OPEN ACCESS
Journal of Innovative Optical Health Sciences
Vol. 16, No. 4 (2023) 2243001 (11 pages)

DOI: 10.1142/S1793545822430015

World Scientific
www.worldscientific.com

LGNet: Local and global representation learning for fast biomedical image segmentation

Guoping Xu*, Xuan Zhang*, Wentao Liao*, Shangbin Chen† and Xinglong Wu*,†,‡

*School of Computer Science & Engineering,
Hubei Key Laboratory of Intelligent Robot,
Wuhan Institute of Technology, Wuhan, Hubei 430205, P. R. China

†Britton Chance Center for Biomedical Photonics,
Wuhan National Laboratory for Optoelectronics-Huazhong
University of Science and Technology, Wuhan, Hubei 430074, P. R. China
‡xwu@wit.edu.cn

Received 16 February 2022
Accepted 28 March 2022
Published 13 September 2022

Medical image segmentation plays a crucial role in clinical diagnosis and therapy systems, yet still faces many challenges. Building on convolutional neural networks (CNNs), medical image segmentation has achieved tremendous progress. However, owing to the locality of convolution operations, CNNs have the inherent limitation in learning global context. To address the limitation in building global context relationship from CNNs, we propose **LGNet**, a semantic segmentation network aiming to learn local and global features for fast and accurate medical image segmentation in this paper. Specifically, we employ a two-branch architecture consisting of convolution layers in one branch to learn local features and transformer layers in the other branch to learn global features. LGNet has two key insights: (1) We bridge two-branch to learn local and global features in an interactive way; (2) we present a novel multi-feature fusion model (MSFFM) to leverage the global contexture information from transformer and the local representational features from convolutions. Our method achieves state-of-the-art trade-off in terms of accuracy and efficiency on several medical image segmentation benchmarks including Synapse, ACDC and MOST. Specifically, LGNet achieves the state-of-the-art performance with Dice's indexes of 80.15% on Synapse, of 91.70% on ACDC, and of 95.56% on MOST. Meanwhile, the inference speed attains at 172 frames per second with 224×224 input resolution. The extensive experiments demonstrate the effectiveness of the proposed LGNet for fast and accurate for medical image segmentation.

Keywords: CNNs; transformers; segmentation; medical image; contextual information.

‡Corresponding author.

1. Introduction

Image segmentation plays an important role in medical image analysis. Particularly, it is widely used for quantified analysis of anatomical structure in clinical diagnosis.[1] With the development of deep learning technology, convolutional neural networks (CNNs) have made substantial progress in medical image segmentation, especially fully convolutional networks (FCNs)[2] and its variants, like U-Net,[3] DeepLab,[4] which has become the *de-facto* choice. Based on these approaches, much progress has been achieved in medical applications such as vessel segmentation for thoracic CT,[5] cardiac segmentation for MRI,[6] and lymph node segmentation.[7]

Although such FCN-based methods have exceptional representational power, their capability to capture explicit global context and long-range dependencies is limited due to the local receptive fields of convolution operation.[8] Such limitations in capture multi-scale contextual information result in sub-optimal segmentation of structures with variable shapes and scales (e.g., pathological lymph nodes with various sizes[7]). Previous works have been tried to migrate this issue by using dilated convolution in DeepLab,[4] feature pyramid pooling in PSPNet,[9] self-attention mechanisms in Attention U-Net[10] etc. However, there, studies still cannot fully extract global contexture features in the task of medical image segmentation.

Transformer-based models[11,12] were proposed for sequence-to-sequence modeling in nature language processing (NLP) domain, and have achieved state-of-the-art results in various tasks. The self-attention mechanism in the transformers enables them to learn its long-range dependencies and build global relations between sequences. Recently, several attempts have been made by the introduction of Vision Transformer[13] into the field of computer vision, and they have achieved state-of-the-art benchmarks in the task of image classification. Later, more works have been proposed based transformer for semantic segmentation such as SETR,[14] Swin Transformer,[15] TransUNet,[8] Swin-UNet,[16] DS-TransUNet,[17] TransFuse,[18] VOLO,[19] and UNETR[20] etc. However, such transformer-based methods suffer from extreme computational and spatial complexities to model the log-range dependency on the extracted feature maps, which impede them to apply in real-time clinical diagnosis such as radiotherapy.

The main contributions are summarized as follows:

- A novel two-branch network is proposed to bridge convolutional neural network and transformer for fast medical image segmentation.
- We present a new multi-feature fusion block to leverage the contexture information from transformer and the local representational features from convolution.
- Extensive experiments demonstrate that our LGNet achieves competitive performance with other state-of-the-art methods in terms of accuracy and efficiency.

We will discuss and compare image segmentation architectures, and focus on light-weight real-time methods with low memory requirements in this section.

1.1. *Fast segmentation architecture*

State-of-the-art semantic segmentation CNNs usually adopt two basic architectures: An encoder–decoder architecture and two or multiple branches architecture. Figure 1 illustrates these two architectures and their variants.

The initial representational work based on the encoder–decoder structure begins from U-Net[3] and SegNet.[21] The encoder extracts contextual information with a stack of coevolution and pooling operations, while the decoder recovers the spatial details from low-resolution features to conduct dense prediction with interpolation or transpose convolution. Some works try to improve the segmentation performance by introducing dilated convolution module,[4] self-attention mechanism,[10] pyramid pooling module,[9] and so on. However, most of them ignore the inference time, which impedes their applications. Inspired by the work of efficient image classification, such as SqueezeNet,[22] MobileNet,[23] and ShuffleNet,[24] some works are conducted for real-time semantic segmentation adopted by encoder–decoder architectures, like ENet,[25] LEDNet,[26] ESPNet, ESPNetv2, ERFNet,[27] etc. More recently, two/multi-branches architecture are introduced for fast semantic segmentation, like BiSeNet,[28,29] Fast-SCNN,[30] ContextNet,[31] DABNet,[32]

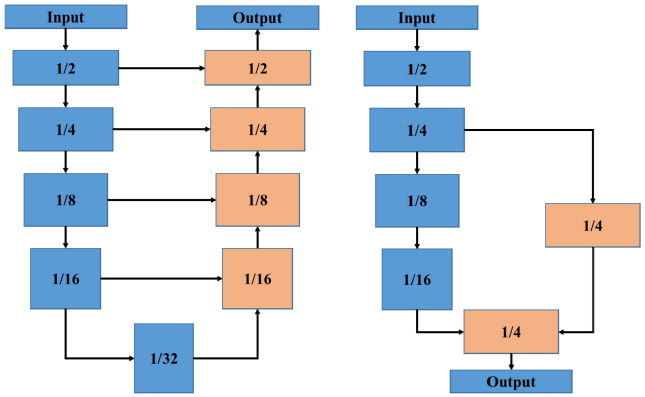

Fig. 1. Schematic comparison of our LGNet (Right) with encoder–decoder and two-branch (and its variant) architectures.

CABiNet,[33] DDRNet,[34] ICNet,[35] etc. The one deep branch learns global context information with downsampled features, while the other shallow branch is used to learn local features, like boundaries and texture, at relative high resolution.

Although much progress has been achieved, how to better address the accuracy and efficiency remains challenging for these state-of-the-art sematic segmentation methods, in which the core is how to learn more distinctive features in a fast way. Inspired by two-branch methods, we re-design a new type of architecture, taking both of the low-level details from one branch with convolution-based operations and high-level semantics information with transformer-based operations into consideration. Meanwhile, we take an iterative way to communicate the features from both branches, resulting in high accuracy without losing its efficiency.

1.2. *Multi-scale feature fusion module*

In recent years, multi-scale context information has proven to be a key factor to improve the segmentation performance.[9,34,36] Atrous Spatial Pyramid Pooling (ASPP)[4] extracts multi-scale context information by a parallel atrous convolution layers with different rates. DenseASPP[37] is proposed to concatenate a set of atrous-convolved features from atrous convolutional layers in a dense way. Due to the complex computation, Pyramid Pooling Moudle (PPM) is proposed in PSPNet[9] to exploit the global context information on multi-scale features. PPM variants are adopted in many fast segmentation architectures such as Fast-SCNN,[30] ICNet,[35] DDRNet[34] etc. All these methods rely on convolution operations to extract multi-scale context information, which is limited to the local reception field of convolution kernels. In this paper, we strengthen the PPM module with integration of the multi-scale features from convolution layers and transformer layers, which aims to improve the segmentation accuracy without the sacrifice of inference speed.

In this paper, we attempt to leverage the power of convolutions in local feature learning and transformers in global feature learning by introducing a novel architecture to do highly efficient semantic segmentation. Specifically, we re-design a two-branch segmentation architecture by integration with convolution blocks and transformer blocks in an interactive way, named as Local and global feature learning network (**LGNet**). The global representations from transformer blocks are merged into convolution blocks. Meanwhile, the extracted features from convolution are integrated with transformer blocks. Owing to this novel two-branch parallel feature learning architecture, the efficiency of segmentation is competitive with other state-of-the-art fast segmentation convolution-based methods. Extensive experiments have validated the segmentation performance of the proposed LGNet on three datasets. Furthermore, we investigate the speed and parameters of LGNet, and demonstrate the competitive performance in comparison to other fast segmentation convolution-based methods. To the best of our knowledge, we are the first to propose a bilateral network by integration features from transformer and convolution block for fast semantic segmentation in medical image analysis.

2. Materials and Methods

2.1. *Datasets: Synapse, ACDC and MOST*

Synapse multi-organ segmentation dataset (Synapse[a]): This dataset includes 30 abdominal CT scans from MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, having 3779 axial contrast-enhanced abdominal clinical CT images in total. Following the splitting way in Ref. 8, we use 18 cases for network training and the other 12 cases for validation. We evaluated the performance with the average dice similarity coefficient (DSC) and the average Hausdorff distance (HD) of each volume on eight abdominal organs, which are aorta, gallbladder (Gall), left kidney (KidL), right kidney (KidR), liver, pancreas (Panc), spleen, and stomach (Stom), respectively.

Automated cardiac diagnosis challenge dataset (ACDC[b]): The automated cardiac diagnosis challenge (ACDC) dataset is collected from 150 patients using cine-MR scanners, splitting into 100 volumes with human annotations and the other 50 volumes which are private for the evaluation purpose. Here, we split the 100 annotated volumes into 80 training samples and 20 testing samples. Each volume is annotated with ground truth for left ventricle (LV), right ventricle (RV) and myocardium (MYO).

Micro-optical sectioning tomography (MOST): The MOST dataset is provided from Huazhong University of Science and Technology,[42] which includes part of mouse brain with micro-optical sectioning tomography. The voxel size of the dataset is $0.35 \times 0.35 \times 1\,\mu$m. The size of a representative image stack is $512 \times 512 \times 1200$ with the human annotations for both soma and vessels within the stack. Here, we randomly split the dataset into training and test set, which consists of 1100 and 100 images with the same resolution.

2.2. *Network architecture*

LGNet is designed to learn more representational features from convolution and transformer layers for medical image segmentation. The architecture of the proposed architecture is shown in Fig. 2. The LGNet has two parallel deep branches. One branch is composed by a series of convolution blocks with the same relatively high-resolution feature maps, which learns local recreational features like boundaries and shapes. The other branch includes three transformer blocks with downsampling operations, which is used to extract rich global contextual information based on self-attention mechanism. The feature maps from convolution and transformer blocks are fused by multiple bilateral connections in an interactive way. Besides, we propose a novel multi-scale feature fusion module named MSFFM (shown in the third row in Fig. 2) which increases contexture information from transformer and local detail information from convolutions.

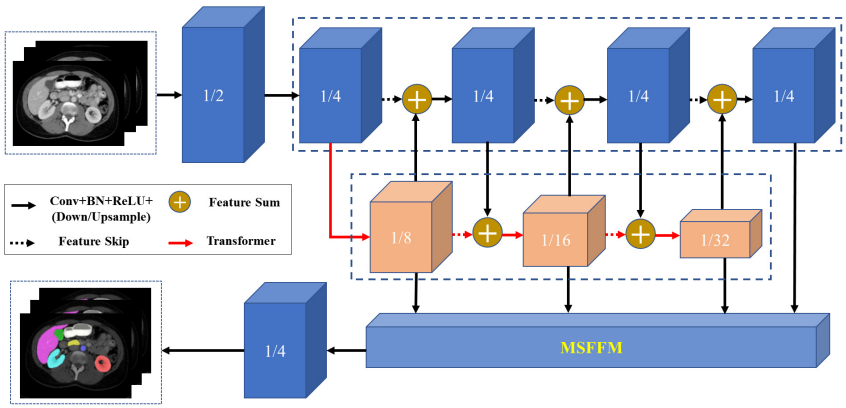

Fig. 2. Architecture of LGNet: CNN blocks (colored in blue), transformer blocks (color in orange) and multi-feature fuse (MSFF) block from top to down.

[a]https://www.synapse.org/#!Synapse:syn3193805/wiki/217789.
[b]https://www.creatis.insa-lyon.fr/Challenge/acdc/.

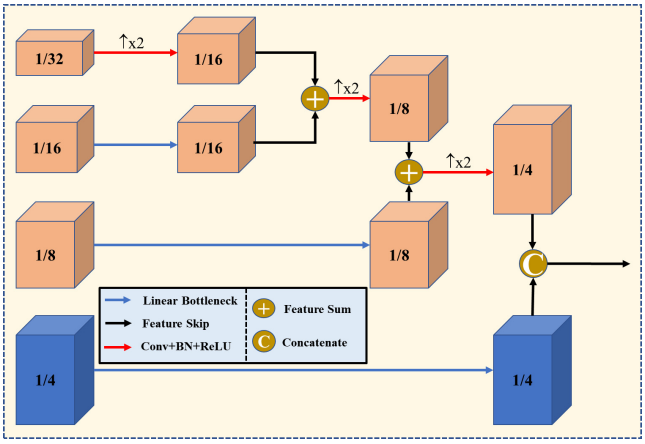

Fig. 3. The diagram of MSFFM, features from transformer colored in orange and features from convolution colored in blue.

CNNs as high-resolution local features extractor

In Fig. 2, we design a stacks of convolution blocks to extract relatively high-resolution features in one branch which colored in blue. The first block is used to learn high-resolution features with setting the stride two of 3×3 convolution kernel considering to trade-off the efficiency and accuracy. The other blocks consist of convolutions to extract features with the size of $1/4$ resolution.

Transformer as long-range relation catcher

Transformer-based module is aimed at capturing the global and long-range context for image segmentation. In contrast to conventional convolution-based two-branch architectures, one branch of LGNet, colored in orange in Fig. 2, is taken as the variant of transformer, named "Outlooker" initially proposed in Ref. 19, to build long-range relationship. The Outlooker module consists of an outlook attention layer for encoding spatial finer-level features and a multi-layer perceptron for inter-channel information interaction. It can be written as follows:

$$\hat{X} = \text{OutlookAtt}(\text{LN}(X)) + X, \quad (1)$$

$$Z = \text{MLP}(\text{LN}(\hat{X})) + \hat{X} \quad (2)$$

Here, X denotes the input token representations and LN refers to LayerNorm.[38]

Multi-scale feature fusion module

Earlier works on object detection[39,40] and segmentation[4,9] have shown their effectiveness via the fusion of multi-scale features. However, it is not known whether the segmentation performance could be further improved by integration of the features from convolution layers and transformer layers. We thus present a novel multi-scale fusion module, named MSFFM. Figure 3 shows the detail of the proposed **MSFFM**. For the low-resolution feature maps, we use a linear bottleneck[41] and interpolation operation, which obtain the same dimension and resolution feature maps, comparing to the previous outputting. For example, feature maps of $1/32$ input image resolution will have the same dimension of feature maps after linear bottleneck, and have the same size after upsampling.

3. Results and Discussion

3.1. *Implementation details*

We run all experiments based on Python 3.8, PyTorch 1.8.0 and Ubuntu 18.04.1 LTS. For all training samples, we apply augmentation strategies such as random flipping and rotations to increase data diversity. All trainings are performed on images size of 224x224 by using a Nvidia 3090 GPU with 24GB memory. For all models are trained by Adam optimizer with learning rate 1e-5 and weight decay of 1e-4. The cross entropy and Dice loss are used as objective function. We set batch size of 8, and the training epochs are set as 350 and 400 for Synapse and ACDC dataset, respectively. Following Ref. 8, all 3D volume datasets are trained by slice and the predicted 2D slice are stacked together to build 3D prediction for evaluation.

3.2. *Speed and accuracy comparisons on synapse dataset*

The results in Table 1 demonstrate that our LGNet achieves the competitive results between accuracy and efficiency. Specifically, comparing to the convolution-based light-weight fast segmentation methods, such as ENet and Fast-SCNN, our LGNet achieves the state-of-the-art performance in terms of mean DSC and HD. In addition, the inference speed of our methods, which is 172 slice per second,

Table 1. Accuracy and speed comparison on Synapse dataset (average DSC % and average HD in mm and DSC % for each organ).

Methods	DSC↑	HD↓	Aorta	Gall	KidL	KidR	Liver	Pancreas	Spleen	Stom	Params(M)	FLOPs(M)	FPS
CGNet[43]	75.08	24.99	83.48	**65.32**	77.91	72.04	91.92	57.37	85.47	67.15	0.49	0.66	124
ContextNet[43]	71.17	36.41	79.92	51.17	77.58	72.04	91.74	43.78	86.65	66.51	0.87	0.16	280
DABNet[43]	74.91	26.39	85.01	56.89	77.84	72.45	93.05	54.39	88.23	71.45	0.75	0.99	221
EDANet[43]	75.43	29.31	84.35	62.31	76.16	71.65	93.20	53.19	85.47	**77.12**	0.69	0.85	213
ENet[43]	77.63	31.83	85.13	64.91	81.10	77.26	93.37	57.83	87.03	74.41	0.36	0.50	141
FPENet[43]	68.67	42.39	78.98	56.35	74.54	64.36	90.86	40.60	78.30	65.35	0.11	0.14	160
FSSNet[43]	74.59	35.16	82.87	64.06	78.03	69.63	92.52	53.10	85.65	70.86	0.17	0.33	213
SQNet[43]	73.76	40.29	83.55	61.17	76.87	69.40	91.53	56.55	85.82	65.24	16.25	18.47	241
FastSCNN[43]	70.53	32.79	77.79	55.96	73.61	67.38	91.68	44.54	84.51	68.76	1.14	0.16	292
LGNet	**80.15**	21.21	**88.06**	66.48	83.98	**81.57**	94.00	**59.99**	**90.90**	76.21	3.06	3.27	**172**

is more competitive comparing with CGNet, ENet, and FPENet on Synapse dataset. In summary, our LGNet can have better trade-off in terms of the segmentation accuracy and the prediction efficiency.

Comparisons with state-of-the-art results

We also conduct experiments on Synapse dataset to demonstrate the capacity of our LGNet by comparing to state-of-the-art offline models, such as U-Net[3] and TransUnet.[8] The experimental results are presented in Table 2. We can see that our method achieves the best performance of DSC with segmentation accuracy of 80.15% and the competitive performance of HD with 21.21 mm. Compared with convolution-based methods like R50 Att-UNet and Deeplabv3+, our method improves the performance of DSC with 4.58% and 4.42%, and the performance of HD with 15.76 mm and 5.72 mm, respectively. Compared with the methods with transformer-based or integration of convolution- and transformer-based, like Swin-Unet, or TransUNet and LeViT-UNet,[43] LGNet could improve the DSC again, while keep the HD in the competitive result. It is noteworthy that our method without any pre-train, yet the Swin-Unet, TransUNet and LeViT-UNet are pretrained on ImageNet dataset to initialize the model parameters.

The qualitative comparing three different methods (TransUNet, UNet, and DeepLabv3+) on the Synapse dataset is demonstrated in Fig. 4. We can see that the other three methods are more likely to under-segment or over segment the organs, for example, the stomach is under-segmented by TransUNet (as indicated by the red arrow in the middle of the third row), and over-segmented by UNet (as indicated by the red arrow in the fourth panel of the second row). Moreover, results in the second row demonstrate that our LGNet outputs are relatively better than those

Table 2. Comparison on the Synapse abdominal CT dataset (average DSC% and average HD in mm, and DSC% for each organ).

Methods	DSC↑	HD↓	Aorta	Gall	KidL	KidR	Liver	Pancreas	Spleen	Stom
V-Net[8]	68.81	—	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR[8]	69.77	—	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
U-Net[16]	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
R50 U-Net[8]	74.68	36.87	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
R50 Att-UNet[8]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-Unet[16]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50-Deeplabv3+	75.73	26.93	86.18	60.42	81.18	75.27	92.86	51.06	88.69	70.19
TransUNet[8]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet[16]	79.13	21.55	85.47	66.53	83.28	79.61	**94.29**	56.58	90.66	**76.60**
LeVit-Unet-128s[45]	73.69	23.92	86.45	66.13	79.32	73.56	91.85	49.25	79.29	63.70
LeVit-Unet-192[45]	74.67	18.86	85.69	57.37	79.08	75.90	92.05	53.53	83.11	70.61
LeVit-Unet-384[45]	78.53	**16.84**	87.33	62.23	**84.61**	80.25	93.11	59.07	88.86	72.76
LGNet	**80.15**	21.21	88.06	66.48	83.98	**81.57**	94.00	**59.99**	**90.90**	76.21

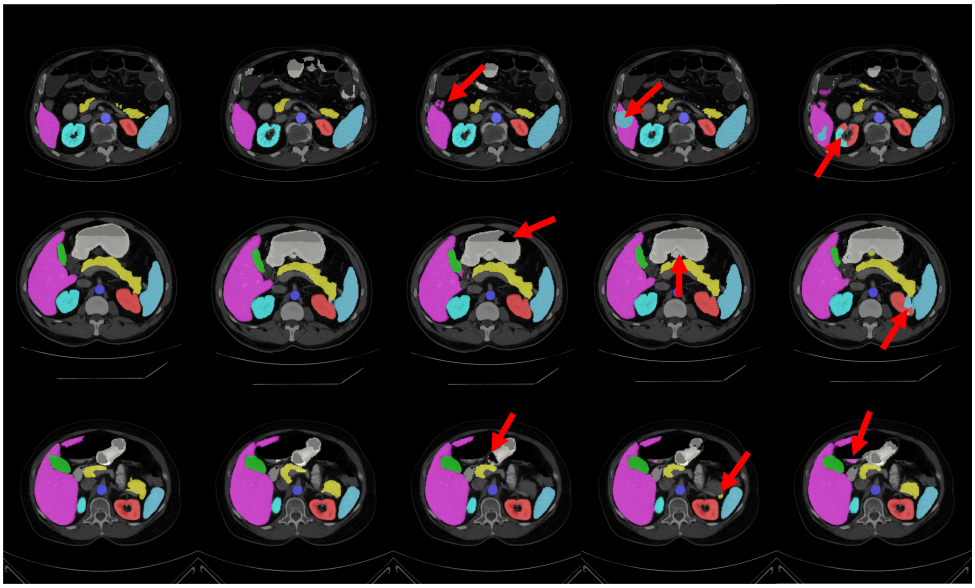

Fig. 4. Qualitative comparison of various methods by visualization From Left to right: Ground Truth, LGNet, TransUNet, UNet, and DeepLabv3+.

from other methods, which indicates that our method has more advantageous in object segmentation prediction.

Ablation study

We perform two kinds of ablation studies to evaluate effectiveness of the proposed LGNet architecture and validate performance of the proposed MSFFM.

Effect of the interaction between convolution block and transformer block: We exploit the influence of interaction by skip-connections between convolution block and transformer block. As is shown in Table 3, we can find that the DSC is boosted to 2.77%, and the HD is improved 7.34 mm

Table 3. Comparison the effectiveness of the interaction of the feature maps from convolution block and transformer block (%).

Methods	DSC↑	HD↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
LGNet	**80.15**	21.21	88.06	66.48	83.98	**81.57**	94.00	59.99	**90.90**	76.21
LGNet-No-Skip	77.38	28.55	88.53	61.09	79.28	75.37	94.08	57.90	90.66	72.13

Table 4. Ablation study w/o MSFFM (%).

Methods	DSC↑	HD↓	Aorta	Gall	KidL	KidR	Liver	Pancreas	Spleen	Stom
LGNet	**80.15**	21.21	88.06	66.48	83.98	**81.57**	94.00	59.99	**90.90**	76.21
LGNet-No-MSFFM	79.97	23.67	**89.28**	**71.91**	83.14	77.61	93.70	**61.27**	89.43	76.83

Table 5. Comparison the effectiveness of the feature maps from convolution block and transformer block for the MSFFM (%). Here, the LGNet-MSFFM-Trans means the feature maps to the MSFFM block are from transformer blocks. The LGNet-MSFFM-Convs represents the feature maps to the MSFFM are from convolution layer.

Methods	DSC↑	HD↓	Aorta	Gall	KidL	KidR	Liver	Pancreas	Spleen	Stom
LGNet	**80.15**	21.21	88.06	66.48	83.98	81.57	94.00	59.99	90.90	76.21
LGNet-MSFFM-Trans	76.06	**16.17**	83.71	58.43	81.82	78.87	92.54	52.56	88.07	72.48
LGNet-MSFFM-Convs	79.59	24.25	88.49	66.98	82.22	79.96	93.48	60.62	89.72	75.24

Table 6. Comparison of different methods on the ACDC dataset (%).

Methods	Dice	RV	MYO	LV
R50-UNet[8]	87.55	87.10	80.63	94.92
R50-Att-UNet[8]	86.55	87.58	79.20	93.47
R50-ViT[16]	87.57	86.07	81.88	94.75
TransUnet[8]	89.71	88.86	84.53	95.73
SwinUNet[16]	90.00	88.55	85.62	**95.83**
LeViT-UNet-128s[45]	89.39	88.16	86.97	93.05
LeViT-UNet-192[45]	90.08	88.86	87.50	93.87
LeViT-UNet-384[45]	90.32	89.55	87.64	93.76
LGNet	**91.70**	**91.48**	**89.12**	94.51

Table 7. Segmentation performance of different methods on the MOST dataset.

Models	Object	Precision	Recall	Dice	FPS
UNet	Vessel	89.75	95.29	92.35	73
	Soma	99.46	94.44	96.87	
SegNet	Vessel	80.87	83.71	82.1	66
	Soma	98.2	93.24	95.63	
FastSCNN	Vessel	69.87	71.59	70.69	**232**
	Soma	83.23	76.03	79.46	
ENet	Vessel	67.85	84.06	74.96	96
	Soma	90.71	86.03	88.3	
CGNet	Vessel	72.02	71.48	71.72	92
	Soma	82.9	77.67	80.19	
ContextNet	Vessel	74.85	72.85	73.79	212
	Soma	87.12	78.38	82.5	
DABNet	Vessel	71.92	72.38	72.13	159
	Soma	83.1	77.33	80.1	
LGNet-Ours	Vessel	**91.07**	82.11	86.26	135
	Soma	**99.72**	91.77	95.56	

by introducing the interactive mechanism between the two branches, which indicates that it is beneficial for segmentation by integration the local and global features from convolution layer and transformer layer.

Effect of the MSFFM: We compare the performance when the MSFFM is employed or not. The results can be found in Table 4, where we can see that the MSFFM could improve the whole performance in terms of mean DSC and HD. Specially, it leads to a performance boost of 0.18% DSC and 2.46 mm HD, respectively.

Effect of the feature maps from transformer and convolution layer: We also compare the effectiveness of the feature maps from convolution block and transformer block for the MSFFM. As we can see in Table 5, the features from three transformer blocks are benefit for improving the HD index. However, inputting the feature maps from the convolution layer will improve the DSC index 3.53% comparing to features from the transformer blocks.

Experiment results on ACDC dataset

Here, we train our model on ACDC dataset for automatic cardiac segmentation in order to demonstrate the generalization ability of the proposed LGNet. Comparing with the Transformer-based methods, such as TransUNet, SwinUnet, and LeViT-UNet, we can see that our proposed LGNet achieved the state-of-the-art results in terms of DSC in Table 6, which are similar to the previous results on the Synapse dataset.

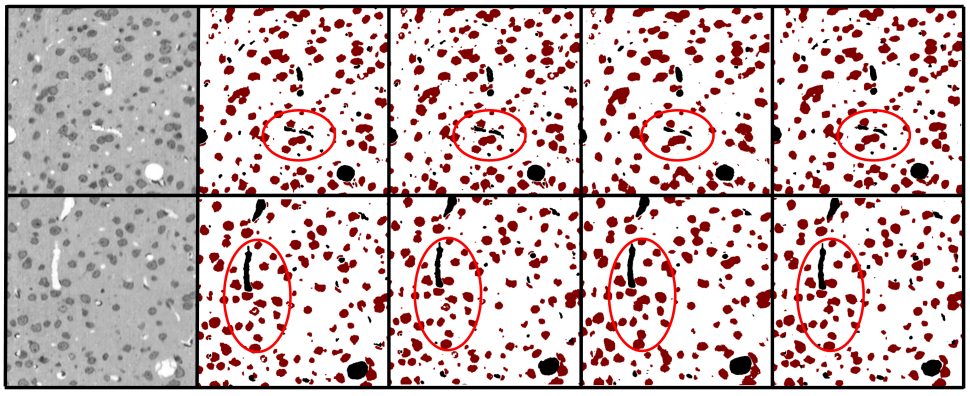

Fig. 5. Visualization of human annotations and prediction results. The first column to the fifth column is original testing images, human annotations, the prediction results from LGNet, ENet, and ContextNet, respectively.

Experiment results on MOST dataset

We also validate our method in the MOST dataset in order to show the generalization on the various modality images. In Table 7, we can see that our LGNet could achieve the competitive results comparing with the SegNet, yet, the inference speed is more than two times faster. Comparing to the light-weight methods, such as FastSCNN, ENet, CGNet, ContextNet, and DABNet, our method could get the best performance in terms of Precision, Recall and Dice with the competitive prediction speed. For example, our LGNet improves the Dice performance about 15.7% and 16.1% for vessel and soma segmentation, respectively. Overall, our method could better trade off the accuracy and efficiency of MOST dataset.

The qualitative comparing of three different methods (LGNet, ENet, and ContextNet) on the MOST 2D dataset is demonstrated in Fig. 5. We can see that the other two methods are more likely to under-segment or over-segment for the soma or vessel, for example, the results which are indicated by the red ellipse in the first and second rows.

4. Conclusion

In this paper, we present a two-branch network named as LGNet, which makes use of both advantage from convolution for local information learning and from transformer for long-range relationship extraction. We show that it is effective to improve the performance by interaction with the feature maps from CNN and transformer layers. Extensive experiments for medical image segmentation indicate that LGNet outperforms CNN-based and transformer-based methods in terms of accuracy. Meanwhile, the proposed LGNet achieves competitive trade-off between efficiency and accuracy when comparing to existing real-time segmentation networks. We expect our methods can be beneficial to other vision applications, like object detection.

Conflicts of Interest

We confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgments

This work was supported by the Open-Fund of WNLO (Grant No. 2018WNLOKF027) and the Hubei Key Laboratory of Intelligent Robot in Wuhan Institute of Technology (Grant No. HBIRL 202003). We thank the Optical Bioimaging Core Facility of WNLO-HUST for providing support in MOST data acquisition.

References

1. G. Xu, J. K. Udupa, Y. Tong *et al.*, "AAR-LN-DQ: Automatic anatomy recognition based disease quantification in thoracic lymph node zones via FDG PET/CT images without nodal delineation," *Med. Phys.* **47**(8), 3467–3484 (2020).
2. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3431–3440 (2015).
3. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015*. Springer, Cham., pp. 234–241 (2015).
4. L. C. Chen, G. Papandreou, I. Kokkinos *et al.*, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2018).
5. R. D. Rudyanto, S. Kerkstra, E. M. van Rikxoort *et al.*, "Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: The VESSEL12 study," *Med. Image Anal.* **18**(17), 1217–1232 (2014).
6. F. Cheng, C. Chen, Y. Wang *et al.*, Learning directional feature maps for cardiac MRI segmentation, *Medical Image Computing and Computer Assisted Intervention — MICCAI 2020*. Springer, Cham. pp. 108–117 (2020).
7. G. Xu, H. Cao, J. K. Udupa *et al.*, "DiSegNet: A deep dilated convolutional encoder-decoder architecture for lymph node segmentation on PET/CT images," *Comput. Med. Imag. Graph.* **88**, 101851 (2021).
8. J. Chen, Y. Lu, Q. Yu *et al.*, Transunet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306 (2021).
9. H. Zhao, J. Shi, X. Qi *et al.*, Pyramid scene parsing network, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2881–2890 (2017).

10. O. Oktay, J. Schlemper, L. L. Folgoc *et al.*, Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999 (2018).
11. A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.* (2017).
12. J. Devlin, M.-W. Chang, K. Lee *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805 (2018).
13. A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv:2010.11929 (2020).
14. S. Zheng, J. Lu, H. Zhao *et al.*, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 6881–6890 (2021).
15. Z. Liu, Y. Lin, Y. Cao *et al.*, Swin transformer: Hierarchical vision transformer using shifted windows, *Proc. IEEE/CVF Int. Conf. Computer Vision*, pp. 10012–10022 (2021).
16. H. Cao, Y. Wang, J. Chen *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," arXiv:2105.05537 (2021).
17. A. Lin, B. Chen, J. Xu *et al.*, "DS-TransUNet: Dual swin transformer U-Net for medical image segmentation," arXiv:2106.06716 (2021).
18. Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, pp. 14–24 (2021).
19. L. Yuan, Q. Hou, Z. Jiang *et al.*, "Volo: Vision outlooker for visual recognition," arXiv:2106.13112 (2021).
20. A. Hatamizadeh, Y. Tang, V. Nath *et al.*, Unetr: Transformers for 3d medical image segmentation, *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, pp. 574–584 (2022).
21. V. Badrinarayanan, A. Kendall, R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017).
22. F. N. Iandola, S. Han, M. W. Moskewicz *et al.*, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," arXiv:1602.07360 (2016).
23. A. G. Howard, M. Zhu, B. Chen *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861 (2017).
24. X. Zhang, X. Zhou, M. Lin *et al.*, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," *Proc. IEEE Conf. Computer Vision Pattern Recognit.* 6848–6856 (2018).
25. A. Paszke, A. Chaurasia, S. Kim *et al.*, "Enet: A deep neural network architecture for real-time semantic segmentation," arXiv:1606.02147 (2016).
26. Y. Wang, Q. Zhou, J. Liu *et al.*, Lednet: A lightweight encoder–decoder network for real-time semantic segmentation, *2019 IEEE Int. Conf. Image Processing (ICIP)*, pp. 1860–1864 (2019).
27. E. Romera, J. M. Álvarez, L. M. Bergasa *et al.*, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.* **19**(1), 263–272 (2018).
28. C. Yu, J. Wang, C. Peng *et al.*, Bisenet: Bilateral segmentation network for real-time semantic segmentation, *Proc. Eur. Conf. Computer Vision (ECCV)*, pp. 325–341 (2018).
29. C. Yu, C. Gao, J. Wang *et al.*, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput Vis.* **129**, 3051–3068 (2021).
30. R. P. Poudel, S. Liwicki, R. Cipolla, "Fast-scnn: Fast semantic segmentation network," arXiv:1902.04502 (2019).
31. R. P. Poudel, U. Bonde, S. Liwicki *et al.*, "Contextnet: Exploring context and detail for semantic segmentation in real-time," arXiv:1805.04554 (2018).
32. G. Li, I. Yun, J. Kim *et al.*, "Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," arXiv:1907.11357 (2019).
33. M. Y. Yang, S. Kumaar, Y. Lyu *et al.*, "Real-time semantic segmentation with context aggregation network," *ISPRS J. Photogramm. Remote Sens.* **178**(8), 124–134 (2021).
34. Y. Hong, H. Pan, W. Sun *et al.*, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," arXiv:2101.06085 (2021).
35. H. Zhao, X. Qi, X. Shen *et al.*, Icnet for real-time semantic segmentation on high-resolution images, *Proc. Eur. Conf. Computer Vision (ECCV)*, pp. 405–420 (2018).
36. L.-C. Chen, Y. Zhu, G. Papandreou *et al.*, Encoder–decoder with atrous separable convolution for semantic image segmentation, *Proc. Eur. Conf. Computer Vision (ECCV)*, pp. 801–818 (2018).
37. M. Yang, K. Yu, C. Zhang *et al.*, Denseaspp for semantic segmentation in street scenes, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3684–3692 (2018).
38. F. Liu, X. Ren, Z. Zhang *et al.*, Rethinking skip connection with layer normalization, *Proc. 28th Int. Conf. Computational Linguistics*, pp. 3586–3598 (2020).

39. K. He, X. Zhang, S. Ren *et al.*, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015).
40. T.-Y. Lin, P. Dollár, R. Girshick *et al.*, Feature pyramid networks for object detection, *Proce. IEEE Conf. Computer Vision and Pattern Recognition.*, pp. 2117–2125 (2021).
41. M. Sandler, A. Howard, M. Zhu *et al.*, Mobilenetv2: Inverted residuals and linear bottlenecks, *Proc. IEEE Conf. Computer Vision and Pattern Recognition.*, pp. 4510–4520 (2018).
42. A. Li, H. Gong, B. Zhang *et al.*, "Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain," *Science*, **330**(6009), 1404–1408 (2010).
43. G. Xu, X. Wu, X. Zhang *et al.*, "Levit-unet: Make faster encoders with transformer for medical image segmentation," arXiv:2107.08623 (2021).