TFNet: Transformer Fusion Network for Ultrasound Image Segmentation

Tao Wang[1], Zhihui Lai[1,2(✉)], and Heng Kong[3]

[1] Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, 518060 Shenzhen, China
lai_zhi_hui@163.com
[2] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
[3] Department of Breast and Thyroid Surgery, Baoan Central Hospital of Shenzhen, The Fifth Affiliated Hospital of Shenzhen University, Shenzhen 518102, Guangdong, China

Abstract. Automatic lesion segmentation in ultrasound helps diagnose diseases. Segmenting lesion regions accurately from ultrasound images is a challenging task due to the difference in the scale of the lesion and the uneven intensity distribution in the lesion area. Recently, Convolutional Neural Networks have achieved tremendous success on medical image segmentation tasks. However, due to the inherent locality of convolution operations, it is limited in modeling long-range dependency. In this paper, we study the more challenging problem on capturing long-range dependencies and multi-scale targets without losing detailed information. We propose a Transformer-based feature fusion network (TFNet), which fuses long-range dependency of multi-scale CNN features via Transformer to effectively solve the above challenges. In order to make up for the defect of Transformer in channel modeling, will be improved by joining the channel attention mechanism. In addition, a loss function is designed to modify the prediction map by computing the variance between the prediction results of the auxiliary classifier and the main classifier. We have conducted experiments on three data sets, and the results show that our proposed method achieves superior performances against various competing methods on ultrasound image segmentation.

Keywords: Ultrasound image segmentation · Transformer · Feature fusion

1 Introduction

An accurate breast lesion segmentation from the ultrasound images helps the early diagnosis of cancer. However, due to the feat that scale of tumor lesions in different periods is significantly different, the intensity distribution of the lesion area is not uniform, and there are fuzzy and irregular boundaries in ultrasound, so it is a challenging task to accurately segment the lesion area from

C. Wallraven et al. (Eds.): ACPR 2021, LNCS 13188, pp. 314–325, 2022.
https://doi.org/10.1007/978-3-031-02375-0_23

ultrasound images. Early attempts, such as [17,21], were mainly based on handmade features to detect the boundaries of breast lesions, but these methods have limit accuracy and are not suitable for more complex situations. Convolutional neural networks (CNN) has been widely used in computer vision, and has achieved excellent performance in medical image segmentation. Especially, UNet [15] based on encoder-decoder structure and its variants networks like [14], UNet++ [25], VNet [11], ResUNet [22] KiUNet [18] and UNet3+ [8] have achieved tremendous success in a wide range of medical applications.

Despite the excellent performance of convolutional neural network, due to the inherent inductive bias of convolution operation, the CNN-based method lacks the modeling ability of long-range dependency. Ultrasound images are often quite different in texture, shape and size, so long-range dependency is necessary. In order to overcome this limitation, some works introduce attention mechanism [12,16,20,23] or use atrous convolution [3,6] to make up for this defect. Recently, transformer [19], designed for sequence-to- sequence prediction, has become an alternative structure of convolutional neural network. Unlike prior CNN-based methods, transformer has strong performance in modeling global context based on self attention mechanism. With large-scale data for training, some pure transformer network models, such as ViT [5], Swin-Transformer [9], have achieved or even surpassed the performance of CNN. However, the pure transformer model often needs large data sets for training in order to achieve better performance, while the number of medical image data is very small. Some researchers are working on combining the CNNs with Transformers to create a hybrid structure. For example, TransUnet [2] utilizes CNNs to extract low-level features, which are then passed through transformer to model global interaction, and finally achieves better performance.

In this paper, we study the more challenging problem of capturing long-range dependencies and multi-scale targets without losing detailed information. We have explored and proposed our method to solve the problem. CNN has a strong ability to capture local details, and because the existing CNN is usually designed as a pyramid structure, the receptive field increases gradually with the deepening of network layers. For small targets, the shallow features contribute more, while for large targets, the deep features contribute more. Therefore, we consider combining the advantages of CNN with transformer and apply transformer to CNN feature fusion, that is to use transformer to model CNN multi-scale features with long-distance dependency, and then perform feature fusion. In addition, the existing transformer structure lacks the ability to weight channel information, so we will improve the transformer structure and add channel attention mechanism. In order to accurately predict the details of the ultrasound images, in addition to fusing the shallow feature information into the up sampling stage by using the jump link, we add an auxiliary classifier, which is different from the deep supervision method to modify the details of the prediction results.

In summary, the contributions of us are as follows: (1) we propose a transformer feature fusion module and designs a novel deep neural network called TFNet, (2) we propose a MultiHead channel attention mechanism and use it in

transformer block to improve its channel modeling ability, (3) this paper designs a loss function based on KL distance to correct the details of the predicted targets.

2 Method

The overall architecture of the proposed method is presented in Fig. 1(a). For an input image, four levels of features $\{f_i, i = 1, ..., 4\}$ can be extracted from CNN backbone network. We divide fifeatures into low-level features (f_1) and high-level features (f_2, f_3, f_4). High-level features contain more image semantic information of multi-scale targets due to the different receptive field. First, high-level features are fused together using our proposed Transformer Fuse Module (TFM). Therefore, high-level features can enhance the long-distance dependency modeling and obtain the features fused with multi-scale targets. Then, the above features are decoded by two-step upsampling using 3×3 convolution and linear interpolation. In the first step, we fuse the low-level features via skip connection to make up for details information. Inspired by [24], we add an auxiliary classifier in f_3 features. The loss function between the prediction results of the auxiliary classifier and the main classifier is established to correct the details of the prediction results. The components in our method: TFM, MCA and $\mathcal{L}_{KL}$ will be elaborated as follows.

2.1 Transformer Fuse Module

As shown in Fig. 1(b), TFM receives three different scales of high-level feature inputs. For each high-level features $f_i \in \mathbb{R}^{H \times W \times C}$, we first divide it into several patches, embed them into vector sequences $z_0^i \in \mathbb{R}^{(H'W') \times D} (where\, W' = W/P_i, H' = H/P_i, P_i$ is the size of the patch. The P_i for f_2, f_3, f_4 were set to 4, 2, 1) and add conditional position encoding, see Eq. (2). Then, z_0^i is transformed by Transformer Block into $z_{sc}^i \in \mathbb{R}^{(H'W') \times D}$ (Eq. (3)) and reshape it to get $f_{sc}^i \in \mathbb{R}^{H' \times W' \times D}$. Finally, the fusion features is as follows:

$$f_{sc} = \sigma(Conv(Concat[f_{sc}^2, f_{sc}^3, f_{sc}^4])) \tag{1}$$

where $f_{sc} \in \mathbb{R}^{H' \times W' \times D}$, $Conv$ is 3×3 convolution operation and σ donates ReLU activation function. Next, we will introduce the design of each part of TFM.

Patch and Position Embedding. We reshape the feature $f \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened patches $x_p \in \mathbb{R}^{P^2 \times C}$ and embed them into a D-dimensional space using a learnable linear projection. In our method, we set $D = 256$ to reduce the amount of subsequent computation.

To encode the patch spatial information, we use the conditional position encoding (CPE) [4] which are added to the patch embeddings to retain positional information as follows:

$$z_0 = [x_p^1 E; x_p^2 E; \cdots; x_p^N E] + E_{pos} \tag{2}$$

where $E \in \mathbb{R}^{(P^2 \times C) \times D}$ is the patch embedding projection, and $E_{pos} \in R^{N \times D}$ denotes the conditional position encoding which is generated by position encoding generator. CPE is dynamically generated and conditioned on the local neighborhood of the patchs. Compared with position encoding, CPE can keep the desired translation-invariance, resulting in improved prediction accuracy.

Transformer Block. The structure of a Transformer Block is illustrated in Fig. 1(c), which consists of two main parts. The first part models the long-range dependency of the input sequences, and the second part learns the weight of the channel. For the sequences z_0^i embedded from high-level features f_i, the transformation of Transformer Block is defined:

$$\begin{aligned} \hat{z}_s^i &= MSA(LN(z_0^i)) + z_0^i \\ z_s^i &= MLP(LN(\hat{z}_s^i)) + \hat{z}_s^i \\ \hat{z}_{sc}^i &= MCA(LN(z_s^i)) + z_s^i \\ z_{sc}^i &= MLP(LN(\hat{z}_{sc}^i)) + \hat{z}_{sc}^i \end{aligned} \tag{3}$$

where $LN(\cdot)$ denotes the layer normalization operator and z_0^i is the encoded representation of f_i. $MSA(\cdot)$ is MultiHead Self-Attention and $MCA(\cdot)$ is our proposed MultiHead Channel-Attention.

MultiHead Self-Attention. We use the vanilla Multi-Head Self-Attention to model long-range dependency of features, which is defined as follows:

$$\begin{aligned} Attention(Q, K, V) &= SoftMax(\frac{QK^T}{\sqrt{d}})V \\ MSA(z) &= Concat(Head_s^i(z), \ldots, Head_s^h(z))W^O \\ Head_s^i(z) &= Attention(zW_i^Q, zW_i^K, zW_i^V) \end{aligned} \tag{4}$$

where queries $Q_i = zW_i^Q$, keys $K_i = zW_i^K$ and values $V_i = zW_i^V$ are all projections computed from the input z. The projection matrices $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{D \times d}$ and $W^O \in \mathbb{R}^{hd \times D}$ are learnable.

MultiHead Channel-Attention. MSA in transformer lacks ability of channel-attention modeling. Based on the inductive bias of MSA, we improved SENet [7] and designed MultiHead Channel-Attention, as follows:

$$\begin{aligned} MCA(z) &= \sigma(W^e(ReLU(Head_c^i(z) + \cdots + Head_c^h(z))))z \\ Head_c^i(z) &= W_i^s(MaxPool(z) + AvgPool(z)) \end{aligned} \tag{5}$$

We first use average pooling and max pooling for each position of all patch sequences and sum them and then map the results to the low dimensional space

via a learnable projection $W^s \in \mathbb{R}^{D \times D/r} (r = 16)$. After calculating multiple such projections and adding the results, the original dimension is restored via another learnable projection $W^e \in \mathbb{R}^{D/r \times D}$. Finally, the learned channel weight vector is multiplied by each patch.

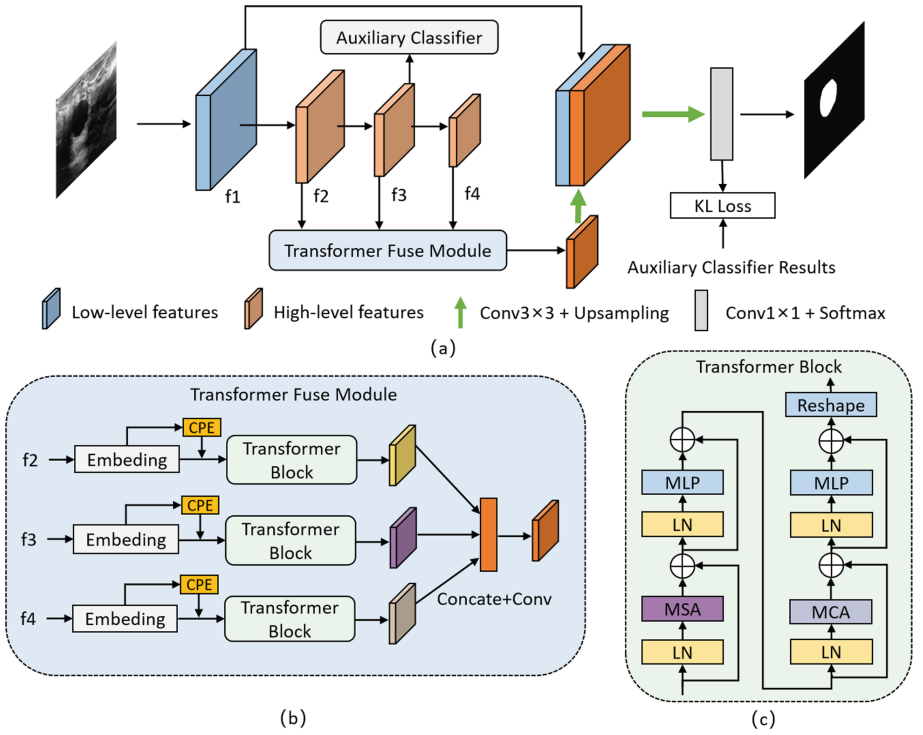

Fig. 1. (a) The main architecture diagram of TFNet. (b) The architecture of TFM. (c) Implementation details of the Transformer Block, which consists of two Transformer blocks. MSA and MCA are multi-head self attention modules and multi-head channel attention modules respectively.

2.2 Loss Function

We add an auxiliary classifier to the f_3 feature. Inspired by [24], the prediction results of the two classifiers are often very different for the targets which are difficult to predict. Therefore, we hope to modify the prediction results of the model through auxiliary classifier. Different from deep supervision, we establish the relationship between the prediction results of the auxiliary classifier and the main classifier by the following loss function:

$$\mathcal{L}_{KL} = \alpha(D_{kl}(p, \overline{p}) + D_{kl}(p_{aux}, \overline{p})) \tag{6}$$

where p is the prediction result of the main classifier, p_{aux} is the prediction result of the auxiliary classifier and $\overline{p} = softmax(p + \gamma \cdot p_{aux}), \gamma = 0.5$. $D_{kl}(P,Q) = \sum_i P(i)\log(P(i)/Q(i))$ donates KL distance. α is the scaling parameter, we set $\alpha = 1/(H \times W)$, H and W are the size of the predicted results.

We hope to reduce the distance between the prediction results of two classifiers by klloss, so as to reduce the area of the uncertain region. Our final loss function is:

$$\begin{aligned} \mathcal{L}_{total} &= \mathcal{L}_{CE} + \lambda \mathcal{L}_{KL} \\ \mathcal{L}_{CE} &= -\sum_{i=1}^{N} y^i log p^i + (1 - y^i) log(1 - p^i) \end{aligned} \tag{7}$$

where $\mathcal{L}_{CE}$ is cross entropy loss function, y is the targets and N is the number of pixels of y. λ is a hyper parameter. In our experiment, we set $\lambda = 0.1$.

3 Experiments

3.1 Datasets

We evaluate our proposed method on three datasets, including two public benchmark datasets BUSI [1], DDTI [13] and our collected dataset. BUSI collected 780 images from 600 female patients, with 437 benign cases, 210 benign masses, and 133 normal cases. In our experiments, we remove images of normal cases as benchmark data, and adopt the three-fold cross-validation to test each model. Another benchmark dataset DDTI contains the analysis of 347 thyroid ultrasound images, performed by two experts in 299 patients with thyroid disorders. We remove some images marked with damage and crop the images containing multiple nodules. We adopt four-fold cross validation on the 637 images to test each model.

Our collected dataset has 982 clinical breast ultrasound images in total from 500 patients. We follow the widely-used annotation procedure of the medical image segmentation for annotating breast lesions. Each image was annotated under the guidance of two experienced physicians. The final ground-truths were reached after several rounds of inspections. We also adopt the four-fold cross-validation to test each segmentation method on this.

3.2 Experimental Settings

Evaluation Metrics. We adopt five metrics to quantitatively compare different methods on the lesion segmentation. They are Dice coefficient (denoted as DSC), Jaccard index (denoted as Jaccard), Recall, Precision and Specificity.

Training Parameters. In order to ensure the fairness of the experiment, each model uses ResNet50 as the backbone network. We initialize the parameters of the backbone network using the pre-trained weights on ImageNet while other parameters of the model are initialized randomly. The input images are uniformly resized to a size of 256×256, and then random clipping, random flip, pad and normalization are used for image enhancement. We use SGD algorithm to optimize the network with a momentum of 0.9, a weight decay of 0.0005 and 40000 iterations. The initial learning rate is set to 0.01, the minimum learning rate is set to 0.0001, and then the polynomial annealing algorithm is used to reduce the learning rate. We implement the network using PyTorch library and train our network on a single NVIDIA RTX 2080Ti GPU with the batch size set to 16.

3.3 Results

We conduct main experiments on three ultrasound image dataset by comparing our TFNet with four previous methods: FCN [10]; UNet [15]; UNet++ [25] and DeepLabV3+ [3].

Table 1 shows the average results of the proposed method based on a cross validation. The experimental results show that our TFNet is superior to other methods in different evaluation metrics. Note that our method performs better on busi dataset. This is because there are a large number of targets with high echo and rugged boundary in BUSI dataset, and the distribution of small, medium and large targets is relatively uniform, which is consistent with the motivation of our proposed method. On the contrary, the boundary of the target in ddti dataset is more smooth, and the scale of target is mostly small. In this case, the improvement effect of our method is relatively less than in other dataset.

Besides, We add a set of vanilla UNet (UNet*) experimental results to compare the effect of the pre-trained backbone network. After replacing the encoder of UNet with ResNet50 network with pre-trained weights, the performance is improved significantly (DSC increases by 2.8%). Therefore, we choose a strong backbone network in the experiment to exclude the influence of CNN feature extraction ability for fair experimental comparison.

3.4 Analysis

Ablation Study. To verify the effectiveness of the principal components of our network: MCA, $\mathcal{L}_{KL}$, and TFM in our network. And the ablation study experiments are conducted on BUSI dataset. The baseline (first row of Table 1) is constructed by removing both components from our network. The first row represents the benchmark architecture of TFNet with TFM and $\mathcal{L}_{KL}$ removed. The second line represents fusion using downsampling feature alignment. MSA donates the TFM only uses MSA, see Fig. 1(c) left. MSA donates the TFM only uses MCA, see Fig. 1(c) right. $\mathcal{L}_{KL}$ donates our KL loss function. The last line shows the result of replacing $\mathcal{L}_{KL}$ with deep supervision.

Table 1. Quantitative results on BUSI, DDTI and our datasets of the TFNet, FCN [10], UNet [15], UNet++ [25] and DeepLabV3+ [3]. Each numerical value represents the average result (%± standard deviation) of k-fold cross validation. UNet* donates the original UNet network without ResNet50.

	Method	DSC %	Jaccard %	Recall %	Precision %	Specificity %
BUSI	UNet*	71.5 ± 5.3	55.8 ± 6.4	77.0 ± 11.0	67.2 ± 3.5	96.7 ± 2.4
	FCN	73.3 ± 5.1	58.0 ± 5.8	78.6 ± 11.4	69.2 ± 2.2	97.9 ± 1.4
	UNet	74.3 ± 5.5	59.3 ± 6.8	78.5 ± 12.8	72.3 ± 2.9	97.7 ± 1.8
	UNet++	75.1 ± 5.5	60.3 ± 6.9	79.2 ± 10.9	71.4 ± 3.2	97.9 ± 1.3
	DeepLabV3+	74.8 ± 6.1	60.0 ± 7.6	77.3 ± 9.8	73.2 ± 8.0	97.5 ± 1.0
	TFNet (Ours)	**77.3 ± 5.5**	**63.0 ± 7.0**	**79.5 ± 11.1**	**75.5 ± 2.3**	**98.1 ± 1.1**
DDTI	FCN	82.7 ± 1.1	70.6 ± 1.6	87.8 ± 1.6	78.3 ± 2.7	**98.1 ± 0.3**
	UNet	83.4 ± 1.0	71.5 ± 1.5	85.7 ± 3.1	81.4 ± 3.0	97.5 ± 0.6
	UNet++	83.0 ± 1.8	71.0 ± 2.6	**87.4 ± 1.2**	79.1 ± 3.0	98.0 ± 0.2
	DeepLabV3+	83.8 ± 1.3	72.2 ± 1.9	84.9 ± 1.2	82.7 ± 2.2	97.5 ± 0.3
	TFNet (Ours)	**84.6 ± 1.7**	**73.2 ± 2.5**	86.6 ± 1.7	**83.2 ± 2.6**	97.8 ± 0.3
Ours	FCN	85.2 ± 1.6	74.2 ± 2.4	85.4 ± 2.2	84.9 ± 1.7	98.0 ± 0.4
	UNet	87.0 ± 1.5	77.1 ± 2.3	88.2 ± 3.4	85.9 ± 1.0	98.2 ± 0.5
	UNet++	86.8 ± 1.6	76.8 ± 2.6	87.9 ± 1.6	85.8 ± 1.9	98.4 ± 0.2
	DeepLabV3+	86.3 ± 2.2	75.9 ± 3.3	85.0 ± 2.1	87.6 ± 4.1	97.9 ± 0.5
	TFNet (Ours)	**87.9 ± 1.4**	**78.4 ± 2.1**	**88.2 ± 1.7**	**87.7 ± 1.6**	**98.5 ± 0.2**

Table 2. Quantitative results of ablation study on BUSI dataset.

Fuse	*MSA*	*MCA*	$\mathcal{L}_{KL}$	Parameters (M)	DSC %	Jaccard %
				23.89	74.2 ± 6.4	59.9 ± 7.4
✓				–	74.8 ± 6.1	60.5 ± 7.2
✓	✓			29.94	76.1 ± 5.1	61.9 ± 6.8
✓	✓	✓		31.53	77.0 ± 4.8	62.6 ± 6.5
✓	✓	✓	✓	33.89	**77.3 ± 5.5**	**63.0 ± 7.0**
✓	✓	✓	DSV	–	77.1±5.5	62.5±7.2

Table 2 shows the comparison results of our method with different components. Compared the first line with the second line, the performance of multi-scale feature fusion strategy will be improved, which shows that feature fusion is useful for dataset with multi-scale targets. When using TFM with MSA only, we can see that learning the long-range dependency has a superior performance(DSC increases by 2.9%). 'MSA + MCA' have better performance than MSA, showing that our proposed MCA introduces channel attention into transformer block, which helps capture the dependence between channels. In addition, $\mathcal{L}_{KL}$ can improve the performance of the model, and the effect is better than deep supervision.

We also compare the number of parameters required by each component. When we add TFM module, the number of parameters increases significantly, which is caused by patch embedding. Note that $\mathcal{L}_{KL}$ increases the number of parameters because it adds an auxiliary classifier, which can be ignored in the model inference stage.

Table 3. Comparison of the parameters and MACs between our model and the existing model. The first row represents the backbone network ResNet50.

Method	Backbone	Parameters (M)	MACs (G)
–	ResNet50	23.52	5.70
FCN		47.12	7.12
UNet		43.91	20.25
UNet++		61.31	69.82
DeepLabV3+		**26.70**	9.51
TFNet		31.53	**6.92**

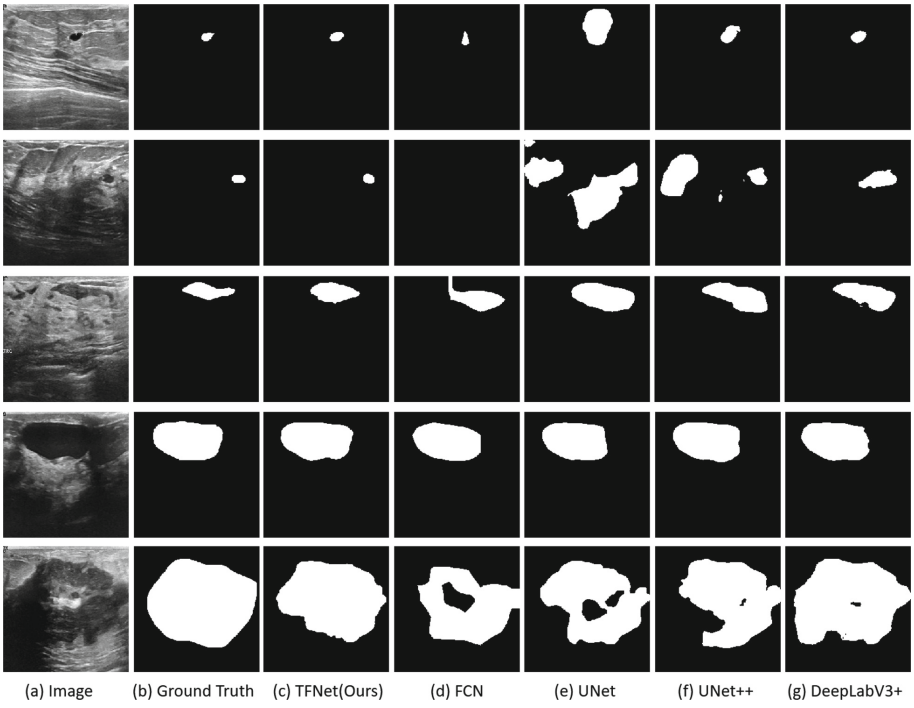

Fig. 2. Visual comparison of the lesion segmentation maps produced by different methods. (a) ultrasound images; (b) ground truths; (c)–(g) are segmentation results by TFNet, FCNN, UNet, UNet++ and DeepLabV3+.

Table 3 shows the comparison of the parameters and MACs between our proposed model and the comparison models. The first row represents the backbone network ResNet50. Our TFNet needs a little more parameters than DeeplabV3+ and less than other models. Our TFNet needs less MACs than other methods.

Visualizations. Figure 2 shows some examples of segmentation results from different cases. We select the small, medium and large scale sample images, as well as the typical images with high echo and serious background influence for visualization. For multi-scale targets, our method shows good segmentation results. Among the comparison method, DeeplabV3+ has better effect on multi-scale target prediction because of the mechanism of ASPP module. UNet, UNet++ and FCN tend to neglect breast lesion details or wrongly classify non-lesion regions as breast lesions into their predicted segmentation maps. For the image with high echo, our method can segment the whole target accurately, while other methods show the phenomenon of internal cavity (the fifth line). This is because our method can capture long-range information. For the images with serious background influence (the second line), the prediction results have a large deviation, and our proposed $\mathcal{L}_{KL}$ can increase the penalty in this case and obtain better results.

4 Conclusion

In this work, in order to solve the problems of large difference in the scale of lesions, uneven intensity distribution in the lesion area, and difficult to distinguish the details in ultrasound images, we propose a feature fusion method using transformer and apply it to ultrasonic image segmentation task. TFNet achieves better performance and faster reasoning speed without using transformer pretraining weights. We design a multi channel attention mechanism to improve the transformer block and enhance its modeling ability on channels. An auxiliary loss function based on KL distance is proposed to correct the detailed information in the prediction results. In our experiments, this loss function is superior to the methods using deep supervision and can transplant to any existing network. In the future, we will study how to improve the multi-head self-attention mechanism to further improve the performance.

Acknowledgement. This work was supported in part by the Natural Science Foundation of China under Grant 61976145 and Grant 61802267, and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grants JCYJ20180305124834854 and JCYJ20190813100801664.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)

2. Chen, J., et al.: TransUnet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018, Part VII. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
4. Chu, X., et al.: Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 (2021)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Gu, Z., et al.: CE-Net: context encoder network for 2D medical image segmentation. IEEE Trans. Med. Imaging **38**(10), 2281–2292 (2019)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
8. Huang, H., et al.: UNet 3+: a full-scale connected UNet for medical image segmentation. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059. IEEE (2020)
9. Liu, Z., Let al.: Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
12. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
13. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: 10th International Symposium on Medical Information Processing and Analysis, vol. 9287, p. 92870W. International Society for Optics and Photonics (2015)
14. Rampun, A., Jarvis, D., Griffiths, P., Armitage, P.: Automated 2D fetal brain segmentation of MR images using a deep U-Net. In: Palaiahnakote, S., Sanniti di Baja, G., Wang, L., Yan, W.Q. (eds.) ACPR 2019, Part II. LNCS, vol. 12047, pp. 373–386. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-41299-9_29
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015, Part III. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018, Part I. LNCS, vol. 11070, pp. 421–429. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_48
17. Shan, J., Cheng, H.D., Wang, Y.: A novel automatic seed point selection algorithm for breast ultrasound images. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4. IEEE (2008)

18. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: KiU-Net: towards accurate segmentation of biomedical images using over-complete representations. In: Martel, A.L., et al. (eds.) MICCAI 2020, Part IV. LNCS, vol. 12264, pp. 363–373. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_36
19. Vaswani, A., et al.: Attention is all you need, pp. 5998–6008 (2017)
20. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
21. Xian, M., Zhang, Y., Cheng, H.D.: Fully automatic segmentation of breast ultrasound images based on breast characteristics in space and frequency domains. Pattern Recognit. **48**(2), 485–497 (2015)
22. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted Res-UNet for high-quality retina vessel segmentation. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), pp. 327–331. IEEE (2018)
23. Xue, C., et al.: Global guidance network for breast lesion segmentation in ultrasound images. Med. Image Anal. **70**, 101989 (2021)
24. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. Int. J. Comput. Vis. **129**(4), 1106–1120 (2021)
25. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1