Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation

Hu Cao[1†], Yueyue Wang[2†], Joy Chen[1], Dongsheng Jiang[3*], Xiaopeng Zhang[3*], Qi Tian[3*], and Manning Wang[2]

[1] Technische Universität München, München, Germany
[2] Fudan University, Shanghai, China
[3] Huawei Technologies, Shanghai, China

Abstract. In the past few years, convolutional neural networks (CNNs) have achieved milestones in medical image analysis. Especially, the deep neural networks based on U-shaped architecture and skip-connections have been widely applied in a variety of medical image tasks. However, although CNN has achieved excellent performance, it cannot learn global and long-range semantic information interaction well due to the locality of convolution operation. In this paper, we propose Swin-Unet, which is a Unet-like pure Transformer for medical image segmentation. The tokenized image patches are fed into the Transformer-based U-shaped Encoder-Decoder architecture with skip-connections for local-global semantic feature learning. Specifically, we use hierarchical Swin Transformer with shifted windows as the encoder to extract context features. And a symmetric Swin Transformer-based decoder with patch expanding layer is designed to perform the up-sampling operation to restore the spatial resolution of the feature maps. Under the direct down-sampling and up-sampling of the inputs and outputs by 4×, experiments on multi-organ and cardiac segmentation tasks demonstrate that the pure Transformer-based U-shaped Encoder-Decoder network outperforms those methods with full-convolution or the combination of transformer and convolution. The codes and trained models will be publicly available at https://github.com/HuCaoFighting/Swin-Unet.

1 Introduction

Benefiting from the development of deep learning, computer vision technology has been widely used in medical image analysis. Image segmentation is an important part of medical image analysis. In particular, accurate and robust medical image segmentation can play a cornerstone role in computer-aided diagnosis and image-guided clinical surgery [1,2].

*Corresponding author
† Work done as an intern in Huawei Technologies

Existing medical image segmentation methods mainly rely on fully convolutional neural network (FCNN) with U-shaped structure [3,4,5]. The typical U-shaped network, U-Net [3], consists of a symmetric Encoder-Decoder with skip connections. In the encoder, a series of convolutional layers and continuous down-sampling layers are used to extract deep features with large receptive fields. Then, the decoder up-samples the extracted deep features to the input resolution for pixel-level semantic prediction, and the high-resolution features of different scale from the encoder are fused with skip connections to alleviate the loss of spatial information caused by down-sampling. With such an elegant structural design, U-Net has achieved great success in a variety of medical imaging applications. Following this technical route, many algorithms such as 3D U-Net [6], Res-UNet [7], U-Net++ [8] and UNet3+ [9] have been developed for image and volumetric segmentation of various medical imaging modalities. The excellent performance of these FCNN-based methods in cardiac segmentation, organ segmentation and lesion segmentation proves that CNN has a strong ability of learning discriminating features.

Currently, although the CNN-based methods have achieved excellent performance in the field of medical image segmentation, they still cannot fully meet the strict requirements of medical applications for segmentation accuracy. Image segmentation is still a challenge task in medical image analysis. Since the intrinsic locality of convolution operation, it is difficult for CNN-based approaches to learn explicit global and long-range semantic information interaction [2]. Some studies have tried to address this problem by using atrous convolutional layers [10,11], self-attention mechanisms [12,13], and image pyramids [14]. However, these methods still have limitations in modeling long - range dependencies. Recently, inspired by Transformer's great success in the nature language processing (NLP) domain [15], researchers have tried to bring Transformer into the vision domain [16]. In [17], vision transformer (ViT) is proposed to perform the image recognition task. Taking 2D image patches with positional embeddings as inputs and pre-training on large dataset, ViT achieved comparable performance with the CNN-based methods. Besides, data-efficient image transformer (DeiT) is presented in [18], which indicates that Transformer can be trained on mid-size datasets and that a more robust Transformer can be obtained by combining it with the distillation method. In [19], a hierarchical Swin Transformer is developed. Take Swin Transformer as vision backbone, the authors of [19] achieved state-of-the-art performance on Image classification, object detection and semantic segmentation. The success of ViT, DeiT and Swin Transformer in image recognition task demonstrates the potential for Transformer to be applied in the vision domain.

Motivated by the Swin Transformer's [19] success, we propose Swin-Unet to leverage the power of Transformer for 2D medical image segmentation in this work. To our best knowledge, Swin-Unet is a first pure Transformer-based U-shaped architecture that consists of encoder, bottleneck, decoder, and skip connections. Encoder, bottleneck and decoder are all built based on Swin Transformer block [19]. The input medical images are split into non-overlapping image

patches. Each patch is treated as a token and fed into the Transformer-based encoder to learn deep feature representations. The extracted context features are then up-sampled by the decoder with patch expanding layer, and fused with the multi-scale features from the encoder via skip connections, so as to restore the spatial resolution of the feature maps and further perform segmentation prediction. Extensive experiments on multi-organ and cardiac segmentation datasets indicate that the proposed method has excellent segmentation accuracy and robust generalization ability. Concretely, our contributions can be summarized as: (1) Based on Swin Transformer block, we build a symmetric Encoder-Decoder architecture with skip connections. In the encoder, self-attention from local to global is realized; in the decoder, the global features are up-sampled to the input resolution for corresponding pixel-level segmentation prediction. (2) A patch expanding layer is developed to achieve up-sampling and feature dimension increase without using convolution or interpolation operation. (3) It is found in the experiment that skip connection is also effective for Transformer, so a pure Transformer-based U-shaped Encoder-Decoder architecture with skip connection is finally constructed, named Swin-Unet.

2 Related work

CNN-based methods : Early medical image segmentation methods are mainly contour-based and traditional machine learning-based algorithms [20,21]. With the development of deep CNN, U-Net is proposed in [3] for medical image segmentation. Due to the simplicity and superior performance of the U-shaped structure, various Unet-like methods are constantly emerging, such as Res-UNet [7], Dense-UNet [22], U-Net++ [8] and UNet3+ [9]. And it is also introduced into the field of 3D medical image segmentation, such as 3D-Unet [6] and V-Net [23]. At present, CNN-based methods have achieved tremendous success in the field of medical image segmentation due to its powerful representation ability.

Vision transformers : Transformer was first proposed for the machine translation task in [15]. In the NLP domain, the Transformer-based methods have achieved the state-of-the-art performance in various tasks [24]. Driven by Transformer's success, the researchers introduced a pioneering vision transformer (ViT) in [17], which achieved the impressive speed-accuracy trade-off on image recognition task. Compared with CNN-based methods, the drawback of ViT is that it requires pre-training on its own large dataset. To alleviate the difficulty in training ViT, Deit [18] describes several training strategies that allow ViT to train well on ImageNet. Recently, several excellent works have been done baed on ViT [25,26,19]. It is worth mentioning that an efficient and effective hierarchical vision Transformer, called Swin Transformer, is proposed as a vision backbone in [19]. Based on the shifted windows mechanism, Swin Transformer achieved the state-of-the-art performance on various vision tasks including image classification, object detection and semantic segmentation. In this work, we attempt to use Swin Transformer block as basic unit to build a U-shaped Encoder-Decoder

architecture with skip connections for medical image segmentation, thus providing a benchmark comparison for the development of Transformer in the medical image field.

Self-attention/Transformer to complement CNNs : In recent years, researchers have tried to introduce self-attention mechanism into CNN to improve the performance of the network [13]. In [12], the skip connections with additive attention gate are integrated in U-shaped architecture to perform medical image segmentation. However, this is still the CNN-based method. Currently, some efforts are being made to combine CNN and Transformer to break the dominance of CNNs in medical image segmentation [2,27,1]. In [2], the authors combined Transformer with CNN to constitute a strong encoder for 2D medical image segmentation. Similar to [2], [27] and [28] use the complementarity of Transformer and CNN to improve the segmentation capability of the model. Currently, various combinations of Transformer with CNN are applied in multi-modal brain tumor segmentation [29] and 3D medical image segmentation [1,30]. Different from the above methods, we try to explore the application potential of pure Transformer in medical image segmentation.

3 Method

3.1 Architecture overview

The overall architecture of the proposed Swin-Unet is presented in Figure. 1. Swin-Unet consists of encoder, bottleneck, decoder and skip connections. The basic unit of Swin-Unet is Swin Transformer block [19]. For the encoder, to transform the inputs into sequence embeddings, the medical images are split into non-overlapping patches with patch size of 4×4. By such partition approach, the feature dimension of each patch becomes to $4 \times 4 \times 3 = 48$. Furthermore, a linear embedding layer is applied to projected feature dimension into arbitrary dimension (represented as C). The transformed patch tokens pass through several Swin Transformer blocks and patch merging layers to generate the hierarchical feature representations. Specifically, patch merging layer is responsible for down-sampling and increasing dimension, and Swin Transformer block is responsible for feature representation learning. Inspired by U-Net [3], we design a symmetric transformer-based decoder. The decoder is composed of Swin Transformer block and patch expanding layer. The extracted context features are fused with multiscale features from encoder via skip connections to complement the loss of spatial information caused by down-sampling. In contrast to patch merging layer, a patch expanding layer is specially designed to perform up-sampling. The patch expanding layer reshapes feature maps of adjacent dimensions into a large feature maps with $2\times$ up-sampling of resolution. In the end, the last patch expanding layer is used to perform $4\times$ up-sampling to restore the resolution of the feature maps to the input resolution ($W \times H$), and then a linear projection layer is applied on these up-sampled features to output the pixel-level segmentation predictions. We would elaborate each block in the following

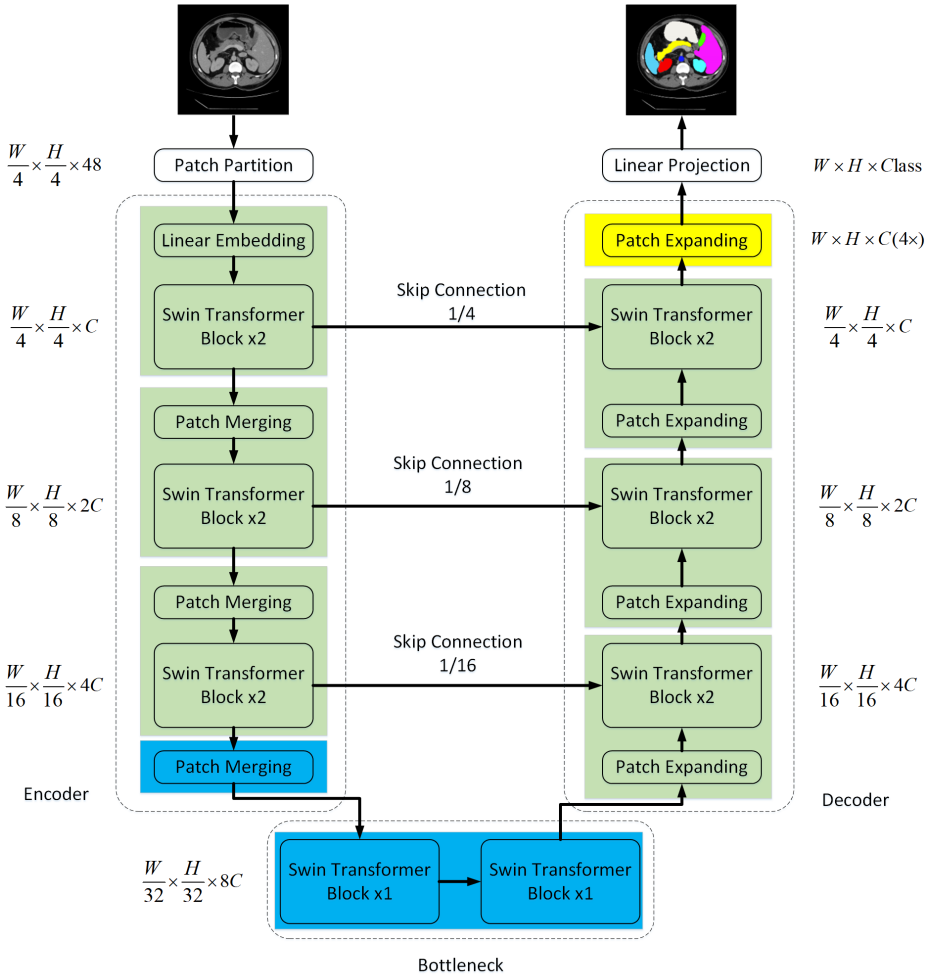

Fig. 1. The architecture of Swin-Unet, which is composed of encoder, bottleneck, decoder and skip connections. Encoder, bottleneck and decoder are all constructed based on swin transformer block.

3.2 Swin Transformer block

Different from the conventional multi-head self attention (MSA) module, swin transformer block [19] is constructed based on shifted windows. In Figure. 2, two consecutive swin transformer blocks are presented. Each swin transformer block is composed of LayerNorm (LN) layer, multi-head self attention module, residual connection and 2-layer MLP with GELU non-linearity. The window-based multi-head self attention (W-MSA) module and the shifted window-based multi-head self attention (SW-MSA) module are applied in the two successive

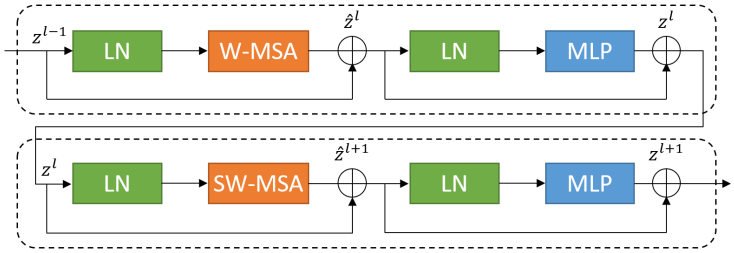

Fig. 2. Swin transformer block.

transformer blocks, respectively. Based on such window partitioning mechanism, continuous swin transformer blocks can be formulated as:

$$\hat{z}^l = W\text{-}MSA(LN(z^{l-1})) + z^{l-1}, \tag{1}$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l, \tag{2}$$

$$\hat{z}^{l+1} = SW\text{-}MSA(LN(z^l)) + z^l, \tag{3}$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}, \tag{4}$$

where $\hat{z}^l$ and z^l represent the outputs of the (S)W-MSA module and the MLP module of the l^{th} block, respectively. Similar to the previous works [31,32], self-attention is computed as follows:

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d}} + B)V, \tag{5}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ denote the query, key and value matrices. M^2 and d represent the number of patches in a window and the dimension of the *query* or *key*, respectively. And, the values in B are taken from the bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M+1)}$.

3.3 Encoder

In the encoder, the C-dimensional tokenized inputs with the resolution of $\frac{H}{4} \times \frac{W}{4}$ are fed into the two consecutive Swin Transformer blocks to perform representation learning, in which the feature dimension and resolution remain unchanged. Meanwhile, the patch merging layer will reduce the number of tokens (2× downsampling) and increase the feature dimension to 2× the original dimension. This procedure will be repeated three times in the encoder.

Patch merging layer : The input patches are divided into 4 parts and concatenated together by the patch merging layer. With such processing, the feature resolution will be down-sampled by 2×. And, since the concatenate operation results the feature dimension increasing by 4×, a linear layer is applied on the concatenated features to unify the feature dimension to the 2× the original dimension.

3.4 Bottleneck

Since Transformer is too deep to be converged [33], only two successive Swin Transformer blocks are used to constructed the bottleneck to learn the deep feature representation. In the bottleneck, the feature dimension and resolution are kept unchanged.

3.5 Decoder

Corresponding to the encoder, the symmetric decoder is built based on Swin Transformer block. To this end, in contrast to the patch merging layer used in the encoder, we use the patch expanding layer in the decoder to up-sample the extracted deep features. The patch expanding layer reshapes the feature maps of adjacent dimensions into a higher resolution feature map (2× up-sampling) and reduces the feature dimension to half of the original dimension accordingly.

Patch expanding layer : Take the first patch expanding layer as an example, before up-sampling, a linear layer is applied on the input features ($\frac{W}{32} \times \frac{H}{32} \times 8C$) to increase the feature dimension to 2× the original dimension ($\frac{W}{32} \times \frac{H}{32} \times 16C$). Then, we use rearrange operation to expand the resolution of the input features to 2× the input resolution and reduce the feature dimension to quarter of the input dimension ($\frac{W}{32} \times \frac{H}{32} \times 16C \rightarrow \frac{W}{16} \times \frac{H}{16} \times 4C$). We will discuss the impact of using patch expanding layer to perform up-sampling in section 4.5.

3.6 Skip connection

Similar to the U-Net [3], the skip connections are used to fuse the multi-scale features from the encoder with the up-sampled features. We concatenate the shallow features and the deep features together to reduce the loss of spatial information caused by down-sampling. Followed by a linear layer, the dimension of the concatenated features is remained the same as the dimension of the up-sampled features. In section 4.5, we will detailed discuss the impact of the number of skip connections on the performance of our model.

4 Experiments

4.1 Datasets

Synapse multi-organ segmentation dataset (Synapse): the dataset includes 30 cases with 3779 axial abdominal clinical CT images. Following [2,34],

Table 1. Segmentation accuracy of different methods on the Synapse multi-organ CT dataset.

Methods	DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net [35]	68.81	-	75.34	51.87	77.10	**80.75**	87.84	40.05	80.56	56.98
DARR [36]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net [2]	74.68	36.87	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
U-Net [3]	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
R50 Att-UNet [2]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-UNet [37]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	**58.04**	87.30	75.75
R50 ViT [2]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUnet [2]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUnet	**79.13**	**21.55**	85.47	66.53	**83.28**	79.61	**94.29**	56.58	**90.66**	**76.60**

18 samples are divided into the training set and 12 samples into testing set. And the average Dice-Similarity coefficient (DSC) and average Hausdorff Distance (HD) are used as evaluation metric to evaluate our method on 8 abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, stomach).

Automated cardiac diagnosis challenge dataset (ACDC): the ACDC dataset is collected from different patients using MRI scanners. For each patient MR image, left ventricle (LV), right ventricle (RV) and myocardium (MYO) are labeled. The dataset is split into 70 training samples, 10 validation samples and 20 testing samples. Similar to [2], only average DSC is used to evaluate our method on this dataset.

4.2 Implementation details

The Swin-Unet is achieved based on Python 3.6 and Pytorch 1.7.0. For all training cases, data augmentations such as flips and rotations are used to increase data diversity. The input image size and patch size are set as 224×224 and 4, respectively. We train our model on a Nvidia V100 GPU with 32GB memory. The weights pre-trained on ImageNet are used to initialize the model parameters. During the training period, the batch size is 24 and the popular SGD optimizer with momentum 0.9 and weight decay 1e-4 is used to optimize our model for back propagation.

4.3 Experiment results on Synapse dataset

The comparison of the proposed Swin-Unet with previous state-of-the-art methods on the Synapse multi-organ CT dataset is presented in Table. 1. Different from TransUnet [2], we add the test results of our own implementations of U-Net [3]and Att-UNet [37] on the Synapse dataset. Experimental results demonstrate that our Unet-like pure transformer method achieves the best performance with segmentation accuracy of 79.13%(DSC↑) and 21.55%(HD↓). Compared with Att-Unet [37] and the recently method TransUnet [2], although our algorithm did not improve much on the DSC evaluation metric, we achieved accuracy improvement of about 4% and 10% on the HD evaluation metric, which

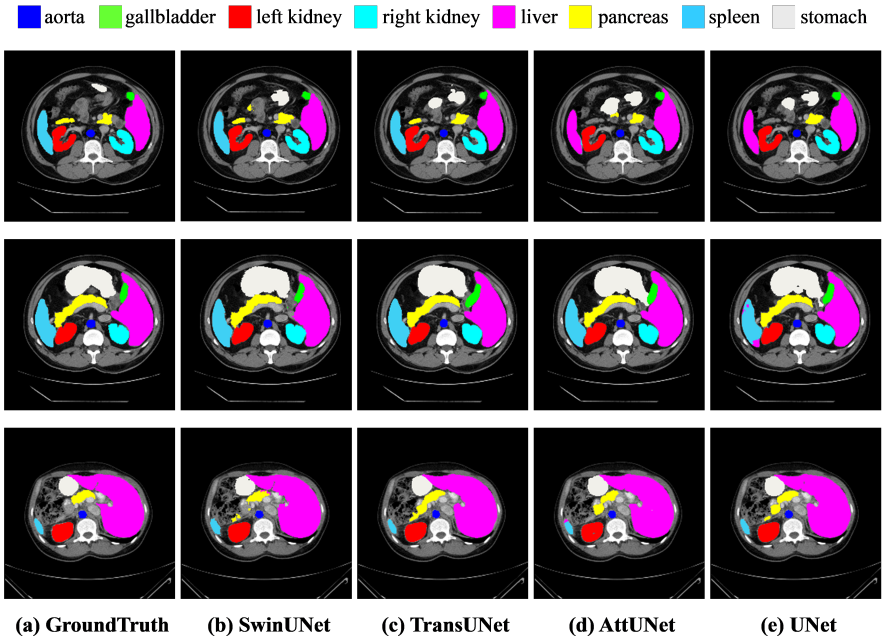

Fig. 3. The segmentation results of different methods on the Synapse multi-organ CT dataset.

Table 2. Segmentation accuracy of different methods on the ACDC dataset.

Methods	DSC	RV	Myo	LV
R50 U-Net	87.55	87.10	80.63	94.92
R50 Att-UNet	86.75	87.58	79.20	93.47
R50 ViT	87.57	86.07	81.88	94.75
TransUnet	89.71	88.86	84.53	95.73
SwinUnet	**90.00**	88.55	**85.62**	**95.83**

indicates that our approach can achieve better edge predictions. The segmentation results of different methods on the Synapse multi-organ CT dataset are shown in Figure. 3. It can be seen from the figure that CNN-based methods tend to have over-segmentation problems, which may be caused by the locality of convolution operation. In this work, we demonstrate that by integrating Transformer with a U-shaped architecture with skip connections, the pure Transformer approach without convolution can better learn both global and long-range semantic information interactions, resulting in better segmentation results.

Table 3. Ablation study on the impact of the up-sampling

Up-sampling	DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
Bilinear interpolation	76.15	81.84	66.33	80.12	73.91	93.64	55.04	86.10	72.20
Transposed convolution	77.63	84.81	65.96	82.66	74.61	**94.39**	54.81	89.42	74.41
Patch expand	**79.13**	**85.47**	**66.53**	**83.28**	**79.61**	94.29	**56.58**	**90.66**	**76.60**

Table 4. Ablation study on the impact of the number of skip connection

Skip connection	DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
0	72.46	78.71	53.24	77.46	75.90	92.60	46.07	84.57	71.13
1	76.43	82.53	60.44	81.36	79.27	93.64	53.36	85.95	74.90
2	78.93	**85.82**	66.27	**84.70**	**80.32**	93.94	55.32	88.35	**76.71**
3	**79.13**	85.47	**66.53**	83.28	79.61	**94.29**	**56.58**	**90.66**	76.60

4.4 Experiment results on ACDC dataset

Similar to the Synapse dataset, the proposed Swin-Unet is trained on ACDC dataset to perform medical image segmentation. The experimental results are summarized in Table. 2. By using the image data of MR mode as input, Swin-Unet is still able to achieve excellent performance with an accuracy of 90.00%, which shows that our method has good generalization ability and robustness.

4.5 Ablation study

In order to explore the influence of different factors on the model performance, we conducted ablation studies on Synapse dataset. Specifically, up-sampling, the number of skip connections, input sizes, and model scales are discussed below.

Effect of up-sampling: Corresponding to the patch merging layer in the encoder, we specially designed a patch expanding layer in the decoder to perform up-sampling and feature dimension increase. To explore the effective of the proposed patch expanding layer, we conducted the experiments of Swin-Unet with bilinear interpolation, transposed convolution and patch expanding layer on Synapse dataset. The experimental results in the Table 3 indicate that the proposed Swin-Unet combined with the patch expanding layer can obtain the better segmentation accuracy.

Effect of the number of skip connections: The skip connections of our Swin-UNet are added in places of the 1/4, 1/8, and 1/16 resolution scales. By changing the number of skip connections to 0, 1, 2 and 3 respectively, we explored the influence of different skip connections on the segmentation performance of the proposed model. In Table 4, we can see that the segmentation performance of the model increases with the increase of the number of skip connections. Therefore, in order to make the model more robust, the number of skip connections is set as 3 in this work.

Table 5. Ablation study on the impact of the input size

Input size	DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
224	79.13	85.47	66.53	83.28	79.61	94.29	56.58	**90.66**	**76.60**
384	**81.12**	**87.07**	**70.53**	**84.64**	**82.87**	**94.72**	**63.73**	90.14	75.29

Table 6. Ablation study on the impact of the model scale

Model scale	DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
tiny	79.13	85.47	66.53	83.28	79.61	**94.29**	56.58	**90.66**	**76.60**
base	**79.25**	**87.16**	**69.19**	**84.61**	**81.99**	93.86	**58.10**	88.44	70.65

Effect of input size: The testing results of the proposed Swin-Unet with 224×224, 384×384 input resolutions as input are presented in Table. 5. As the input size increases from 224×224 to 384×384 and the patch size remains the same as 4, the input token sequence of Transformer will become larger, thus leading to improve the segmentation performance of the model. However, although the segmentation accuracy of the model has been slightly improved, the computational load of the whole network has also increased significantly. In order to ensure the running efficiency of the algorithm, the experiments in this paper are based on 224×224 resolution scale as the input.

Effect of model scale: Similar to [19], we discuss the effect of network deepening on model performance. It can be seen from Table. 6 that the increase of model scale hardly improves the performance of the model, but increases the computational cost of the whole network. Considering the accuracy-speed trade off, we adopt the Tiny-based model to perform medical image segmentation.

4.6 Discussion

As we all known, the performance of Transformer-based model is severely affected by model pre-training. In this work, we directly use the training weight of Swin transformer [19] on ImageNet to initialize the network encoder and decoder, which may be a suboptimal scheme. This initialization approach is a simple one, and in the future we will explore the ways to pre-train Transformer end-to-end for medical image segmentation. Moreover, since the input images in this paper are 2D, while most of the medical image data are 3D, we will explore the application of Swin-Unet in 3D medical image segmentation in the following research.

5 Conclusion

In this paper, we introduced a novel pure transformer-based U-shaped encoder-decoder for medical image segmentation. In order to leverage the power of Transformer, we take Swin Transformer block as the basic unit for feature representation and long-range semantic information interactive learning. Extensive ex-

periments on multi-organ and cardiac segmentation tasks demonstrate that the proposed Swin-Unet has excellent performance and generalization ability.

References

1. A. Hatamizadeh, D. Yang, H. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," 2021.
2. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021.
3. O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.
4. K. S. P. J. M.-H. K. Isensee F, Jaeger PF, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat Methods*, vol. 18(2):203-211, 2021.
5. Q. Jin, Z. Meng, C. Sun, H. Cui, and R. Su, "Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 1471, 2020.
6. Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9901. Springer, Oct 2016, pp. 424–432.
7. X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 327–331, 2018.
8. Z. Zhou, M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation." Springer Verlag, 2018, pp. 3–11.
9. H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," 2020.
10. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
11. Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.
12. J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.
13. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
14. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.

15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
16. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020.
17. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
18. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *CoRR*, vol. abs/2012.12877, 2020.
19. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021.
20. A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Transactions on Medical Imaging*, vol. 22, no. 2, pp. 137–154, 2003.
21. K. Held, E. Kops, B. Krause, W. Wells, R. Kikinis, and H.-W. Muller-Gartner, "Markov random field segmentation of brain mr images," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 878–886, 1997.
22. X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
23. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.
24. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423
25. W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *CoRR*, vol. abs/2102.12122, 2021. [Online]. Available: https://arxiv.org/abs/2102.12122
26. K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *CoRR*, vol. abs/2103.00112, 2021. [Online]. Available: https://arxiv.org/abs/2103.00112
27. J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *CoRR*, vol. abs/2102.10662, 2021.
28. Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," *CoRR*, vol. abs/2102.08005, 2021. [Online]. Available: https://arxiv.org/abs/2102.08005
29. W. Wang, C. Chen, M. Ding, J. Li, H. Yu, and S. Zha, "Transbts: Multimodal brain tumor segmentation using transformer," *CoRR*, vol. abs/2103.04430, 2021. [Online]. Available: https://arxiv.org/abs/2103.04430

30. Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging CNN and transformer for 3d medical image segmentation," *CoRR*, vol. abs/2103.03024, 2021. [Online]. Available: https://arxiv.org/abs/2103.03024
31. H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3588–3597.
32. H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3463–3472.
33. H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," *CoRR*, vol. abs/2103.17239, 2021. [Online]. Available: https://arxiv.org/abs/2103.17239
34. S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, and A. Yuille, "Domain adaptive relational reasoning for 3d multi-organ segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 2020, pp. 656–666.
35. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.
36. S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, and A. Yuille, "Domain adaptive relational reasoning for 3d multi-organ segmentation," Germany, 2020, pp. 656–666.
37. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *IMIDL Conference*, 2018.