electronics

MDPI

Article

OcularSeg: Accurate and Efficient Multi-Modal Ocular Segmentation in Non-Constrained Scenarios

Yixin Zhang [1,2], Caiyong Wang [1,2,*], Haiqing Li [1,2], Xianyun Sun [1,2], Qichuan Tian [1,2] and Guangzhe Zhao [1,2]

[1] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; zhangyixin@stu.bucea.edu.cn (Y.Z.); haiqing_li@stu.bucea.edu.cn (H.L.); sunxianyun@stu.bucea.edu.cn (X.S.); tianqichuan@bucea.edu.cn (Q.T.); zhaoguangzhe@bucea.edu.cn (G.Z.)
[2] Beijing Key Laboratory of Robot Bionics and Function Research, Beijing 100044, China
* Correspondence: wangcaiyong@bucea.edu.cn

Abstract: Multi-modal ocular biometrics has recently garnered significant attention due to its potential in enhancing the security and reliability of biometric identification systems in non-constrained scenarios. However, accurately and efficiently segmenting multi-modal ocular traits (periocular, sclera, iris, and pupil) remains challenging due to noise interference or environmental changes, such as specular reflection, gaze deviation, blur, occlusions from eyelid/eyelash/glasses, and illumination/spectrum/sensor variations. To address these challenges, we propose OcularSeg, a densely connected encoder–decoder model incorporating eye shape prior. The model utilizes Efficientnetv2 as a lightweight backbone in the encoder for extracting multi-level visual features while minimizing network parameters. Moreover, we introduce the Expectation–Maximization attention (EMA) unit to progressively refine the model's attention and roughly aggregate features from each ocular modality. In the decoder, we design a bottom-up dense subtraction module (DSM) to amplify information disparity between encoder layers, facilitating the acquisition of high-level semantic detailed features at varying scales, thereby enhancing the precision of detailed ocular region prediction. Additionally, boundary- and semantic-guided eye shape priors are integrated as auxiliary supervision during training to optimize the position, shape, and internal topological structure of segmentation results. Due to the scarcity of datasets with multi-modal ocular segmentation annotations, we manually annotated three challenging eye datasets captured in near-infrared and visible light scenarios. Experimental results on newly annotated and existing datasets demonstrate that our model achieves state-of-the-art performance in intra- and cross-dataset scenarios while maintaining efficient execution.

Keywords: ocular segmentation; iris segmentation; sclera segmentation; biometric recognition; shape prior

Citation: Zhang, Y.; Wang, C.; Li, H.; Sun, X.; Tian, Q.; Zhao, G. OcularSeg: Accurate and Efficient Multi-Modal Ocular Segmentation in Non-Constrained Scenarios. *Electronics* **2024**, *13*, 1967. https://doi.org/10.3390/electronics13101967

Academic Editor: Hyunjin Park

Received: 23 April 2024
Revised: 13 May 2024
Accepted: 14 May 2024
Published: 17 May 2024

1. Introduction

As a form of single-modal ocular biometrics, iris recognition has gained widespread recognition as a reliable authentication method due to its unique, stable, accurate, and noninvasive characteristics [1]. It has found extensive applications across various domains, including public safety, border control, mobile payment, and the metaverse. Additionally, in recent years, research has indicated that other ocular modalities (as illustrated in Figure 1), such as sclera and periocular [2,3], can effectively complement the iris recognition, substantially enhancing its suitability, accuracy, and security [4–7]. To exploit multi-modal ocular traits for recognition, the initial step is to perform multi-modal ocular segmentation on the input eye image. As depicted in Figure 2, this study concentrates on the simultaneous segmentation of the periocular (as a background class), sclera, iris, and pupil regions. As a result, the segmented region of interest (ROI) images of the periocular, sclera, and iris are further fed into their corresponding feature extractors to extract multi-modal identity features for fusion recognition. As for the pupil, it can serve various purposes, such as fatigue

detection and gaze estimation [8]. Given that multi-modal ocular segmentation occurs during the pre-processing stage, any inaccuracies in segmentation could result in the loss of identity-related modality information or the introduction of distracting textures. Such inaccuracies can substantially impair the accuracy of multi-modal ocular recognition [9,10].

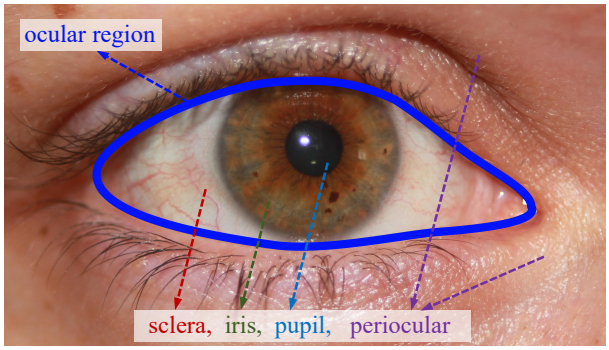


Figure 1. Periocular and ocular components (sclera, iris, and pupil). The eye image is from the SBVPI [11] dataset.

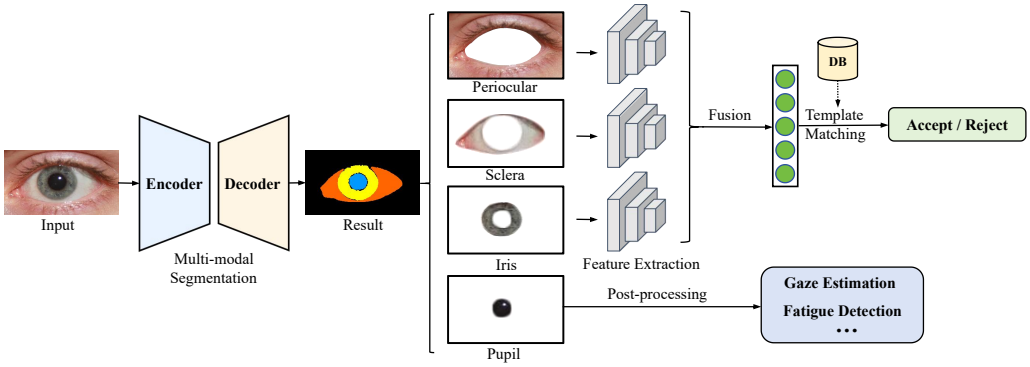


Figure 2. A standard multi-modal ocular recognition pipeline typically incorporates multi-modal segmentation to identify the regions of interest (ROIs) for fusion recognition.

Biometric recognition usually occurs in constrained scenarios. However, with the deepening of its application in our daily lives, it has become a trend to develop ocular biometrics in non-constrained scenarios (e.g., at-a-distance, on-the-move, and visible illumination), which would greatly reduce the constraints for user cooperation and imaging conditions. However, under these conditions, the segmentation process is highly susceptible to various noise factors such as specular reflection, gaze deviation, motion/defocus blur, occlusions from eyelid/eyelash/glasses, as well as environmental changes such as variations in illumination/spectrum/sensor. In addition, as a pre-processing operation, it should also be computationally efficient for practical deployment. Overall, achieving accurate and efficient multi-modal ocular segmentation is inherently challenging.

Some efforts have been made in the literature to enhance the accuracy of multi-modal ocular segmentation. Early approaches, such as EyeNet [12] and MinENet [13], primarily utilized classic CNN architectures like ResNet [14] to assess the feasibility of multi-modal ocular segmentation tasks. Subsequently, several models based on improved encoder–decoder architecture, such as RITnet [15], SCN [16], and Eye-UNet [17], were proposed to elevate segmentation accuracy further. In recent years, to enhance segmentation performance with limited annotated datasets, methods based on semi-supervised learning have been proposed [18]. Hassan et al. introduced a new framework named SIPFormer [19], which integrates transformer architecture for this multi-modal segmentation task. At the same time, it improves segmentation accuracy and introduces many parameters, reducing model efficiency. Additionally, most current methods are trained and evaluated on datasets from a single collection environment with limited samples, such as OpenEDS [20], which

may not adequately reflect the performance in real-world ocular recognition scenarios. Therefore, substantial challenges remain in developing efficient, accurate, and generalizable multi-modal ocular segmentation models.

More specifically, several challenges are highlighted in the multi-modal ocular segmentation task: (i) Current models exhibit inadequate adaptability to environmental fluctuations, encompassing factors like illumination, resolution, and contrast, among others, coupled with heightened computational complexity, constraining their applicability in resource-constrained settings like embedded devices. (ii) Most current models utilize an end-to-end pixel-wise semantic segmentation strategy for multi-modal ocular segmentation. However, they often fail to effectively leverage prior knowledge concerning the overall eye shape and the spatial distribution of different modalities. Consequently, this limitation hinders the extraction of contextual features, rendering the models prone to segmentation errors. (iii) The availability of finely annotated multi-modal ocular datasets in real-world open environments is scarce. Furthermore, certain datasets mentioned in the literature, like OpenEDS [20], exhibit limited accessibility or are not specifically collected for ocular biometrics.

This paper proposes a novel multi-modal ocular segmentation approach, named OcularSeg, to address the challenges above. The proposed approach is an encoder–decoder model like U-Net [21]. Specifically, the encoder employs the lightweight Efficientnetv2 [22] as the backbone for extracting multi-level visual features while mitigating computational complexity. Since the initially extracted hierarchical features are relatively coarse and lack discrimination for multi-modal ocular traits, we further introduce the Expectation–Maximization Attention (EMA) [23] module to alleviate this problem. Unlike certain mechanisms that require generating large attention maps, resulting in high computational complexity and consuming significant GPU memory—such as the self-attention mechanism in Transformer [24]—the EMA module is designed based on the Expectation–Maximization [25] algorithm. This allows it to dynamically adjust attention weights within the neural network and integrate spatial information. Such a design enhances the model's perceptual and discriminative abilities in noisy environments, thereby achieving coarse aggregation of ocular features and effectively improving the accuracy of prediction results.

In the decoder, we focus more on the information differences between different levels and consider that the semantic features at the deep level are richer and more likely to capture intricate ocular parts; hence, we propose a bottom-up densely connected subtraction module. It starts from the deepest level and applies subtraction units to the feature maps at all scales larger than the current one. This facilitates the exchange of cross-resolution feature information while accentuating useful distinctions between features, thereby eliminating interference from redundant components. Moreover, we incorporate the prior knowledge of ocular by integrating the boundary-guided prior and semantic-guided prior as supplementary constraints within the model structure and training procedure. This optimization enhances the model's predictive capabilities across various modalities and diminishes mis-segmentation. Lastly, we manually annotate three diverse multi-modal ocular datasets collected under visible and near-infrared light conditions with noise to assess the model's accuracy and generalization in real-world open environments. Experimental findings on self-collected and publicly available datasets show that our model achieves state-of-the-art performance in intra- and cross-dataset scenarios while maintaining low computation costs.

In summary, our main contributions can be summarized as follows:

- We present OcularSeg, a highly efficient and accurate, densely connected encoder–decoder model tailored for multi-modal ocular segmentation. This model integrates a lightweight EfficientNetv2 as its backbone, an EMA module for aggregating features from different modalities, and a bottom-up dense subtraction module to refine prediction results through feature exchange across different levels.
- We incorporate prior knowledge of eye shape, including boundary-guided and semantic-guided priors, to offer additional and refined guidance for the model's predictions

regarding shape, position, and structural relationships between different modalities. This inclusion substantially enhances the accuracy of our method.

- We manually annotate three diverse eye image datasets collected under various environmental conditions, encompassing illumination, resolution, and spectrum differences. These datasets are meticulously categorized for periocular, sclera, iris, and pupil pixels. Combining these datasets with existing ones demonstrates our method's effectiveness, superiority, and efficiency for multi-modal ocular segmentation across intra- and cross-dataset scenarios.

The structure of this paper is as follows: Section 2 reviews related work, while Section 3 elaborates on our multi-modal ocular segmentation framework. The experimental settings, including datasets and evaluation protocols, are detailed in Section 4. In Section 5, we present and analyze the experimental results quantitatively and qualitatively. Finally, Section 6 concludes the paper and discusses future work.

2. Related Work

Few studies have concentrated on multi-modal ocular segmentation, particularly for delineating multiple ocular regions from images using a single segmentation model. Rot et al. [26] pioneered the training of a convolutional encoder–decoder network based on SegNet [27], categorizing pixels into six classes: pupil, iris, sclera, eyelashes, canthus, and periocular (as listed in Table 1). Subsequently, Ref. [28] manually annotated 500 eye images from the NICE. I competition dataset [29], expanding on NICE. I's two-category real iris segmentation mask to encompass 10 semantic categories. Their work achieved comparable segmentation accuracy utilizing FCN networks [30]. In another approach, Ref. [31] designed a miniature multi-scale segmentation network (Eye-MS) founded on multi-scale interconnected convolutional modules. They also developed a lightweight variant named Eye-MMS, containing only 80k parameters, to preserve performance while reducing parameters.

The eye segmentation challenge organized by Facebook research was conducted on the OpenEDS dataset [20]. Kansal et al. proposed Eyenet [32] to address this challenge, employing residual connections, multi-scale supervision, a squeeze-and-excitation module [33], and a convolutional block attention module [34]. MinENet [13] was introduced to streamline model complexity by removing redundancy within the central layer of ENet [35], which utilizes a dilated and asymmetric convolution design. Chaudhary et al. proposed the RITnet architecture [15], amalgamating DenseNet [36] and U-Net [21], integrating pre-processing enhancement operations and boundary-aware loss functions to produce clear regional boundaries.

Table 1. Comparison of the proposed method with other multi-class ocular segmentation methods.

Method	Backbone	Dataset	Spectrum		Modality	Weakness
			NIR	VIS		
Rot et al. [26]	SegNet	MASD [37]	-	✓	S/I/P/E/C/PO	The dataset is relatively small, consisting of 120 samples.
D. et al. [28]	FCN	NICE.I [29]/MobBIO [38]	-	✓	S/I/P/E/C/PO/SR/EB/H/GF	The computational demands and rough annotation.
Eye-MMS [31]	-	OpenEDS [20]	✓	-	S/I/P/PO	The model is simple and the accuracy is poor.
EyeNet [12]	ResNet50	OpenEDS [20]	✓	-	S/I/P/PO	Large number of parameters and additional optimization.
MinENet [13]	ENet	OpenEDS [20]	✓	-	S/I/P/PO	Only the modifications in the model architecture.

Table 1. *Cont.*

Method	Backbone	Dataset	Spectrum		Modality	Weakness
			NIR	VIS		
RITnet [15]	U-Net/DenseNet	OpenEDS [20]	✓	-	S/I/P/PO	It is computationally intensive and includes pre-processing.
iBUG [39]	VGG-16/ResNet101	iBUG (Proprietary) [39]	-	✓	S/I	Utilized for iris-only and sclera-only segmentation, including pre-processing and post-processing operations.
Eyenet [32]	ResNet	OpenEDS [20]	✓	-	S/I/P/PO	Contains extensive post-processing.
EyeSeg [40]	-	OpenEDS [20]	✓	-	S/I/P/PO	Contains redundant processing.
SCN [16]	SegNet	Proprietary [16]	-	✓	S/I	Introduced a large number of parameters.
Ocular-Net [41]	lite-residual	NICE-II [42]/SBVPI [11]	✓	-	S/I	Trains each region individually, working on one region at a time.
SSL [18]	-	OpenEDS [20]	✓	-	S/I/P/PO	Poor for low-quality images.
SIPFormer [19]	Transformer	CASIA [43]	✓	-	S/I/P/PO	Contains a large number of parameters, as well as a large number of pre-processing and post-processing.
Eye-UNet [17]	ResNet	OpenEDS [20]	-	✓	S/I/PO	Inference time is not ideal.
OcularSeg (Ours)	Efficientnetv2	Proprietary	✓	✓	S/I/P/PO	Rigorous training is required.

S = Sclera, I = Iris, P = Pupil, E = Eyelashes, C = Canthus, PO = Periocular (background), SR = Specular reflections, EB = Eyebrows, H = Hair, GF = Glass frames, NIR = Near-infrared, VIS = Visible light.

Subsequently, Ref. [39] introduced a low-resolution ocular segmentation dataset, offering two types of annotations: 30 keypoints and pixel-level annotations. They conducted preliminary eye segmentation investigations using deformable model-based methods and DeepLab with Atrous CNN+CRF, respectively. EyeNet [12] tackled multiple heterogeneous tasks related to gaze estimation and user semantic understanding. In addition, the feature encoding layer utilized ResNet50 as the backbone and integrated feature pyramid (FPN) [44] to capture the information across different scales. Similarly, EyeSeg [40] incorporated two key components: residual connections and dilated convolutional layers. This combination substantially improved performance without considerably increasing computational complexity. Furthermore, Luo et al. [16] introduced a shape-constrained network (SCN), which first uses VAE-GAN [45] to learn shapes, and employed pre-trained networks to regularize the training of SegNet. They curated and annotated a dataset comprising 8882 ocular images from 4461 facial images with varying resolutions, lighting conditions, and head poses. Subsequently, Naqvi et al. proposed Ocular-Net [41], a deep-learning-based lite-residual network. Additionally, Kothari proposed EllSeg [46], a simple three-category full ellipse segmentation framework, to extend the traditional encoder–decoder architecture. The results demonstrated that predicting pupil and iris centers and directions yielded superior performance compared to pixel-level segmentation models.

Hassan et al. [19] introduced SIPFormer, a novel framework comprising encoder, decoder, and transformer modules designed for joint segmentation. Their approach includes a pre-processing stage to enhance eye features while suppressing information from the periocular regions. By leveraging transformer modules, SIPFormer demonstrates improved feature learning capabilities, resulting in high accuracy in segmenting multi-modal features. Similarly, Eye-UNet [17] tackled the segmentation challenge posed by low-quality human

eye images captured outdoors. Initially, they curated a dataset of 5000 low-quality eye images with varying lighting conditions, occlusions, and motion blur. Based on U-Net architecture, their proposed network replaces the backbone network with Resnet18 and integrates an attention mechanism and a deep supervision module [47].

The field of ocular segmentation is dynamically evolving, witnessing the emergence of novel architectures and fusion strategies for multi-model data. Early studies predominantly focused on a single spectrum, such as visible light, employing traditional deep-learning techniques. While these methods succeeded under specific conditions, their limitations become apparent with increasing data volume and task complexity. To address these challenges more effectively, we advocate for a deep learning architecture amalgamating traditional and modern prior knowledge methods. By doing so, we aim to accurately capture subtle features and structures in eye images, offering a comprehensive solution for ocular segmentation.

3. Methodology

3.1. Overview

The proposed OcularSeg model is designed to perform joint segmentation of the periocular, sclera, iris, and pupil from ocular scans. A high-level overview of the model is depicted in Figure 3. This architecture begins by extracting features using a lightweight network, yielding five feature embeddings $E_i, i \in \{1, 2, 3, 4, 5\}$ through five distinct feature extraction stages. Subsequently, a convolution filter having size (3×3) pixels is applied individually to each feature embedding to reduce the channels to 64 and further minimize parameter redundancy. The resulting features are passed through the EMA module for feature aggregation. Following this, different levels of features are directed to the dense subtraction module, producing the decoder feature map $D_i, i \in \{1, 2, 3, 4, 5\}$. Ultimately, each D_i progressively contributes to the decoder and is combined with eye shape prior to generate the final predictions.

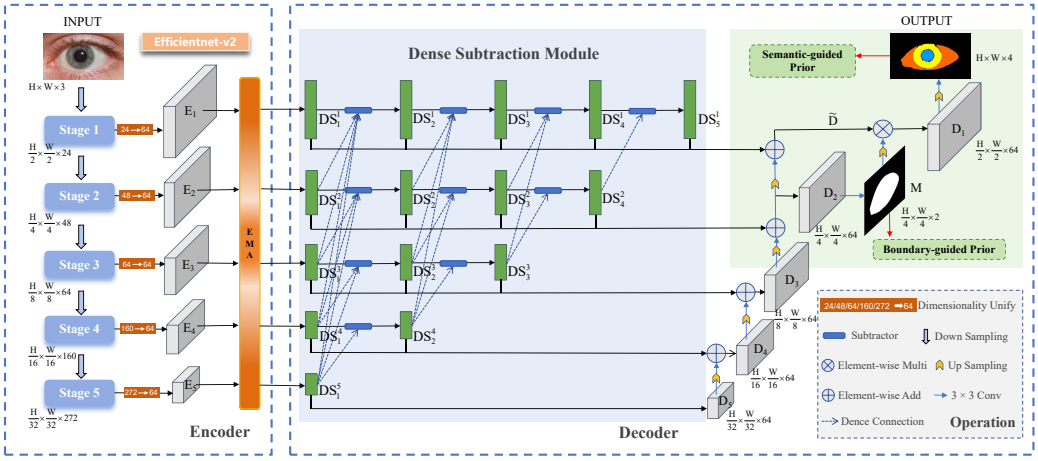


Figure 3. Overview of the proposed OcularSeg model.

3.2. Feature Extraction

We adopt the lightweight Efficientnetv2 [22] as the backbone to extract five levels of features. Then, a convolution filter having size (3×3) pixels is applied to the feature map output by each encoder block, standardizing the number of channels to 64. This facilitates subsequent operations and reduces the number of parameters. The resulting latent features E_i are then fed into the EMA [23] network, as shown in Figure 4. The attention mechanism is rethought from the perspective of the Expectation–Maximization (EM) [25] algorithm. Specifically, the EM algorithm is employed to identify a more compact base set μ and then operate the attention on this set. Through dynamic learning and adjustment of attention weights, we can obtain discriminative feature representations, enabling the model to focus more on key regions and specific semantic categories. Details are described as follows:

Starting with an input feature map E_i of size ($C \times H \times W$) pixels, where C, H, and W denote the number of channels, height, and width, respectively, it is reshaped into an $N \times C$ matrix by flattening it along the spatial dimensions ($N = H \times W$ for simplicity). Herein, $E_i^j \in \mathbb{R}^C$ represents the C-dimensional feature vector at pixel j. The EMA module comprises three key operations: responsibility estimation (A_E), likelihood maximization (A_M), and data re-estimation (A_R). Given the input $E_i \in \mathbb{R}^{N \times C}$ and initial bases $\mu \in \mathbb{R}^{K \times C}$, A_E estimates latent variables $Z \in \mathbb{R}^{N \times K}$, resembling the Expectation (E) step in the EM algorithm. Subsequently, A_M utilizes this estimation to update the bases μ, resembling the Maximization (M) step in the EM algorithm. The A_E and A_M steps alternate for a fixed number of iterations (here, we empirically set it to 3). Finally, with the converged μ and Z, A_R reconstructs the original E_i as Y.

This algorithm treats the construction base μ as the learnable parameter and the attention map Z as the latent variable. The objective is to find the maximum likelihood estimate of the parameters. It has been demonstrated that the complete data likelihood $\ln p(E_i, Z)$ monotonically increases with the iteration of EM steps. The E step computes the expectation value for each position, leveraging the current attention weights and feature representations to estimate the significance associated with each position. The M step utilizes the estimated E values to adjust the attention weights, thereby directing the model's focus towards features deemed more critical for the current ocular segmentation task. This module is integrated into the feature extraction stage of the segmentation model, called the EMA unit. By alternating the E step and the M step, the EMA module dynamically learns and adjusts the attention weights. Consequently, the updated Z and μ better reconstruct the original ocular data E_i, reducing intra-class feature differences and rendering features more compact.

In summary, the operation of A_E in the t-th iteration is formulated as

$$Z^{(t)} = \text{softmax}\left(\lambda E_i \left(\mu^{(t-1)}\right)^\top\right), \; i \in \{1, 2, 3, 4, 5\}, \tag{1}$$

where λ is a hyperparameter that controls the distribution of Z.

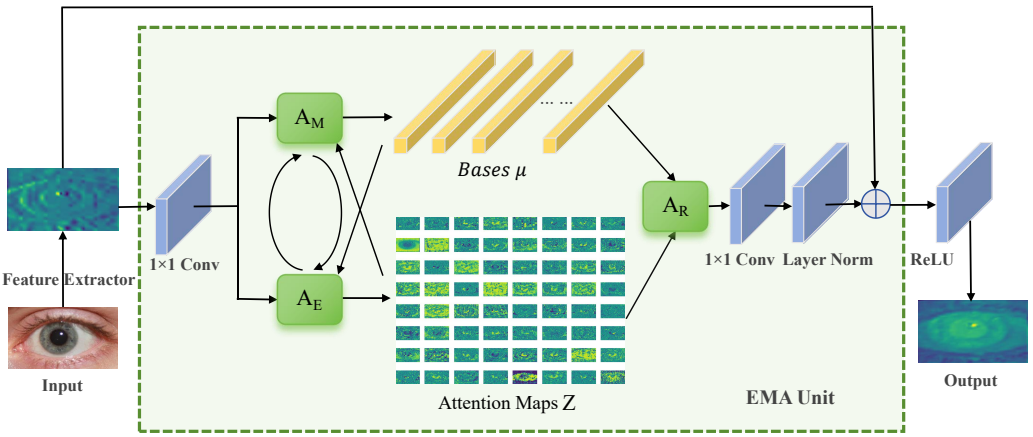


Figure 4. Overall structure of the EMA unit, where A_E and A_M execute alternately.

3.3. Dense Subtraction Module

In the classic U-shaped segmentation architecture, various feature levels undergo gradual fusion within the decoder via element-wise addition or concatenation. Nonetheless, these conventional operations often lead to substantial redundancy, undermining the complementary relationship between features at disparate levels and resulting in inaccurate segmentation. To alleviate this problem, we propose a bottom-up dense subtraction module (DSM), where each subtractor module is designed based on the multi-scale subtraction unit (MSU) from M^{2}SNet [48]. Concretely, the subtractor employs all-one convolutional filters

of size (1 × 1), (3 × 3), and (5 × 5) pixels to compute detail and structure differences based on pixel–pixel and regional patterns, defined as follows:

$$\begin{aligned} Subtractor(DS_h\ ,\ DS_l) = Conv(\\ \mid F(DS_h)_{1\times1} \ominus F(DS_l)_{1\times1} \mid + \\ \mid F(DS_h)_{3\times3} \ominus F(DS_l)_{3\times3} \mid + \\ \mid F(DS_h)_{5\times5} \ominus F(DS_l)_{5\times5} \mid), \end{aligned} \tag{2}$$

where $\ominus$ is the element-wise subtraction operation, and $F(\cdot)_{n\times n}$ represents a convolutional filter of size $(n \times n)$ pixels. The subtractor captures complementary information from DS_h and DS_l, representing high-level and low-level semantic features, respectively, and emphasizes their differences, thereby enriching information for the decoder.

As depicted in Figure 3, to obtain high-level complementary information at multiple feature levels, we horizontally and vertically connect multiple subtractors to compute a series of differential features with varying orders and receptive fields. Subsequently, we aggregate scale-specific features DS_1^i and cross-scale differential features $DS_{n\neq1}^i$ between the corresponding and higher levels. This process can be expressed as follows:

$$DS_i^k = \sum_{j=k+1}^{7-i} \left| Subtractor(DS_{i-1}^j\ , DS_{i-1}^k) \right|, \tag{3}$$

where i and k represent the row and column indices of the feature maps in DSM, respectively. Notably, the channel number of each feature map is kept uniform. This process aids in further restructuring semantic information after feature extraction and aggregation from the original backbone, thereby facilitating efficient segmentation of the eye region.

3.4. Decoder with Eye Shape Priors

Finally, the results generated by the DSM are input into the decoder for feature enhancement and up-sampling. The decoder feature map D_i is obtained as

$$D_i = \sum_{k=1}^{6-i} DS_k^i + UP[Conv(D_{i+1})], \quad i = 2,3,4,5. \tag{4}$$

where $Conv$ denotes the convolution filter of size 3 × 3 pixels and UP represents the up-sampling operation.

Due to the fixed positional relationship between each modality imposed by the ocular region's physiological structure and functional requirements, a specific aggregation of eye features exists. Considering this property, we incorporate the convex hull [49] of the ocular region, design the eye shape within the decoder part to constrain the ocular region, and guide the model to focus more effectively on key regions of the eye image. As illustrated in Figure 5a, we integrate the boundary-guided prior into the high-resolution layer D_1. Initially, a convolution filter having size (3 × 3) pixels is applied to D_2, followed by a softmax operation to generate a binary supervised signal M. Subsequently, M undergoes a slicing operation to acquire a probability-valued feature map, which is then element-wise multiplied with the result of the up-sampling operation on D_2 to obtain D_1. The formulation is as follows:

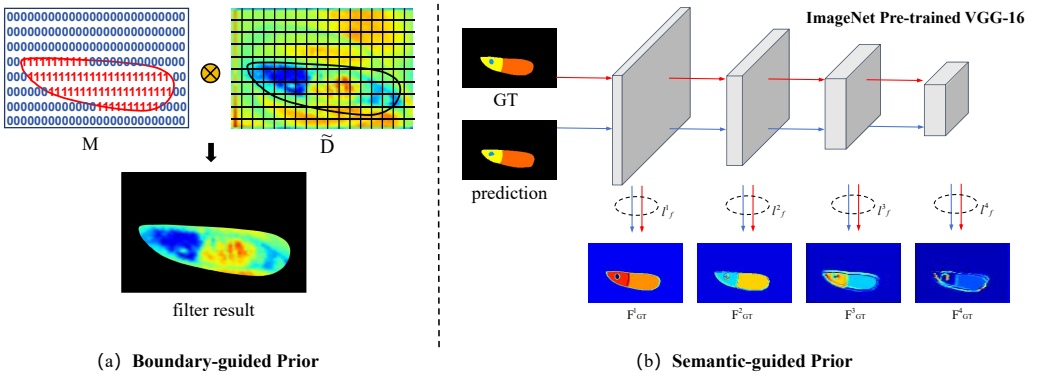
$$M = softmax[Conv(D_2)], \tag{5}$$

$$\widetilde{D} = \sum_{i=1}^{5} DS_i^1 + UP[Conv(D_2)], \tag{6}$$

$$D_1 = \widetilde{D} \otimes UP(M\ [:,1,:,:]), \tag{7}$$

where $Conv$ denotes the convolution filter of size 3 × 3 pixels, UP represents the up-sampling operation, and the notation $\otimes$ signifies element-wise multiplication.

By augmenting visual features, the boundary-guided eye shape prior can furnish more accurate and reliable information in ocular image analysis. This constraint mechanism is anticipated to elevate performance and accuracy in processing and analyzing eye images, particularly in tasks necessitating refined modeling of visual attention regions.



(a) **Boundary-guided Prior** (b) **Semantic-guided Prior**

Figure 5. An illustration of the proposed eye shape priors.

When observing a person's eyes, discernible features encompass the ocular regions and the positional relationships among the periocular, sclera, iris, and pupil areas. The pupil typically manifests as a black circular region positioned centrally within the eye. The iris, between the black pupil and the white sclera, presents as a colored circular region, while the sclera encompasses the iris. The periocular region, comprising the skin around the eyes, may exhibit various textures surrounding the above ocular modalities. These components collectively contribute to the distinct semantic features of each individual's eye. We aim to fine-tune a pre-trained VGG network to correct and enhance feature discrepancies between prediction results and ground truth progressively, from shallow to deep layers. This process, termed semantic-guided prior, captures the semantic relationships between different modalities by supervising the generation of 4-classes prediction results.

It is evident that low-level feature maps harbor abundant information, whereas high-level feature maps encapsulate more location information, as depicted in Figure 5b. We extract multi-scale features from the prediction and ground truth, respectively. Subsequently, the feature difference between them is computed as loss L_f:

$$L_f = l_f^1 + l_f^2 + l_f^3 + l_f^4, \tag{8}$$

where l_f^i is described as

$$l_f^i = \left\| F_P^i - F_G^i \right\|_2, \quad i = 1, 2, 3, 4, \tag{9}$$

where F_P^i and F_G^i represent the feature maps extracted from prediction results and ground-truth labels at layer i from shallow to deep, respectively. Thus, l_f^i is calculated as the Euclidean distance (L2-Loss) between two feature maps at the same level.

3.5. Training Objectives

During the training process, we utilize three loss functions to optimize the entire model: boundary-guided prior loss, semantic-guided prior loss, and regular semantic segmentation loss. Specifically, for the boundary-guided prior loss, we employ a combination of cross-entropy loss [50] and dice loss [51] for joint optimization to learn the ocular masks of the eye regions. Mathematically, it is formulated as L_B:

$$L_B = \alpha \cdot L_{CE}^b + \beta \cdot L_{Dice}^b. \tag{10}$$

The semantic-guided prior loss is expressed as L_f, as described in Section 3.4. For the regular semantic segmentation loss L_S, we employ structure loss, consisting of cross-

entropy loss and IoU loss, to learn multi-modal (4-classes) ocular segmentation results. Mathematically, the function of L_S is formulated as

$$L_S = \delta \cdot (L^s_{IoU} + L^s_{CE}). \tag{11}$$

Overall, these loss functions are jointly optimized as follows:

$$L_{total} = L_B + L_S + \gamma \cdot L_f. \tag{12}$$

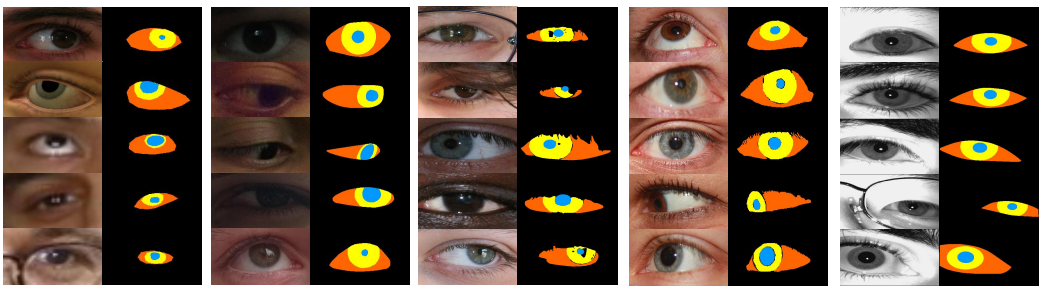
In the loss functions above, the terms α, β, γ, and δ are tunable hyperparameters. The L_{CE} and L_{Dice} terms are commonly employed in training models for semantic segmentation tasks. While the gradient calculation of L_{CE} is more intuitive, facilitating easier optimization, L_{Dice} effectively addresses pixel category imbalance, making it suitable for our eye images. Additionally, the L_{IoU} loss aids in suppressing false positives by quantifying the intersection-over-union ratio between the prediction masks and the ground-truth masks.

4. Experimental Settings

This section outlines the experimental setup for the ocular segmentation task. We commence by introducing the dataset collected and annotated for this purpose. Subsequently, we describe the performance metrics utilized and discuss recent state-of-the-art baseline methods for comparative evaluation. Finally, we provide detailed insights into implementing the proposed method and any reproduced methods used in the study.

4.1. Datasets

We are dedicated to constructing comprehensive and accurate datasets to address the existing scarcity data, enabling the model to perform effectively across various real-world application scenarios. Our experiments gathered five standard datasets, each potentially containing different ocular modalities and image qualities. For instance, MOBIUS is a dataset of low quality, encompassing annotations for all four required modalities, whereas SBVPI is of high quality but only includes sclera and periocular modalities. The specifics of these datasets are outlined in Table 2; each of them is divided into two subsets, where 80% of the images are randomly selected for training and the remaining 20% for testing. We augmented the annotations based on the original datasets, ensuring the availability of all necessary annotations for experimental training, as depicted in Figure 6.



(a) ESLD (b) MOBIUS (c) UBIRIS.v2 (d) SBVPI (e) MMU

Figure 6. Example images and corresponding ground-truth segmentation masks of five datasets.

Table 2. Summary of the datasets used in our method.

Subset	Spectrum	Quality	Images	Training Set	Testing Set	Input Size	Original	Supplement
ESLD [52]	VIS	low	2353	1882	471	256 × 128	S/I/P/PO	-
MOBIUS [53]	VIS	low	3500	2800	700	288 × 160	S/I/P/PO	-
UBIRIS.v2 [54]	VIS	low	919	735	184	288 × 160	I	S/P/PO
SBVPI [11]	VIS	high	1233	986	247	288 × 160	S/PO	I/P
MMU [55]	NIR	low	993	794	199	288 × 160	-	S/I/P/PO

S = Sclera, I = Iris, P = Pupil, PO = Periocular (background), NIR = Near-infrared, VIS = Visible light.

It is important to note that within these five datasets in Figure 6, (a), (b), and (e) are categorized under coarse labeling, while (c) and (d) fall under elaborate labeling. Coarse labeling provides complete information without noise, simulating real-world scenarios where labeling may lack detail and comprise only basic structural information. By incorporating these coarse annotations, we aim to enhance the model's robustness, enabling it to handle better challenges, such as missing or incomplete annotations that may arise in real-world environments. In contrast, elaborate annotations prioritize offering high-quality and accurate information, potentially including noise, occlusions, or other complexities, to ensure the model receives adequate guidance when learning crucial features. This is particularly crucial for tasks demanding high data accuracy, such as medical image recognition, where capturing subtle structures accurately is paramount. By leveraging both coarse and elaborate annotations in our experiments, we seek to further enhance the model's ability to handle diverse modalities, balance its performance across different annotation levels, and enable it to adapt flexibly to various application scenarios. The relevant information for each dataset is as follows:

ESLD [52] is a multi-type ocular structure dataset comprising ocular images captured by standard cameras under natural lighting conditions and synthetic ocular images. The dataset is obtained through three primary methods: (i) capturing facial images of users during computer usage; (ii) obtaining facial images from public datasets captured with standard cameras under natural lighting conditions; (iii) generating synthetic eye images using the UnityEye software (https://www.cl.cam.ac.uk/research/rainbow/projects/unityeyes/tutorial.html (accessed on 16 May 2024)). These acquisition methods yielded 1386, 804, and 1600 eye images. Subsequently, 40 feature points within the ocular region are annotated on the original images, and ocular images of varying sizes are normalized to dimensions of 256 $\times$ 128 pixels. This dataset serves as valuable support for researchers investigating changes in users' emotional and psychological states through the analysis of ocular images. The ground-truth iris/sclera/pupil/periocular segmentation masks were manually labeled by the dataset owner.

MOBIUS [53] was developed for mobile ocular biometrics in uncontrolled environments. It comprises 16,717 RGB images of 200 eyes from 100 Caucasian subjects, with an image resolution of 3000 $\times$ 1700 pixels. Images in the dataset were captured under various gaze directions (left, right, straight, and up) using three different mobile phones, i.e., Sony Xperia Z5 Compact (made by Sony, in Tokyo, Japan), Apple iPhone 6s (made by Apple Inc., in Cupertino, CA, USA), and Xiaomi Pocophone F1 (made by Xiaomi, in Beijing, China), resulting in considerable variability in image quality. We utilize a subset of this dataset specifically for ocular segmentation research, where annotations for the four modalities were manually created by the dataset owner.

UBIRIS.v2 [54] was originally collected for less-constrained iris recognition. It includes 11,102 RGB images from 261 subjects captured on the move and at a distance with a Canon EOS 5D camera (made by Canon Inc., in Tokyo, Japan). In the experiment, we utilize the subset from the NICE. I competition [29], where each image was manually labeled with iris mask. We also manually annotated the ground-truth masks of sclera, pupil, and periocular regions.

SBVPI [11] is a high-quality ocular dataset tailored for sclera recognition, but it is also suitable for iris and periocular recognition research. It consists of 1858 RGB images of 110 eyes from 55 Caucasian subjects, captured with a DSLR camera in a controlled laboratory setting, with a resolution of 3000 $\times$ 1700 pixels. Similar to MOBIUS, images in SBVPI were captured under four different gaze directions (left, right, straight, and up). We employ a subset of this dataset for our experiments, and manually annotate the iris and pupil masks for each image based on the sclera and periocular segmentation annotations previously provided by the dataset owner. Notably, the image quality in SBVPI is substantially higher than that of all other datasets.

MMU [55] is provided by the Malaysian Multimedia University and captured under near-infrared light conditions. The dataset comprises two subsets, MMU1 and MMU2,

categorized based on noise exposure and image quality. MMU1 contains 450 iris images with less noise, while MMU2 contains 995 images captured at a distance, with a 320×238 pixels resolution. These images exhibit various types of noise, such as eyelashes, eyelids, occlusions, specular reflections, uneven lighting, nonlinear deformation, and low contrast. We manually labeled a subset with ground-truth iris/sclera/pupil/periocular segmentation masks for our experiments.

Our experimental design aimed to ensure the model's robust performance across a wide range of data qualities and complexities. This comprehensive evaluation validates the model's efficacy and adaptability in real-world applications, enabling it to be generalizable and handle data with diverse modalities and annotation levels.

4.2. Evaluation Metrics

To measure the performance of multi-modal ocular segmentation, we compute the Precision (P), Recall (R), F_1-score (F_1), Intersection over Union (IoU), and Dice score (Dice) for each modality, respectively, and then take the mean value of all modalities as the whole evaluation metrics. The single-modal performance metrics are defined as follows:

- Precision (P): It measures the proportion of correctly predicted pixels to the total number of predicted pixels, calculated as $\frac{TP}{TP+FP}$.
- Recall (R): It measures the proportion of correctly predicted pixels relative to the total number of ground-truth pixels, formulated as $\frac{TP}{TP+FN}$.
- F_1-score (F_1): Defined as the harmonic mean between precision and recall, given by $2 \cdot \frac{P \cdot R}{P+R}$. It is a balance metric between precision and recall and is considered as the prior metric for comparing different methods.
- Intersection over Union (IoU): Represents the ratio between (i) the size of the intersection of the predicted and ground-truth regions and the size of their union, calculated as (ii) $\frac{TP}{TP+FN+FP}$.
- Dice score (Dice): Another measure of the overlap between predicted results and ground-truth labels, commonly used in segmentation tasks. It is calculated as $\frac{2TP}{2TP+FN+FP}$.

TP represents the number of true positives, indicating correctly predicted pixels; *FP* stands for false positives, representing background pixels incorrectly predicted as pixels; and *FN* denotes false negatives, indicating pixels incorrectly predicted as background pixels. These metrics are bounded in 0 and 1, where a higher value indicates a better segmentation result. In addition, the multi-class receiver operating characteristic (ROC) and precision–recall (PR) curves are generated by varying the decision threshold to yield different binary segmentation masks, thereby evaluating the overall segmentation performance.

4.3. Implementation Details

The proposed model is implemented in PyTorch (Version: 1.8.0) and initialized with the Efficientnetv2 [22] pretrained on ImageNet. Throughout the experiment, we standardized the image resolutions of different datasets to a fixed size using bilinear interpolation, as outlined in Table 2, to facilitate batch processing. We employed the SGD optimizer to optimize our model, with a batch size of 8, momentum of 0.9, and weight decay of 1×10^{-4}. Our learning rate policy followed the polynomial decay, where the learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with the power of 0.9. We set the initial learning rate to 0.1 and the maximum iteration to 30,000. The hyperparameters α, β, γ, and δ were configured to 1, 1, 0.1, and 1, respectively. All experiments were conducted using a single NVIDIA RTX 3090 GPU (made by NVIDIA Corporation, in Santa Clara, CA, USA).

5. Experimental Results

5.1. Comparison with State-of-the-Art

In this section, we assess the performance of multi-modal ocular segmentation on the collected datasets. To ensure a fair comparison, we evaluate not only classical CNN-based semantic segmentation models such as U-Net [21], DeepLabv3+ [56], and transformer-based

methods like TransUNet [57] but also recent ocular segmentation methods like EyeSeg [40] and Eye-UNet [17]. As the base model of our proposed OcularSeg, M^{2}SNet [48] is also used for comparison. We retrain these baseline methods on the same training datasets used for our proposed method. As outlined in Section 4.3, we utilized the proposed method to predict segmentation results for four categories, subsequently computing Precision, Recall, F1, IoU, and Dice metrics for different methods. The comparison results are presented in Table 3.

Table 3. Comparison of the proposed OcularSeg method with existing methods on five datasets. The bold values represent the best performances.

Dataset	Method	Precision (%) ↑	Recall (%) ↑	F_1(%) ↑	IoU (%) ↑	Dice(%) ↑
ESLD	U-Net [21]	87.3481	83.8864	85.9616	76.2965	82.8186
	DeepLabv3+ [56]	87.6251	86.5164	87.0468	77.8608	83.6691
	TransUNet [57]	**87.8762**	86.3939	87.1206	77.9557	84.3116
	M^{2}SNet [48]	87.3153	87.6695	87.4766	78.4824	84.6058
	EyeSeg [40]	87.2658	84.0316	85.5464	75.7187	82.0045
	Eye-UNet [17]	87.6377	86.8893	87.2518	78.1396	84.3350
	OcularSeg (ours)	87.4155	**88.2361**	**87.8214**	**78.9869**	**84.9862**
MOBIUS	U-Net [21]	91.5314	91.1735	91.3513	84.6038	90.3752
	DeepLabv3+ [56]	90.7674	92.5963	91.6383	84.9759	90.6610
	TransUNet [57]	**92.2497**	90.7409	91.4758	84.7980	90.3963
	M^{2}SNet [48]	92.1401	91.8056	91.9716	85.5650	90.9405
	EyeSeg [40]	90.6816	88.9764	89.8077	82.1804	88.2383
	Eye-UNet [17]	92.2172	90.3344	91.2222	84.2480	89.6927
	OcularSeg (ours)	92.1548	**92.6767**	**92.4134**	**86.2826**	**91.4060**
UBIRIS.v2	U-Net [21]	92.2269	93.5845	92.8452	87.0086	90.1322
	DeepLabv3+ [56]	92.7268	93.8245	93.2649	87.6711	90.8170
	TransUNet [57]	93.2628	93.7707	93.5077	88.0714	91.2042
	M^{2}SNet [48]	91.8245	**95.0000**	93.3620	87.8614	91.0248
	EyeSeg [40]	91.4751	93.7088	92.5234	86.4000	89.2864
	Eye-UNet [17]	92.9044	93.9907	93.4320	87.9459	90.7327
	OcularSeg (ours)	**94.2200**	93.1717	**93.6832**	**88.3299**	**91.3085**
SBVPI	U-Net [21]	95.4764	96.8177	96.1145	92.5744	95.6842
	DeepLabv3+ [56]	95.2275	97.1899	96.1734	92.6796	95.7601
	TransUNet [57]	95.9070	96.9274	96.4031	93.0974	96.0489
	M^{2}SNet [48]	95.8556	96.7834	96.2979	92.9049	95.8969
	EyeSeg [40]	94.6019	96.2444	95.3999	91.2714	94.7392
	Eye-UNet [17]	**96.4776**	95.7972	96.1274	92.5865	95.7257
	OcularSeg (ours)	95.2585	**97.7663**	**96.4817**	**93.2409**	**96.1035**
MMU	U-Net [21]	95.2909	95.4245	95.3484	91.2482	95.1574
	DeepLabv3+ [56]	95.0387	95.6079	95.2998	91.1421	95.1575
	TransUNet [57]	96.0650	94.7250	95.3441	91.2222	95.1254
	M^{2}SNet [48]	95.3193	94.1703	94.6143	89.9287	94.3720
	EyeSeg [40]	**96.2267**	93.5702	94.8473	90.3489	94.6760
	Eye-UNet [17]	93.4111	95.0539	94.2011	89.3189	94.0891
	OcularSeg (ours)	95.2293	**95.9017**	**95.5616**	**91.6152**	**95.4080**

It can be seen that our proposed method demonstrates the best performance in most metrics across all datasets, particularly low-quality ones. However, it is noteworthy that our method does not consistently achieve optimality in terms of the Precision metric. Our analysis indicates that this phenomenon may arise due to the model's tendency to become more aggressive in predicting positive (target) categories during iteration, thereby increasing false positives. Several specific issues contribute to this behavior, including the following: (i) Imbalanced category distribution: In scenarios where the ocular region occupies a relatively small portion of the image compared to background categories, the model may overpredict the target category to ensure a higher Recall. Consequently, this behavior can increase false positives in the background, resulting in lower Precision. (ii) Model prediction bias: Semantic segmentation models are often biased toward larger objects or more prominent image regions. This bias can cause the model to overpredict target categories, consequently increasing Recall. However, due to this bias, some predictions may

lack accuracy, leading to lower Precision. (iii) The trade-off between Recall and Precision: A trade-off exists between Recall and Precision, wherein increasing Recall typically leads to a decrease in Precision and vice versa. This trade-off reflects the model's inherent balance between accurately capturing all relevant instances of the target category (Recall) and minimizing false positives (Precision). To comprehensively evaluate the model's performance, it is essential to consider multiple metrics rather than focusing solely on a single metric. Evaluating the model using a combination of metrics provides a more holistic understanding of its performance across various aspects.

Furthermore, we conducted a detailed analysis of the OcularSeg performance by examining the receiver operating characteristic (ROC) and precision–recall (PR) curves for each category on the MOBIUS dataset, as illustrated in Figure 7. In the ROC curves depicted in Figure 7a, we observed that our proposed method accurately predicts the positive labels in each category after careful training. Comparatively, the precision–recall curves in Figure 7b are more sensitive to the unbalanced categories, making them particularly suitable for our ocular segmentation task and demonstrating the superiority of our algorithm more intuitively. Additionally, the area-under-the-curve (AUC) values, represented as the area in the figure, serve as quantitative indicators of the model's superior predictive performance. Among these, the micro-average (https://sklearn-evaluation.ploomber.io/en/latest/classification/micro_macro.html (accessed on 16 May 2024)) performs excellently in addressing category imbalance and effectively reflects the overall performance, especially when the sample size of some categories substantially outweighs others. Conversely, the macro-average is suitable for scenarios where each category is treated equally and remains unaffected by category imbalance.

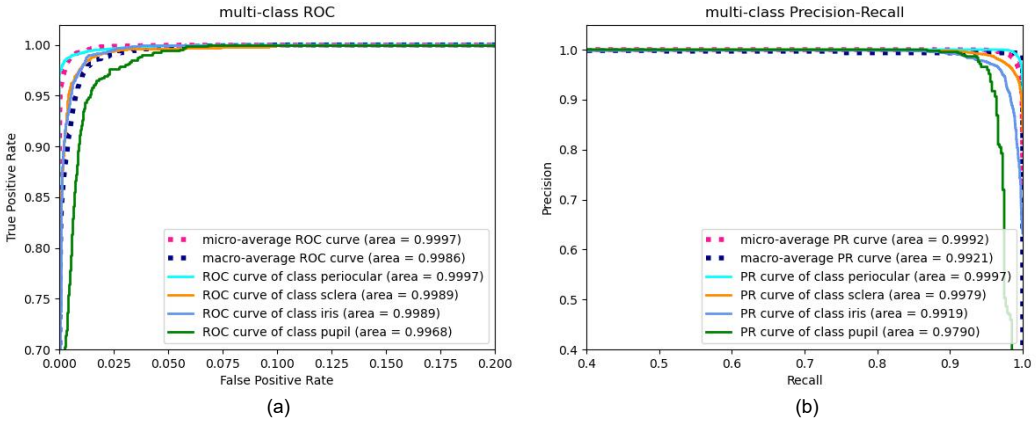


Figure 7. (**a**) Receiver operating characteristic (ROC) and (**b**) precision–recall (PR) curves generated by our proposed method, OcularSeg, for each class on the MOBIUS dataset.

5.2. Qualitative Evaluation

Here, a qualitative evaluation is performed to further analyze the proposed model. To this end, we select several representative and challenging eye images from all datasets for experiments, and different segmentation models are used for comparisons. The visualization results are shown in Figure 8. It can be seen that our OcularSeg model outperforms other baseline models in accurately segmenting four ocular modalities across the majority of samples. Nevertheless, we acknowledge that there is still room for improvement in existing methods when dealing with lower-quality datasets such as the ESLD and MOBIUS. This could be attributed to the presence of serious noise interference in the dataset such as blur, occlusions, specular reflection, and uneven illumination, coupled with potential errors in annotations.

In non-constrained scenarios, apart from normally captured eye images, a variety of variables such as changes in head pose, errors in eye movement tracking, or illumination variations often result in the occurrence of almost closed, fully closed, or misaligned eyes in the captured images. These extreme cases are of significant interest to the research

community as they reflect the robustness of the model in real-world applications. For this reason, we select several representative samples from the low-quality UBIRIS.v2 [54] dataset for testing. Figure 9 shows their multi-modal ocular segmentation results using the proposed OcularSeg. As can be seen, our model is still able to accurately segment the multi-modal ocular structures in the face of these adverse factors. Therefore, visual results demonstrate that the opening and closing state or position of the eyes does not significantly affect the accuracy of our model in most cases.

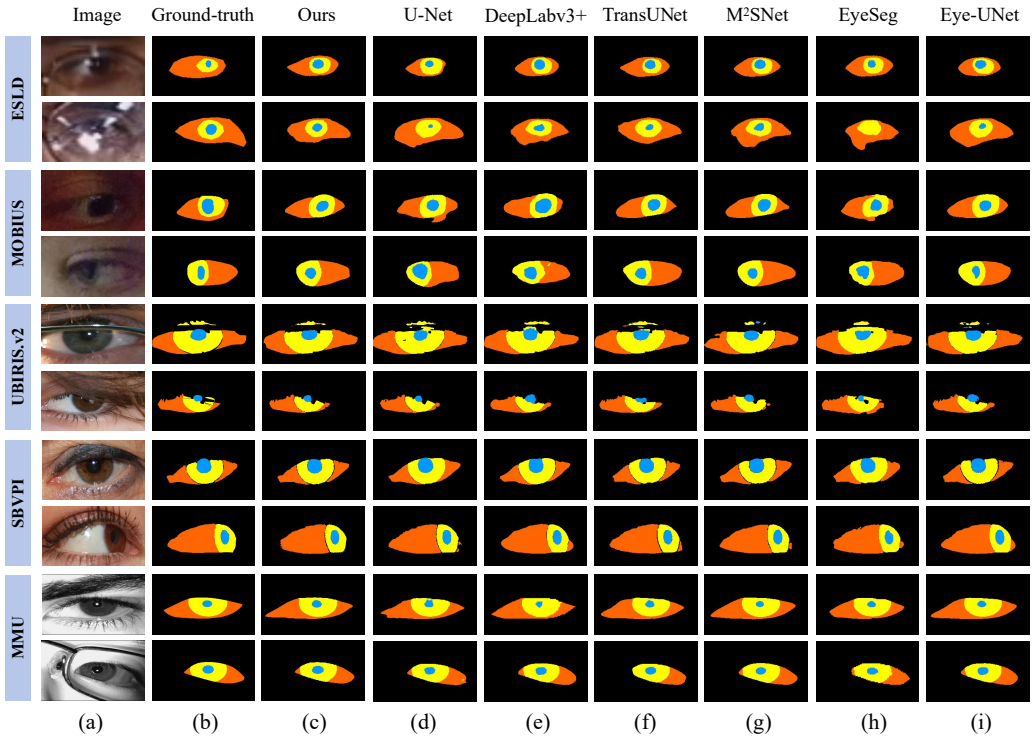


Figure 8. Multi-modal ocular segmentation results of challenging samples on multiple datasets. (**a**) Original images, (**b**) Ground-truth labels, (**c**) OcularSeg (ours), (**d**) U-Net, (**e**) DeepLabv3+, (**f**) TransUNet, (**g**) M^{2}SNet, (**h**) EyeSeg, (**i**) Eye-UNet.

Figure 9. Multi-modal ocular segmentation results of extreme cases using the proposed OcularSeg, including (**a**) almost closed eye images, (**b**) fully closed eye images, and (**c**) misaligned eye images, which are from the UBIRIS.v2 [54] dataset.

5.3. Comparison with Single-Modal Segmentation Techniques

In this section, we compare the performance of the multi-modal OcularSeg model to the performance of the OcularSeg model trained only for a single modal by leveraging the low-quality ESLD and MOBIUS datasets. The single-modal OcularSeg model is trained by removing two eye shape priors and modifying the final number of segmentation categories to two (modality and background). As a result, both OcularSeg variants have approximately the same set of parameters. The experimental results are shown in Tables 4 and 5.

As we can see from the results, the multi-modal OcularSeg model consistently outperforms the single-modal OcularSeg model in terms of F_1, IoU, and Dice metrics across different datasets for segmenting each individual modality. For example, on the ESLD dataset, the multi-modal OcularSeg model achieves the F_1 metric improvements by 0.0303%, 0.8639%, 0.8884%, and 1.2606%, respectively, for the periocular, sclera, iris, and pupil categories. Besides, corresponding improvements on the MOBIUS dataset are 0.0615%, 0.5704%, 0.8595%, and 2.1062%, respectively. Figure 10 also visually shows the improvements of the multi-modal OcularSeg model over the single-modal OcularSeg model in terms of the IoU metric. These results suggest the potential advantages of multi-modal ocular segmentation in enhancing the performance of single-modal ocular segmentation.

Table 4. Comparison of multi-modal and single-modal segmentation results on ESLD with the proposed OcularSeg. The bold values represent the best performances.

Dataset	Category	Precision (%) ↑	Recall (%) ↑	F_1 (%) ↑	IoU (%) ↑	Dice (%) ↑
ESLD	Periocular (single-modal)	98.5216	**99.1404**	98.8300	97.6871	98.8114
	Periocular (multi-modal)	**98.9457**	98.7750	**98.8603**	**97.7462**	**98.8373**
	Sclera (single-modal)	**85.0790**	80.9840	82.9810	70.9124	80.1791
	Sclera (multi-modal)	83.7979	**83.8919**	**83.8449**	**72.1836**	**81.1347**
	Iris (single-modal)	**88.0152**	85.9134	86.9516	76.9154	84.1478
	Iris (multi-modal)	86.7830	**88.9230**	**87.8400**	**78.3167**	**85.1080**
	Pupil (single-modal)	**80.7008**	78.2368	79.4497	65.9059	73.5095
	Pupil (multi-modal)	80.1353	**81.3545**	**80.7403**	**67.7012**	**74.8650**

Table 5. Comparison of multi-modal and single-modal segmentation results on MOBIUS with the proposed OcularSeg. The bold values represent the best performances.

Dataset	Category	Precision (%) ↑	Recall (%) ↑	F_1 (%) ↑	IoU (%) ↑	Dice (%) ↑
MOBIUS	Periocular (single-modal)	98.8133	**99.181**	98.9968	98.0135	98.9827
	Periocular (multi-modal)	**99.0999**	99.0168	**99.0583**	**98.1343**	**99.0455**
	Sclera (single-modal)	**93.9194**	91.9747	92.9369	86.8057	92.3554
	Sclera (multi-modal)	93.5948	**93.4199**	**93.5073**	**87.8063**	**92.9326**
	Iris (single-modal)	91.0376	90.8156	90.9265	83.3625	89.7834
	Iris (multi-modal)	**91.4323**	**92.1424**	**91.7860**	**84.8189**	**90.6959**
	Pupil (single-modal)	**84.5237**	81.9092	83.1959	71.2269	80.7107
	Pupil (multi-modal)	84.4922	**86.1277**	**85.3021**	**74.3711**	**82.9499**

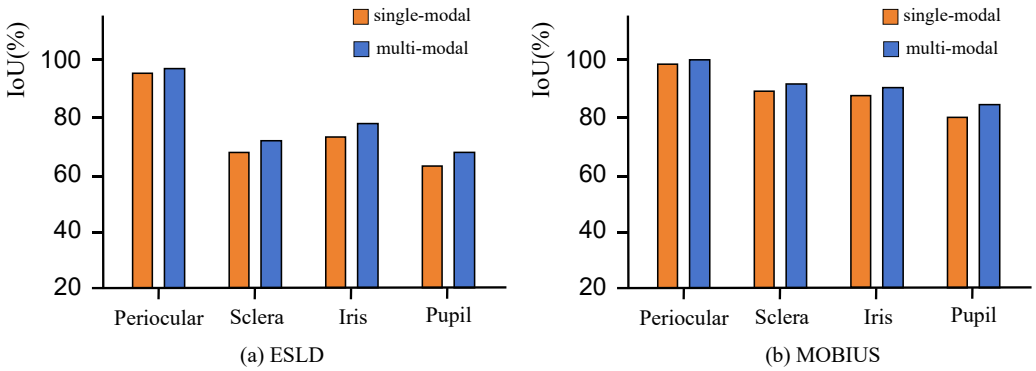


Figure 10. Performance comparison of single-modal and multi-modal segmentation on (**a**) ESLD and (**b**) MOBIUS datasets with the proposed OcularSeg.

For the observed improvements in performance, we analyze that the reason may be as follows: (1) The richer information and stronger complementary provided by multi-modal segmentation enable the model to robustly resist the interference of noise frequently encountered in single-modal scenarios. (2) Multi-modal segmentation alleviates the problem of foreground–background category imbalance during segmentation. These experimental findings further highlight the advantages of our multi-modal ocular segmentation model and offer valuable insights for enhancing single-modal segmentation methods.

5.4. Cross-Domain and Cross-Spectrum Evaluation

Addressing the generalization challenge of multi-modal ocular segmentation, we endeavor to assess performance across domains and spectral ranges using the datasets provided. Leveraging the diversity of label annotations in these datasets, we conduct a cross-domain evaluation using UBIRIS.v2 and SBVPI for visible light scenarios. At the same time, we utilize MOBIUS and MMU datasets for the cross-spectral problem. The experimental results are summarized in Tables 6 and 7.

Table 6. Cross-domain performance comparison on UBIRIS.v2 and SBVPI.

Training	Testing	Precision (%) ↑	Recall (%) ↑	F_1 (%) ↑	IoU (%) ↑	Dice (%) ↑
UBIRIS.v2	SBVPI	93.4298	88.2239	90.5547	83.1643	90.4116
SBVPI	UBIRIS.v2	79.7851	90.7899	83.7912	73.5187	80.2446

Table 7. Cross-spectral performance comparison on MOBIUS and MMU.

Training	Testing	Precision (%) ↑	Recall (%) ↑	F_1 (%) ↑	IoU (%) ↑	Dice (%) ↑
MOBIUS	MMU	93.6960	89.0114	91.0423	84.3484	90.7386
MMU	MOBIUS	84.7607	60.2717	66.2934	54.2809	62.6713

In evaluating visible light cross-domain performance, we assess the model's ability to generalize to a new domain by learning visible light features. This analysis aids in understanding the model's adaptability across diverse visible light datasets and offers valuable insights for real-world applications. As illustrated in Figure 11, the visualization highlights the model's robust generalization performance in the visible light domain. Notably, the model demonstrates effective transfer learning between the UBIRIS.v2 and SBVPI datasets, indicating its proficiency in capturing common visible light features. Consequently, the model achieves satisfactory segmentation across different data sources, bolstering the feasibility of ocular biometrics applications in varied visible light conditions.

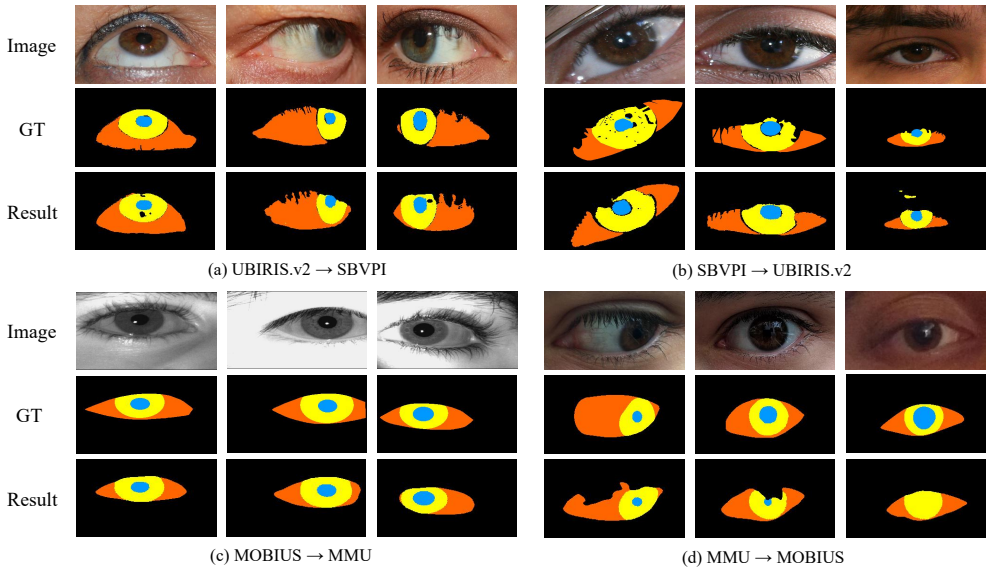


Figure 11. Visualization of cross-domain (**a**,**b**) and cross-spectral (**c**,**d**) segmentation results using the proposed OcularSeg.

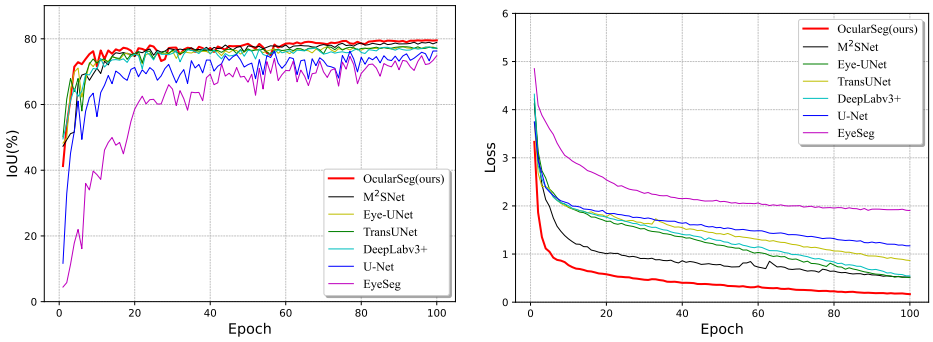
The assessment of cross-spectral performance introduces additional complexities compared to cross-domain evaluation. In our experiments using MOBIUS for visible light and MMU for near-infrared light, we observed a disparity in performance. Specifically, the model exhibited better performance when generalized from the visible to the near-infrared spectrum, whereas the reverse, near-infrared to visible light generalization yielded

comparatively lower results. This observation underscores the inherent challenges of adapting models to eye images captured across different spectral ranges.

Our experimental findings underscore the persistent challenges encountered in cross-domain and cross-spectral segmentation tasks. While we have made strides in achieving favorable outcomes on specific datasets, extending model performance to diverse conditions remains intricate. These challenges stem from fundamental differences in image features, resolutions, and lighting conditions inherent to each dataset, making accurate generalization across varied environments difficult. Consequently, achieving satisfactory performance on one dataset does not guarantee similar results on others. Moreover, addressing the cross-spectral problem amplifies the complexity of model adaptation to diverse spectral bands. Effectively coping with this challenge necessitates the model's sensitivity and adaptability to comprehend and capture the distinct information in different spectral ranges. Successfully navigating these complexities demands a deeper understanding of the intricate relationships between specific spectral bands and robust adaptation strategies.

5.5. Computational Complexity Analysis

In this section, we analyze the computational complexity of the proposed model during both the training and testing phases. First, we compare the convergence speed and performance of different segmentation models in the training. Therefore, we draw their corresponding IoU curves (Figure 12a) and loss curves (Figure 12b) with respect to the epochs, which are generated on the validation set and training set of the ESLD dataset, respectively. It should be noted that the original training set is partitioned into a final validation set comprising 10% of the data and a final training set consisting of the remaining 90% during the model development.



(**a**) (**b**)

Figure 12. IoU curves (**a**) and loss curves (**b**) of different segmentation models.

By observing the overall trend of the curves, we have the following findings: (1) For all methods, the IoU metric gradually increases while the loss value decreases over time until they all reach stability. (2) Our OcularSeg model exhibits the fastest convergence speed at the initial stages and maintains the highest performance throughout. Hence, we can conclude that compared to several baselines, our OcularSeg model does not impose a heavier training burden; instead, it demonstrates higher accuracy in most of the training time.

Secondly, we evaluate the computational complexity of different segmentation models in the testing. The model parameter amount, FLOPs, running time, and frames per second (FPS) are calculated in the consistent simulation environment, where the latter three metrics are with respect to the input of 288×160 pixels. From the results in Table 8, we can observe that (1) our OcularSeg model exhibits a more compact computational load regarding model parameters than other majority methods and (2) our OcularSeg model can process 24.08 FPS and requires 6.55 GFLOPs and 41.53 ms to process a single image. Although there are some gaps between our method and the most simplified one regarding inference efficiency, it is crucial to note that while the most simplified method may be computationally more efficient, it sacrifices segmentation accuracy. Overall, thanks to the strate-

gies employed for optimizing the model structure, such as lightweight feature extraction and cross-layer connections in DSM, our OcularSeg model achieves a good balance between performance and efficiency, rendering it feasible for real-time biometric applications.

Table 8. Computational complexity analysis. The bold and underlined values represent the best and second-best performances, respectively.

Metrics	OcularSeg (Ours)	U-Net [21]	DeepLabv3+ [56]	TransUNet [57]	M^{2}SNet [48]	EyeSeg [40]	Eye-UNet [17]
Params (M) ↓	<u>22.65</u>	39.40	59.34	93.19	27.7	**0.25**	31.04
FLOPs (G) ↓	<u>6.55</u>	61.70	18.76	128.68	6.92	**3.28**	41.92
Runtime (ms) ↓	41.53	**10.70**	34.66	48.57	27.74	<u>11.79</u>	12.95
FPS ↑	24.08	**93.46**	28.85	20.59	36.05	<u>84.82</u>	77.22

5.6. Ablation Study

We refine our research through ablation studies conducted on the challenging ESLD [52] dataset to validate the effectiveness of the core components of our method. The experimental results are detailed in Table 9, where the symbol ✓ denotes the inclusion of the module, whereas ✗ indicates its absence. Among the evaluation metrics, Precision, Recall, and F_1 are notably influenced by data imbalance. Therefore, we prioritize IoU and Dice for their ability to comprehensively and accurately capture the degree of overlap between prediction results and ground truth. Starting with Efficientnetv2 as the baseline, with simple subtraction cells as setting a, we analyze the contribution of each component in detail.

Table 9. Ablation experiments of the four parts in our proposed method. Here, Prior1 represents the boundary-guided prior and Prior2 is the semantic-guided prior. The bold values represent the best performances.

Settings	EMA	DSM	Prior1	Prior2	Precision (%) ↑	Recall (%) ↑	F_1 (%) ↑	IoU (%) ↑	Dice (%) ↑
a	✗	✗	✗	✗	85.9952	85.9889	85.6497	76.7358	82.5455
b	✓	✗	✗	✗	86.4852	87.9172	86.1581	76.9980	83.4262
c	✓	✓	✗	✗	85.8027	87.9969	86.7555	77.3918	84.3962
d	✓	✓	✓	✗	**88.2327**	86.8165	87.5091	78.5481	84.8176
e (ours)	✓	✓	✓	✓	87.4155	**88.2361**	**87.8214**	**78.9869**	**84.9862**

In setting b, we enhance the original baseline model by integrating the EMA algorithm, which aggregates the features of ocular modalities in the spatial domain after feature extraction. Incorporating this module yields gains of 0.26% and 0.88% on the evaluation metrics IoU and Dice, respectively. Moving to set c, we interconnect DSM across layers to accentuate feature disparities between adjacent layers, effectively mitigating the interference of redundant features. This enhancement encourages the model to extract richer semantic information, and the quantitative metrics in the experimental results corroborate the effectiveness of this component, with improvements of 0.39% and 0.97% in IoU and Dice, respectively.

Finally, in settings d and e, the position, shape, and internal topological relationship among modalities are supervised by the boundary-guided prior and the semantic-guided prior. Compared with the previous modules, the eye shape prior proves more effective in performance enhancement, underscoring the efficacy of our proposed eye shape prior constraints. Especially, in d, we introduce the convex hull as an extra boundary supervision. This step is akin to using manually annotated labels, providing more accurate guidance for the ocular boundary by leveraging external forces. Compared to internal annotations of the eye, convex hull is often smoother and relatively accurate. Therefore, by utilizing such prior knowledge for auxiliary supervision, precise boundary information of the target is provided, enabling the model to better understand the spatial location and shape of the ocular during learning. Additionally, this prior weakens noise around the ocular, helping to reduce errors and noise in prediction results, thus enhancing the accuracy

and generalization capability of the segmentation model. Overall, this represents a global optimization, whereas improvements in other modules often only optimizes specific feature representations locally, leading to limited performance gains. Quantitative results also demonstrate the effectiveness of the boundary-guided prior in segmentation performance compared to other modules.

6. Conclusions and Future Work

This study investigates the multi-modal ocular segmentation task in non-constrained scenarios, including near-infrared and visible light illumination conditions. To tackle this challenge comprehensively, we have annotated multiple challenging datasets and developed an effective segmentation model (OcularSeg). Our model encompasses lightweight feature extraction, feature aggregation, bottom-up dense connection layers, and guidance from eye shape priors, ultimately achieving state-of-the-art performance. We particularly emphasize performance enhancement compared to single-modal approaches and explore feasibility issues in cross-domain and cross-spectral contexts.

Future research directions will prioritize further optimization of the model to enhance its generalization ability and adaptability, especially for ocular images spanning different domains and spectral ranges. We will introduce more advanced domain adaptation techniques and inter-domain normalization methods, and enhance the model's ability to perceive spectral differences. Despite the progress made, there remains a pressing need to expand the size of available datasets, which currently hampers the training of more sophisticated segmentation models. Additionally, we will focus on algorithm optimization to improve efficiency and actively explore hardware acceleration solutions to ensure real-time inference.

Author Contributions: Conceptualization, Y.Z. and C.W.; methodology, Y.Z.; software, Y.Z.; validation, H.L.; formal analysis, Y.Z. and X.S.; investigation, C.W. and H.L.; resources, C.W.; data curation, Y.Z. and H.L.; writing—original draft preparation, Y.Z.; writing—review and editing, C.W., Q.T. and G.Z.; visualization, Y.Z. and X.S.; supervision, C.W.; project administration, C.W.; funding acquisition, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62106015), the Beijing Natural Science Foundation (4242018), the Young Elite Scientist Sponsorship Program by BAST (BYESS2023130), and the Pyramid Talent Training Project of BUCEA (JDYC20220819).

Data Availability Statement: The UBIRIS.v2 dataset is publicly available at http://iris.di.ubi.pt/ubiris2.html (accessed on 16 May 2024). The SBVPI and MOBIUS datasets are publicly available at https://sclera.fri.uni-lj.si/datasets.html (accessed on 16 May 2024). The ESLD dataset is publicly available at http://www.cjig.cn//html/jig/2022/8/20220802.htm (accessed on 16 May 2024). The MMU dataset is publicly available at https://www.kaggle.com/datasets/naureenmohammad/mmu-iris-dataset (accessed on 16 May 2024). Our code and ocular segmentation annotations for UBIRIS.v2, SBVPI, and MMU are publicly available via https://github.com/koala0623/OcularSeg (accessed on 16 May 2024).

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

References

1. Nguyen, K.; Proença, H.; Alonso-Fernandez, F. Deep Learning for Iris Recognition: A Survey. *ACM Comput. Surv.* **2024**, *56*, 1–35. [CrossRef]
2. Evangeline, D.; Parkavi, A.; Bhutaki, R.; Jhawar, S.; Pujitha, M.S. Person Identification using Periocular Region. In Proceedings of the 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India, 24–25 January 2024; pp. 1–6. [CrossRef]
3. Li, H.; Wang, C.; Zhao, G.; He, Z.; Wang, Y.; Sun, Z. Sclera-TransFuse: Fusing Swin Transformer and CNN for Accurate Sclera Segmentation. In Proceedings of the 2023 IEEE International Joint Conference on Biometrics (IJCB), Ljubljana, Slovenia, 25–28 September 2023; pp. 1–8. [CrossRef]
4. Nigam, I.; Vatsa, M.; Singh, R. Ocular biometrics: A survey of modalities and fusion approaches. *Inf. Fusion* **2015**, *26*, 1–35. [CrossRef]

5. Umer, S.; Sardar, A.; Dhara, B.C.; Rout, R.K.; Pandey, H.M. Person identification using fusion of iris and periocular deep features. *Neural Netw.* **2020**, *122*, 407–419. [CrossRef] [PubMed]
6. Gragnaniello, D.; Poggi, G.; Sansone, C.; Verdoliva, L. Using iris and sclera for detection and classification of contact lenses. *Pattern Recognit. Lett.* **2016**, *82*, 251–257. [CrossRef]
7. Oh, K.; Oh, B.S.; Toh, K.A.; Yau, W.Y.; Eng, H.L. Combining sclera and periocular features for multi-modal identity verification. *Neurocomputing* **2014**, *128*, 185–198. [CrossRef]
8. Xiong, J.; Zhang, Z.; Wang, C.; Cen, J.; Wang, Q.; Nie, J. Pupil localization algorithm based on lightweight convolutional neural network. *Vis. Comput.* **2024**, 1–17. [CrossRef]
9. He, Z.; Tan, T.; Sun, Z.; Qiu, X. Toward accurate and fast iris segmentation for iris biometrics. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 1670–1684.
10. Lucio, D.R.; Laroca, R.; Severo, E.; Britto, A.S.; Menotti, D. Fully convolutional networks and generative adversarial networks applied to sclera segmentation. In Proceedings of the 2018 IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), Redondo Beach, CA, USA, 22–25 October 2018; pp. 1–7.
11. Vitek, M.; Rot, P.; Štruc, V.; Peer, P. A comprehensive investigation into sclera biometrics: A novel dataset and performance study. *Neural Comput. Appl.* **2020**, *32*, 17941–17955. [CrossRef]
12. Wu, Z.; Rajendran, S.; van As, T.; Zimmermann, J.; Badrinarayanan, V.; Rabinovich, A. EyeNet: A multi-task network for off-axis eye gaze estimation and user understanding. *arXiv* **2019**, arXiv:1908.09060.
13. Perry, J.; Fernandez, A. Minenet: A dilated cnn for semantic segmentation of eye features. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3671–3676.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Chaudhary, A.K.; Kothari, R.; Acharya, M.; Dangi, S.; Nair, N.; Bailey, R.; Kanan, C.; Diaz, G.; Pelz, J.B. Ritnet: Real-time semantic segmentation of the eye for gaze tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3698–3702.
16. Luo, B.; Shen, J.; Cheng, S.; Wang, Y.; Pantic, M. Shape constrained network for eye segmentation in the wild. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 1952–1960.
17. Wang, Y.; Wang, J.; Guo, P. Eye-UNet: A UNet-based network with attention mechanism for low-quality human eye image segmentation. *Signal Image Video Process.* **2023**, *17*, 1097–1103.
18. Chaudhary, A.K.; Gyawali, P.K.; Wang, L.; Pelz, J.B. Semi-supervised learning for eye image segmentation. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, 25–29 May 2021; pp. 1–7.
19. Hassan, B.; Hassan, T.; Ahmed, R.; Werghi, N.; Dias, J. SIPFormer: Segmentation of Multiocular Biometric Traits With Transformers. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 1–14. [CrossRef]
20. Garbin, S.J.; Shen, Y.; Schuetz, I.; Cavin, R.; Hughes, G.; Talathi, S.S. Openeds: Open eye dataset. *arXiv* **2019**, arXiv:1905.03702.
21. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
22. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 10096–10106.
23. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9167–9176.
24. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. [CrossRef]
25. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–22. [CrossRef]
26. Rot, P.; Emeršič, Ž.; Struc, V.; Peer, P. Deep multi-class eye segmentation for ocular biometrics. In Proceedings of the IEEE International Work Conference on Bioinspired Intelligence (IWOBI), San Carlos, Costa Rica, 18–20 July 2018; pp. 1–8.
27. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
28. Osorio-Roig, D.; Rathgeb, C.; Gomez-Barrero, M.; Morales-González, A.; Garea-Llano, E.; Busch, C. Visible wavelength iris segmentation: A multi-class approach using fully convolutional neuronal networks. In Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 26–28 September 2018; pp. 1–5.
29. Proença, H.; Alexandre, L.A. The NICE. I: noisy iris challenge evaluation-part I. In Proceedings of the IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), Crystal City, VA, USA, 27–29 September 2007; pp. 1–4.
30. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA , 7–12 June 2015; pp. 3431–3440.

31. Boutros, F.; Damer, N.; Kirchbuchner, F.; Kuijper, A. Eye-mms: Miniature multi-scale segmentation network of key eye-regions in embedded applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Seoul, Republic of Korea, 27–28 October 2019.
32. Kansal, P.; Devanathan, S. Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3688–3693.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
34. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
35. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
36. Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv* **2014**, arXiv:1404.1869.
37. Das, A.; Pal, U.; Ferrer, M.A.; Blumenstein, M.; Štepec, D.; Rot, P.; Emeršič, Ž.; Peer, P.; Štruc, V.; Kumar, S.A.; et al. SSERBC 2017: Sclera segmentation and eye recognition benchmarking competition. In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, 1–4 October 2017; pp. 742–747.
38. Sequeira, A.F.; Monteiro, J.C.; Rebelo, A.; Oliveira, H.P. MobBIO: A multimodal database captured with a portable handheld device. In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, 5–8 January 2014; Volume 3, pp. 133–139.
39. Luo, B.; Shen, J.; Wang, Y.; Pantic, M. The iBUG Eye Segmentation Dataset. In Proceedings of the 2018 Imperial College Computing Student Workshop (ICCSW), London, UK, 20–21 September 2018; Volume 66, pp. 7:1–7:9. [CrossRef]
40. Perry, J.; Fernandez, A.S. EyeSeg: Fast and Efficient Few-Shot Semantic Segmentation. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW), Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 570–582.
41. Naqvi, R.A.; Lee, S.W.; Loh, W.K. Ocular-net: Lite-residual encoder decoder network for accurate ocular regions segmentation in various sensor images. In Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp), Busan, Republic of Korea, 19–22 February 2020; pp. 121–124.
42. Bowyer, K.W. The results of the NICE. II iris biometrics competition. *Pattern Recognit. Lett.* **2012**, *33*, 965–969. [CrossRef]
43. Test, B.I. CASIA.v4 Database. Available online: http://www.idealtest.org/dbDetailForUser.do?id=4 (accessed on 16 May 2024).
44. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA , 21–26 July 2017; pp. 2117–2125.
45. Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of the International Conference on Machine Learning (ICML), New York, NY, USA, 20–22 June 2016; pp. 1558–1566.
46. Kothari, R.S.; Chaudhary, A.K.; Bailey, R.J.; Pelz, J.B.; Diaz, G.J. Ellseg: An ellipse segmentation framework for robust gaze tracking. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 2757–2767. [CrossRef]
47. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In Proceedings of the International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
48. Zhao, X.; Jia, H.; Pang, Y.; Lv, L.; Tian, F.; Zhang, L.; Sun, W.; Lu, H. M^2SNet: Multi-scale in Multi-scale Subtraction Network for Medical Image Segmentation. *arXiv* **2023**, arXiv:2303.10894.
49. Seidel, R. Convex hull computations. In *Handbook of Discrete and Computational Geometry*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2017; pp. 687–703.
50. Pihur, V.; Datta, S.; Datta, S. Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach. *Bioinformatics* **2007**, *23*, 1607–1615. [CrossRef] [PubMed]
51. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
52. Zhang, J.J.; Sun, G.M.; Zheng, K.; Li, Y.; Fu, X.H.; Ci, K.Y.; Shen, J.J.; Meng, F.C.; Kong, J.P.; Zhang, Y. ESLD: Eyes segment and landmark detection in the wild. *J. Image Graph.* **2022**, *27*, 2329–2343.
53. Vitek, M.; Das, A.; Pourcenoux, Y.; Missler, A.; Paumier, C.; Das, S.; De Ghosh, I.; Lucio, D.R.; Zanlorensi, L.A.; Menotti, D.; et al. Ssbc 2020: Sclera segmentation benchmarking competition in the mobile environment. In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 28 September–1 October 2020; pp. 1–10.
54. Proença, H.; Filipe, S.; Santos, R.; Oliveira, J.; Alexandre, L.A. The UBIRIS. v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1529–1535. [CrossRef]
55. Teo, C.C.; Neo, H.F.; Michael, G.; Tee, C.; Sim, K. A robust iris segmentation with fuzzy supports. In Proceedings of the International Conference on Neural Information Processing: Neural Information Processing. Theory and Algorithms, Sydney, Australia, 21–25 November 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 532–539.

56. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
57. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.