Rethinking Dual-Stream Super-Resolution Semantic Learning in Medical Image Segmentation

Zhongxi Qiu, Yan Hu, Xiaoshan Chen, *Member, IEEE*, Dan Zeng, Qingyong Hu,
and Jiang Liu, *Senior Member, IEEE*

***Abstract*—Image segmentation is fundamental task for medical image analysis, whose accuracy is improved by the development of neural networks. However, the existing algorithms that achieve high-resolution performance require high-resolution input, resulting in substantial computational expenses and limiting their applicability in the medical field. Several studies have proposed dual-stream learning frameworks incorporating a super-resolution task as auxiliary. In this paper, we rethink these frameworks and reveal that the feature similarity between tasks is insufficient to constrain vessels or lesion segmentation in the medical field, due to their small proportion in the image. To address this issue, we propose a DS2F (Dual-Stream Shared Feature) framework, including a Shared Feature Extraction Module (SFEM). Specifically, we present Multi-Scale Cross Gate (MSCG) utilizing multi-scale features as a novel example of SFEM. Then we define a proxy task and proxy loss to enable the features focus on the targets based on the assumption that a limited set of shared features between tasks is helpful for their performance. Extensive experiments on six publicly available datasets across three different scenarios are conducted to verify the effectiveness of our framework. Furthermore, various ablation studies are conducted to demonstrate the significance of our DS2F.**

***Index Terms*—Dual-stream learning, medical image segmentation, shared feature, super-resolution.**

I. Introduction

MEDICAL image segmentation, which aims to automatically identify and delimit regions of interest (RoI) within medical images, is a widely applied technique for facilitating automatic diagnosis in the medical field. A deep neural networks can provide high-performance analysis, they have been obtaining increasing attention in a variety of medical demand, including vessel segmentation [1], [2], [3], lesion segmentation [4], [5], [6], tumor segmentation [7], [8], [9]. The use of high-resolution representation, which offer rich semantic and spatial details, is particularly desirable for boundary recognition and object localization [10], and is also a requirement for many deep-learning-based segmentation algorithms. However, in practical medical applications, limited by the computational ability of imaging capturing or operating devices, it is not often possible to obtain high-resolution segmentation results for diagnosis.

To achieve high-accuracy segmentation results despite limited computational resources, researchers have explored a variety of approaches. A typical way is to reduce the computational demands of the algorithms themselves, such as reducing input image size or lightweight models. The size of the input image is often reduced by downsampling [5], [6], [11] or patch splitting [12], [13], [14], which may cause segmentation results with low-resolution or with chessboard effect, which can negatively impact accuracy. The resolution of input images has been shown to have a significant impact on the accuracy of segmentation results generated by lightweight models [15], [16], [17].Low-resolution or noisy images usually exist in medical scenes, which results in unsatisfactory segmentation results.

Single segmentation networks alone may not be able to extract sufficient features from low-resolution input to achieve high accuracy. To address this issue, researchers have proposed utilizing a single image super-resolution (SISR) network as an auxiliary to enhance the resolution of the segmentation results [18], [19], [20], [21], which is expected to meet the demands of medical practitioners. Most of these approaches have aligning the segmentation features with the SISR branch through a feature transform module, and minimizing the distance between the features through a feature affinity [18], [19]. The shared decoder is explored to extract the shared features, whose similarity is constrained by the structural similarity loss. For example, the structure similarity loss is adopted to constrain the features [20], and the L1 regularization constraint is introduced in CogSeg to minimize the distance between the features of the decoders for the task [21].

The objectives of the super-resolution and semantic segmentation tasks are distinct, with the former aimed at producing high-resolution images, and the latter focused on identifying regions of interest within images. However, many existing

Manuscript received 3 February 2023; revised 26 September 2023; accepted 4 October 2023. Date of publication 9 October 2023; date of current version 5 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 82102189 and 82272086, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515012195, in part by Guangdong Provincial Key Laboratory under Grant 2020B121201001, and in part by Shenzhen Stable Support Plan Program under Grant 20220815111736001. Recommended for acceptance by P. Arbelaez *(Zhongxi Qiu and Yan Hu (co-first authors) contributed equally to this work). (Corresponding authors: Jiang Liu; Yan Hu.)*

Zhongxi Qiu, Yan Hu, Xiaoshan Chen, Dan Zeng, and Jiang Liu are with the Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: qiuzhongxi@163.com; huy3@sustech.edu.cn; 11910233@mail.sustech.edu.cn; zengd@sustech.edu.cn; liuj@sustech.edu.cn).

Qingyong Hu is with the Department of Computer Science, Oxford University, OX1 2JD Oxford, U.K. (e-mail: qingyong.hu@cs.ox.ac.uk).

We share our project at https://github.com/Qsingle/imed_vision.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3322735, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3322735

dual-stream algorithms often adopt feature similarity between super-resolution and semantic segmentation to constrain feature learning. Such mandatory feature similarity constraints can lead to suboptimal model optimization or collapse. Moreover, the proportion of vessels or lesions in medical images is often relatively small, making it difficult to effectively constrain the targets using similarity loss of the whole image features alone. Such algorithms cannot dig out the target-related shared features between the medical image semantic segmentation and the super-resolution task. In other words, the auxiliary super-resolution task is not effectively contributing to region of interest-related feature learning when only relying on features similarity constraints. To overcome this limitation, we rethink the dual-stream learning framework and find new ways to extract shared features related to the RoI.

As illustrated by Argyriou et al. [22], multiple tasks share a small set of features, which is also applicable to our dual-stream framework in the medical field. We believe that the more shared characteristics are related to the RoI, the higher the segmentation accuracy will be. The paper mainly considers how to focus these small number of shared features on our area of interest as much as possible. Specifically, we propose a novel high-resolution medical image semantic segmentation framework, named Dual-Stream Shared Feature (DS2F) framework, exploring the RoI-related shared features between segmentation and super-resolution. The DS2F framework consists of a semantic image segmentation network, a super-resolution network, and a shared feature extraction module (SFEM). We propose a novel feature extraction and supervision way in the SFEM.

Due to the small proportion of RoI in medical images, their corresponding features for segmentation are dispersed or sparse. Existing feature integration methods, such as concatenation or convolution 1×1, which treat all the features equally, are not effective in assigning higher importance to RoI features. We consider that there is spatial structure correspondence of features between segmentation and super-resolution tasks, such as vessels or lesion areas. Thus, we first propose a new way to bestow RoI features with higher weights based on the consideration of channel selection and spatial structure correspondence. Second, for the supervision ways, as it is too difficult to obtain the ground truth of share features between two tasks, we cannot adopt supervision ways to extract shared information in SFEM. As the shared features are supposed to improve the performance of both tasks, we propose to convert the supervision of shared information extraction into the problem of how to improve the performance of both tasks based on the shared information. Thus, we define a proxy task to extract the shared features in SFEM.

The DS2F framework presented in this paper is an extension and improvement of our previous works [23], [24]. First, we have enhanced the theoretical foundations of shared information between tasks. Second, we further generalize the structures of extracting shared information and define a proxy task. Third, we conduct experiments with different medical scenarios and verify the effectiveness of our DS2F framework on the cityscape dataset. Finally, the comprehensive ablation studies further prove the effectiveness of the structure design. Our contributions are summarized as follows:

1) We rethink the dual-stream segmentation and super-resolution framework and identified that the main limitation of existing dual-stream networks when applied to medical image segmentation is that the similarity loss of global features cannot effectively constrain the small proportion of RoI. Therefore, we propose a shared feature extraction method, which can focus on the region of interest as much as possible.
2) The proposed Dual-Stream Shared Feature (DS2F) framework incorporates a semantic segmentation branch, a super-resolution branch, and a shared feature extraction module (SFEM). For SFEM, we propose a novel feature extraction and supervision way. Specifically, we propose a new instance of SFEM, named multi-scale cross gate (MSCG), and a proxy loss for module constraint. The extracted features mainly focus on the RoI, such as the vessels and lesions.
3) The proposed DS2F framework has been evaluated on five publicly available datasets across two distinct medical scenarios. The results of an ablation study demonstrate the superiority of our proposed module. We implement the algorithm by Pytorch framework, which is publicly available at https://github.com/Qsingle/imed_vision

II. Related Work

Semantic Segmentation: As illustrated by Horwath et al. [25], high-resolution feature representation is critical for medical image segmentation. However, learning or retaining high-resolution feature representation is a challenging problem, and the collaboration of medical image segmentation increases the problem's difficulty. To learn high-resolution feature representation, researchers have explored different algorithms, such as atrous convolution [26], dense atrous convolution (DAC) blocks [1], and scale-aware feature aggregation (SFA) modules [2]. Moreover, attention mechanisms are also used to retain important features for representation. For example, CS-Net uses channel-and-spatial attention [25] for segmentation. However, to obtain high-resolution segmentation results, the existing methods are often computationally expensive in both training and test phases, which limits their applications to resource-constrained devices in the medical field.

To degrade the computation costs, researchers have also explored lightweight models. For example, SA-UNet [16] uses spatial attention to retain important information while reducing the number of filters. ESPNets [27], [28] adopts the reduce-split-transform-merge strategy, which accelerates the convolutional neural network and optimizes for the edge devices. Mobilenets [29], [30], [31] use the depthwise separable convolution and the inverted bottleneck to reduce the computational cost for the model. However, lightweight models with relatively low computation costs often provide limited segmentation performance as they may not obtain rich feature support. To achieve high-accuracy results without increasing computational cost, we propose a framework based on the dual-stream learning framework using low-resolution inputs. The framework takes a super-resolution stream as an auxiliary task, which provides extra features for the segmentation task during training and is deleted during the test, without increasing the computation costs of medical image segmentation.

Dual-Stream Super-Resolution Semantic Learning: Single-image super-resolution networks can extract high-resolution features outputting high-resolution results only based on low-resolution input. Based on the characteristic, researchers have proposed dual-stream super-resolution semantic learning frameworks to degrade the computation costs without decreasing the segmentation accuracy [18], [19], [20], [32], [33]. For natural scenes, these frameworks focus on the whole counterparts in images. For example, DSRL [18] adopts a feature affinity (FA) module to constraint the network to extract similar features from two tasks, and ColSeg [20] uses structural affinity block to constraint features from two streams. For the medical field, Yu et al. [19] proposed CogSeg for CT segmentation guided by super-resolution learning, in which the L1 losses of decoder layers from two tasks are adopted. Wang et al. [32], [33] adopted a spatial similarity matrix to constrain the features from two streams and a selective cropping strategy for guidance. We analyze such dual-stream frameworks and identify their limitations. First, from the task-specific level, the features extracted from different tasks cannot be strictly similar. Second, for medical segmentation regions (vessels or lesions), their proportions in the whole image are relatively small, so the features from the segmentation stream are supposed to be mostly different from those from the super-resolution stream. Thus, purely feature similarity between two streams cannot provide reasonable constraints on medical RoI segmentation. To solve these limitations, the shared features extracted by our proposed framework mainly focus on the medical region of interest. We also propose a novel supervision way to optimize the shared feature extraction.

III. Method

A. Problem Preliminary

For the dual-stream super-resolution learning of the semantic segmentation model, we adopt one shared encoder mapping the input x to the features F_{en}, and two task-specific decoders dealing with the features F_{en} to two task-dependent parts F_{seg} and F_{sr}. Then Seg Head (the task head for semantic segmentation) outputs the segmentation results O_{seg} based on features F_{seg}, and SR Head (the task head for super-resolution) provides the corresponding high-resolution image O_{sr} mapped by features F_{sr}. We define the process as follows:

$$F_{en} = \text{Encoder}(x) \tag{1}$$

$$F_{seg} = \text{Decoder}_{seg}(F_{en}) \tag{2}$$

$$F_{sr} = \text{Decoder}_{sr}(F_{en}) \tag{3}$$

$$O_{seg} = \text{Head}_{seg}(F_{seg}) \tag{4}$$

$$O_{sr} = \text{Head}_{sr}(F_{sr}) \tag{5}$$

where Encoder is the shared encoder, Decoder$_{seg}$ and Decoder$_{sr}$ are the decoders for segmentation and super-resolution respectively, Head$_{seg}$ and Head$_{sr}$ are the task head for segmentation and super-resolution separately.

Several papers [18], [19], [20] adopt a loss of feature similarity between two tasks for constraint. In other words, they try to minimize the distance between features F_{seg} and F_{sr}. This may produce a good performance for natural scenarios, whose proportions of ROIs are large. For medical images, the proportions of ROIs are often very small. The feature similarity loss of the whole image may not efficiently work for such medical image segmentation. We analyze the existing dual-stream learning framework applied in the blood vessel segmentation, as shown in Fig. 1. The model with feature similarity loss cannot extract sufficient features related to our target vessel in Fig. 1(b). Super-resolution network reconstructs amounts of high-frequency details from low-resolution input, including RoI and other areas. As the loss or model cannot constrain the dual-stream learning framework to extract RoI-related features, the auxiliary task does not work for medical image segmentation. As its targets are often tiny, the whole feature constraints cannot be helpful, which may lead to constraining in the wrong direction or collapsing. Therefore, we propose to fully adopt the information captured from different tasks instead of enforcing feature similarity constraints.

As there is no ground truth to supervise the extraction of shared features, how to oversee a module to extract these features becomes challenging. The features extracted by each task are formulated as

$$F_t = \sum_{i=0}^{d} F_i \tag{6}$$

where i is the index of task, d is the number of tasks, $F_t \in \mathbb{R}^d$ is the summation of task features, and F is the features from the corresponding task. For instance, in our paper, as the framework includes two tasks, d equals to 1, $i = 0, 1$, F_t is the concat of segmentation features and super-resolution features (as shown in Fig. 2(b)). As multiple tasks only share a small set of features, we expect that the small set of features is able to focus on the segmentation RoI as much as possible, providing detailed features for segmentation and super-resolution. In other words, the objective of our framework is to extract a small set of shared features from F_t to improve the performance of medical image segmentation. Thus, we propose a simple but efficient idea minimizing the distances of corresponding task results, which are generated by the shared features and task targets, formulated as:

$$\begin{aligned} &Min(\alpha(Head_{Seg}(A_{seg} \odot F_{seg}), target_{Seg}) \\ &+ \beta(Head_{SR}(A_{sr} \odot F_{sr}), target_{SR})) \end{aligned} \tag{7}$$

where α and β are the coefficients to adjust the weights for each task, $Head_{Seg}$ is the head to generate the segmentation results, $Head_{SR}$ is the head to generate the super-resolution results. A_{seg} and A_{sr} are the weights of the segmentation and super-resolution features, respectively. F_{seg} and F_{sr} are the segmentation and super-resolution features, respectively.

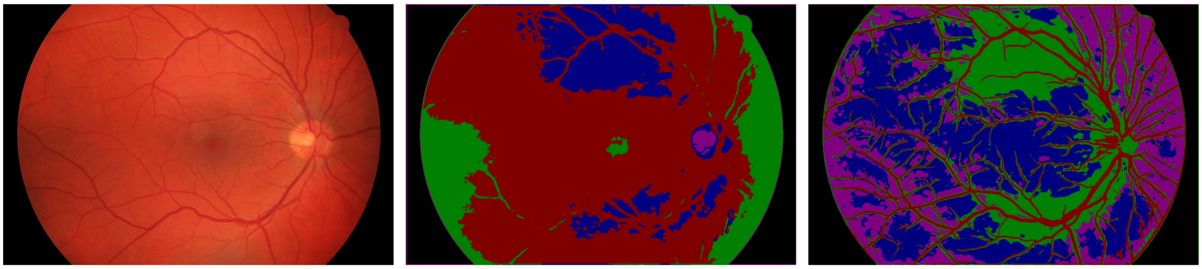
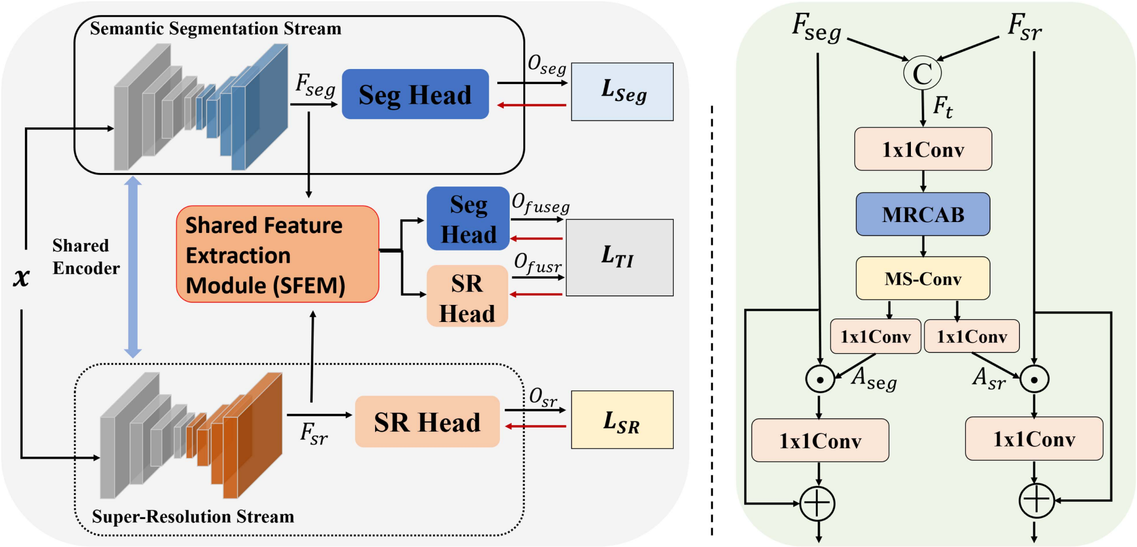

(a) (b) (c)

Fig. 1. Visualization of the features from segmentation branches. KMeans are used to produce the results, in which the number of clusters is set as 5. (a) The input image; (b) The extracted segmentation features by DSRL [18]; (c) The extracted segmentation features by our DS2F framework. The rich features in Fig. (c) are helpful to discriminate the vessel structure. The vessel features in Fig. (b) are destroyed, and the vessels are hardly classified around the macular and optic. Our DS2F framework mainly focuses on the features of RoI, such as vessels.



(a) Framework (b) MSCG

Fig. 2. Pipeline of our medical image dual-stream framework. (a) The pipeline of our Dual-Stream Shared (DS2F) framework. (b) The structure of the multi-scale cross gate (MSCG), which is a novel instance of SFEM. The © is concatenation operation, $\odot$ is the Hadamard product, and $\oplus$ is element-wise addition.

$target_{Seg}$ and $target_{SR}$ are the targets of segmentation and super-resolution tasks. As shown in Fig. 1(c), our framework extracts more features focusing on the vessels, which are supposed to be the shared target between vessel segmentation and super-resolution task.

We have explored several existing feature interaction operations, which can improve the results for the shared feature extraction module (SFEM) in our DS2F framework. For example, the previously proposed modules [23], [24] can increase the segmentation accuracy of the blood vessels and lesions, respectively. A simple feature intersection is also efficient, and we will explain and prove it in the following section. Moreover, we propose another efficient feature extraction module, which can be suitable for various medical scenarios.

B. Dual-Stream Shared Feature (DS2F) Framework

As shown in the Fig. 2(a), a down-sampled image x with size $W/n \times H/n$ is fed into the shared encoder, which produces the encoded features F_{en}, formulated as (1). The segmentation decoder Decoder$_{Seg}$ and super-resolution decoder Decoder$_{sr}$ deal with the F_{en} to output decoded features F_{seg} and F_{sr}, formulated as (2) and (3). As formulated by (4) and (5), $SegHead$ and $SRHead$ deal with the decoded features to output the final targets, segmented target O_{seg} and high-resolution images O_{sr}, respectively. The losses $\mathcal{L}_{Seg}$ and $\mathcal{L}_{SR}$ are adopted to constrain the semantic segmentation and super-resolution streams, respectively. Then we propose a Shared Feature Extraction module (SFEM) to extract our defined small set of shared features between two tasks and use the task interaction loss $\mathcal{L}_{TI}$ to constraint the module learning.

In our DS2F framework, the objective of SFEM is to extract the shared features between two tasks. It first integrates information from two tasks, then extracts shared features based on the attention weights. The reweighted features are input into Seg Head and SR Head to generate the corresponding results for proxy tasks. The proposed task interaction loss $\mathcal{L}_{TI}$ constraints the SFEM learning. The components of the SFEM can be various. For example, simply one 1×1 Conv and the existing channel or other attention mechanisms can be used

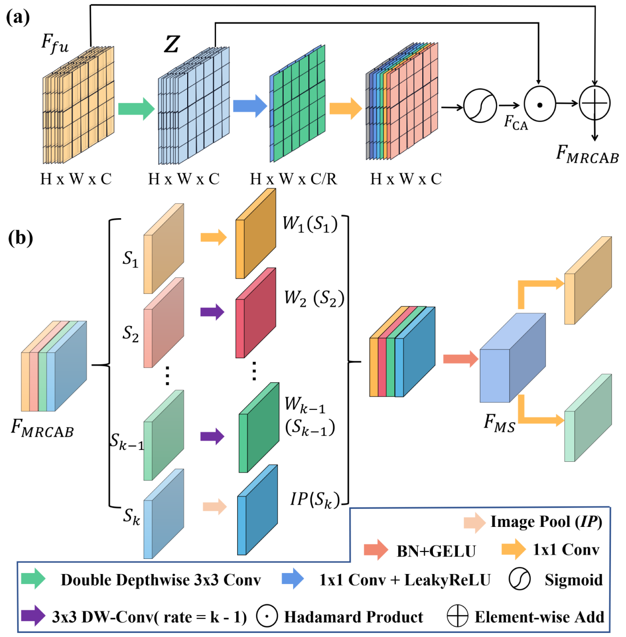


Fig. 3. Our proposed modules. (a) The structure of our proposed Modified Residual Channel Attention Block (MRCAB); (b) The structure of Multi-Scale Convolution (MS-Conv).

to integrate the features from two streams. 1 × 1 Conv can be directly applied to generate the attention weights based on the fused features. Our previously proposed modules [23], [24] also can be suitable for specific medical image segmentation. Here, we present another novel instance to extract the shared features named the Multi-Scale Cross Gate (MSCG) module.

C. Multi-Scale Cross Gate (MSCG) Module

We propose a Multi-Scale Cross Gate (MSCG) module as a new example of the SFEM. Its construction is shown in Fig. 2(b). We utilize one 1×1 Conv to fuse features F_{fu}, our modified residual channel attention block (MRCAB), and a proposed multi-scale convolution (MS-Conv) to integrate the features from two tasks. Then we merely use two 1×1 Conv to generate the attention weights for two tasks. We will introduce the details of our proposed MRCAB and MS-Conv in the following.

Modified Residual Channel Attention Block (MRCAB) As the low-resolution input images contain lots of redundant information, we propose a modified residual channel attention block (MRCAB) to focus on more specific components related to the necessary small set of shared features, which is inspired by RCAB [34]. As shown in Fig. 3(a), we adopt two 3×3 depthwise convolutions to capture the local pattern from the integrated features. Instead of a global average to statistic the global spatial information, we propose to use two linear layers to directly squeeze and expand the features, capturing the channel relationship. Then a Sigmoid function generates the weights along the channel axis based on the above features. The weights re-weight the above local pattern by Hadamard product. Finally, the input F_{fu} is added to the re-weighted features. The process is formulated as:

$$F_{fu} = W(C(F_{seg}, F_{sr})) \tag{8}$$

$$z = Conv_{dw}(Conv_{dw}(F_{fu})) \tag{9}$$

$$F_{CA} = \sigma(W(\delta(W(z)))) \odot z \tag{10}$$

$$F_{MRCAB} = F_{fu} + F_{CA} \tag{11}$$

where F_{fu} is the input features, C is the concatenate operation, $Conv_{dw}$ is the depthwise convolution, δ and σ are LeakyReLU and Sigmoid, respectively, W is the weights of Conv 1×1. The proposed MRCAB further fuses the decoded features, and its channel-wise statistics enhance the discriminative ability of features from different tasks. We adopt the depthwise convolution in the MRCAB to reduce the computation, which also mixes the information in spatial space.

Multi-Scale Convolution (MS-Conv) As discussed above, the features F_{MRCAB} dig out the helpful information that improves the performance of both segmentation and super-resolution tasks. We propose one Multi-Scale Convolution (MS-Conv) that is a multi-scale strategy using the information from different scales. Its structure is shown in Fig. 3(b). We divide the features F_{MRCAB} into k groups according to the channel size. Then we adopt different processing ways for various groups of information. A 1×1 Conv is applied to the first group to keep the current scale. For the second to $(k-1)th$ group, we adopt 3×3 depthwise convolution with different dilation rates to capture various levels of information, in which the dilation rates are set as the group index minus 1. For the kth group, image pooling [35] is adopted to statistic the global spatial information. Batch normalization is used to integrate the information of the groups' combinations. Finally, GELU is used as the nonlinear activation function for the layer. The MRCAB extracts the most significant shared features related to our RoI.

Based on the above description, the procedure of the MS-Conv is formulated as:

$$F_{MS} = \delta(N(C(W_1(S_1), W_2(S_2), \ldots, W_{k-1}(S_{k-1}), IP(S_k)) \tag{12}$$

where F_{MS} is the features extracted by our proposed MS-Conv, S is one group of features (including $S_1, S_2, \ldots, S_{k-1}, S_k$), which are split from F_{MRCAB} based on the channel. k is the index of the group. $W_1, W_2, \ldots, W_{k-1}$ are the weights of the convolutional layer for every features in the group. C is the concatenation operation, IP is the image pool operation, N and δ are the normalization operation and GELU activation function respectively.

In the MSCG module, we use two 1×1 Conv mapping the output of our MS-Conv to the space of segmentation decoded features F_{seg} and super-resolution decoded features F_{sr}, respectively. Then, we use the Hadamard product to separately combine the outputs with F_{seg} and F_{sr}. After processing by 1×1 Conv, the results are added with F_{seg} and F_{sr}. Then we can obtain the re-scaled decoded features for segmentation and super-resolution tasks. Finally, the outputs of segmentation and super-resolution tasks O_{fuseg} and O_{fusr} are processed by task heads Head$_{seg}$ and Head$_{sr}$, respectively. The process is formulated as:

$$O_{fuseg} = \mathrm{Head}_{seg}(\delta(W_2(\delta(W_1(F_{MS})) \odot F_{seg}))) \quad (13)$$

$$O_{fusr} = \mathrm{Head}_{sr}(\delta(W_2(\delta(W_1(F_{MS})) \odot F_{sr}))) \quad (14)$$

where δ is GELU, W_1 and W_2 is the weights for two 1×1 Conv as shown in MSCG module (Fig. 2(b)).

D. Objective Function

As shown in Fig. 2, our objective function of the DS2F framework includes three parts: $\mathcal{L}_{Seg}$ for segmentation task, $\mathcal{L}_{SR}$ for super-resolution task, and $\mathcal{L}_{TI}$ for the task interaction to constrain our shared feature extraction module. We introduce them one by one in the following.

For the segmentation task, we employ a common cross-entropy loss function, formulated as:

$$\mathcal{L}_{Seg} = \frac{1}{C} \sum_{i=0}^{C} -y_i \log(z_i) \quad (15)$$

where C represents the number of classes, y_i is the ground truth of class i, and z is the softmax result of the output O_{Seg}. To simplify the implementation of the code, we take the binary segmentation task as the segmentation task of two classes, the target RoI and the background.

We employ a mean square error (MSE) function for super-resolution task, described as:

$$\mathcal{L}_{SR} = \frac{1}{N} \sum_{i=0}^{N} (O_{SR_i} - HR_i)^2 \quad (16)$$

where N is the number of pixels of the image, O_{SR} is the output of the super-resolution task, HR is the high-resolution target image, and i represents the index of the pixel.

The interaction part plays a vital role in extracting shared features in our DS2F framework, whose objective function is one of our major concerns. As there is no exact definition or ground truth for the shared features, we cannot directly adopt supervised feature extraction. Here, we propose a proxy-loss way to get the supervision implicit. As illustrated, the small set of shared features is supposed to improve the performance of both tasks. Based on this property, we deduce that the segmentation or super-resolution results predicted by the combination with shared features should be better than those only by single-task features. Thus, we propose to use the following objective function as one proxy objective of our DS2F framework, so that the shared features are mined implicitly. The formulation of $\mathcal{L}_{TI}$ for our SFEM is defined as:

$$\mathcal{L}_{TI} = \mathcal{L}_{ProxySeg}(O_{fuseg}, Y) + \mathcal{L}_{ProxySR}(O_{fusr}, HR) \quad (17)$$

where Y is the ground truth of the segmentation task, and HR is the target high-resolution image.

For the proxy-task losses, there are two strategies, the same as the other two streams (such as cross-entropy or MSE), or higher strength of constraint. The former can improve the results but may not explore the enormous set of shared features caused by the same constraint strength. The latter often gains better supervision to explore the shared features. We will explore and discuss this in the experiment Section IV-B1.

TABLE I
THE ILLUSTRATION OF DATASETS FOR THE EXPERIMENTS

Dataset	Task	Size	Proportion(%)
HRF [36]	Vessel	45	7.71
PRIME-FP20 [37]	Vessel	15	2.54
FIVES [38]	Vessel	800	7.46
IDRID [39]	Lesion	81	2.09
DDR [40]	Lesion	757	0.79
Cityscapes [41]	Cityscape	5000	97.38

Proportion represents the percentage of RoI areas in the whole dataset.

IV. EXPERIMENTS

A. Experiments Settings

1) Datasets: We conduct our experiments on six publicly available datasets, including three for retinal vessel segmentation (two different image modalities), two for retinal lesion segmentation (multiple targets), and one for cityscape segmentation. As listed in Table I, the proportions of vessels or lesions in the medical images are remarkably small, less than 8%, such as the lesion only accounts for 0.79% in the DDR dataset. The proportion in the Cityscapes dataset is about 97.38%, much higher than that in medical image datasets.

HRF: The HRF (High-Resolution Fundus) [36] dataset includes 45 fundus images with the size of 3504×2336 in total, of which 15 are from healthy patients, 15 are from patients with glaucoma, and 15 have Diabetic Retinopathy (DR). We conduct the five-fold cross-validation experiments at this dataset and set 1752×1168 as the target resolution for the super-resolution task. We set the batch size and epoch number as 2 and 300, respectively.

PRIME-FP20: It provides 15 high-resolution ultra-widefield (UWF) fundus photography (FP) images using Optos 200Tx camera, and their resolution is 4000×4000. We use the official mask to remove the invalid area in the images, then the minimal and max height for the images are 2444 and 2631, and the minimal and max-width for the images are 2817 and 2932. The five-fold cross-validation is applied to this dataset. Considering the computation ability of our devices, the output size for the super-resolution task is set as 1408×1296. We set the batch size and epoch number as 2 and 300, respectively.

FIVES: The FIVES [38] (Fundus Image Vessel Segmentation) dataset consists of 800 high-resolution (2048×2048) multi-disease color fundus with acceptable vessel pixel annotation. The dataset contains train (600 images) and test (200 images) sets. We set the batch size and epoch number as 4 and 128, respectively.

IDRID: The IDRiD (India Diabetic Retinopathy Image Dataset) [39] dataset consists of 81 color fundus images with adequate pixel-level annotation of four types of retinal lesions: microaneurysms(MA), soft exudates(SE), hard exudates(EX), and hemorrhages(HE). The dataset is split into a training set with 54 images and a testing set with 27 images. We run the training for 300 epochs with batch size 2.

DDR: The DDR [40] is another dataset applied for lesion segmentation. It provides 757 color fundus images with acceptable pixel-level annotation. The images are split into three sets for training, validation, and testing with a ratio of 5:2:3. We set 1024×1024 as the target resolution for the super-resolution task. On this dataset, we set the batch size to 2 and the training epoch to 128.

Cityscapes: The Cityscapes [41] dataset consists of 5000 images with fine-grained annotation for urban visual scene understanding. There are 2975 images for training, 500 for validation, and 1525 for testing. The images are collected from 50 cities in different seasons with image size 2048×1024. Following previous works, we also do the 19 categories of segmentation. We set the batch size to 4 and the training epoch to 108. We resize the image to 1024×512 as the model input and set the upscale rate to 2.

2) Evaluation Metrics: For the vessel segmentation task, we adopt the intersection over union (IoU), bookmaker informedness (BM), Matthews correlation coefficient (MCC), and dice score (Dice) to evaluate the performance of the models. For the lesion segmentation task, IoU, precision and recall area under the curve (PR-AUC), and dice score are adopted to evaluate the performance of the model, due to the MCC and BM are not suitable for the evaluation of the multi-categories classification task. We use the most common evaluation metric mean intersection over union (mIoU) for the cityscape segmentation task. The formulations are as follows:

$$IoU = \frac{TP}{TP + FN + FP} \tag{18}$$

$$BM = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \tag{19}$$

$$Dice = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{20}$$

where $TP, TN, FP,$ and FN are the true positive, true negative, false positive, and false negative respectively. More details are in Appendix C, available online.

3) Implementation Details: We implement the models by Pytorch [42] framework, and all experiments are run on the machine with one NVIDIA RTX A6000 graphics card. The mini-batch stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0001 is applied to optimize the model. Poly learning rate adjusts strategy [43] is adopted to set the learning rate dynamically during training, which sets the learning rate according to $lr = init\_lr \times (1 - \frac{iter}{max\_iter})^{power}$, and we set $init\_lr = 0.01, power = 0.9$.

B. Ablation Study

1) Ablation for Proxy Strategy: In this section, we explore two aspects of the proxy strategy, 1) what kinds of losses would be better? 2) do we need to take two sub-tasks for the proxy task? To prove the effectiveness, we conduct the ablation experiments on two modalities of vessel segmentation datasets, including the HRF and PRIME-FP20 datasets. We set the upscale rate as 2. That is to say, the input size for the two datasets is 876×584 and 704×648 respectively, and the output size is 1752×1168 and 1408×1296.

TABLE II
EXPERIMENT RESULTS OF DIFFERENT STRATEGIES FOR THE PROXY LOSS BASED ON HRF AND PRIME-FP20 DATASETS (MEAN ± STD)

Dataset	Loss		IoU(%)	MCC(%)	BM(%)
	SR	Seg			
HRF	N/A	N/A	61.12 ± 2.07	74.21 ± 1.57	69.64 ± 1.99
	MSE	CE	68.62 ± 2.89	80.17 ± 1.93	75.36 ± 3.75
	SSIM	CE	69.66 ± 2.04	80.86 ± 1.41	76.82 ± 2.10
	SSIM	GDice	69.09 ± 2.08	80.51 ± 1.39	75.72 ± 2.39
	SSIM	RMI	$\mathbf{71.26} \pm 1.59$	$\mathbf{81.88} \pm 1.12$	$\mathbf{80.07} \pm 1.12$
PRIME-FP20	N/A	N/A	26.72 ± 3.91	41.78 ± 4.69	34.26 ± 5.20
	MSE	CE	36.15 ± 5.02	52.91 ± 5.26	44.10 ± 5.84
	SSIM	CE	36.98 ± 4.40	53.76 ± 4.39	45.32 ± 5.42
	SSIM	GDice	35.41 ± 5.01	52.18 ± 5.06	43.30 ± 6.31
	SSIM	RMI	$\mathbf{41.16} \pm 3.41$	$\mathbf{57.58} \pm 3.38$	$\mathbf{52.37} \pm 4.12$

Higher strength of constraint produces better results.

TABLE III
RESULTS FOR ABLATION STUDY OF PRETEXT LOSS FUNCTIONS OF OUR PROPOSED MODEL ON HRF AND PRIME-FP20 DATASETS (MEAN ± STD)

Dataset	Loss		IoU(%)	MCC(%)	BM(%)
	SR	Seg			
HRF	x	x	61.12 ± 2.07	74.21 ± 1.57	69.64 ± 1.99
	✓	x	68.79 ± 1.92	80.33 ± 1.36	75.12 ± 1.73
	x	✓	70.56 ± 1.98	81.35 ± 1.39	79.73 ± 1.78
	✓*	✓	$\mathbf{71.26} \pm 1.59$	$\mathbf{81.88} \pm 1.12$	$\mathbf{80.07} \pm 1.12$
PRIME-FP20	x	x	26.72 ± 3.91	41.78 ± 4.69	34.26 ± 5.20
	✓	x	35.14 ± 5.07	51.88 ± 5.16	43.00 ± 6.62
	x	✓	40.98 ± 3.24	57.48 ± 3.19	51.56 ± 3.94
	✓	✓	$\mathbf{41.16} \pm 3.41$	$\mathbf{57.58} \pm 3.38$	$\mathbf{52.37} \pm 4.12$

The * represents a p-value< 0.05, which indicates significantly different results are obtained. More details are shown in appendix D.1, available online.

For the first aspect, we explore two different strategies for the proxy loss to make an implicate supervision: 1) the same loss as that of sub-tasks; 2) more restrictive losses, such as region-based loss. For the second strategy, we use the structure similarity (SSIM) loss, region mutual information (RMI) [44] loss or generalized dice loss (GDice) for task interaction module. As described in Table II, both strategies can significantly improve the performance of the model, compared with the baseline model U-Net [45] without proxy tasks. For our first strategy, when MSE and CE are used as the losses of the proxy task, the IoUs of HRF and PRIME-FP20 are 7.5% and 9.5% higher. After using more restrictive losses, SSIM and RMI, for the proxy task, the improvement of IoUs on HRF and PRIME-FP20 datasets are more than 10% and 14%. The evaluation metrics of MCC and BM see the same rise. The view field of PRIME-FP20 is ultra-wide, whose vessel proportion is only 2.54% relatively small compared with other vessel segmentation datasets, so it can be supposed that the GDice loss may excessively punish the background. Thus, we adopt SSIM and RMI for the proxy loss in the following experiments.

For the second aspect, we use one sub-task loss or two sub-task losses for proxy loss. As shown in Table III, x means without using the corresponding proxy loss, ✓means using the corresponding proxy loss. Compared with the first rows based on two datasets, adding one or two losses to constrain the task interaction significantly improves the performance. Adding

TABLE IV
ABLATION STUDY FOR THE MODIFIED RESIDUAL CHANNEL ATTENTION BLOCK BASED ON HRF DATASET (MEAN ± STD), WE ADOPT THE RCAB TO REPLACE OUR MRCAB

Model	IoU(%)	MCC(%)	BM(%)
RCAB	71.12 ± 1.72	81.75 ± 1.24	$\mathbf{80.31} \pm 0.95$
Ours(w/ gpool)	70.76 ± 1.31	81.60 ± 0.92	78.40 ± 0.94
Ours(w/o gpool)*	$\mathbf{71.26} \pm 1.59$	$\mathbf{81.88} \pm 1.12$	80.07 ± 1.12

The gpool means global average pooling. The * represents a p-value< 0.05, which indicates significantly different results are obtained. More details are shown in appendix D.2, available online.

TABLE V
ABLATION STUDY FOR MULTI-SCALE FEATURE EXTRACTION MODULES BASED ON HRF DATASET (MEAN ± STD)

Module	IoU(%)	MCC(%)	BM(%)
ASPP	70.97 ± 1.63	81.66 ± 1.12	$\mathbf{80.13} \pm 2.08$
Self-Att	70.06 ± 2.17	81.12 ± 1.44	77.71 ± 2.96
MS-Conv(Ours)*	$\mathbf{71.26} \pm 1.59$	$\mathbf{81.88} \pm 1.12$	80.07 ± 1.12

Atrous spatial pyramid pooling (ASPP), self-attention module (self-att) are used to compare with our multi-spatial convolution (ms-conv). The * represents a p-value< 0.05, which indicates significantly different results are obtained. More details are shown in appendix D.2, available online.

TABLE VI
EXPERIMENTS OF THE MS-CONV COMPONENTS

	Counterpart				
1 × 1 Conv	Image pooling	DW	IoU(%)	MCC(%)	BM(%)
x	x	✓	71.00 ± 2.20	81.67 ± 1.59	79.94 ± 1.50
✓	x	✓	70.63 ± 2.18	81.44 ± 1.47	79.41 ± 3.01
x	✓	✓	71.06 ± 1.98	81.75 ± 1.33	79.77 ± 2.53
✓	✓	x	70.85 ± 1.72	81.60 ± 1.23	79.33 ± 1.05
✓	✓	✓	$\mathbf{71.26} \pm 1.59$	$\mathbf{81.88} \pm 1.12$	$\mathbf{80.07} \pm 1.12$

We gradually evaluate their affects in the MS-conv layer (mean ± std).

SSIM for proxy loss improves the segmentation performance, as the segmentation task can learn more information from the backward of the feature interaction module. The IOU after adding the RMI loss is more than 9.5% or 14.2% higher than that without proxy loss or just adding the SSIM loss, as the shared feature extraction module can be guided to extract more features focusing on the segmentation target, which intuitively achieves better performance. Moreover, applying proxy loss for both segmentation and super-resolution tasks further improves the segmentation performance, BM with 18% higher on the PRIME-FP20 dataset. Thus, the proxy task loss for our proposed shared feature extraction module improves the segmentation performance. More statistical analysis is added in Appendix D, available online.

2) Ablation Study for the Proposed MSCG: The main components of the MSCG are our proposed MRCAB and MS-Conv. The ablation studies about the two elements are analyzed in this section.

MRCAB: To verify the validation of our MRCAB, we use the original RCAB to replace it and do the experiments on the HRF dataset, as shown in Table IV. Our module can get a slight performance improvement, which means our module may extract the channel relation better than the original structure. We also conduct the experiments that add the global average pooling to the block to statistic the global information, as shown in the second line of Table IV. When adding the global pooling, the performance dropped, which means only capturing the channel relationship by the squeeze and excitation operation is enough when using the depthwise convolution. The metric BM is slightly dropped when using our method. The possible reason is that our algorithm may trend to classify some background to the foreground target, but from the MCC we can see that our approach gets more accurate results compared to RCAB.

MS-Conv: The proposed MS-Conv extracts the spatial correlation of the features by utilizing the multiple-scale context. Thus, we analyze two aspects of the MS-Conv, including its multi-scale feature extraction ability and the effectiveness of depthwise convolution. *Multi-scale feature extraction ability:* To verify the effectiveness of our MS-Conv, we use two other similar modules to replace the MS-Conv in MSCG, including ASPP [46], [47] and self-attention module that extracts spatial information (abbreviated as Self-Att). As shown in Table V, our MS-Conv outperforms the ASPP and Self-Att on the HRF dataset, and its standard deviation values are relatively smaller. That is to say, our MS-Conv extracts multi-scale features helpful for target segmentation with relatively higher stability.

Effectiveness of Components: MS-Conv adopts 1 × 1 Conv, image pooling, and depthwise convolution to extract multi-scale features from integrated information. We analyze their effectiveness one by one. As shown in Table VI, DW is short for depthwise convolution, and we analyze MS-Conv with/without DW. MS-Conv with depthwise gives out a better performance than only using convolution without DW. As the MS-Conv adopts 1 × 1 Conv and image pooling to extract features from different groups, we demonstrate their necessity by using depthwise convolution substitution. The results tell that MS-Conv without 1 × 1 Conv or image pooling degrades the segmentation performance, and their standard deviation values increase, which means they effects the performance of stability. Thus, our MS-Conv, using 1 × 1 Conv, image pooling, and depthwise convolution, is helpful to dig multi-scale features to improve the accuracy and stability performance of segmentation.

3) Ablation Study for the Framework: Our framework consists a semantic segmentation stream, a shared feature extraction module (MSCG as an instance), and a super-resolution stream. U-Net is adopted as the baseline of segmentation. We compare U-Net, U-Net with extra interpolation (U-Net+Inter), U-Net+Inter with super-resolution (U-Net+Inter+SR), and our proposed framework (U-Net+Inter+SR+MSCG). For the segmentation stream, we adopt an extra interpolate operation to produce the same size of the output as that of the high-resolution target. We conduct the ablation study on these components and different scales. The ablation study is based on HRF and PRIME-FP20 datasets, whose output resolutions are set as 1752 × 1168 and 1408 × 1296, respectively. In the experiments, different upscale rates are conducted. That is to say, the input size equals $W/upscale\_rate \times H/upscale\_rate$ (W and H are output resolutions). As shown in Fig. 4, IoU, BM, and MCC of different upscale rates by different combinations. Only adding interpolation cannot bring a gain for the segmentation performance

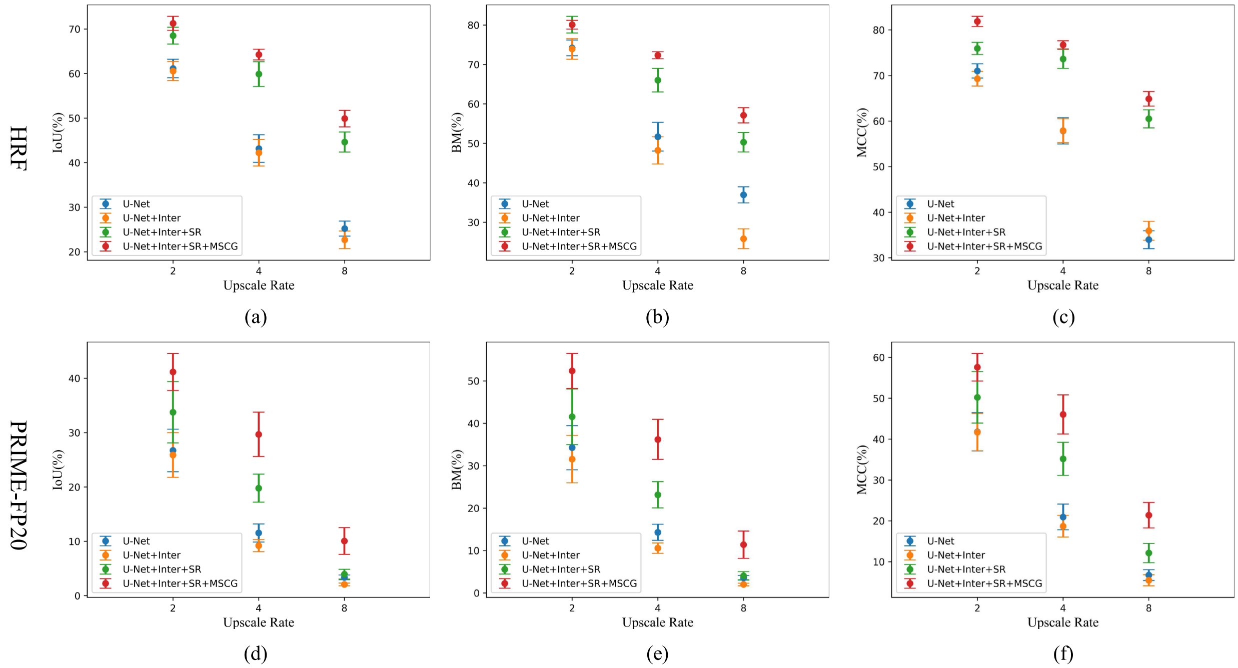


Fig. 4. Ablation study of framework components with different upscale rates based on HRF and PRIME-FP20. The first column is the trends for the IoU, second column is the trends for BM, and the last column is the results for MCC. (a) The trend of the IoU on HRF dataset. (b) The trend of the BM on HRF dataset; (c) The trend of the MCC on HRF dataset; (d) The trend of the IoU on PRIME-FP20 dataset; (e) The trend of the BM on PRIME-FP20 dataset; (f) The trend of the MCC on PRIME-FP20 dataset.

or even cause a drop, as there is no extra useful information for segmentation. The segmentation performance after adding a super-resolution stream improves, as super-resolution can provide some information for target segmentation. After adding our MSCG, the segmentation accuracy further improves, as the shared features extracted by our MSCG are helpful for the target segmentation. Moreover, the higher upscale rates produce larger improvements in segmentation. The standard deviations for the PRIME-FP20 datasets seem slightly large, limited by the small number of images in the dataset with only 15 images.

C. Comparison Experiments

To evaluate the effectiveness of our framework, we conduct comparison experiments based on 6 datasets in three different scenarios, including vessel segmentation, lesion segmentation, and natural image segmentation.

1) Vessel Segmentation Task: We employ U-Net [45] as the base model to build our framework. We compare our method with other 9 state-of-the-art methods including 6 single-dual segmentation methods (U-Net, SCS-Net [2], SA-UNet [16], DE-DCGCN-EE [48], SkelCon [3], and Little W-Net [15]) and 3 dual-stream learning methods (SuperVessel [23], CogSeg [21] and SS-MAF [24]). The experiments are based on three datasets, including HRF, PRIME-FP20, and FIVES. We conduct the experiments five times, and the results are listed with mean ± std of metrics Dice, IoU, MCC, and BM. We also list the floating-point operations per second (FLOPs) to compare the computation cost.

As shown in Table VII, a dual-stream learning framework with a feature interaction module produces the best segmentation accuracy for three datasets. Among them, the SuperVessel and SS-MAF are also proposed by our group according to this idea. Compared with single-stream segmentation algorithms, all the dual-stream learning frameworks improve the segmentation accuracy greatly, for example, the IoU of our framework is about 10% higher on HRF, 15% higher on PRIME-FP20 and 12% higher on FIVES than that of U-Net. Compared with other dual-stream learning frameworks, ours provides higher accuracy and lower standard deviation, which means that our algorithm can be more stable. For the PRIME-FP20 dataset, the image number is very small with only 15 images, and the IoU of CogSeg is only about 26%, about 15% lower than ours, which illustrates that the performance of CogSeg is affected by the size of the dataset, but our framework can overcome this problem to some extent.

The qualitative results of the three datasets are shown in Fig. 5. We can observe that our framework segments the vessels more accurately, and precisely locate the vessel edges. Compared with the single-stream segmentation methods (U-Net, SCSNet, SA-UNet, and DE-DCGCN-EE), dual-stream frameworks obtain more accurate and smooth boundaries. But the methods like CogSeg, which optimizes the similarity distance between segmentation features and super-resolution features, opt to misclassify the vessels, especially for the tiny vessels. Shared feature extraction integrated dual-stream frameworks (SuperVessel, SS-MAF, and our proposed MSCG-integrated framework) segment the vessel edge more accurately and alleviate the misclassification problem caused by vessel similarity. As shown in Fig. 5(b), our framework discriminates the vessels better. For example, in the PRIME-FP20 dataset with large view field images, the proportion of vessels is extremely small, and our framework segments tiny vessels precisely.

TABLE VII
COMPARISON RESULTS FOR VESSEEL SEGMENTATION TASK

Dataset	Model	Dice	IoU	MCC	BM	GFLOPs
HRF	U-Net	75.85 ± 1.57	61.12 ± 2.07	74.21 ± 1.57	69.64 ± 1.99	384.49
	SCSNet	74.15 ± 1.77	58.95 ± 2.28	72.44 ± 1.79	67.39 ± 1.87	342.22
	SA-UNet	75.73 ± 1.78	60.98 ± 2.35	73.91 ± 1.77	71.11 ± 2.58	51.15
	DE-DCGCN-EE	71.55 ± 2.25	55.75 ± 2.78	69.56 ± 2.13	65.48 ± 3.50	721.31
	SkelCon	-	-	79.15	-	-
	Little W-Net	81.03	-	79.09	-	-
	CogSeg	80.14 ± 1.07	66.87 ± 1.48	79.07 ± 1.03	72.45 ± 1.82	384.49
	SuperVessel	81.62 ± 1.49	68.98 ± 2.13	80.42 ± 1.43	75.70 ± 2.49	388.23
	SS-MAF	82.87 ± 0.90	70.76 ± 1.31	81.60 ± 0.92	78.40 ± 0.94	384.49
	Ours*	**83.21** ± 1.08	**71.26** ± 1.59	**81.88** ± 1.12	**80.07** ± 1.12	384.49
PRIME-FP20	U-Net	42.05 ± 4.94	26.72 ± 3.91	41.78 ± 4.69	34.26 ± 5.20	346.23
	SCSNet	55.97 ± 2.85	38.92 ± 2.75	55.17 ± 2.79	50.69 ± 3.37	342.22
	SA-UNet	45.60 ± 20.13	31.43 ± 14.54	43.81 ± 22.09	41.57 ± 21.65	51.15
	DE-DCGCN-EE	49.45 ± 2.71	32.89 ± 2.41	48.65 ± 2.67	43.33 ± 2.76	721.31
	CogSeg	41.55 ± 2.36	26.25 ± 1.87	42.90 ± 2.06	32.46 ± 4.99	346.23
	SuperVessel	52.74 ± 4.47	35.94 ± 4.11	52.77 ± 4.36	43.78 ± 4.63	345.67
	SS-MAF	57.86 ± 2.82	40.77 ± 2.80	57.45 ± 2.52	50.78 ± 4.27	346.23
	Ours*	**58.25** ± 3.40	**41.16** ± 3.41	**57.58** ± 3.38	**52.37** ± 4.12	346.23
FIVES	U-Net	83.86 ± 0.08	72.21 ± 0.11	82.66 ± 0.08	81.40 ± 0.28	196.54
	SCSNet	82.77 ± 0.15	70.61 ± 0.21	81.50 ± 0.17	80.03 ± 0.23	85.55
	SA-UNet	78.83 ± 0.85	65.06 ± 1.16	77.27 ± 0.84	81.37 ± 1.05	12.79
	DE-DCGCN-EE	83.61 ± 0.07	71.84 ± 0.11	82.42 ± 0.07	80.54 ± 0.29	295.45
	CogSeg	86.68 ± 1.05	76.51 ± 1.63	85.73 ± 1.09	83.55 ± 1.72	198.77
	SuperVessel	**92.10** ± 0.04	**85.36** ± 0.07	**91.52** ± 0.04	90.02 ± 0.10	198.46
	SS-MAF	92.07 ± 0.05	85.30 ± 0.08	91.47 ± 0.05	**90.59** ± 0.13	198.77
	Ours*	91.83 ± 0.03	84.90 ± 0.04	91.22 ± 0.03	90.15 ± 0.14	198.77

On HRF and PRIME-FP20 datasets, the input images size of SCSNet, SA-UNet and DE-DCGCN-EE are set as 1024 × 1024, 1024 × 1024 and 800 × 800, respectively (according to the original paper). On fives dataset, glops are calculated based on the input size of 512 × 512. The $*$ represents a p-value$<$ 0.05, which indicates significantly different results are obtained. More details are shown in appendix D.3, available online.

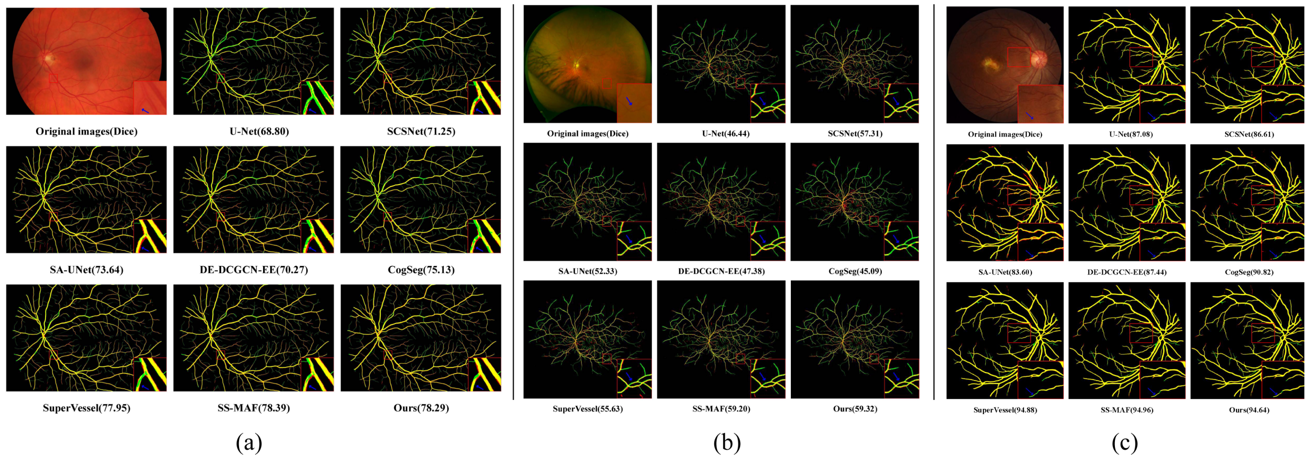


Fig. 5. Visualization results of our proposed method and other state-of-the-art methods on HRF, PRIME-FP20, and FIVES datasets. Green and red markings denote ground truth and segmentation output, respectively. The yellow marking represents the correct prediction of the retinal vessel. (a) Visualization of the samples on the HRF dataset. (b) Visualization of the examples on the PRIME-FP20 dataset. (c) Visualization of the examples on the FIVES dataset.(**Please zoom in for a best view.**).

2) Lesion Segmentation Task: We conduct lesion segmentation tasks based on two multi-lesion segmentation datasets, including IDRID [39] and DDR [40] datasets. We employ U-Net [45] and DeepLabV3+[47] as the backbone of our framework, which also proves that our framework can be suitable for different backbones. The comparison methods include U-Net [45] and DeepLabV3+[47] as the backbone of our framework. The U-Net++[49], DenseUNet [50], DeepLabV3+, FCRN [51], CASENet [52], L-Seg [53], PMCNet [54] and SS-MAF [24]. We use mDice, mIoU, and mAUC as the evaluation metrics.

As shown in Table VIII, dual-stream learning frameworks give out much higher accuracy than single-stream lesion segmentation algorithms. For example, compared with the DeepLabV3+, our framework obtains about 10% and 5% higher mIoU for IDRiD and DDR datasets, respectively. SS-MAF is another of our proposed dual-stream learning frameworks with feature interaction, proving our proposed thinking of shared feature

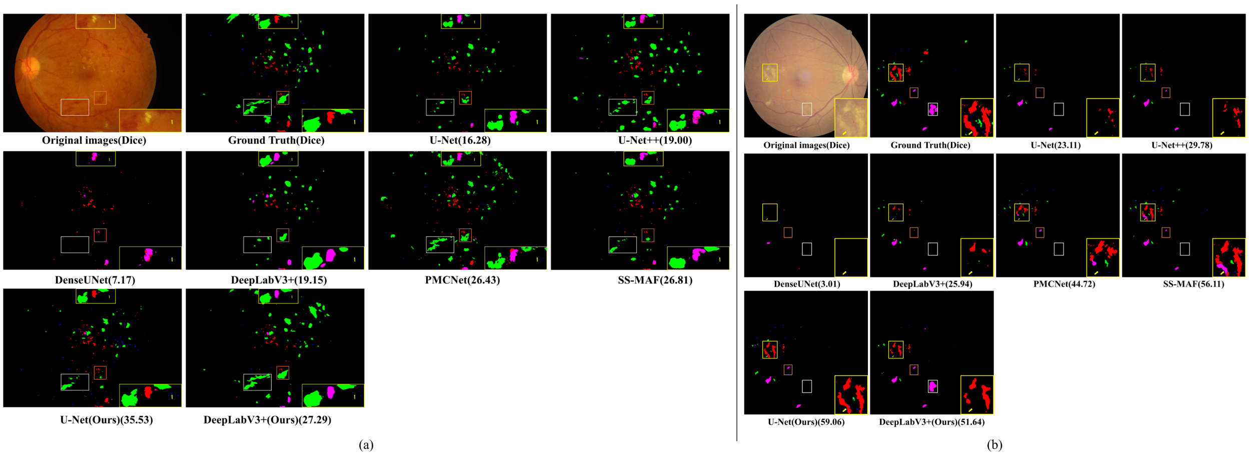


Fig. 6. Visualization of the results on IDRID and DDR dataset. Red, green, blue, and pink markings denote Hard Exudate(EX), Haemorrhages (HE), Microaneurysms, and Soft Exudate (SE), respectively. (a) Visualization of the IDRID dataset. (b) Visualization of the DDR dataset. (**Please zoom in for a best view.**).

TABLE VIII
COMPARISON RESULTS FOR LESION SEGMENTATION TASK

Dataset	Model	mDice(%)	mIoU(%)	mAUC(%)
IDRiD	U-Net	33.44 ± 3.97	22.39 ± 3.49	41.34 ± 2.83
	U-Net+	38.81 ± 1.60	27.14 ± 1.27	47.86 ± 4.27
	DenseUNet	18.41 ± 2.86	11.67 ± 2.21	22.84 ± 3.12
	DeepLabV3+	40.99 ± 1.35	28.56 ± 1.34	48.98 ± 2.28
	FCRN	-	-	45.52
	CASENet	-	-	48.23
	PMCNet	38.39 ± 0.61	27.24 ± 0.47	55.50 ± 0.74
	SS-MAF	48.63 ± 1.95	33.96 ± 1.51	51.19 ± 2.31
	U-Net(Ours)*	50.70 ± 2.34	35.42 ± 2.08	56.02 ± 3.33
	DeepLabV3+(Ours)*	$\mathbf{54.13} \pm 1.53$	$\mathbf{38.82} \pm 1.32$	$\mathbf{59.08} \pm 2.14$
DDR	U-Net	29.40 ± 2.08	18.16 ± 1.38	31.71 ± 2.83
	U-Net++	29.73 ± 1.55	18.44 ± 1.00	32.88 ± 0.91
	DenseUNet	19.53 ± 4.36	11.55 ± 2.71	22.65 ± 0.92
	DeepLabV3+	32.80 ± 0.79	20.91 ± 0.66	34.12 ± 0.67
	FCRN	-	-	9.60
	CASENet	-	-	19.28
	L-Seg	-	-	32.08
	PMCNet	22.70 ± 0.30	15.89 ± 0.25	29.28 ± 3.20
	SS-MAF	$\mathbf{40.05} \pm 0.33$	25.59 ± 0.27	$\mathbf{36.56} \pm 1.20$
	U-Net(Ours)*	37.53 ± 1.67	23.52 ± 1.35	35.13 ± 0.86
	DeepLabV3+(Ours)	39.69 ± 1.27	$\mathbf{25.76} \pm 0.98$	35.60 ± 1.61

The * represents a p-value< 0.05, which indicates significantly different results are obtained. More details are shown in appendix d.4, available online.

TABLE IX
COMPARISON RESULTS FOR THE CITYSCAPES DATASET

Model	Val(%)	Test(%)	GFLOPs
FCN [56]	-	65.3	1335.60
ENet [57]	-	58.3	7.24
ESPNet [27]	-	60.3	8.86
ERFNet [58]	-	68.0	25.60
PSPNet(ResNet18(1.0)) [59]	-	67.6	512.80
ESPNetV2 [28]	64.5	65.1	5.85
ESPNetV2(DSRL) [18]	66.5	65.9	5.85
ESPNetV2(Ours)	66.8	64.4	5.85
DeeplabV3+ [47]	70.0	67.1	565.35
DeeplabV3+(DSRL) [18]	72.0	69.3	568.53
DeeplabV3+(Ours)	73.6	70.8	565.83

The GFLOPs is calculated when the input size is 1024 × 512.

extraction is right. The categories of the DDR dataset are extremely imbalanced, which often causes the optimization to be difficult and easy to misclassify. This is the reason that the accuracy of the DDR dataset is relatively lower.

We visualize the results of two benchmarks and show some samples in Fig. 6. The figure tells that our method discriminates the boundaries better. Compared to the base model U-Net, the models trained by our framework classify the lesion more precisely with smooth edges. For example, on the IDRID dataset, the base model U-Net trends to misclassify the hard exudate (EX) as the soft exudate (SE), but the U-Net trained in our framework overcomes this problem and provides the correct classification. One interesting phenomenon is that the trained DeepLabV3+ based on our framework seems to inherit the misclassification for the HE, but the accuracy of our segmented lesion edges is better than that of the base DeepLabV3+. The possible reason is that the super-resolution brings the shape or geometry information, which may enhance the boundary of the lesions, but can not provide a rich semantic context for the classification. The structure of the model determines that U-Net fuses the semantic context by the skip connection of high-level and low-level features, but DeeplabV3+ obtains less context information in the decoder.

3) Cityscape Segmentation Task: To evaluate the generalization of our framework, we conduct the comparison experiment on Cityscapes [41] dataset, whose proportion of segmentation target is considerable. We choose the DSRL [18] as the comparison framework based on dual-stream super-resolution semantic learning. We use ESPNetV2 and DeeplabV3+ as the base model to build our framework. The GDice [55] and SSIM are used as the proxy loss for semantic segmentation and super-resolution, respectively. For DeeplabV3+, we use ResNet101 as the backbone to extract the features, and the weights trained on the ImageNet to initialize the backbone for DeeblabV3+ and ESPNetV2. We list the accuracy of validation and test and the GFLOPs. The GFLOPs are calculated when the input size is 1024×512. The quantitative results are shown in Table IX. We can see that our framework can work well on the cityscape scene. Compared with the DeeplabV3+ baseline and DSRL-integrated framework, our

TABLE X
RESULTS FOR OUT OF DISTRIBUTION EXPERIMENTS

Train	Test	Model	Dice	MCC
FIVES	HRF	U-Net	45.52 ± 0.16	42.06 ± 0.14
		CogSeg	50.00 ± 0.65	46.73 ± 0.82
		SuperVessel	54.35 ± 0.07	**51.71** ± 0.10
		SS-MAF	**54.37** ± 0.08	51.58 ± 0.06
		Ours	54.26 ± 0.18	51.50 ± 0.21
HRF	FIVES	U-Net	14.31 ± 3.76	24.99 ± 3.46
		CogSeg	28.27 ± 5.53	37.72 ± 4.31
		SuperVessel	36.11 ± 1.74	44.46 ± 1.33
		SS-MAF	19.59 ± 4.14	30.28 ± 3.39
		Ours	**46.04** ± 2.70	**52.02** ± 2.03
FIVES	DRHAGIS	U-Net	61.62 ± 0.12	60.32 ± 0.15
		CogSeg	61.94 ± 1.34	61.55 ± 1.08
		SuperVessel	67.65 ± 0.21	66.56 ± 0.22
		SS-MAF	**67.73** ± 0.20	**66.71** ± 0.22
		Ours	67.65 ± 0.10	66.61 ± 0.09
HRF	DRHAGIS	U-Net	45.98 ± 4.48	45.11 ± 4.99
		CogSeg	31.86 ± 4.32	32.40 ± 3.92
		SuperVessel	53.43 ± 6.19	52.50 ± 6.22
		SS-MAF	**53.97** ± 4.67	**52.90** ± 4.79
		Ours	53.69 ± 6.46	52.69 ± 6.33
IDRID	DDR	U-Net	13.86 ± 7.41	-
		SS-MAF	24.96 ± 1.19	-
		U-Net(ours)	**25.67** ± 1.43	-
		DeeplabV3+	21.38 ± 1.93	-
		DeeplabV3+(Ours)	23.33 ± 1.16	-
DDR	IDRID	U-Net	18.96 ± 1.09	-
		SS-MAF	32.70 ± 2.49	-
		U-Net(ours)	30.26 ± 3.11	-
		DeeplabV3+	26.44 ± 1.97	-
		DeeplabV3+(Ours)	**34.95** ± 1.24	-

The train means the training dataset, test means the testing dataset. (mean ± std).

framework obtains 3.6% and 1.6% for the validation set, 3.7% and 1.5% for the test set, with degrading GFLOPs.

D. Out of Distribution Experiments

We try to conduct experiments under the condition of out-of-distribution (OOD), which also reflects the robustness of the trained model based on our proposed framework. For the vessel segmentation task, we introduce another dataset named DRHAGIS [60] as the extra dataset to evaluate the model trained on HRF and FIVES. The performance in the OOD scenery is significant for clinical applications, and the cross-dataset experiments are to simulate the OOD scenery that the data from different clinics. We report the Dice and MCC for the vessel segmentation task. For the lesion segmentation task, we only report the Dice, as the MCC is not suitable to evaluate the performance of lesion segmentation. For comparison methods, we adopt U-Net, CogSeg, SuperVessel, and SS-MAF for the vessel segmentation task, U-Net, SS-MAF, and DeeplabV3+ for the lesion segmentation task.

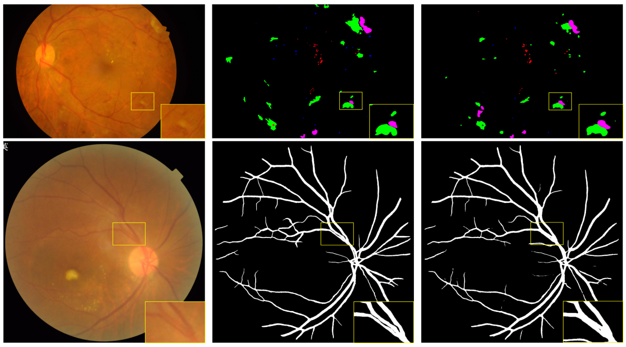

(a) Input (b) Ground Truth (c) Output

Fig. 7. Failure examples. Input image, ground-truth, and our results.

As shown in Table X, the first column is the dataset we train the models, and the second column is the dataset used for the test. The FIVES dataset holds a large number of images (800 images in total), and all the trained models work relatively robustly on other test datasets. For the HRF dataset with less of images, our framework provides robust performance. For example, U-Net and SS-MAF trained on the HRF dataset only produce about 14% and 19% Dice on the FIVES dataset, but our framework gives about 46% Dice. The Dices of CogSeg, which trains on HRF and tests on FIVES or DRHAGIS, are about 18% and 22% lower than those of our framework. For the lesion segmentation task, our framework with U-Net or DeeplabV3+ as backbones produces the highest accuracy. Therefore, compared with single-stream lesion segmentation models, dual-stream models provide higher robustness for OOD problems.

V. DISCUSSION

The experiments based on 6 publicly available datasets for 3 types of tasks show that our method can work on both medical image and natural image scenarios. The RoI proportion of the former is relatively small, and that of the latter is very large. As the resolution of input images is relatively low, our framework still achieves a promising performance. But during experiments, we find several limitations in our framework. The first is about the standard deviation. The values of our framework in this paper are a little larger than our previous model SS-MAF, which means the stability of our structure is a little inferior to that of SS-MAF. We hypothesize that the SSIM loss function giving one strong supervision signal for super-resolution may disturb parts of segmentation results. As the Fig. 7 shows, if the area is vague, the two targets may adhesion due to the information brought by the super-resolution.

Moreover, the proposed way is to guide learning the shared features between tasks with optimization methods, such as maximizing the mutual information between tasks. In the future, we can explore more effective structures to capture shared information, such as the self-attention mechanism at the multi-axis or the combination of global and local information.

VI. CONCLUSION

As the proportions of target areas in medical image segmentation are relatively small, the existing dual-stream framework based on the similarity loss may collapse or cannot achieve the desired performance. After rethinking the segmentation ability in the dual-stream framework, we identified its limitations applied to medical image segmentation. We proposed a Dual-Stream Shared Feature (DS2F) framework based on the hypothesis that a small set of features is shared between tasks. We proposed a novel shared feature extraction module and defined proxy tasks to constrain the module learning in the DS2F framework. Extensive experiments on six publicly available datasets, including medical and nature scenes, verify the effectiveness of our proposed framework.

References

[1] Z. Gu et al., "Ce-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[2] H. Wu, W. Wang, J. Zhong, B. Lei, Z. Wen, and J. Qin, "SCS-Net: A scale and context sensitive network for retinal vessel segmentation," *Med. Image Anal.*, vol. 70, 2021, Art. no. 102025.

[3] Y. Tan, K.-F. Yang, S.-X. Zhao, and Y.-J. Li, "Retinal vessel segmentation with skeletal prior and contrastive loss," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2238–2251, Sep. 2022.

[4] A. He, K. Wang, T. Li, W. Bo, H. Kang, and H. Fu, "Progressive multiscale consistent network for multiclass fundus lesion segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3146–3157, Nov. 2022.

[5] C. Xue et al., "Global guidance network for breast lesion segmentation in ultrasound images," *Med. Image Anal.*, vol. 70, 2021, Art. no. 101989.

[6] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag.*, 2019, pp. 683–687.

[7] N. Heller et al., "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the kits19 challenge," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101821.

[8] V. Oreiller et al., "HEad and neCK TumOR segmentation in PET/CT: The HECKTOR challenge," *Med. Image Anal.*, vol. 77, 2022, Art. no. 102336.

[9] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, "Cross-modality deep feature learning for brain tumor segmentation," *Pattern Recognit.*, vol. 110, 2021, Art. no. 107562.

[10] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[11] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2020.

[12] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, 2019.

[13] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki, "Iter-Net: Retinal image segmentation utilizing structural redundancy in vessel networks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 3645–3654.

[14] D. Wang, A. Haytham, J. Pottenburgh, O. Saeedi, and Y. Tao, "Hard attention net for automatic retinal vessel segmentation," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3384–3396, Dec. 2020.

[15] A. Galdran, A. Anjos, J. Dolz, H. Chakor, H. Lombaert, and I. B. Ayed, "State-of-the-art retinal vessel segmentation with minimalistic models," *Sci. Rep.*, vol. 12, no. 1, pp. 1–13, 2022.

[16] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "SA-UNet: Spatial attention u-net for retinal vessel segmentation," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 1236–1242.

[17] T. M. Khan, A. Robles-Kelly, and S. S. Naqvi, "T-Net: A resource-constrained tiny convolutional neural network for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 644–653.

[18] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3773–3782.

[19] Y. Sang, J. Sun, S. Wang, H. Qi, and K. Li, "Super-resolution and infection edge detection co-guided learning for COVID-19 CT segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 1665–1669.

[20] Q. Zhang, G. Yang, and G. Zhang, "Collaborative network for super-resolution and semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4404512.

[21] Y. Sang, J. Sun, S. Wang, H. Qi, and K. Li, "Super-resolution and infection edge detection co-guided learning for COVID-19 CT segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 1665–1669.

[22] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 41–48.

[23] Y. Hu, Z. Qiu, D. Zeng, L. Jiang, C. Lin, and J. Liu, "SuperVessel: Segmenting high-resolution vessel from low-resolution retinal image," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2022, pp. 178–190.

[24] J. Zhang, X. Chen, Z. Qiu, M. Yang, Y. Hu, and J. Liu, "Hard exudate segmentation supplemented by super-resolution with multi-scale attention fusion module," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2022, pp. 1375–1380.

[25] J. P. Horwath, D. N. Zakharov, R. Mégret, and E. A. Stach, "Understanding important features of deep learning models for segmentation of high-resolution transmission electron microscopy images," *NPJ Comput. Mater.*, vol. 6, no. 1, 2020, Art. no. 108.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[27] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 561–580.

[28] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9182–9192.

[29] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv: 1704.04861*.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[31] A. Howard et al., "Searching for mobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[32] H. Wang et al., "Patch-free 3D medical image segmentation driven by super-resolution technique and self-supervised guidance," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 131–141.

[33] H. Wang et al., "Super-resolution based patch-free 3D medical image segmentation with self-supervised guidance," 2022, *arXiv:2210.14645*.

[34] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[35] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv: 1706.05587*.

[36] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," *Int. J. Biomed. Imag.*, vol. 2013, 2013, Art. no. 154860.

[37] L. Ding, A. E. Kuriyan, R. S. Ramchandran, C. C. Wykoff, and G. Sharma, "Weakly-supervised vessel detection in ultra-widefield fundus photography via iterative multi-modal registration and learning," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2748–2758, Oct. 2021.

[38] K. Jin et al., "FIVES: A fundus image dataset for artificial intelligence based vessel segmentation," *Sci. Data*, vol. 9, 2022, Art. no. 475.

[39] P. Porwal et al., "Indian diabetic retinopathy image dataset (IDRID): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, 2018. [Online]. Available: https://www.mdpi.com/2306-5729/3/3/25

[40] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci.*, vol. 501, pp. 511–522, 2019.

[41] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[42] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[43] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*.

[44] S. Zhao, Y. Wang, Z. Yang, and D. Cai, "Region mutual information loss for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11115–11125.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi Eds., Berlin, Germany: Springer, 2015, pp. 234–241.

[46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[48] Y. Li, Y. Zhang, W. Cui, B. Lei, X. Kuang, and T. Zhang, "Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 8, pp. 1975–1989, Aug. 2022.

[49] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer, 2018, pp. 3–11.

[50] Y. Cao, S. Liu, Y. Peng, and J. Li, "DenseuNet: Densely connected UNet for electron microscopy image segmentation," *IET Image Process.*, vol. 14, no. 12, pp. 2682–2689, 2020.

[51] J. Mo, L. Zhang, and Y. Feng, "Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks," *Neurocomputing*, vol. 290, pp. 161–171, 2018.

[52] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "CaseNet: Deep category-aware semantic edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5964–5973.

[53] S. Guo, T. Li, H. Kang, N. Li, Y. Zhang, and K. Wang, "L-Seg: An end-to-end unified framework for multi-lesion segmentation of fundus images," *Neurocomputing*, vol. 349, pp. 52–63, 2019.

[54] A. He, K. Wang, T. Li, W. Bo, H. Kang, and H. Fu, "Progressive multi-scale consistent network for multi-class fundus lesion segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3146–3157, Nov. 2022.

[55] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, M. J. Cardoso Eds., Berlin, Germany: Springer, 2017, pp. 240–248.

[56] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[57] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[58] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[59] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2599–2608.

[60] S. Holm, G. Russell, V. Nourrit, and N. McLoughlin, "DR HAGIS–a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients," *J. Med. Imag.*, vol. 4, no. 1, 2017, Art. no. 014503.

Zhongxi Qiu received the BS degree from the Department of Internet of Things, University of South China, Hengyang, China, in 2020. His research interests include medical image analysis, computer vision, and adversarial learning.

Yan Hu received the PhD degree from the Department of Information Science and Technology, the University of Tokyo, Japan. She is working now in the Southern University of Science and Technology, China. Her research interests include medical image analysis, surgery video processing, and computer-aided surgery.

Xiaoshan Chen (Member, IEEE) is currently working toward the undergraduation degree with the Department of Computer Science and Technology, Southern University of Science and Technology. Her current research interest is medical image analysis.

Dan Zeng received the BE and PhD degrees in computer science and technology from Sichuan University, in 2013 and 2018. From 2018 to 2020, she worked as a post-doc research fellow in the Data Management and Biometrics Group with the University of Twente, the Netherlands. She is currently a research assistant professor in the Department of Computer Science and Engineering at Southern University of Science and Technology. Her main research topics lie in machine learning for biometrics, including face super-resolution, 2D face recognition, 3D face modelling, and facial expression recognition.

Qingyong Hu received the BS degree and MS degeree from China. He is currently working toward the PhD degree with the Department of Computer Science, University of Oxford, U.K. His current research interest include 3D computer vision, machine learning, and robotics.

Jiang Liu (Senior Member, IEEE) received the BS degree from the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China, MS degree and PhD degree from the Department of Computer Science and Technology, National University of Singapore, Singapore. He is working now in the Southern University of Science and Technology, China. His research interest include medical image analysis and surgical robots.