X-Net: A Dual Encoding-Decoding Method In Medical Image Segmentation

Yuanyuan Li
College of Automation
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
liyy@cqupt.edu.cn

Ziyu Wang
College of Automation
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
s200332016@stu.cqupt.edu.cn

Zhiqin Zhu
College of Automation
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
zhuzq@cqupt.edu.cn

Xinghua Huang*
College of Automation
Chongqing University
Chongqing 400044, China
huangxh1980@126.com

Guanqiu Qi
Computer Information Systems Department
Buffalo State College
Buffalo, NY 14222, USA
qig@buffalostate.edu

Baisen Cong
Data Scientist, Diagnostics Digital DH(Shanghai) Diagnostics Co, Ltd, a Danaher company
Tel: +86 13594157666
bcong@dhdiagnostics.com

***Abstract*—Medical image segmentation has a priori guiding significance for clinical diagnosis and treatment. In the past ten years, a large number of experimental facts have proved that deep convolutional neural networks have achieved great success in various medical image segmentation tasks. However, the convolutional network seems to focus too much on the local part of the image, while ignoring the long-range dependence on the whole image. The Transformer structure can encode long-range dependencies in images and learn high-dimensional image information through the self-attention mechanism. But this structure currently depends on the scale of the datasets to give full play to its excellent performance, which limits its application in medical images with limited datasets size. In this paper, we combine the characteristics of Transformer and CNNs to propose a dual encoding-decoding structure of the X-shaped network (X-Net). It can serve as a good alternative to the traditional purely convolutional medical image segmentation network. In the encoding phase, the local and global features will be simultaneously extract by two types of encoders, convolutional downsampling and Transformer, then merge through jump connection. And in the decoding phase, a Variational Auto-Encoder branch has been added to reconstruct the input image itself in order to weaken the impact of insufficient data. The results show that X-Net have realized the organic combination of Transformer and CNNs.**

Keywords—Medical image segmentation, Transformer, Variational Auto-Encoder, Dual encoding-decoding

I. INTRODUCTION

The segmentation task of medical images is a significant work in the field of medical images. Segmentation methods based on convolutional neural networks (CNNs) have shown extraordinary performance in numerous subtasks, such as nuclei segmentation [1], brain tumor segmentation [2] and multi-organs segmentation [3]. Since the excellent segmentation performance of U-Net on most medical image datasets, U-Net [4] which is an encoder-decoder network entirely based on CNN has become a representative of the current neural network structure for medical image segmentation tasks. It obtains coarse-grained deep features through continuous downsampling in the encoding stage, and gradually upsampling to obtain fine-grained shallow features in the decoding stage. At the same time of encoding-decoding, the deep features and shallow features are merged by skip connection, so as to integrate the global context and high-resolution details. Researchers have developed and designed various versions of U-Net by adding multi-scale dilated convolutions [5], residual module [6], attention mechanism [7] or other methods. These new ideas have improved the accuracy, stability, and computational speed on the original basis.

However, it is still challenging for U-Net based on full convolution mode on further improvement of the segmentation efficiency. The network of full convolution mode is lack of learning long-range dependencies. Each convolution kernel focuses on a local subset of pixels in the whole image, then forces the network to focus on local areas instead of global context. The judgments are necessary during the segmentation prediction process, which informs each pixel in the segmented image is closer to the segmentation mask pixel or the background pixel. Learning long-term dependencies can help the network avoid misclassification of mask pixels and background pixels. Thus, long-range dependency is vital for medical image segmentation. Related researches have explored the use of attention mechanism [8], image pyramids [9], atrous convolution [10] and other methods to achieve long-term dependent modeling, but the computational complexity of these methods tends to grow exponentially relative to the size of the input images, so many methods are only suitable for low-resolution images.

The issue of long-term dependence does not only exist in the field of images, but it is more prominent in the field of natural language processing (NLP) due to the more complicated contextual relationship. In recent years, a large number of works aimed at global context relations, also firstly appeared in tasks related to NLP, such as GPT [11] and BERT [12]. These Transformer structures have completely subverted machine translation [13], question answering [14] and document analysis [15] and other research directions. Transformer was quickly introduced and applied to the field of computer vision, with the great breakthroughs made by Transformer in the field of NLP. A. Dosovitskiy et al. [16] took the lead in trying and exploring, and successfully designed a vision transformer (ViT). This structure can disperse the two-dimensional image into patches. Then the patches is transformed into the input sequence through a position embedding method to perform a self-attention mechanism similar to the language text sequence. The results shows that ViT has equipped an image recognition function in large-scale datasets that is exceed the pure convolutional network. S. Zheng et al. [17] then proposed SETR. It initially combined CNNs

and Transformer by replacing the encoder in the traditional encoding-decoding structure network with Transformer, thus successfully achieving SOTA performance in the natural image segmentation task.

It is obviously that Transformer has a great potential in the medical image segmentation tasks. But the recent works [16][18] present that the Transformer-based model can perform better than CNNs when trained on a larger dataset. The main reason for this phenomenon is that the small-scale data sets will make it more difficult to learn the position coding of the image. Therefore, it is negative for Transformer to applied in medical image datasets with scarcity and low pixel resolution. The problem of how to take advantage of the self-attentive performance provided by Transformer while taking into account the location encoding for small datasets is of research interest.

Based on the idea above, we proposed a dual encoding-decoding X-shaped network (X-Net) structure as shown in the Fig. 1. The dual feature encoder is composed of two different coding branch structures, CNNs and Transformer. Convolution and self-attention mechanism are used to effectively extract local and global contextual semantic information, respectively. While progressive upsampling is used to generate segmentation predictions, we also use skip connection to interactively fuse local and global information with the same resolution to achieve precise positioning. Besides, we added a Variational Auto-Encoder (VAE) [19] branch to reconstruct the input image in the decoding process. The purpose is to standardize the shared decoder to impose additional constraints on training. The results show that compared with the previous self-attention mechanism based on CNN, the transformer-based structure provides a better way to utilize self-attention. Due to the parallel structure, the X-shaped design does not increase much in model size compared with the traditional U-Net.

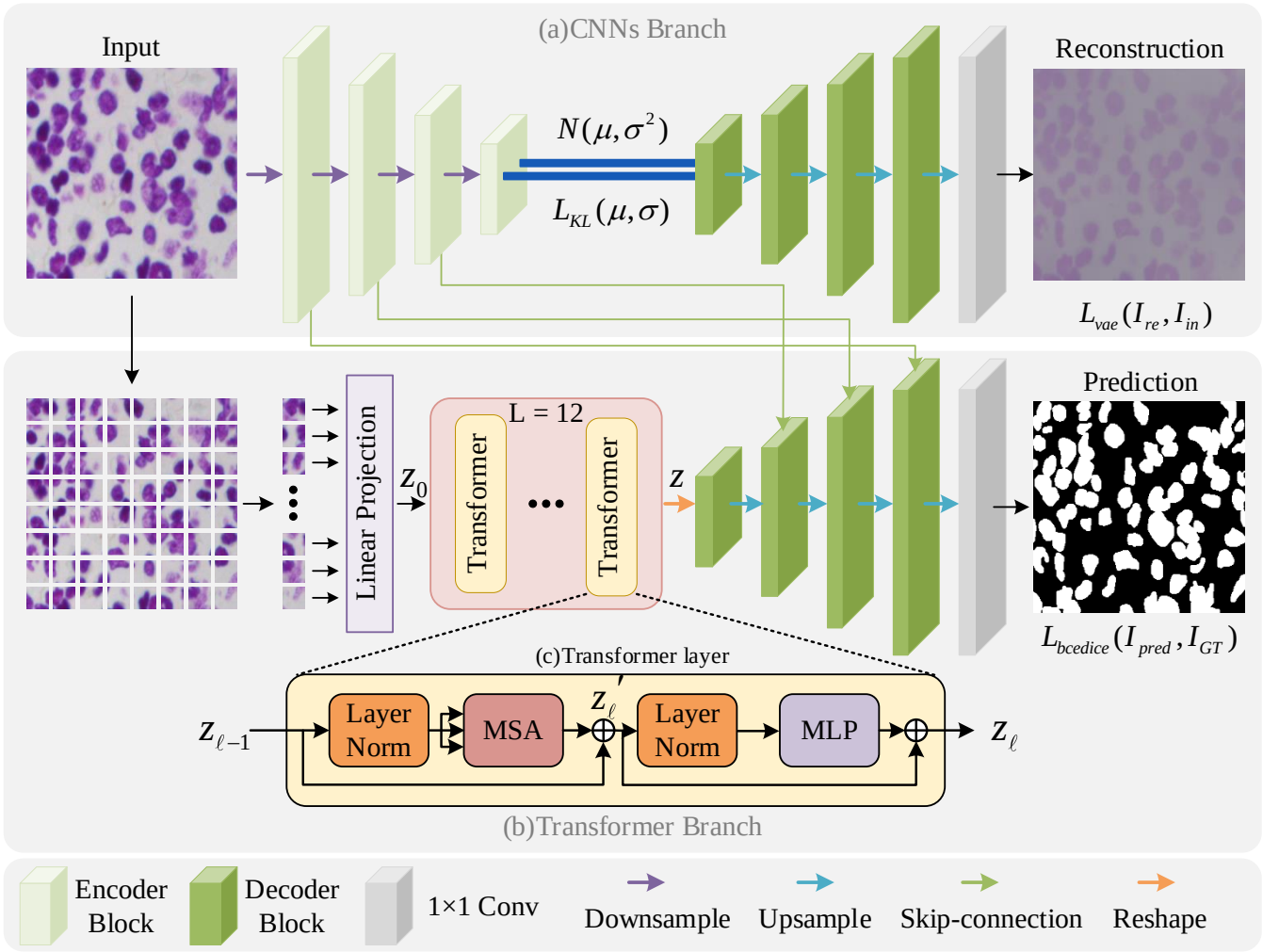

Fig. 1. The dual encoding-decoding X-shaped network (X-Net) structure. (a) architecture of the CNNs branch,(b) architecture of the Transformer branch, (c) schematic of the Transformer layer.

II. METHOD

A. Overrall

In this work, we proposed a X-Net structure consisting of CNNs branch and Transformer branch. It based on the U-Net structure and combining the advantages and characteristics of Transformer and variable auto-encoder. For focusing on the acquisition of local and global features at the same time, the two branches implement parallel training, using the same input but different encoding methods. Then the local and global information are interacted by skip connection, so as to realize the precise positioning of the segmentation target. Next, the characteristics and design methods of the X-Net structure will be explained in detail.

B. CNNs Branch

Encoder Part:

In the encoding stage of the CNNs branch, a conventional encoding structure is used to extract image features. In order to obtain an image information as deep as possible, and more global relationships, we perform feature extraction on the input image by downsampling. Specifically, the ResNet50 [6] is used to downsample the input image size in a double mode. After downsampling four times. The output characteristics of each sample block are:

$$\text{1th: } g_1 \in \mathbb{R}^{C_1 \times \frac{H}{2} \times \frac{W}{2}}; \quad \text{2th: } g_2 \in \mathbb{R}^{C^2 \times \frac{H}{4} \times \frac{W}{4}};$$

$$\text{3th: } g_3 \in \mathbb{R}^{C^3 \times \frac{H}{8} \times \frac{W}{8}}; \quad \text{4th: } g_4 \in \mathbb{R}^{C^4 \times \frac{H}{16} \times \frac{W}{16}}.$$

Finally a high-dimension feature image of size $C_4 \times \frac{H}{16} \times \frac{H}{16}$ is obtained. The output of each sample block will be passed to the stage with the same resolution in the Transformer branch decoding stage by skip connection. So as to merge the high and low dimension features.

Decoder Part:

The decoding part of the CNNs branch is a framework similar to an auto-encoder. In the case of a small number of training datasets (a common phenomenon in medical images), this setting can add additional guidance and regularization to the encoder part, so that better cluster and group the characteristics of the encoder endpoints. Specifically, starting from the end-point output of the encoder, we first simplify the input to a low-dimensional space of 512 (256 represents the average value and 256 represents the std). Then, a sample is drown from the Gaussian distribution of the given mean and std, and reconstructed into the input image according to the same architecture as the decoder, but we do not use the inter-layer skip connection from the encoder.

C. Transformer Branch

Encoder Part:

In the Transformer branch, the image Transformer will be used as an operation layer in the encoding process through the method provided by ViT, thereby adding a local self-attention mechanism to the network.

Specifically, the input image $X \in^{H \times W \times C}$ will be divided into N patches of sizes $P \times P$. These patches will be reshaped into one-dimensional vectors $(N \times P^2 C)$, and then each vector x_p^i is compressed into D dimension by a trainable linear projection E. Then the output becomes a patch embedding. In order to add position information to such an image sequence, a trainable variable E_{pos} is introduced instead:

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

where z_0 is embedded sequence as the input of the Transformer layer.

The Transformer encoder is composed of L layers (the structure of each layer is shown in Fig 1(3)) multi-head self-attention (MSA) and MLP blocks. Layernorm (LN) is applied before each block, and residual connection is applied after each block. MLP contains two fully connected layers with GELU sub-linearity.

$$z_\ell' = \mathrm{MSA}(\mathrm{LN}(z_{\ell-1})) + z_{\ell-1}, \tag{2}$$

$$z_\ell = \mathrm{M\,LP}(\mathrm{LN}(z_\ell')) + z_\ell', \tag{3}$$

where LN(.) represents the layer normalization operator, which is the image representation of feature coding.

$$z = \mathrm{LN}(z_\ell) \tag{4}$$

At the end of the Transformer layer, z is the output by layer normalization again.

Decoder Part:

Before the decoding of Transformer branch starts, it is necessary to reconstruct the output sequence z of the encoding part back to the three-dimensional form:

$$t \in \mathbb{R}^{D \times \frac{H}{16} \times \frac{W}{16}},$$

which t can be regarded as a two-dimensional feature image with channels. Later, the conventional progressive upsampling (PUP) method is used to gradually restore the spatial resolution.

Each upsampling stage will get the output as:

$$\text{1th: } t_4 \in \mathbb{R}^{D_4 \times \frac{H}{8} \times \frac{W}{8}}; \quad \text{2th: } t_3 \in \mathbb{R}^{D_3 \times \frac{H}{4} \times \frac{W}{4}};$$

$$\text{3th: } t_2 \in \mathbb{R}^{D_2 \times \frac{H}{2} \times \frac{W}{2}}; \quad \text{4th: } t_1 \in \mathbb{R}^{D_1 \times H \times W}.$$

The upsampling process will merge the high-dimension information passed by the CNNs branch in a way of dimensional merging.

D. Loss Function

The entire network uses weighted loss for end-to-end training, and the loss function is composed of:

$$L = L_{bcedice} + 0.1 * L_{vae} + 0.1 * L_{KL}, \tag{5}$$

$L_{bcedice}$ is the main loss function of the segmentation network used to match the segmentation prediction of Transformer branch and the ground truth (GT). It is a combination of Binary Cross Entropy (BCE) loss and soft Dice loss. It can be expressed as:

$$L_{bcedice} = 0.5 * BCE(I_{pred}, I_{GT}) + Dice(I_{pred}, I_{GT}) \tag{6}$$

L_{vae} is an L2 loss used to match the VAE reconstructed image I_{re} with the input image I_{in}:

$$L_{vae} = \left\| I_{re} - I_{in} \right\|_2^2, \tag{7}$$

L_{KL} is the standard VAE penalty item, used to estimate the KL dispersion between the normal distribution $N(\mu, \sigma^2)$ and the prior distribution $N(0,1)$, and can be expressed as:

$$L_{KL} = \frac{1}{N} \sum \mu^2 + \sigma^2 - \log \sigma^2 - 1, \tag{8}$$

where N is the total number of pixels in the image. μ and σ are the mean and standard deviation extracted by Gaussian distribution, respectively.

After a lot of experiments, we find that setting the weight hyperparameter to 0.1 is a better solution that can balance the relationship between the three losses.

III. EXPERIMENTS

A. Datasets details

In the experiment, we implemented the segmentation task on two public nuclear datasets with different resolution sizes.

DATA-SCIENCE-BOWL-2018(DSB18) [20]: This dataset comes from the kaggle challenge in 2018. Since the test set of it does not contain GT images, in the experiment we only used the training set (including 670 images with annotations) in the dataset, and randomly separated it into two parts at a ratio of 0.8/0.2 used for training and testing respectively. For image preprocessing, we did not use too many complicated methods to enhance the quality of the original image. This is to highlight the performance advantages of the proposed network under simple preconditions. Specifically, since the resolution type of the original data is not unique, we will uniformly scale all images to 256×256 size. Before sending the image to the network for training, a simple normalization and standardization process was done.

Triple Negative Breast Cancer (TNBC) [21]: This dataset is provided by the authors in work [21]. The dataset was collected from 11 triple negative breast cancer (TCBN) patients and contained a total of 50 annotated images. Since the size of each image has been unified to 512×512, we only used the normalization and standardization before training to process the original image.

B. Evaluation details

In order to have a unified index parameter when evaluating the performance of the network, we choose to use two parameters IoU and F1-Score to judge the similarity between the segmentation prediction of the network output and the ground truth.

IoU is the intersection ratio between the prediction and GT.

$$\mathrm{IoU} = \frac{Overlap}{Union}, \tag{9}$$

where *Overlap* is the overlapping part of the two, and *Union* is the joint area of the two.

F1-Score is the weighted harmonic average of Precision and Recall.

$$\mathrm{F1} = \frac{2PR}{P+R}, \tag{10}$$

where P represents the accuracy rate and R represents the recall rate.

C. Setting dedails

We completed the experiment on the Pytorch program architecture, using two Tesla P100GPUs to provide accelerated computing for network training. In the experiment, we used a training batch of 16(DSB18) and 4(TNBC), and the optimizer chose SGD, where the learning rate, momentum, and decay rate were set to 1e-3, 0.9, and 1e-5, respectively. The entire network is trained for 400 epochs.

D. Comparison study

In order to verify the segmentation effect of our proposed X-Net, we conducted comparative experiments with some network structures that have shown excellent performance (including classic segmentation networks such as FCN [22], U-Net [4], Res-UNet [23] and TransUnet [24] which also combines the Transformer structure). In order to qualitatively analyze the results of our experiments, all experiments are carried out on a unified benchmark and evaluation index.

The experimental results are shown in TABLE I. . For a dataset such as DSB18 with a relatively large number of images, the segmentation effect of X-Net with a Transformer structure has been significantly improved. On the smaller TNBC dataset, X-Net shows that its performance improvement is not obvious. The reason mentioned above is that it is currently difficult to train a full attention model due to less data. In the comparative experiment, our proposed network model achieved 0.8250/ 0.9036 (DSB18) and 0.6437/0.7822 (TNBC) IoU scores/F1 scores on the DSB18 and TNBC datasets, respectively.

TABLE I. QUANTITATIVE COMPARISONS OF DIFFERENT METHODS ON TWO DATASETS.

Dataset	DSB 18		TNBC	
Network	IoU	F1-Score	IoU	F1-Score
FCN	0.7408	0.8503	0.5889	0.7407
U-Net	0.8031	0.8903	0.6200	0.7644
Res-UNet	0.8028	0.8901	0.6318	0.7735
TransUNet	0.8179	0.8992	0.6161	0.7609
Ours(X-Net)	**0.8250**	**0.9036**	**0.6437**	**0.7822**

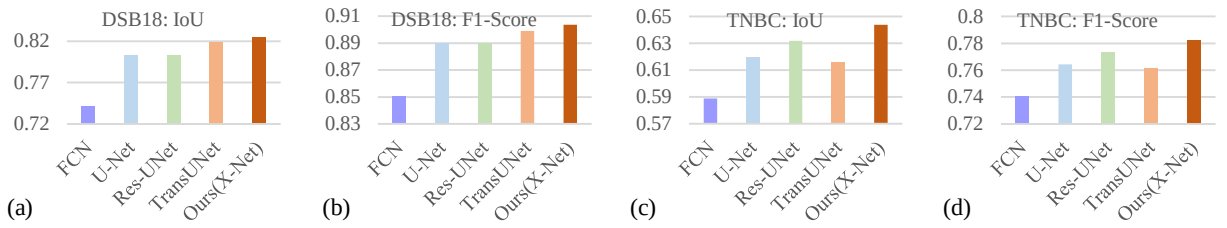

Fig. 2. The objective evaluation index results of different methods on the two datasets of DSB18 and TNBC. The higher the IoU and F1-Score values, the better the segmentation effect. (a) IoU of the DSB18 dataset, (b) F1-score of the DSB18 dataset, (c) IoU of the TNBC dataset, (d) F1-score of the TNBC dataset.

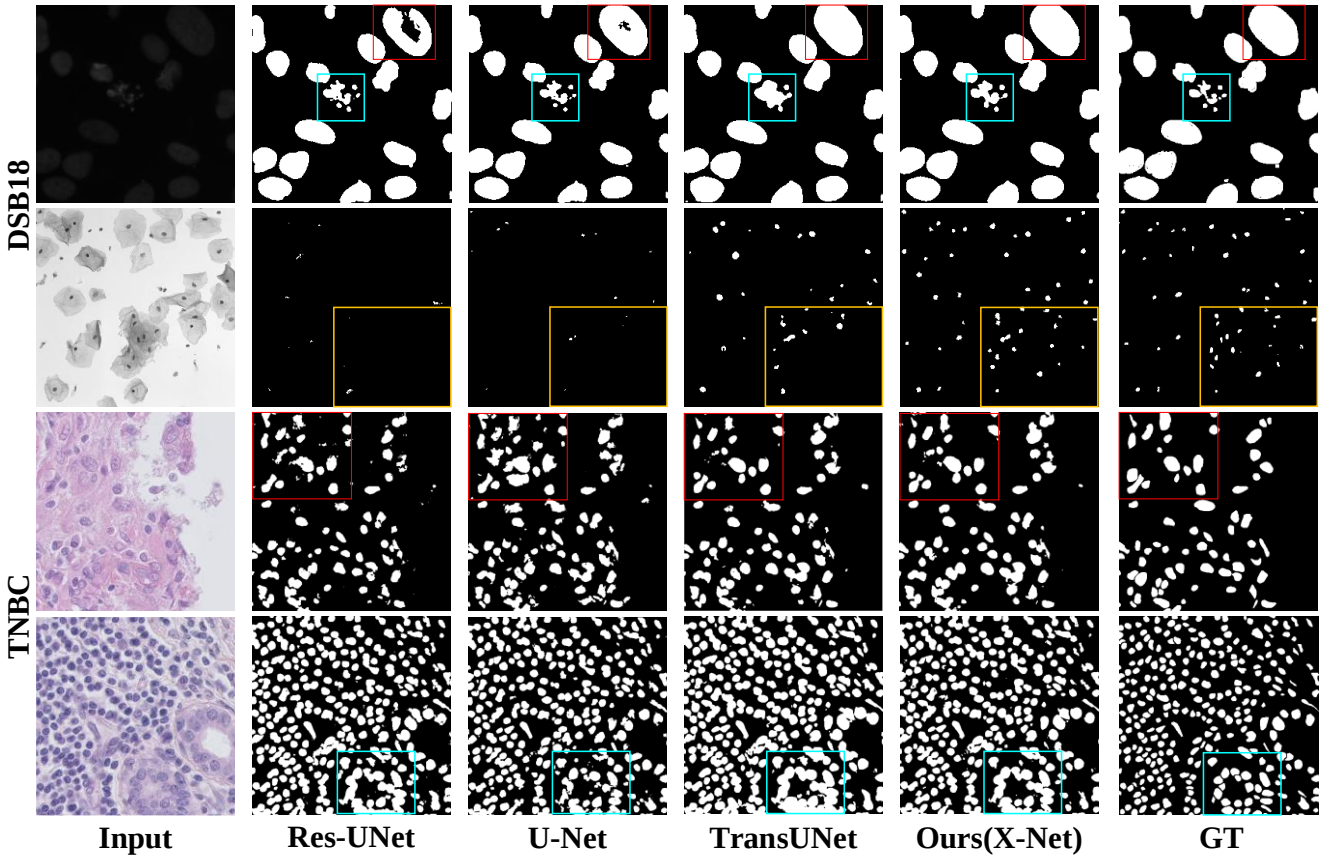

Fig. 3. Qualitative results on sample images from DSB18 and TNBC datasets. The color boxes highlight regions where exactly X-Net performs better than the other methods.

For observing and comparing the actual segmentation effect of various network models in the experiment more intuitively, we respectively selected some samples from the two datasets DBS18 (1, 2 rows) and TNBC (3, 4 rows) for visualization, as shown in the Fig. 3. It can be seen in the comparative example shown: the traditional U-Net architecture of convolutional network performs well for segmentation details (the blue boxes in the first and fourth rows). However, there are some over-segmentation conditions (red boxes in the first and third rows), which easily lead to misjudgments in the uneven part of a single cell or the overlapping junctions of multiple cells. The network combined with the Transformer structure captures long-term dependencies, integrates global contextual information, and avoids appeals. In the second row of results, the positioning of the Transformer structure for the global segmentation position is more prominent. At the same time, the dual encoding method prompts the network to pay more attention to more local details (orange box), which achieves the segmentation prediction closest to the real image.

E. Ablation study

We continue to conduct ablation experiments on the DSB18 dataset to further verify the contribution and role of each component in X-Net. We start from the basic structure form with only convolutional up-downsampling branches (similar to the U-Net structure with jump connections). Then, the structure of the network is reconstructed by gradually increasing the Transformer encoding branch (the initial point of the decoder is the Transformer encoding branch output endpoint) and the VAE decoding branch. Since each structural variant is an end-to-end network structure, there is no deviation caused by model integration. From the results shown in E, the segmentation effect achieved by the basic CNNs encoding and decoding structure is similar to that of the U-Net series. The VAE branch can slightly improve the segmentation performance. The addition of the Transformer encoding branch to form a dual-stream encoding structure can significantly improve the overall network performance due to the effective introduction of global information. Ablation studies

have proved that each individual component of X-Net provides a useful contribution to improving performance.

TABLE II. COMPARISONS ON DSB18 FOR DIFFERENT CONFIGURATIONS

Dataset	DSB 18	
Network	IoU	F1-Score
Baseline	0.7881	0.8809
Baseline+VAE	0.7963	0.8945
Baseline+Transformer	0.8217	0.9016
X-Net	**0.8250**	**0.9036**

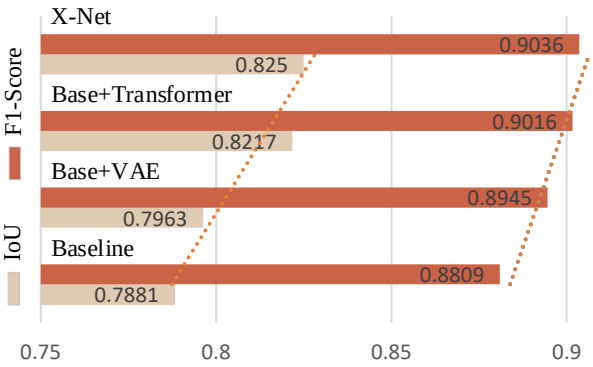

Fig. 4. Contributions of Transformers and VAE on a baseline basis.

IV. CONCLUSION

In this paper, we have conceived a medical image segmentation idea that realizes the organic combination of Transformer structure and traditional convolutional network. Specifically, starting from a dual encoding strategy, the convolutional network and Transformer structure are used to form a dual encoding branch feature extraction path, and local and global features are obtained in parallel. The two features will interact benignly through the skip connection to realize the remote connection of the dual-stream context information. We added a VAE branch in the network decoding part to make up for the performance shortcomings of the current Transformer structure on small-scale datasets. The method of reconstructing the input image strengthens the encoder's ability to cluster features, and at the same time adds constraints to the shared decoder. Therefore, the encoding fineness of the dual encoding form is further refined. Experimental results show that X-Net has achieved excel lent results beyond the classic convolutional network structure in cell segmentation at different resolutions. It reflects Transformer's excellent modeling capabilities for long-term dependencies and CNNs' ability to accurately locate spatial correlations. These can effectively complement each other in medical image segmentation tasks. In future work, we will further study and explore the potential connection between the Transformer structure and the convolutional network structure to improve the performance of the segmentation network.

ACKNOWLEDGMENT

This work is jointly supported by the National Natural Science Foundation of China under Grant No. 61803061, 61906026; Innovation research group of universities in Chongqing; the Chongqing Natural Science Foundation under Grant cstc2020jcyj-msxmX0577, cstc2020jcyj-msxmX0634; "Chengdu-Chongqing Economic Circle" innovation funding of Chongqing Municipal Education Commission KJCXZD2020028; the Science and Technology Research Program of Chongqing Municipal Education Commission grants KJQN202000602; Ministry of Education China Mobile Research Fund (MCM 20180404); Special key project of Chongqing technology innovation and application development: cstc2019jscx-zdztzx0068.

REFERENCES

[1] Naylor, P., Laé, M., Reyal, F., & Walter, T. (2017, April). Nuclei segmentation in histopathology images using deep neural networks. In 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017) (pp. 933-936). IEEE.

[2] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., ... & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. Medical image analysis, 35, 18-31.

[3] Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., ... & Barratt, D. C. (2018). Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE transactions on medical imaging, 37(8), 1822-1834.

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[5] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[7] Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. arXiv preprint arXiv:1406.6247.

[8] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., & Liu, W. (2019). Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 603-612).

[9] Qi, D., Guo, F., & Yu, L. (2007, August). Medical image edge detection based on omni-directional multi-scale structure element of mathematical morphology. In 2007 IEEE International Conference on Automation and Logistics (pp. 2281-2286). IEEE.

[10] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[11] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[13] Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018). Scaling neural machine translation. arXiv preprint arXiv:1806.00187.

[14] Shao, T., Guo, Y., Chen, H., & Hao, Z. (2019). Transformer-based neural network for answer selection in question answering. IEEE Access, 7, 26146-26156.

[15] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019, December). Hierarchical transformers for long document classification. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 838-844). IEEE.

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[17] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2020). Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. arXiv preprint arXiv:2012.15840.

[18] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877.

[19] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[20] Caicedo, J. C., Goodman, A., Karhohs, K. W., Cimini, B. A., Ackerman, J., Haghighi, M., ... & Carpenter, A. E. (2019). Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nature methods, 16(12), 1247-1253.

[21] Naylor, P. , M Laé, Reyal, F. , & Walter, T. . (2018). Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Transactions on Medical Imaging.

[22] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[23] Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162, 94-114.

[24] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

[25] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. arXiv preprint arXiv:2102.10662.

[26] Zhang, Y., Liu, H., & Hu, Q. (2021). Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:2102.08005.

[27] Myronenko, A. (2018, September). 3D MRI brain tumor segmentation using autoencoder regularization. In International MICCAI Brainlesion Workshop (pp. 311-320). Springer, Cham.