch normalization layer and one ReLU activation layer. After each convolutional unit, a 2×2 max pooling layer with stride 2 is adopted for downsampling. As the network goes deeper, the number of channels gradually increases while the size of feature maps gradually decreases. To recover the spatial information lost in pooling layers of the contracting path and meanwhile reduce the number of channels, the expansive path adopts a series of bilinear upsampling operations, followed by two 3×3 convolutional units. Then, high resolution features from the contracting path and the upsampled output from the expansive path are concatenated via skip connections for more precise localization. Besides, the central bottleneck part between the contracting path and the expansive path could encode the most powerful and discriminative semantic features. Finally, a 1×1 convolutional layer and a sigmoid activation function are used to output the probability map of sclera segmentation, which has the same size as the original input.

2.2. Bottleneck architecture

As discussed in the former section, there is a central bottleneck part (highlighted in pink in Figure 2) between the contracting path and the expansive path. In the original U-Net [17], the bottleneck part consists of several convolutional units, which contain high-level semantic information collected from the contracting path, and these representative semantic information is then propagated to the later expansive path. Therefore, the bottleneck part has a far-reaching influence on the final predicted segmentation mask.

In general, informative features in the bottleneck part could be decomposed spatial-wise or channel-wise. Spatial-wise features encode the most important location information associated with the segmentation object, while channel-wise features focus on the semantic categories about the segmentation object [15, 20]. In order to enable the bottleneck part to extract more representative features and make the network focus on the most important information, several necessary steps are adopted, including re-estimating the spatial distribution of feature maps and adaptively recalibrating channel-wise feature responses.

In the following section, we will introduce four types of attention modules, which are embedded in the bottleneck part to achieve the goal mentioned above. As a baseline, we also introduce the bottleneck part of the original U-Net. The detailed bottleneck architectures are illustrated in Figure 3.

Figure 3 (a) shows the baseline architecture. More specifically, for the give input feature map, a 2×2 max pooling operation with stride 2 is firstly applied to downsample the size for further feature extraction. Then, the pooled feature map is split into two parts, one is followed by a bilin-

ear interpolation operation and two convolutional units, the other is an identity mapping. Finally, these two parts are combined together by channel-wise concatenation.

Other bottleneck architectures we concern differ in types of embedded attention modules, which are illustrated in Figure 3 (b), (c), (d) and (e). Although detailed compositions of these bottleneck networks are different, they share the same overall architecture. To be speciifc, given the input feature map $F \in R^{512 \times H \times W}$, a 3×3 max pooling operation with stride 1 is firstly applied to F to get the refined feature map $P \in R^{512 \times H \times W}$ which keeps the feature size unchanged, then the final discriminative feature map F' is computed as:

$$F' = P \oplus \{P \otimes M(P)\} \tag{1}$$

where $M(P) \in R^{512 \times H \times W}$ is the inferred 3D attention map, and $\otimes$ and $\oplus$ represents element-wise multiplication and channel-wise concatenation, respectively. From the equation, we see that pooled feature map is adaptively updated via pixel-wise multiplication with the 3D attention map. Besides, the original pooled feature map is also stored via concatenation with the updated ones to keep other valuable information in the original input signal. Such design makes the original feature further refined and more discriminative. In addition, the only difference among all mentioned bottleneck architectures is the specific architecture of $M(P)$, which is further introduced and compared in the following sections.

2.2.1 Channel attention module

Channel Attention Module (CAM) is firstly introduced in the SENet [9], then developed in BAM [15]. It is expected to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. Channel attention module contains a squeeze block, which takes global average pooling on the feature map P to produce a channel vector F_c, then followed by a excitation block, which uses a multi-layer perceptron (MLP) with one hidden layer to estimate attention across channel from the channel vector F_c. More precisely, given the pooled feature map P, the channel attention module is computed as:

$$\begin{aligned} M(P) &= \sigma(M_c(P)) \\ &= \sigma(\text{MLP}(\text{GAP}(P))) \\ &= \sigma(W_1(W_0(\text{GAP}(P)) + b_0) + b_1) \end{aligned} \tag{2}$$

where $W_0 \in R^{256 \times 512}$, $b_0 \in R^{256}$, $W_1 \in R^{512 \times 256}$, $b_1 \in R^{512}$, GAP is global average pooling along the spatial axis, σ is a sigmoid function which normalizes the output range of $M_c(P)$ to $[0, 1]$. Note that the initially produced channel attention map $M(P) \in R^{512 \times 1}$ needs to be broadcasted along the spatial dimension to match with the dimension of the original input, *i.e.*, $R^{512 \times H \times W}$.
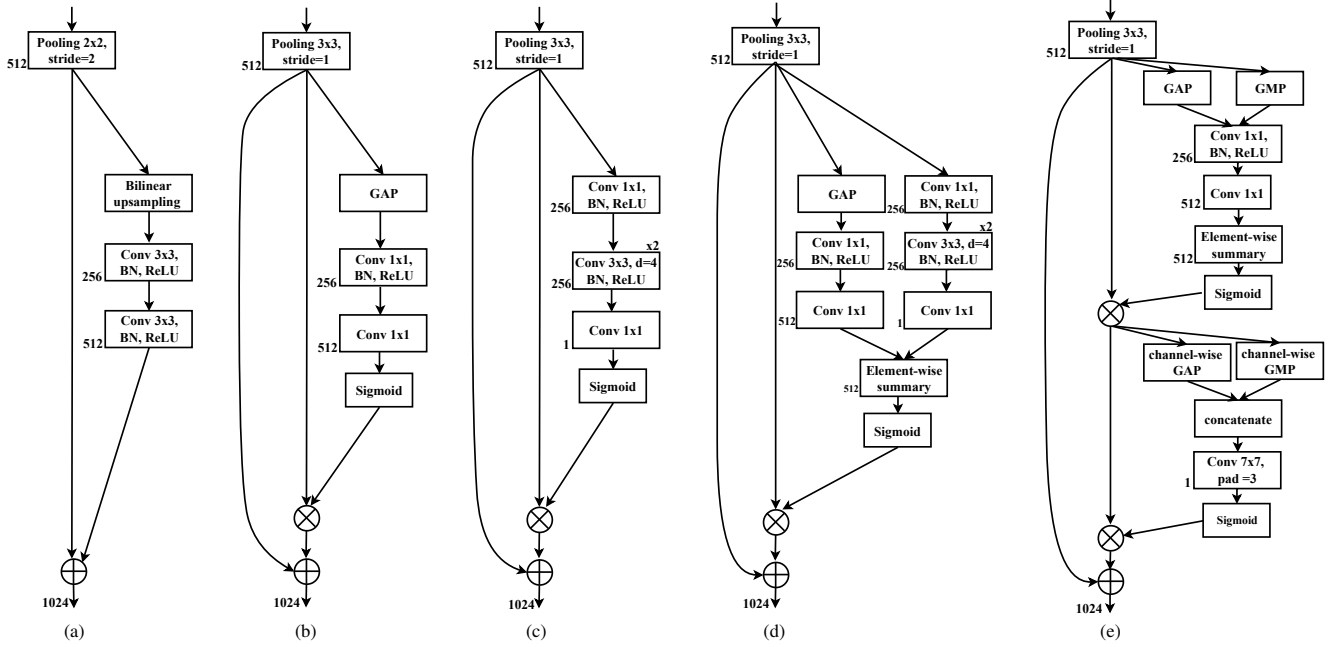
Figure 3. Different bottleneck architectures: (a) simple feature concatenation from U-Net [17]; (b) channel attention module (CAM) from [9]; (c) spatial attention module (SAM) from [15]; (d) parallel channel attention and spatial attention module (BAM) from [15]; (e) Sequential channel attention and spatial attention module (CBAM) from [20]. GAP and GMP represent global average pooling and global max pooling, respectively.

2.2.2 Spatial attention module

Spatial Attention Module (SAM) is introduced in BAM [15]. It is expected to learn a spatial attention map to emphasize or suppress features in different spatial locations. Mathematically, the function of SAM is formulated as:

$$M(P) = \sigma(M_s(P))$$
$$= \sigma(f^{1\times1}(F_{d=4}^{3\times3}(F_{d=4}^{3\times3}(F^{1\times1}(P))))) \quad (3)$$

More specifically, the channel dimension of feature $P \in R^{512\times H\times W}$ is reduced to 256 using a 1×1 convolution unit $F^{1\times1}$. Then two 3×3 dilated convolutions ($F_{d=4}^{3\times3}$) with dilation value of 4 are applied to enlarge the receptive fields for effectively leveraging contextual information. Finally, the feature map is compressed into $M_s(P) \in R^{1\times H\times W}$ using a single 1×1 convolution operation ($f^{1\times1}$). Besides, we also use a sigmoid function (σ) to normalize the output range to $[0, 1]$. Note that the initially produced spatial attention map $M(P) \in R^{1\times H\times W}$ needs to be replicated by 512 times along the channel dimension to match the dimension of the original input, *i.e.*, $R^{512\times H\times W}$.

2.2.3 Parallel channel attention and spatial attention module

Recently, Park *et al*. propose Bottleneck Attention Module (BAM) [15], a parallel integration of channel attention mod-

ule and spatial attention module. For the given pooled feature map $P \in R^{C\times H\times W}$, BAM infers a 3D attention map $M(P) \in R^{C\times H\times W}$ as follows:

$$M(P) = \sigma(M_c(P) + M_s(P)) \quad (4)$$

where $M_c(P)$ and $M_s(P)$ are described as in Section 2.2.1 and Section 2.2.2, respectively. Note that outputs of both branches are resized to $R^{512\times H\times W}$ before addition.

2.2.4 Sequential channel attention and spatial attention module

Different from BAM, Woo *et al*. propose Convolutional Block Attention Module (CBAM) [20], a sequential connection of channel attention module and spatial attention module. The channel attention module of CBAM adds max-pooled features in addition to average-pooled features, hence the new channel attention module is computed as:

$$M_c(P) = \sigma(\text{MLP}(\text{GAP}(P)) + \text{MLP}(\text{GMP}(P)))$$
$$= \sigma(W_1(W_0(\text{GAP}(P))) + W_1(W_0(\text{GMP}(P)))) \quad (5)$$

where GMP is global max pooling along the spatial axis.

Given the channel attention map, the channel attention process is computed as:

$$P' = M_c(P) \otimes P \quad (6)$$

4

The spatial attention module, following the channel attention module, aggregates average-pooled features and max-pooled features along the channel axis, which could be formulated as follows:

$$M_s(P') = \sigma(f^{7 \times 7}(\text{GAP}_c(P') \oplus \text{GMP}_c(P'))) \quad (7)$$

where $\text{GAP}_c(P')$ and $\text{GMP}_c(P')$ represent global average pooling and global max pooling along the channel axis, respectively. We first apply global averag pooling and global max pooling operations along the channel axis and concatenate them to generate an efficient feature descriptor. Then, a 7×7 convolution operation with padding size of 3 followed by a sigmoid function are applied on the concatenated feature descriptor to generate a spatial attention map $M_s(P') \in R^{H \times W}$.

The final spatial attention process is computed as:

$$P'' = M_s(P') \otimes P' \quad (8)$$

Note that same as Section 2.2.1 and Section 2.2.2, during the element-wise multiplication of Equation (6) and Equation (8), the attention map is firstly broadcasted or copied accordingly.

2.3. Network training and testing

Since sclera segmentation could be regarded as a pixel-wise binary classification task, a binary cross-entropy loss function is used for training.

Once the model is trained, it takes an eye image of arbitrarily size as input and outputs a probability map of sclera of the same size as the original input image. To generate the final segmentation result, we need to threshold the predicted probability map to get a binary mask using a certain threshold. More specifically, for those pixels of the probability above the selected threshold, the corresponding pixels of the binary mask are assigned to 1, otherwise the corresponding pixels are assigned to 0.

3. Experiments

3.1. Datasets

In this section, we present detailed descriptions of three datasets used in our experiments: UBIRIS.v2 [16], MICHE-I [7] and MASD.v1 [6]. Among these datasets, UBIRIS.v2 and MICHE-I are used to train and evaluate the proposed model. Inspired by [14], each of them is divided into three subsets, where 40% of the images are used for training, 20% for validation, and 40% for testing. MASD.v1 is not used for model training or fine-tuning, but directly for testing. Detailed information of these datasets are summarized in Table 1.

UBIRIS.v2 [16] was originally developed for iris recognition in less constrained conditions. The dataset consists of

Dataset	Resolution	No. of training	No. of testing	No. of validation
UBIRIS.v2	400×300	120	120	60
MICHE-I	Various	400	400	200
MICHE-GS4	Various	133	133	67
MICHE-IP5	Various	138	138	68
MICHE-GT2	640×480	129	129	65
MASD.v1	Various	N/A	119	N/A

Table 1. Summary of the datasets used in this work. Each of these is a subset of the corresponding original database.

11,102 images from 261 subjects. The ground-truth sclera segmentation masks are manually labeled by [14].

MICHE-I [7] was originally developed for mobile iris recognition. Images in MICHE-I were captured by three mobile devices: iPhone5(IP5), Samsung Galaxy S4(GS4), and Samsung Galaxy Tab2(GT2) (1262, 1297 and 632 images, respectively) in uncontrolled conditions. Same as U-BIRIS.v2, the ground-truth sclera segmentation masks are also manually labeled by [14].

MASD.v1 [6] was collected for sclera segmentation benchmarking competition(SSBC). For each eye, images of 4 gaze angles (looking straight, left, right and up) are captured. In SSBC 2015, a subset of 120 sclera images and corresponding ground-truth masks were provided to the academic community. However, there were only 119 images with ground truths available from the organizers of the competition. Since the amount of images in the dataset is small, they are only used for testing.

3.2. Evaluation metrics

To quantitatively evaluate the proposed method, precision(P), recall (R) and F-measure(F) are computed in a pixel-wise comparison manner between the ground truth and the predicted binary mask image. Among the above metrics, precision measures the percentage of correctly retrieved sclera pixels. Recall gives the percentage of sclera pixels in the ground truth which are correctly retrieved. F-measure is defined as the harmonic mean of precision and recall to balance the two metrics.

Besides the fixed P/R/F values due to the fixed threshold, the complete precision-recall curves (PR-curve) could be generated by varying the decision threshold to evaluate the overall segmentation performance. In this context, the F-measure is obtained under the optimal threshold over the whole dataset.

3.3. Implementation Details

The proposed architecture is implemented based on the openly available caffe [10] framework and initialized by using the VGG-16 model pretrained on ImageNet [18]. Other hyper-parameters and corresponding values are: optimization method: stochastic gradient descent(SGD), mini-batch
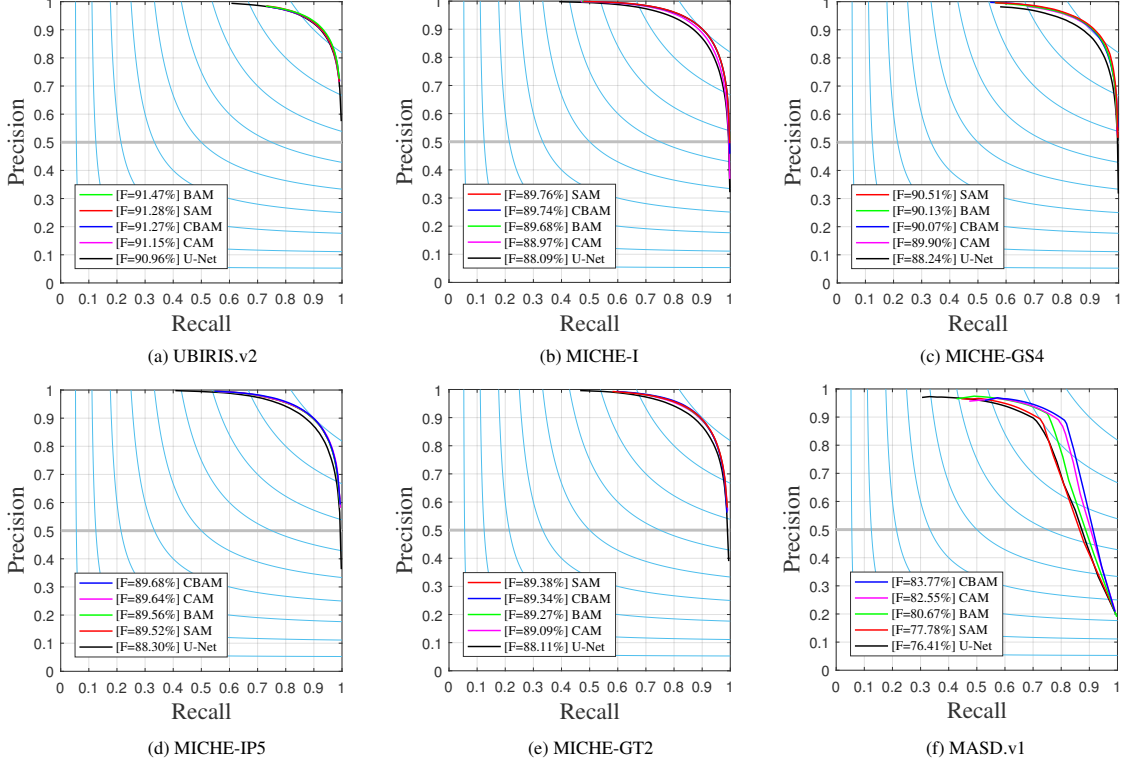
5

Figure 4. Average precision-recall curves generated by U-Net and the improved U-Nets with different attention modules on six datasets.

size (4), base learning rate (10^{-3}), learning rate adjustment method: "poly" policy with *power* set to 0.9, momentum (0.9), weight decay (0.0005), and maximal iteration (30000).

During the experiments, we augment the training dataset by randomly resizing (0.5, 0.75, 1, 1.25, 1.5), translation (x,y [-30,30]), rotation ([-60,60]), blurring (mean filter, gaussian blur, median blur, bilateral filter, box blur), horizontal flipping, and cropping (321×321 as input size) on the fly.

3.4. Experimental results

3.4.1 Evaluation of different attention modules

In this section, we evaluate the segmentation performance of the proposed ScleraSegNet. Firstly, we implement the original U-Net model as the baseline model. Then, four types of improved U-Nets equipped with different attention modules are compared with the baseline model. With the exception of the network structure, all other aspects of the network are the same, such as data augmentation and training process, as introduced in Section 3.3. Besides, all the models are trained on UBIRIS.v2 and MICHE-I as well as three subsets of the MICHE-I, then tested on the them. As for MASD.v1, we directly use it to test the model trained on UBIRIS.v2, as UBIRIS.v2 is similar with MASD.v1 in illumination, noise distribution, and the position and proportion

of periocular region occupying the whole of image, *etc*.

PR-curve and F-measure are the evaluation metrics used to compare the performance of segmentation algorithms, and results are shown in Figure 4. As can be seen, the proposed improved U-Nets with different attention modules outperform the baseline model with a significant margin on MICHE-I, MICHE-GS4, MICHE-IP5, MICHE-GT2, and MASD.v1. For UBIRIS.v2, there is very little difference on PR-curves and F-measure values obtained by the baseline model and its improvements. The reason for the large performance gap on UBIRIS.v2 and MICHE may be that the images in UBIRIS.v2 are relatively concentrated, *i.e.*, containing only the periocular region, while for MICHE, we need to enhance the feature expression ability of the original U-Net to suppress the influence of other facial parts, such as nose, ears, forehead, cheeks, *etc*. The final experiment results validate the effectiveness of proposed attention modules on improving the performance of the original U-Net. The results on MASD.v1 further suggest that attention modules also benefit the generalization ability of the original U-net.

It is worth noting that, although the improved U-Nets with different attention modules show similar PR-curves and F-measure values on UBIRIS.v2 and MICHE-I as well as their subsets, they output completely different segmentation results on MASD.v1. The improved U-Net with

CBAM achieves the best segmentation performance with the F-measure value of 83.77%, followed by CAM, BAM and SAM, achieving F-measure values of 82.55%, 80.67%, 77.78%, respectively. More careful observation suggests that channel-wise attention is more important than spatial-wise attention for the accuracy of segmentation tasks, and this is why many other segmentation networks, such as [21, 22], also adopt channel attention modules to improve the performance of the network.

3.4.2 Evaluation of the effectiveness of data augmentation

Data augmentation is a simple yet effective way to enrich training data for accelerating the network to converge and helping the network avoid over-fitting. Besides, we have further demonstrated that data augmentation also improves the generalization ability of the network. We use the original U-Net as the experiment model (*The results obtained using the improved U-Net are also similar*) and train U-Net with and without data augmentation on UBIRIS.v2 dataset. Then, we test the trained model to the MASD.v1 dataset. The experiment results are illustrated in Figure 5. From the result, we could conclude that the U-Net trained with data augmentation significantly outperforms that trained without data augmentation by a large margin of 27.97% under the F-measure. As we do not train or fine tune the model on MASD.v1 dataset, such a large performance gain demonstrates the effectiveness of data augmentation in improving the generalization ability of the network.
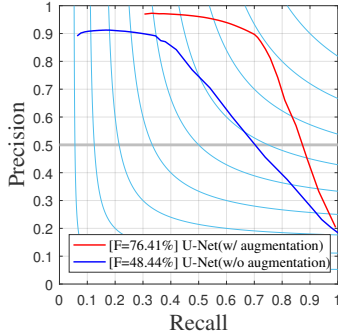

Figure 5. Average precision-recall curves generated by U-Net on MASD.v1 dataset. The U-Net is trained on UBIRIS.v2 dataset with and without data augmentation, respectively.

3.4.3 Comparison with other methods

We also compare the proposed method with the best performing sclera segmentation algorithm in [14], *i.e.*, FCN. In fact, FCN is the segmentation part of Multinet [19] and adopts the similar network structure as the earliest FCN8 in [13]. Table 2 lists a summary of Recall, Precision, and F-measure, for the original U-Net, the proposed improved U-Nets with different attention modules, and FCN. Here, we empirically set the threshold to 0.5 to get the final binary mask for simplicity. Better threshold could be obtained by cross validation on the validation set. Results show that in all the datasets except UBIRIS.v2, original U-Net does not outperform FCN for F-measure, but the proposed improved U-Nets with attention modules consistently outperform FCN with larger mean values and smaller or comparable standard deviations for F-measure. Besides, the improved U-Net also outperforms FCN considerably in terms of Recall and Precision values in most cases. In summary, the improved U-Net with CBAM achieves a leading segmentation performance in most settings, which is consistent with the conclusion drawn in Section 3.4.1. Besides, the improved U-Net with CBAM also has a high segmentation efficiency. For a $400 \times 300 \times 3$ input image, it takes merely 0.05 second on a NVIDIA TITAN Xp GPU with 12GB memory.

Dataset	Method	Recall %	Precision %	F-measure %
UBIRIS .v2	FCN[14]	87.31(06.68)	88.45(06.98)	87.48(03.90)
	U-Net[17]	90.51(06.53)	91.81(05.05)	90.89(03.77)
	SAM	91.14(05.66)	91.77(04.78)	91.25(03.32)
	CAM	91.13(06.04)	91.58(04.79)	91.13(03.52)
	BAM	**91.24(06.20)**	**92.11(04.76)**	**91.43(03.75)**
	CBAM	91.42(06.38)	91.59(05.03)	91.26(04.09)
MICHE -I	FCN[14]	87.59(11.28)	89.90(09.82)	88.32(09.80)
	U-Net[17]	86.05(09.67)	90.60(05.98)	87.83(06.56)
	SAM	87.87(07.90)	91.91(05.80)	89.53(05.33)
	CAM	87.85(08.44)	90.81(07.15)	88.90(06.14)
	BAM	87.34(08.56)	92.24(05.16)	89.37(05.49)
	CBAM	**88.11(08.26)**	**91.71(05.42)**	**89.54(05.37)**
MICHE -GS4	FCN[14]	88.24(12.03)	88.65(10.62)	88.12(10.56)
	U-Net[17]	86.65(08.98)	90.42(09.71)	87.87(07.80)
	SAM	**90.24(06.17)**	**91.12(06.46)**	**90.45(04.87)**
	CAM	90.14(07.30)	90.40(08.53)	89.86(06.64)
	BAM	89.05(07.44)	91.73(07.93)	89.95(06.28)
	CBAM	88.71(07.46)	91.82(06.90)	89.90(05.59)
MICHE -IP5	FCN[14]	87.51(11.61)	89.32(05.22)	87.80(08.24)
	U-Net[17]	84.77(08.72)	91.67(05.00)	87.73(05.30)
	SAM	86.15(07.44)	92.61(05.47)	88.95(04.64)
	CAM	**87.16(07.00)**	**92.10(05.73)**	**89.28(04.63)**
	BAM	86.69(06.67)	92.26(05.10)	89.16(04.19)
	CBAM	86.32(07.80)	92.76(04.56)	89.13(04.62)
MICHE -GT2	FCN[14]	87.86(12.23)	88.50(12.68)	87.94(11.59)
	U-Net[17]	86.28(09.99)	90.47(05.97)	87.89(06.81)
	SAM	88.20(09.06)	91.09(05.47)	89.29(06.16)
	CAM	88.81(09.16)	90.06(05.94)	89.07(06.43)
	BAM	88.69(09.24)	90.48(05.63)	89.24(06.40)
	CBAM	**89.07(09.27)**	**90.38(05.93)**	**89.34(06.51)**

Table 2. Performance comparison between FCN[14] and the proposed models. The values in parentheses represent standard deviations. F-measure is considered as the prior measure for ranking the methods.

4. Conclusions

This paper introduces a improved U-Net model, namely ScleraSegNet, for accurate sclera segmentation in an end-to-end manner. The improved U-Net model is combined with different attention modules and could be trained using effective data augmentation techniques. Extensive experiments are carried out on three public datasets, and results show the proposed model is able to accurately segment the sclera region with high robustness.

Acknowledgment

References

[1] S. Alkassar, W. L. Woo, S. S. Dlay, and J. A. Chambers. Robust sclera recognition system with novel sclera segmentation and validation techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(3):474–486, 2017.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[3] A. Das, U. Pal, M. A. Ferrer, and M. Blumenstein. Ssrbc 2016: sclera segmentation and recognition benchmarking competition. In *International Conference on Biometrics*, pages 1–6. IEEE, 2016.

[4] A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Štepec, P. Rot, Ž. Emeršič, P. Peer, V. Štruc, S. A. Kumar, et al. Sserbc 2017: Sclera segmentation and eye recognition benchmarking competition. In *International Conference on Biometrics*, pages 742–747. IEEE, 2017.

[5] A. Das, U. Pal, M. A. Ferrer, M. M. Blumenstein, D. Stepec, P. Rot, Z. Emersic, P. Peer, and V. Struc. Ssbc 2018: Sclera segmentation benchmarking competition. In *International Conference on Biometrics*, pages 303–308. IEEE, 2018.

[6] A. Dasa, U. Palb, M. A. Ferrerc, and M. Blumensteina. Ssbc 2015: Sclera segmentation benchmarking competition. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–6, 2015.

[7] M. De Marsico, C. Galdi, M. Nappi, and D. Riccio. Firme: Face and iris recognition for mobile engagement. *Image and Vision Computing*, 32(12):1161–1172, 2014.

[8] N. Guliani, M. K. Shukla, A. K. Dubey, and Z. A. Jaffery. Analysis of multimodal biometrie recognition using iris and sclera. In *International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, pages 472–475. IEEE, 2017.

[9] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[11] S. Lee and J. Kim. Vessel pattern enhancement based on weber's law for sclera recognition. In *International Conference on Electronics, Information, and Communication*, pages 1–2. IEEE, 2018.

[12] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[14] D. R. Lucio, R. Laroca, E. Severo, A. S. B. Jr, and D. Menotti. Fully connected networks and generative neural networks applied to sclera segmentation. *CoRR*, abs/1806.08722, 2018.

[15] J. Park, S. Woo, J. Lee, and I. S. Kweon. Bam: Bottleneck attention module. In *British Machine Vision Conference*, page 147, 2018.

[16] H. Proenca, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre. The ubiris.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1529–1535, 2010.

[17] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[19] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *IEEE Intelligent Vehicles Symposium*, pages 1013–1020. IEEE, 2018.

[20] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision*, pages 3–19, 2018.

[21] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1857–1866, 2018.

[22] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[23] Z. Zhou, E. Y. Du, N. L. Thomas, and E. J. Delp. A new human identification method: sclera recognition. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(3):571–583, 2012.