Global and Local Feature Reconstruction for Medical Image Segmentation

Jiahuan Song, Xinjian Chen, *Senior Member, IEEE*, Qianlong Zhu, Fei Shi, *Member, IEEE*, Dehui Xiang, *Member, IEEE*, Zhongyue Chen, Ying Fan, Lingjiao Pan, and Weifang Zhu, *Member, IEEE*

***Abstract*— Learning how to capture long-range dependencies and restore spatial information of down-sampled feature maps are the basis of the encoder-decoder structure networks in medical image segmentation. U-Net based methods use feature fusion to alleviate these two problems, but the global feature extraction ability and spatial information recovery ability of U-Net are still insufficient. In this paper, we propose a Global Feature Reconstruction (GFR) module to efficiently capture global context features and a Local Feature Reconstruction (LFR) module to dynamically up-sample features, respectively. For the GFR module, we first extract the global features with category representation from the feature map, then use the different level global features to reconstruct features at each location. The GFR module establishes a connection for each pair of feature elements in the entire space from a global perspective and transfers semantic information from the deep layers to the shallow layers. For the LFR module, we use low-level feature maps to guide the up-sampling process of high-level feature maps. Specifically, we use local neighborhoods to reconstruct features to achieve the transfer of spatial information. Based on the encoder-decoder architecture, we propose a Global and Local Feature Reconstruction Network (GLFRNet), in which the GFR modules are applied as skip connections and the LFR modules constitute the decoder path. The proposed GLFRNet is applied to four different medical image segmentation tasks and achieves state-of-the-art performance.**

Manuscript received 18 November 2021; revised 10 February 2022; accepted 20 March 2022. Date of publication 24 March 2022; date of current version 31 August 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701700; and in part by the National Natural Science Foundation of China under Grant U20A20170, Grant 61622114, and Grant 62001196. *(Corresponding author: Weifang Zhu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols for choroidal atrophy segmentation was granted by the Institutional Review Board of Shanghai General Hospital under Approval No. 2016KY126.

Jiahuan Song, Qianlong Zhu, Fei Shi, Dehui Xiang, Zhongyue Chen, and Weifang Zhu are with the MIPAV Laboratory, School of Electronic and Information Engineering, Soochow University, Jiangsu 215006, China (e-mail: jiahuansong@qq.com; 1592404966@qq.com; shifei@suda.edu.cn; xiangdehui@suda.edu.cn; chenzy@suda.edu.cn; wfzhu@suda.edu.cn).

Xinjian Chen is with the MIPAV Laboratory, School of Electronic and Information Engineering, and the State Key Laboratory of Radiation Medicine and Protection, Soochow University, Jiangsu 215006, China (e-mail: xjchen@suda.edu.cn).

Ying Fan is with the First People's Hospital Affiliated to Shanghai Jiao Tong University, Shanghai 200940, China (e-mail: mdfanying@sjtu.edu.cn).

Lingjiao Pan is with the School of Electrical and Information Engineering, Jiangsu University of Technology, Jiangsu 213000, China (e-mail: jsjshedy@jsut.edu.cn).

Digital Object Identifier 10.1109/TMI.2022.3162111

***Index Terms*— Medical image segmentation, deep learning, convolutional neural network, global feature reconstruction module, local feature reconstruction module.**

I. Introduction

AUTOMATIC segmentation of medical images is a crucial step in the quantitative pathological assessment and diagnosis of many diseases such as colorectal polyp segmentation in colonoscopy images [1]–[4], choroidal atrophy segmentation in fundus images [5], [6], retinal fluid segmentation in optical coherence tomography (OCT) images [7], [8], and multi-organ segmentation in computed tomography(CT) images [9]–[12].

Modeling long-range dependencies and recovering spatial information of feature maps are critical for the encoder-decoder structure networks represented by U-Net [13] in medical image segmentation. Although U-Net and it variations have achieved state-of-the-art performances in many medical image segmentation tasks, they still suffer from the following problems.

First, the ability of global context feature extraction is insufficient which only depends on feature fusion at different levels. Although the down-sampling of feature maps allows the network to capture longer dependencies in the deep layer, the empirical receptive field of CNN is much smaller than the theoretical one especially on high-level layers [14]. Recently, some multi-scale feature fusion or attention mechanism based fully convolutional networks (FCNs) [15] have been proposed to capture long-range dependencies. For example, DeepLab [16], [17], CE-Net [18] and PSPNet [19] combine feature maps generated by different dilated convolution and pooling operations to extract multi-scale context features. Limited by the parameter sharing of convolution, these networks may still lack spatial awareness to deal with different positions, which does not satisfy the requirement that different pixels need different contextual dependencies. Attention mechanism has also been employed to exploit long-range dependencies in some networks [20]–[22]. For example, non-local network [20] models the features at any two locations in the feature map, leading to generate more powerful pixel-wise representation. SENet [21] uses global pooling to collect a global feature of the entire space, and then distributes it to each location. However, non-local module is computationally heavy, and SENet treats pixels with different semantic information equally which is not robust to pixelwise dense prediction tasks.

Second, the simple skip connection combines local information with different levels indiscriminately and ignores semantic information. On one hand, low-level features are too noisy to provide sufficient high-resolution semantic guidance. On the other hand, due to the lack of spatial information in high-level feature maps, direct concatenation or addition cannot solve the misalignment of semantic information between feature maps. In order to fuse features efficiently and eliminate the interference of irrelevant noise in low-level features, Attention U-Net [23], AG-Net [24], ACNet [25] and CPFNet [26] use gating mechanism to emphasize or suppress features with different semantic information, which makes the feature fusion more flexible. However, none of these methods solve the problem of semantic misalignment between high-level features and low-level features.

Third, feature up-sampling for spatial information recovery is very important in semantic segmentation, especially in medical image segmentation. The most widely used feature up-sampling operators are the nearest neighbor interpolation and bilinear interpolation, which only depend on the distance between pixels. Transposed convolution based up-sampling applies the same convolution kernel across the entire image, regardless the semantic information of different locations in the image. SFNet [27] predicts semantic flow to align low-level and high-level feature maps and achieves good performance in natural image segmentation. Sub-pixel convolution [28] is widely used for semantic segmentation, which is based on the assumption that spatial information is embedded in channels. For example, the idea of DUpsampling [29] is to use a linear transformation to approximate the structural information of the label. Similar to transposed convolution, DUpsampling applies the same parameter to the entire space. CARAFE [30] reassembles the neighbor of each location to achieve up-sampling, but it cannot integrate the rich spatial information of low-level feature maps.

Motivated by the above discussions and attention mechanism [20], [31], [32], we propose two novel feature reconstruction modules to solve the above problems in the encoder-decoder structure network in this paper, named as Global Feature Reconstruction (GFR) module and Local Feature Reconstruction (LFR) module respectively. The GFR module introduces global features from high-level features to low-level features and reassembles cross-level global features to increase the receptive field of the network and reduce the semantic gap between features at different levels. The LFR module uses low-level feature map to guide the up-sampling process of high-level feature map, and can adaptively reconstruct local features at different locations to achieve spatial information restoration. The main contributions of this work are summarized as follows:

1) We propose a novel Global and Local Feature Reconstruction Network (GLFRNet) equipped with GFR and LFR modules, which can effectively capture global context features and restore the spatial information in high-level features, respectively.
2) From the viewpoint of feature fusion, the proposed GFR and LFR modules solve the imbalance between the semantic information and spatial information of feature maps from the global and local perspectives respectively.
3) The proposed GLFRNet is applied in four challenging tasks including colorectal polyp segmentation, choroidal atrophy segmentation, multi-class retinal fluid segmentation and multi-organ segmentation. The state-of-the-art segmentation performances show the good generalization of the proposed GLFRNet.

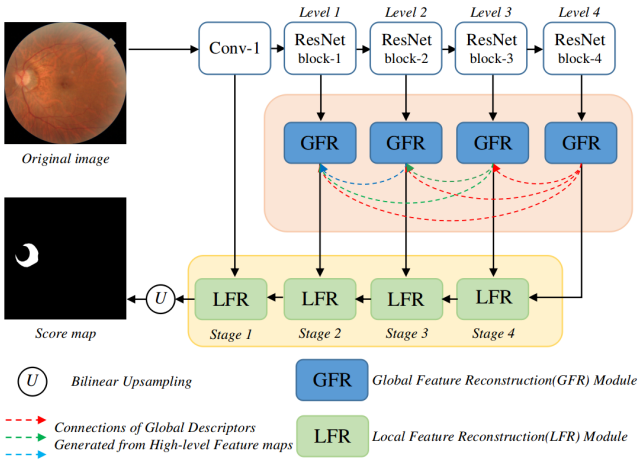

Fig. 1. Illustration of the proposed GLFRNet.

The remainder of this paper is organized as follows. In Section II, the proposed GLFRNet is described in detail. Section III presents the relevant experimental results on four different medical image segmentation tasks. The discussion and analysis about the proposed GFR and LFR modules and conclusion are given in Section IV.

II. Method

In this section, we first present the general framework of the proposed GLFRNet and then introduce the two feature reconstruction modules which capture global semantic information and recover local spatial information, respectively.

A. Overview

Fig.1 shows the proposed GLFRNet which is based on the encoder-decoder architecture. The pre-trained ResNet34 is used as the backbone network to extract hierarchical feature representations. Multiple GFR modules are placed between the encoder and the decoder as skip connections to increase the receptive fields and enrich the semantic information of the low-level features. The LFR modules act as decoder, adaptively up-sampling and fusing features step by step.

B. Global Feature Reconstruction Module

The overall architecture of the proposed GFR module is shown in Fig.2. For each level feature map, the GFR module consists of three steps: (1) Generate a set of global descriptors representing each class. (2) Combine the global descriptors of this level with all higher-level global descriptors to obtain cross-level global descriptors. (3) Predict the reconstruction

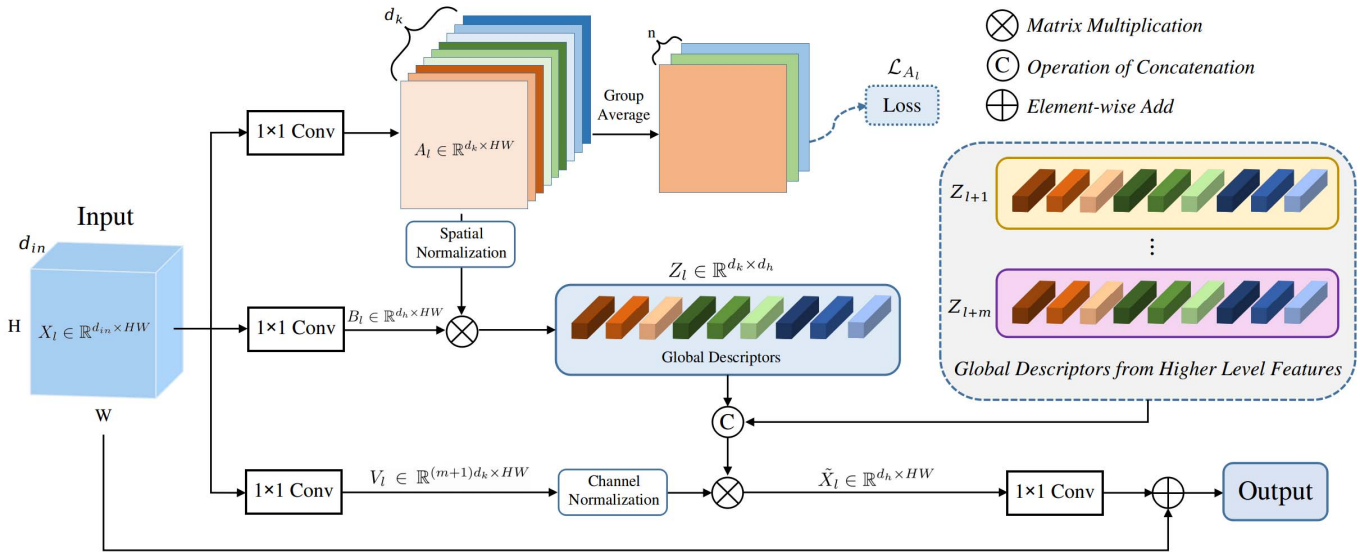

Fig. 2. The illustration of global feature reconstruction (GFR) module.

weights for each location and use the cross-level global descriptors to reconstruct each pixel. Therefore, each pixel can not only aggregate all the features of this level, but also capture features at other levels.

1) Global Fescriptor Generator: Given a flattened feature map $X_l \in \mathbb{R}^{d_{in} \times HW}$, where l denotes the level of this feature map, d_{in} denotes the number of channels, H and W are its spatial dimensions. We first feed X_l into a convolution layer to generate a attention map $A_l \in \mathbb{R}^{d_k \times HW}$ and a embedding feature $B_l \in \mathbb{R}^{d_h \times HW}$, where d_h and d_k denote the dimension and the number of global descriptors, respectively. Considering that the feature map of each level has different channel dimensions, we set d_h as 64 in our experiments to reduce the dimension of weights and computational cost. The global descriptors are generated as follows:

$$Z_l = [z_l^1, z_l^2, \ldots, z_l^{d_k}] = \rho(A_l)B_l^\top \in \mathbb{R}^{d_k \times d_h} \tag{1}$$

where ρ denotes the operation of applying $softmax$ normalization in space.

In order to get discriminative global descriptors, the learning of attention map A_l is deeply supervised. Specifically, the attention map A_l is divided into n groups by channel, where n is the number of segmentation classes. Then the features of each group in A_l are averaged in the channel dimension to obtain the prediction of deep supervision. Note that the s-th global descriptor $z_l^s \in \mathbb{R}^{d_h}, (s = 1, 2, \ldots, d_k)$ depends on the s-th channel in the attention map $A_l \in \mathbb{R}^{d_k \times HW}$. Therefore, corresponding to different groups of A_l, z_l^s can aggregate discriminative features of different classes and each class has d_k/n global descriptors at each level.

2) Cross-Level Global Descriptors: For the l-th level feature map, if the global descriptors Z_l extracted at this level are used for feature reconstruction, the pixel at each location will be associated with each other through Z_l. In addition, we combine the global descriptors of this level with all higher-level global descriptors to obtain a set of cross-level global descriptors, which can be formulated as:

$$Z_l' = concat(Z_l, Z_{l+1}, \ldots, Z_{l+m}) \in \mathbb{R}^{(m+1)d_k \times d_h} \tag{2}$$

where $Z_{l+1}, \ldots, Z_{l+m}$ are the global descriptors generated from $(l+1)$-th, $\ldots$, and $(l+m)$-th level feature maps.

In this way, low-level feature maps can be reconstructed by using descriptors from high-level features with strong semantic information. So the reconstructed feature map will be both rich in spatial details and semantic information. In other words, high-level feature maps use a few global descriptors to efficiently transfer the semantic information to low-level feature maps.

3) Global Feature Reconstruction: The next step is to use the cross-level global descriptors Z_l' to reconstruct the features of each location. We use 1×1 convolution to predict global reconstruction weights $V_l \in \mathbb{R}^{(m+1)d_k \times HW}$ based on the current feature map X_l, where m represents the number of connections of global descriptors generated from higher level feature maps, seen in Fig.1. $softmax$ function in channel dimension is used to normalize the reconstruction weights and enhance the ability of global descriptors selection. The reconstruction process of the feature map X_l can be formulated as:

$$\tilde{X}_l = Z_l'^\top softmax(V_l) \tag{3}$$

where $\tilde{X}_l \in \mathbb{R}^{d_h \times HW}$ represents the reconstructed feature map. Finally, to prevent the degradation of network training, through a 1×1 convolution, $\tilde{X}_l$ is added with the input feature X_l to obtain the final global reconstructed feature.

C. Local Feature Reconstruction Module

In order to transfer the spatial information in low-level features to high-level features, we propose the LFR module. The LFR module uses low-level features to guide the local feature reconstruction of high-level features so that the feature up-sampling and semantic alignment can be achieved.

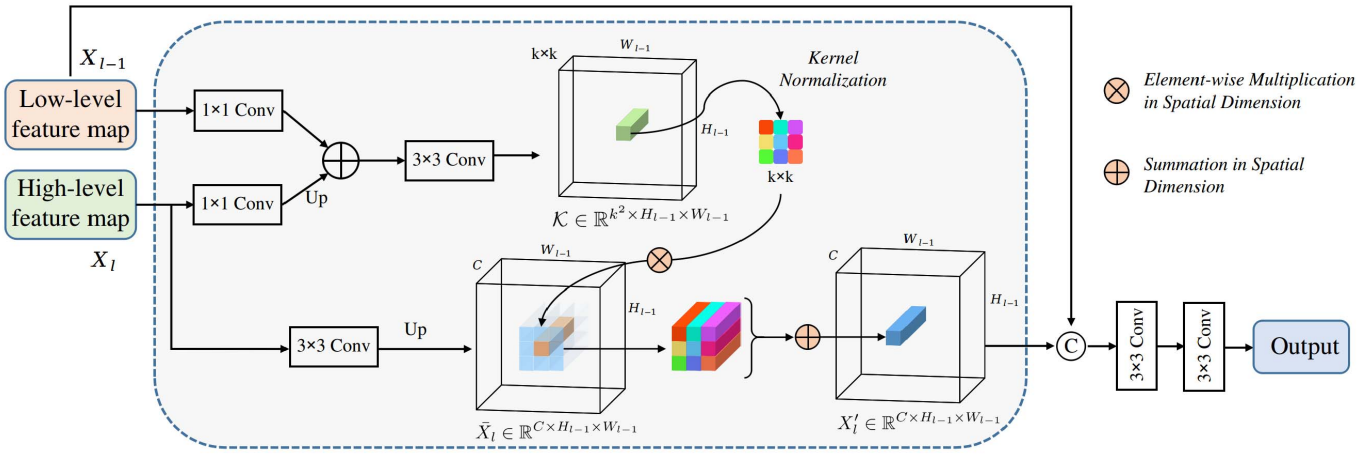

Fig. 3. The illustration of local feature reconstruction (LFR) module.

As shown in Fig.3, the proposed LFR module takes two adjacent feature maps X_{l-1} and X_l as inputs and outputs a new feature map X_{out}, which combines the spatial information and semantic information of both X_{l-1} and X_l.

1) Reconstruction Kernel Prediction: First, in order to reduce parameters and computational cost, we adopt two 1×1 convolutions to compress the channels of X_{l-1} and X_l respectively. Bilinear interpolation is used to up-sample the high-level feature map. Next, these two feature maps are fused by element-wise summation. Then, the fused feature map is fed into a 3×3 convolution to get the prediction of reconstruction kernel $\mathcal{K} \in \mathbb{R}^{k^2 \times H_{l-1} \times W_{l-1}}$, where k represents the neighborhood size of local feature reconstruction. So the prediction of the reconstruction kernel $\mathcal{K}$ can be formulated as:

$$\mathcal{K} = softmax\Big(conv_f\Big(Up\big(\theta(X_l)\big) + \psi(X_{l-1})\Big)\Big) \tag{4}$$

where $\theta(\cdot)$ and $\psi(\cdot)$ represent 1×1 convolutions with parameters $\mathcal{W}_\theta$ and $\mathcal{W}_\psi$ respectively, and $Up(\cdot)$ represents the bilinear interpolation. $conv_f(\cdot)$ is the 3×3 convolution. $softmax$ function is used to normalize the predicted reconstruction kernel.

2) Local Feature Reconstruction: To preserve the relative location information during the reconstruction process and obtain the semantic-rich up-sampled feature, local feature reconstruction linearly assembles the $k \times k$ neighborhood of each location. We use another 3×3 convolution $conv_l(\cdot)$ to reduce the channel dimension of X_l, up-sample it to the same size as X_{l-1} and get $\bar{X}_l$:

$$\bar{X}_l = Up\big(conv_l(X_l)\big) \tag{5}$$

For the pixel at position $[i, j]$ and its corresponding reconstruction kernel $\mathcal{K}_{[i,j]} \in \mathbb{R}^{k \times k}$, the reconstructed local feature $X'_l[i, j]$ is can be formulated as Equation (6), where $r = \lfloor k/2 \rfloor$:

$$X'_l[i, j] = \sum_{n=-r}^{r} \sum_{m=-r}^{r} \mathcal{K}_{[i,j]}[n, m] \cdot \bar{X}_l[i + n, j + m] \tag{6}$$

Local feature reconstruction is more flexible than other up-sampling operations such as bilinear interpolation and transposed convolution, because it can predict the kernel by integrating spatial information and semantic information in X_{l-1} and X_l, enabling the relevant points in the local region get more attention. This benefits from $softmax$ function shown in Equation (4), which makes the local reconstruction kernels more sharp. The phenomenon of inconsistent categories and features often occurs at the edge of objects, that is, pixel $[i, j]$ and its neighbor pixel $[i + n, j + m]$ may have the same category and different features. LFR module can make the categories and features consistent, thereby improving the recognition of the edge of the object. For example, if the prediction $\mathcal{K}_{[i,j]}[i + n, j + m]$ is close to 1 and the weights of other positions are close to 0, the position $[i + n, j + m]$ will get almost all attention of position $[i, j]$. According to Equation (6), the reconstructed feature will be $X'_l[i, j] \approx \bar{X}_l[i+n, j+m]$, which makes the features of the two positions consistent. At last, the reconstructed high-level feature map X'_l and the low-level feature map X_{l-1} are concatenated and fed into a sub-network with two 3×3 convolution layers to get the final output feature map X_{out}.

D. Implementation Details

Our experiments are implemented on the public PyTorch platform and NVIDIA RTX 3090 GPU with 24GB memory. SGD with a momentum of 0.9 and weight decay of 0.0001 is used as the optimizer. We adopt the poly learning rate policy to schedule learning rate $lr = baselr\big(1 - \frac{iter}{max\_iter}\big)^{power}$, where basic learning rate $baselr$ is set to 0.01, $power$ is set to 0.9, $iter$ and max_iter represent the current iteration number and the maximum iteration. Batch size varies according to the dataset. We replace the encoder in U-Net with the pre-trained ResNet34 and take it as the Baseline. We have release our codes on Github.[1]

The number of global descriptors d_k in the GFR module is set to 8 times the number of categories. The reconstruction kernel size k in the LFR module is set to 5. We employ a joint loss function consisting of Dice loss $\mathcal{L}_{Dice}$ and cross-entropy

[1] https://github.com/blue88blue/GLFR

loss $\mathcal{L}_{CE}$ as the segmentation loss $\mathcal{L}_{seg}$:

$$\mathcal{L}_{seg} = \mathcal{L}_{Dice} + \mathcal{L}_{CE} \tag{7}$$

In the GFR module, the deeply supervised loss $\mathcal{L}_{A_l}$ for the learning of the attention map A_l also uses the Dice loss and cross-entropy loss based joint loss function. The total loss function for the training of GLFRNet can be defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \sum_{l=1}^{4} \mathcal{L}_{A_l} \tag{8}$$

where $l = 1, 2, 3, 4$ correspond to the four GFR modules in Fig.1. λ is a trade-off parameter between segmentation loss and deep supervision loss of attention map, which is set to 0.2 in all our experiments.

III. Experiments

The proposed framework has been validated with four applications: (1) colorectal polyp segmentation in colonoscopy images, (2) choroidal atrophy segmentation in fundus images, (3) multi-class retinal fluid segmentation in optical coherence tomography (OCT) images, and (4) multi-organ segmentation in abdominal computed tomography (CT) scans. The evaluation metrics used in the experiments include Dice coefficient (Dice), intersection over union (IoU), accuracy(Acc), sensitivity (Sen) and Hausdorff distance (HD), which are defined as follows.

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{9}$$

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

$$Sen = \frac{TP}{TP + FN} \tag{11}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

$$HD = \max\left(\max_{p\in P}\min_{y\in Y} d(p, y), \max_{y\in Y}\min_{p\in P} d(p, y)\right) \tag{13}$$

where TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives respectively. P and Y represent the predicted pixel set and the target pixel set respectively. Function $d(\cdot)$ calculates the Euclidean distance between two pixels. Hausdorff distance measures the similarity between two sets of points.

In all tasks, we take Dice as the major evaluation metric and conduct Wilkerson signed rank test on it to verify the statistical significance of improvement.

A. Colorectal Polyp Segmentation

Colorectal cancer has high morbidity and mortality, which is a serious threat to human health [33]. Colorectal polyp is believed as one of the early symptoms of colorectal cancer. The automatic segmentation of colorectal polyp from colonoscopy images is very important, since it can help the clinicians accurately locate polyp areas for the further diagnosis or surgeries. Due to the varying appearances and similar color to the background, colorectal polyp segmentation is very challenging.

TABLE I
THE RESULT OF COMPARISON EXPERIMENTS AND ABLATION STUDIES ON COLORECTAL POLYP SEGMENTATION TASK (W/O MEANS WITHOUT THE FOLLOWING COMPONENT)

Methods	Dice(%)	IoU(%)	Acc(%)	GFLOPs	p-value
U-Net [13]	79.15	69.30	93.82	27.33	<0.001
Attention U-Net [23]	79.63	70.51	93.77	36.83	<0.001
UNet++ [35]	82.28	73.52	94.57	121.33	<0.001
CE-Net [18]	88.02	81.50	96.18	31.21	<0.001
PSPNet [19]	88.15	81.57	96.26	96.15	<0.001
CPFNet [26]	88.67	82.37	96.40	28.18	<0.001
GCN [36]	88.75	82.63	96.36	25.42	<0.001
DeepLabV3+ [17]	88.96	82.68	96.15	112.49	<0.001
SFNet [27]	89.03	82.89	96.58	31.32	<0.001
PraNet [37]	89.30	82.85	96.73	**18.68**	<0.001
EMANet [32]	89.53	82.87	96.56	89.57	<0.001
HarDNet-MSEG [38]	89.98	83.90	96.57	21.08	0.001
TransFuse-S [39]	90.05	83.88	96.80	40.26	0.039
TransFuse-L [39]	90.19	84.35	96.76	128.81	0.032
Baseline	87.54	81.16	95.94	29.52	<0.001
Baseline+GFR_w/o_DS	89.22	82.80	96.48	33.74	<0.001
Baseline+GFR_w/o_Conn	89.45	83.31	96.65	33.68	<0.001
Baseline+GFR	90.19	84.15	96.78	33.74	0.010
Baseline+1GFR(level4)	88.41	82.20	96.42	29.56	<0.001
Baseline+2GFR(level3-4)	89.83	83.81	96.80	30.37	0.011
Baseline+3GFR(level2-4)	90.09	84.09	96.66	31.46	0.009
Baseline+LFR_w/o_LG	88.24	81.97	96.35	27.96	<0.001
Baseline+LFR	88.69	82.77	96.61	29.37	0.022
Baseline+1LFR(stage4)	87.74	81.32	96.20	28.90	<0.001
Baseline+2LFR(stage3-4)	88.16	81.79	96.21	28.33	<0.001
Baseline+3LFR(stage2-4)	88.44	82.33	96.14	27.97	<0.001
Baseline(ResNet50)	87.23	80.36	95.96	196.56	<0.001
GLFRNet(ResNet50)	89.61	83.36	96.54	167.56	0.014
GLFRNet	**91.06**	**85.33**	**97.05**	32.54	-

Kvasir-SEG [34] is a large-scale challenging dataset which contains 1000 colonoscopy images with polyp regions. We randomly divide the 1000 images into 525 for training, 175 for validation and 300 for testing. The resolution of the images varies from 332 × 487 to 1920 × 1072 pixels. For simplicity, the images are resized to 512 × 448 with maintaining the average aspect ratio. Online random contrast transformation, brightness transformation, left-right and top-down flipping are applied for data augmentation. Dice coefficient (Dice), intersection over union (IoU) and accuracy (Acc) are adopted as evaluation metrics.

The proposed GLFRNet is compared with state-of-the-art algorithms, including U-Net [13], Attention U-Net [23], UNet++ [35], CE-Net [18], PSPNet [19], CPFNet [26], GCN [36], SFNet [27], DeepLabV3+ [17], PraNet [37], EMANet [32]. In addition, in order to verify the effectiveness of the proposed GFR and LFR modules, we conduct ablation experiments of two modules respectively.

As shown in Table I, our GLFRNet achieves the best performance. Compared with the Baseline (replace the encoder in

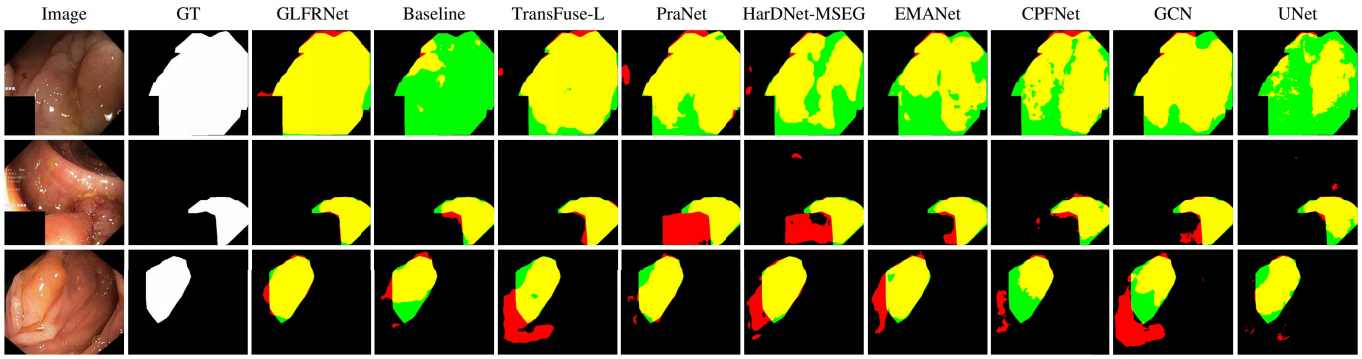

Fig. 4. Visual comparison between GLFRNet and state-of-the-art networks for polyp segmentation. For each row, we show an input image and its ground truth, and the corresponding output of each network. Green, red and yellow regions represent the false negative, false positive and true positive, respectively.

U-Net with the pre-trained ResNet34), the performance of the proposed GLFRNet gets overall improvement (3.52% for Dice coefficient, 4.17% for IoU and 1.11% for Accuracy). Dilated convolutional structure networks such as DeepLabV3+ [17] and EMANet [32] achieve similar performances compared with encoder-decoder structure networks such as PraNet [37] and SFNet [27]. However, the use of dilated convolution keeps the feature map at a high resolution, which increases the memory consumption and computational cost. PraNet [37] was proposed for real-time colorectal polyp segmentation, but its performances in all metrics are far behind our proposed GLFRNet. HarDNet-MSEG [38] used HarDNet [40] as the backbone. Multi-scale convolution, dilated convolution and dense aggregation were used in the decoder stage. Its overall encoder-decoder structure is similar to our GLFRNet, but it still does not outperform the proposed GLFRNet on Kvasir-SEG dataset. We think there are two possible reasons: (1) The element-wise multiplication based dense feature aggregation in the decoder of HarDNet-MSEG treats all levels of features equally, while our LFR module takes into account the relationship between feature maps efficiently, which can preserve the relative location information during the reconstruction process and obtain the semantic-rich up-sampled feature. (2) The self-attention mechanism used in GFR module is more flexible than the convolution and dilated convolution used in RFB. In order to combine the advantages of CNN and attention mechanism, TransFuse-S [39] and TransFuse-L innovatively used TransFormer [41] and CNN as dual encoders, and fused features in a way similar to CBAM [42] module in the decoder stage. Although both TransFuse-S and TransFuse-L perform well on Kvasir-SEG dataset, this structure, especially TransFuse-L, introduces a lot of computational cost (see the GFLOPs in TABLE I). The proposed GLFRNet gets an efficient combination of CNN and attention mechanism, and achieves the best trade-off between performance and efficiency.

To evaluate if our improvement is statistically significant, the Wilcoxon signed-rank test is conducted on Dice coefficient in both comparison experiments and ablation studies. It can be seen from Table I that all p-values are less than 0.05, indicating that our method has achieved a significant improvement compared to other methods. We give a few examples for visual comparison in Fig.4, which demonstrate the powerful global context feature aggregation capability of GLFRNet.

1) Ablation Study for GFR: As shown in Table I, the addition of the GFR module without deep supervision to the Baseline (Baseline+GFR_w/o_DS) makes the network comprehensively improve on all three evaluation metrics. Using deep supervision (Baseline+GFR) to perform feature selection on global descriptors further improves the performance, which has completely outperformed other state-of-the-art methods. In order to verify the validity of the global descriptor connection, we remove all the global descriptor connections from the high-level feature maps (Baseline+GFR_w/o_Conn). This results in the decrease in all three metrics compared with the complete GFR module (Baseline+GFR) and proves the necessity of the global descriptor connection in turn. As can be seen from Table I, the computational cost of the global descriptor connection is negligible, which shows that it is very efficient in semantic guidance. In addition, we also conduct ablation experiments on the number of GFR modules (Baseline+1GFR, Baseline+2GFR, Baseline+3GFR), as shown in Table I. It shows that one GFR module per level is suitable for the effective obtainment of the global receptive field and the reduction of the semantic gap between hierarchical features.

Fig.5 presents some visual comparisons between the output feature maps of simple skip-connections and our GFR modules, in which the feature maps are averaged according to the channel dimension and normalized to 0-255 to get the visualizations. As shown in Fig.5, our GFR module highlights the response to the segmentation targets and suppresses the response to the irrelevant background noise, indicating that it can indeed reduce the gap in the semantic information of features at all levels.

2) Ablation Study for LFR: As shown in Table I, The embedding of LFR module into Baseline (Baseline+LFR) also helps to improve the performance. Compared with the Baseline, the Dice increases 1.15% and reaches 88.69%. And the other two metrics also show significant improvements. This is benefited from the LFR module's recovery of the spatial information of the deep feature map. GFLOPs shows that the LFR module is more efficient than the Baseline's original decoder due to

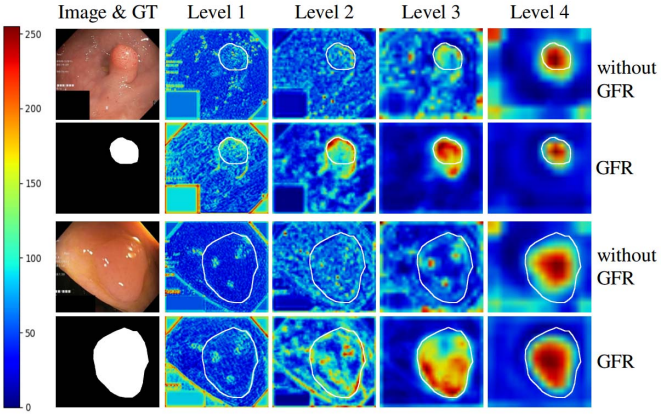

Fig. 5. Visual comparison of feature maps transferred by the skip connection before and after insertion of GFR module. The first column shows two input images and their corresponding ground truth. Columns 2 to 5 show different levels of feature maps, where the white curves represent the boundaries of the ground truth.

channel compression. In order to verify that the low-level feature maps have guiding effect on the up-sampling process of the high-level feature maps, we remove the guidance of the low-level feature maps (Baseline+LFR_w/o_LG), that is, the low-level feature maps do not participate in the prediction of the local reconstruction kernel. This leads to 0.45% decrease of Dice than that of the complete LFR module(Baseline+LFR), which implies that the guidance for spatial information recovery is necessary. The ablation experiments of the number of LFR modules are also given in Table I, which show that one LFR module per stage is appropriate for the recovery of spatial information and feature alignment to the features of adjacent stages.

3) Computational Complexity Analysis: Our GFR module can be regarded as an improvement of the self-attention mechanism. Compared with non-local [20] block, GFR module has much lower computational complexity and can be embedded anywhere in the network. The non-local block needs to calculate the similarity between all pixels, so its computational complexity is $O(N^2)$, where $N = H \times W$ represents the size of feature map. The GFR module uses global descriptors to compress the features in space, reducing the computational complexity to $O(d_k N)$, where d_k is the number of global descriptors and $d_k \ll N$. The LFR module only uses pixels in $k \times k$ neighborhood for reconstruction, and its computational complexity is $O(k^2 N)$, where $k \ll N$. In our experiments, k is set to 5 and d_k is set to 8 times the number of categories. In addition, because medical image datasets are usually small, it is not the best choice to use deeper ResNet50 as backbone, as shown in Table I.

B. Choroidal Atrophy Segmentation

Pathologic myopia and its complications are major causes of visual impairment and even blindness [43]. Choroidal atrophy is one of the earliest pathological changes of pathologic myopia and is also an important clinical manifestation in the diagnosis of pathological myopia. Therefore, the automatic segmentation of choroidal atrophy is important for the early diagnosis and treatment of pathologic myopia. Segmentation of choroidal atrophy on fundus images is still challenging because the shape and size of the atrophy vary greatly in different stages of pathologic myopia and the boundary is blurred.

TABLE II
THE RESULT OF COMPARISON EXPERIMENTS AND ABLATION STUDIES ON CHOROIDAL ATROPHY SEGMENTATION TASK

Methods	Dice(%)	IoU(%)	Sen(%)	Acc(%)	p-value
U-Net [13]	84.55	74.76	82.43	98.22	<0.001
UNet++ [35]	84.57	74.77	82.86	98.26	<0.001
EMANet [32]	85.32	76.00	82.42	98.39	<0.001
GCN [36]	85.52	76.31	84.53	98.34	<0.001
PSPNet [19]	85.72	76.50	84.20	98.30	<0.001
DeepLabV3+ [17]	86.16	77.19	85.10	98.41	<0.001
CPFNet [26]	86.27	77.30	84.33	98.31	<0.001
PraNet [37]	86.48	77.65	84.65	98.38	<0.001
CE-Net [18]	86.66	77.96	84.74	98.41	0.003
SFNet [27]	86.85	78.02	84.16	98.43	0.003
Baseline	84.46	74.82	80.43	98.25	<0.001
Baseline+GFR_w/o_Conn	85.32	76.29	82.80	98.22	<0.001
Baseline+GFR_w/o_DS	86.27	77.37	84.36	98.33	<0.001
Baseline+GFR	86.97	78.31	84.98	98.38	0.021
Baseline+LFR_w/o_LG	85.87	76.86	83.55	98.38	<0.001
Baseline+LFR	86.48	77.58	84.58	98.39	<0.001
GLFRNet	**87.61**	**79.28**	**86.12**	**98.50**	-

The dataset is provided by Shanghai General Hospital which contains 600 fundus images with pathologic myopia (2032 × 1934). The collection and analysis of image data were approved by the Institutional Review Board of Shanghai General Hospital and adhered to the tenets of the Declaration of Helsinki. Informed consent was obtained from all subjects. We center-crop each image and downsample the image to 512×512. The dataset is randomly divided into 320 images for training, 80 images for validation and 200 images for testing. Multiple online random augmentation methods are used for data augmentation, including random contrast and brightness transformation, left-right and up-down flipping and rotations from -60 degree to 60 degree.

The results of comparison experiments and ablation experiments are shown in Table II. The proposed GLFRNet achieves 87.61% in Dice, 79.28% in IoU, 98.50% in Accuracy and 86.12% in Sensitivity. Compared with the Baseline, GLFRNet improves significantly in Dice, IoU and Sensitivity with 3.15%, 4.46%, and 5.69%, respectively. GLFRNet also significantly outperforms all the other state-of-the-art networks with all p-values < 0.05 (Dice index, Wilcoxon signed-rank test). The ablation experiments demonstrate the effectiveness of the proposed GFR and LFR modules. The visual comparisons are shown in Fig.6. Although the shape and size of choroidal atrophy vary greatly in different stages of pathologic myopia, GLFRNet uses GFR and LFR modules to process large and small targets from both global and local perspectives and achieves good segmentation results.

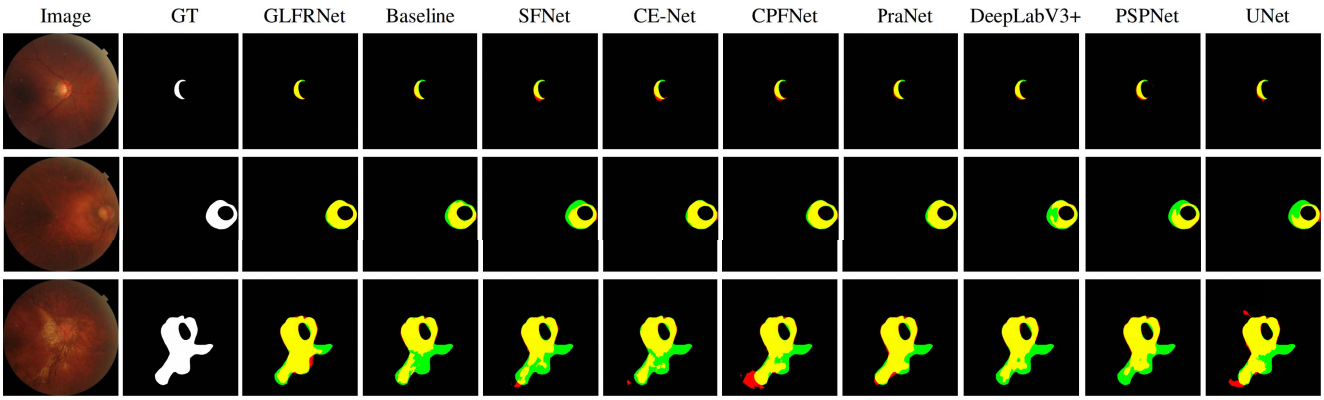

Fig. 6. Visual comparison between GLFRNet and state-of-the-art networks for choroidal atrophy segmentation. Green, red and yellow regions represent the false negative, false positive and true positive, respectively.

TABLE III
THE RESULT OF COMPARISON EXPERIMENTS AND ABLATION STUDIES ON MULTI-CLASS RETINAL FLUID SEGMENTATION TASK

	Dice(%)				IoU(%)				Sen(%)				Acc(%)	p-value
Methods	Ave	PED	SRF	IRF	Ave	PED	SRF	IRF	Ave	PED	SRF	IRF	Glob	Dice
EMANet [32]	70.78	68.67	74.74	68.92	57.60	55.99	63.04	53.78	69.27	67.61	70.83	69.37	99.10	<0.001
PSPNet [19]	71.26	68.76	76.13	68.90	58.07	56.03	64.39	53.80	70.29	67.08	75.12	68.66	99.13	<0.001
CENet [18]	71.40	68.22	76.36	69.62	58.60	55.59	65.06	55.15	70.64	67.73	75.61	68.59	99.17	<0.001
U-Net [13]	71.69	66.38	76.88	71.81	58.93	53.76	65.65	57.39	70.87	65.64	76.49	70.47	99.17	<0.001
Attention U-Net [23]	72.17	67.33	77.09	72.10	59.38	54.68	65.72	57.74	71.27	67.29	76.04	70.47	99.17	<0.001
GCN [36]	72.50	70.21	75.70	71.58	59.63	57.52	64.46	56.90	73.04	70.44	76.01	72.66	99.15	<0.001
CPFNet [26]	73.01	69.99	76.59	72.45	60.12	57.12	65.41	57.83	72.38	67.79	76.24	73.12	99.17	<0.001
DeepLabV3+ [17]	73.77	71.78	78.48	71.05	60.71	58.69	67.09	56.33	76.29	76.20	79.05	73.62	99.15	<0.001
SFNet [27]	73.83	71.22	78.24	72.04	61.31	59.14	67.16	57.63	72.27	71.13	75.20	70.48	99.24	<0.001
UNet++ [35]	74.10	71.38	78.01	72.90	61.38	58.53	67.02	58.59	74.42	71.87	78.29	73.10	99.21	<0.001
Baseline	73.05	70.02	76.67	72.45	60.24	57.08	65.47	58.17	71.31	67.61	74.30	72.02	99.21	<0.001
Baseline+GFR	75.02	73.35	79.25	72.47	62.50	60.97	68.13	58.41	75.47	**76.41**	79.02	70.98	99.24	0.045
Baseline+LFR	74.55	71.99	78.95	72.73	62.01	59.70	67.83	58.50	74.03	71.76	79.42	70.90	99.23	0.003
GLFRNet	**76.11**	**74.63**	**79.78**	**73.92**	**63.73**	**62.25**	**69.05**	**59.90**	**76.85**	76.02	**80.65**	**73.87**	**99.28**	-

C. Multi-Class Retinal Fluid Segmentation

Retinal fluid refers to the accumulation of the leaked fluid within the intercellular space of the retina due to the disruptions in blood-retinal barrier, which mainly includes three types: intra-retinal fluid (IRF), sub-retinal fluid (SRF) and pigment epithelial detachment (PED). Retinal fluid is the pathological manifestation of many fundus diseases in macula, such as diabetic retinopathy (DR), age-related macular degeneration (AMD), and retinal vein occlusion (RVO).

The multi-class retinal fluid segmentation task is performed on the public dataset from the MICCAI2017 RETOUCH Challenge [44], which consists of 70 OCT volumes (6936 B-scan slices) with three types of retinal fluid. We resize each image and crop out a 256×512 region-of-interest (ROI) in each image according to the pixel intensity distribution. The 70 OCT volumes are vendor evenly divided into 23, 23 and 24. A 3-fold cross validation strategy is performed both in ablation experiments and comparison experiments. In the test phase, each slice in the volume is processed separately and then recombined into one volume to calculate 3D evaluation metrics including Dice, IoU, Accuracy (Acc) and Sensitivity (Sen). Online random contrast, brightness transformation, left-right flipping and Gaussian blur are applied for data augmentation.

As shown in Table III, the proposed GLFRNet is compared with ten other excellent networks. Since the fluid targets are often small in the OCT images, the networks without skip connections such as EMANet [32] and PSPNet [19] show poor performances. UNet++ [35] achieves a better performance, due to the dense connections of the features in skip connections. It is worth noting that although DeepLabV3+ [17] achieves comparable results with the proposed GLFRNet in terms of sensitivity (Sen), its IoU and Dice are far behind those of GLFRNet, indicating that Deeplabv3+ has over-segmentation problems in all three types of fluids. The proposed GLFRNet still significantly outperforms other excellent networks on this task. Relying on the GFR and LFR modules, our method has achieved stable improvements in the segmentation of three types of retinal fluids, which proves

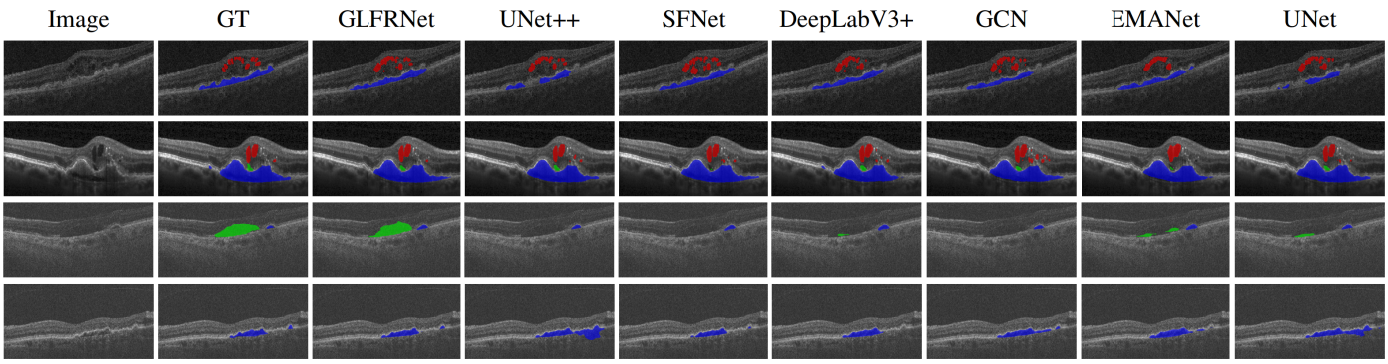

Fig. 7. Visual comparison between GLFRNet and state-of-the-art networks for multi-class retinal fluid segmentation, where the red, green and blue regions denote the IRF, SRF and PED, respectively. Best view in color and zoom in.

TABLE IV
THE RESULT OF COMPARISON EXPERIMENTS AND ABLATION STUDIES ON MULTI-ORGAN SEGMENTATION TASK

	Dice(%)									HD(mm)	*p*-value
Methods	Ave	spleen	right kidney	left kidney	gallbladder	liver	stomach	aorta	pancreas	Ave	Dice
U-Net [13]	75.09	88.64	86.19	87.22	40.80	93.92	69.66	86.42	47.89	21.17	<0.001
Attention U-Net [23]	76.58	89.60	84.43	87.88	46.31	93.95	72.78	87.09	50.58	21.68	<0.001
UNet++ [35]	76.59	89.79	88.65	89.23	33.17	95.12	73.93	88.13	54.67	21.98	<0.001
EMANet [32]	77.76	90.15	87.09	88.31	41.26	94.19	77.76	87.86	55.42	13.21	<0.001
DeepLabV3+ [17]	77.83	89.42	87.51	90.06	44.93	94.51	75.54	88.10	52.57	16.73	<0.001
CENet [18]	78.16	90.57	87.60	89.47	42.03	94.81	76.22	87.48	57.14	19.96	<0.001
PSPNet [19]	78.92	90.64	87.69	89.07	46.08	94.22	77.21	88.13	58.36	15.17	<0.001
CPFNet [26]	80.09	89.62	88.98	89.59	51.07	95.04	77.86	89.18	59.40	12.02	<0.001
SFNet [27]	82.94	91.60	**89.98**	90.01	62.91	95.57	78.94	89.65	64.90	10.72	<0.001
Baseline	80.63	90.37	88.60	90.56	53.28	94.63	77.11	88.53	61.96	14.41	<0.001
Baseline+GFR	83.33	90.84	89.93	**91.44**	60.61	95.71	81.57	89.93	66.61	9.14	0.004
Baseline+LFR	83.40	91.68	89.18	91.33	60.22	95.61	81.11	**90.43**	67.69	11.43	0.007
GLFRNet	**84.79**	**92.59**	89.20	90.58	**67.11**	**95.89**	**83.91**	90.28	**68.72**	**8.55**	-

the effectiveness of the two proposed modules. As can be seen from Fig.7, GLFRNet can adapt to fluids with various shapes and low contrast better and achieve better segmentation performance.

D. Multi-Organ Segmentation

The segmentation of abdominal organs is important for clinical diagnosis and treatment planning of related diseases [45]. Recently, the performances of deep learning based methods in multi-organ segmentation are greatly improved compared with the classical methods which are based on statistical shape models [46] or multi-atlas segmentation [12].

We apply the proposed GLFRNet on 30 abdominal CT scans (3779 axial slices in total) from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge with 8 types of organ targets including spleen, right kidney, left kidney, gallbladder, liver, stomach, aorta and pancreas [47]. In order to use the contextual information in 3D space, we convert the 3D CT slices to 2.5D data to train the proposed GLFRNet. Specifically, we combine each axial slice and its two adjacent slices into a 3-channel image, which is taken as the input of GLFRNet. The output of GLFRNet corresponds to the prediction result of the middle slice. Same as in the multi-class retinal fluid segmentation task, we combine the prediction of each slice into one volume to calculate the 3D evaluation metrics. The 2.5D data processing strategy is also adopted for all the networks in the comparison and ablation experiments, which are shown in Table IV. The 3-fold cross validation strategy is adopted both in comparison and ablation experiments. Online data augmentation is performed including random contrast enhancement and random brightness enhancement.

Table IV shows the results of comparison experiments between the proposed GLFRNet and other excellent networks and ablation experiments of GFR and LFR modules. Evaluation metrics include Dice and Hausdorff distance (HD). Relying on the semantic flow to achieve semantic alignment between feature maps, SFNet [27] has achieved good performance in this task and the above three tasks, while our LFR module can achieve better spatial information recovery (Baseline+LFR). With the equipping of GFR and LFR modules, our GLFRNet outperforms other methods in the segmentation of each organ, and the average Dice is improved by 4.16% compared to the Baseline. The p-values calculated

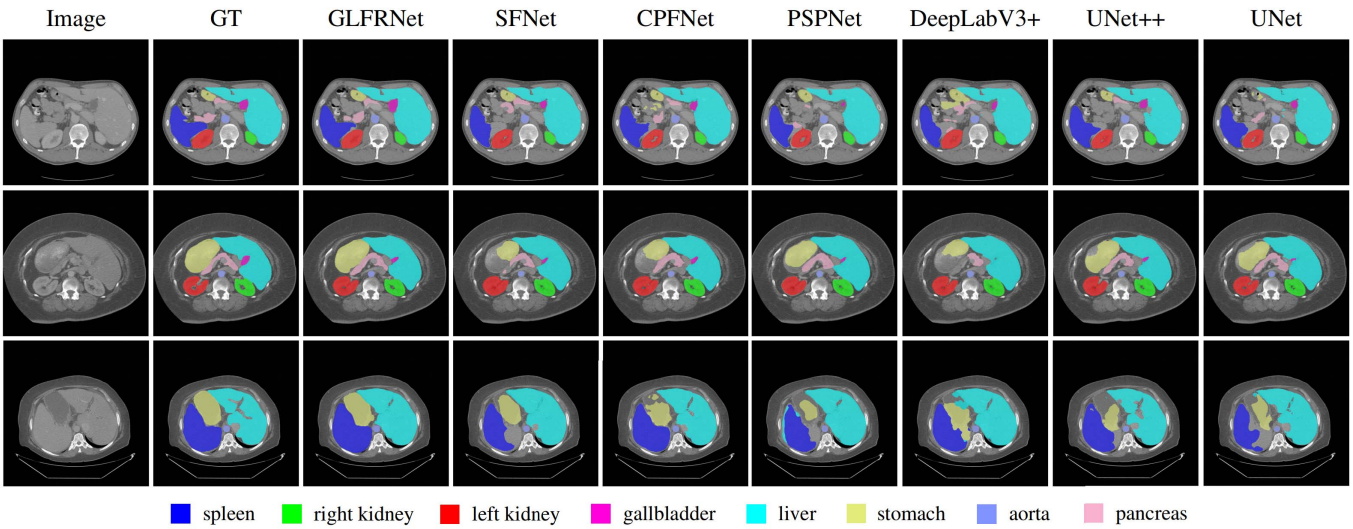

Fig. 8. Visual comparison between GLFRNet and state-of-the-art networks for multi-organ segmentation. Best view in color and zoom in.

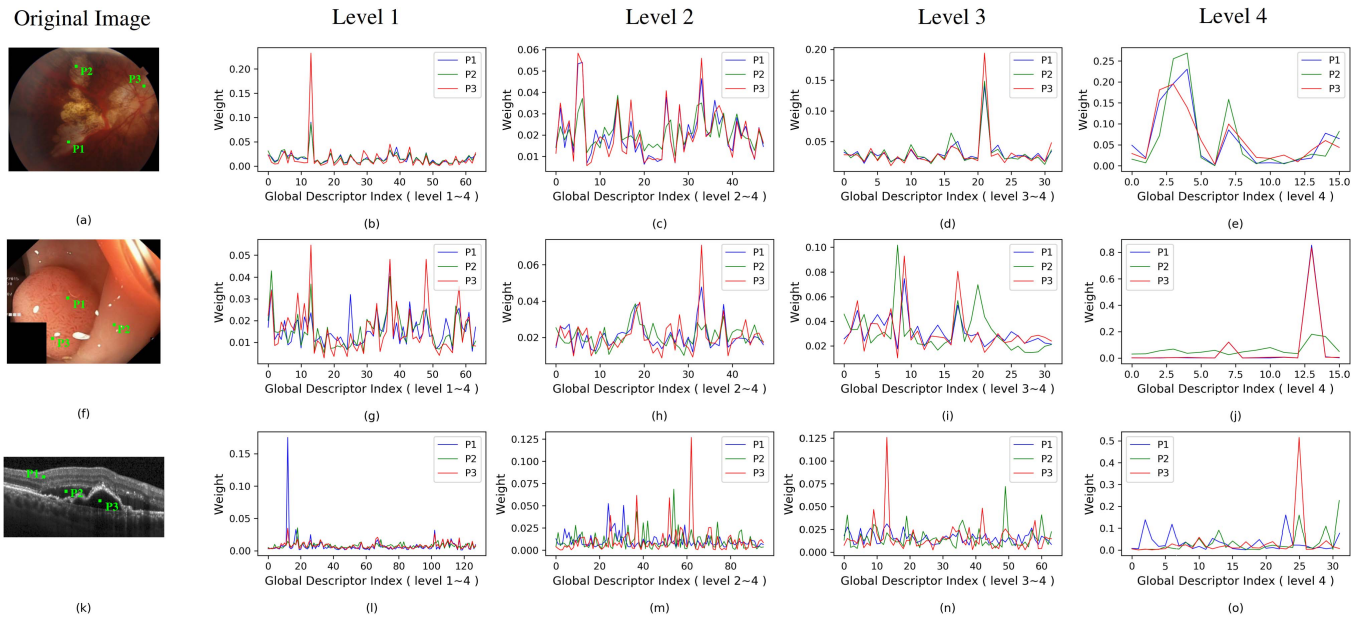

Fig. 9. The reconstruction weight visualization results of each global descriptor in the GFR module at some specific locations. The horizontal axis represents the index of the cross-level global descriptors, and the vertical axis represents the corresponding reconstruction weight. According to the different number of feature levels and target categories in different segmentation tasks, the numbers of cross-level global descriptors are different. P1, P2 and P3 indicate different locations of the images. In the GFR module, pixels of the same category tend to choose consistent global descriptors for feature reconstruction, while pixels with different semantic information adaptively choose different global descriptors. Best view in color and zoom in.

based on the average Dice shows the improvements are statistically significant. As can be seen from Fig.8, GLFRNet can segment both small organs such as gallbladder and large organs such as stomach more accurately.

IV. Discussion and Conclusion

A. Analysis of GFR Module

To explain how our proposed GFR module works, we visualize the reconstruction weights at different locations (P1, P2 and P3 shown in Fig.9) in the input images. As shown in Fig.9, except the fourth-level feature that only uses its own global descriptor for reconstruction, features at levels 1, 2 and 3 are all reconstructed by adaptively selected cross-level global descriptors. As can be seen from Fig.9(a), P1, P2 and P3 belong to choroidal atrophy, though they are far away from each other. As can be seen from Fig.9(b-e), the features of P1, P2 and P3 capture consistent global features at all levels, making the intra-class features more compact. In Fig.9(f), P1 and P3 are located in the colorectal polyp regions, and P2 is located in the background. In the low-level feature reconstruction, since the local features of the three locations

TABLE V
DICE COMPARISON OF THE LFR MODULE AND OTHER UP-SAMPLING METHODS ON FOUR TASKS(%). TASK1 TO TASK4 REPRESENT COLORECTAL POLYP SEGMENTATION, CHOROIDAL ATROPHY SEGMENTATION, MULTI-CLASS RETINAL FLUID SEGMENTATION AND MULTI-ORGAN SEGMENTATION, RESPECTIVELY

Methods	Task 1	Task 2	Task 3	Task 4
Bilinear	87.54	84.46	73.05	80.63
Transposed Conv	88.05	85.81	72.81	79.26
Sub-pixel Conv [28]	88.22	86.27	73.36	76.78
CARAFE [30]	87.96	86.22	74.36	81.83
LFR	**88.69**	**86.48**	**74.55**	**83.40**

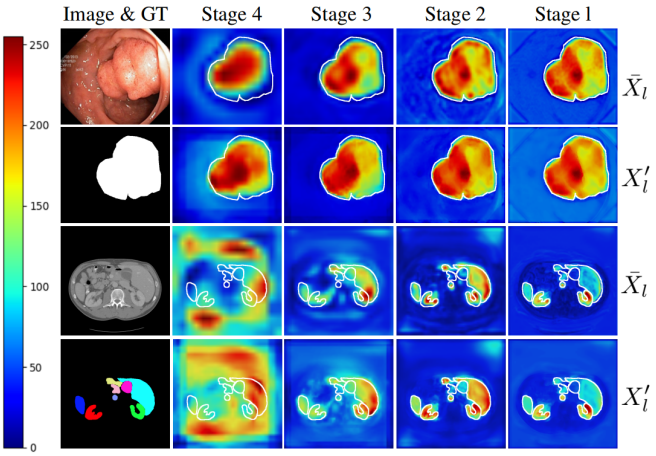

Fig. 10. Visual comparison of feature maps before and after the local feature reconstruction. $\bar{X}_l$ and X'_l represent the visualized feature maps before and after reconstruction in each stage of LFR module. Stage 1 to stage 4 correspond to the decoding process of the feature maps of level 1 to level 4 in the encoder, where the white curves represent the boundaries of the ground truth.

are similar, the reconstruction weights do not show obvious differences, as shown in the Fig.9(g) and (h). As the feature layer gets deeper, the difference between P2 and the other two locations become more obvious, while the reconstruction weights of P1 and P3 are more consistent, as shown in the Fig.9(i) and (j). As shown in Fig.9(k), P1, P2 and P3 belong to the background, SRF and PED respectively. The reconstruction weights shown in Fig.9(l)-(o) indicate that the GFR module can enable features at each location adaptively capture the global features. Based on the above analysis, it is shown that our GFR module can adaptively capture distinguishable global features according to the local features.

B. Analysis of LFR Module

The LFR module acts like an up-sampling function in the decoder stage. In order to verify that our LFR achieves a more flexible and robust feature up-sampling, we adopt other four up-sampling methods including bilinear interpolation, transposed convolution, sub-pixel convolution [28] and CARAFE [30] in the Baseline and compare the Dice coefficients.

As shown in the Table V, although transposed convolution gets better performances on Task1 (colorectal polyp segmentation) and Task2 (choroidal atrophy segmentation), it is worse than bilinear interpolation on Task3 (multi-class retinal fluid segmentation) and Task4 (multi-organ segmentation). Similarly, sub-pixel convolution exceeds bilinear interpolation on three of the tasks, while it is worse on Task4. CARAFE and our proposed LFR module have achieved stable improvements on all four tasks. Although CARAFE has achieved comparable results to LFR on Task2 and Task3, the overall improvement of the LFR module is better.

Fig.10 shows the visualized features of each level before($\bar{X}_l$, bilinearly upsampled) and after the local reconstruction (X'_l). As can be seen from Fig.10, with the guidance of low-level features, the misalignment of semantic boundaries caused by down-sampling is gradually corrected and the intra-class features become more consistent.

C. Conclusion

In this paper, we propose an end-to-end deep learning framework named GLFRNet for medical image segmentation. Two novel modules including global feature reconstruction (GFR) module and local feature reconstruction (LFR) module are designed to solve the problem of insufficient global context feature extraction and spatial information restoration in encoder-decoder structure network respectively. To validate our approach, we conduct experiments on four different segmentation tasks including colorectal polyp segmentation, choroidal atrophy segmentation, multi-class retinal fluid segmentation and multi-organ segmentation. The proposed GLFRNet achieves excellent performances on these four different segmentation tasks, which indicates that the proposed GLFRNet is more practical and universal than other state-of-the-art methods. Our future research direction is to try to apply the proposed global and local feature reconstruction mechanism to multi-modal data analysis, which may provide a flexible and efficient way for multi-modal data registration and fusion.

REFERENCES

[1] M. Akbari *et al.*, "Polyp segmentation in colonoscopy images using fully convolutional network," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 69–72.
[2] Y. Fang, C. Chen, Y. Yuan, and K.-Y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2019, pp. 302–310.
[3] X. Guo, C. Yang, Y. Liu, and Y. Yuan, "Learn to threshold: Thresholdnet with confidence-guided manifold mixup for polyp segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 4, pp. 1134–1146, Apr. 2021.
[4] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive context selection for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2020, pp. 253–262.
[5] A. Septiarini, R. Pulungan, A. Harjoko, and R. Ekantini, "Peripapillary atrophy detection in fundus images based on sectors with scan lines approach," in *Proc. 3rd Int. Conf. Informat. Comput. (ICIC)*, Oct. 2018, pp. 1–6.
[6] Y. Guo *et al.*, "Lesion-aware segmentation network for atrophy and detachment of pathological myopia on fundus images," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1242–1245.

[7] T. J. N. Rao, G. N. Girish, A. R. Kothari, and J. Rajan, "Deep learning based sub-retinal fluid segmentation in central serous chorioretinopathy optical coherence tomography scans," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 978–981.

[8] R. Tennakoon, A. K. Gostar, R. Hoseinnezhad, and A. Bab-Hadiashar, "Retinal fluid segmentation in OCT images using adversarial loss based convolutional neural networks," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1436–1440.

[9] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Drinet for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2453–2462, Nov. 2018.

[10] Y. Zhao, H. Li, S. Wan, A. Sekuboyina, X. Hu, G. Tetteh, M. Piraud, and B. Menze, "Knowledge-aided convolutional neural network for small organ segmentation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1363–1373, Jul. 2019.

[11] X. Fang and P. Yan, "Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3619–3629, Nov. 2020.

[12] T. Tong *et al.*, "Discriminative dictionary learning for abdominal multi-organ segmentation," *Med. Image Anal.*, vol. 23, no. 1, pp. 92–104, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841515000651

[13] 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. Cham, Switzerland: Springer, 2015, pp. 234–241.

[14] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," 2014, *arXiv:1412.6856*.

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

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

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

[18] 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.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jul. 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[20] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[22] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[23] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[24] S. Zhang *et al.*, "Attention guided network for retinal image segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2019, pp. 797–805.

[25] J. Fu *et al.*, "Adaptive context network for scene parsing," in *2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6747–6756.

[26] S. Feng *et al.*, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[27] X. Li *et al.*, "Semantic flow for fast and accurate scene parsing," in *Proc. ECCV*, 2020, pp. 775–793.

[28] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[29] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3121–3130.

[30] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3007–3016.

[31] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-Nets: Double attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 350–359.

[32] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9166–9175.

[33] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, 2018. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21492

[34] D. Jha *et al.*, "Kvasir-SEG: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2020, pp. 451–462.

[35] Z. Zhou *et al.*, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2020.

[36] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jan. 2017, pp. 1743–1751.

[37] D.-P. Fan *et al.*, "PraNet: Parallel reverse attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2020, pp. 263–273.

[38] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "HarDNet-MSEG: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS," 2021, *arXiv:2101.07172*.

[39] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," 2021, *arXiv:2102.08005*.

[40] P. Chao, C.-Y. Kao, Y. Ruan, C.-H. Huang, and Y.-L. Lin, "HarDNet: A low memory traffic network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3551–3560.

[41] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers distillation through attention," 2020, *arXiv:2012.12877*.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 3–19.

[43] K. Ohno-Matsui, T. Y. Y. Lai, C. C. Lai, and C. M. G. Cheung, "Updates of pathologic myopia," *Prog. Retinal Eye Res.*, vol. 52, pp. 156–187, 2016.

[44] H. Bogunović *et al.*, "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1858–1874, Aug. 2019.

[45] E. Gibson *et al.*, "Automatic multi-organ segmentation on abdominal CT with dense v-networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1822–1834, Aug. 2018.

[46] T. Okada *et al.*, "Automated segmentation of the liver from 3D CT images using probabilistic atlas and multi-level statistical shape model," in *Medical Image Computing and Computer Assisted Intervention*, N. Ayache, S. Ourselin, and A. Maeder, Eds. Berlin, Germany: Springer, 2007, pp. 86–93.

[47] *MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge*. [Online]. Available: https://www.synapse.org/#!Synapse:syn3193805/wiki/217789