DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation

Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, *Senior Member, IEEE*, Guangming Lu, *Member, IEEE*, and David Zhang

***Abstract*— Automatic medical image segmentation has made great progress owing to powerful deep representation learning. Inspired by the success of self-attention mechanism in transformer, considerable efforts are devoted to designing the robust variants of the encoder–decoder architecture with transformer. However, the patch division used in the existing transformer-based models usually ignores the pixel-level intrinsic structural features inside each patch. In this article, we propose a novel deep medical image segmentation framework called dual swin transformer U-Net (DS-TransUNet), which aims to incorporate the hierarchical swin transformer into both the encoder and the decoder of the standard U-shaped architecture. Our DS-TransUNet benefits from the self-attention computation in swin transformer and the designed dual-scale encoding, which can effectively model the non-local dependencies and multiscale contexts for enhancing the semantic segmentation quality of varying medical images. Unlike many prior transformer-based solutions, the proposed DS-TransUNet adopts a well-established dual-scale encoding mechanism that uses dual-scale encoders based on swin transformer to extract the coarse and fine-grained feature representations of different semantic scales. Meanwhile, a well-designed transformer interactive fusion (TIF) module is proposed to effectively perform multiscale information fusion through the self-attention mechanism. Furthermore, we introduce the swin transformer block into the decoder to further explore the long-range contextual information during the up-sampling process. Extensive experiments across four typical tasks for medical image segmentation demonstrate the effectiveness of DS-TransUNet, and our approach significantly outperforms the state-of-the-art methods.**

***Index Terms*— Hierarchical swin transformer, long-range contextual information, medical image segmentation, transformer interactive fusion (TIF) module.**

Manuscript received December 29, 2021; revised March 30, 2022; accepted May 8, 2022. Date of publication May 30, 2022; date of current version June 16, 2022. This work was supported in part by the NSFC Fund under Grant 62176077 and Grant 61906162, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515120055, in part by the Shenzhen Key Technical Project under Grant 2020N046, in part by the Shenzhen Fundamental Research Fund under Grant JCYJ20210324132210025 and Grant JCYJ20210324132212030, in part by the Medical Biometrics Perception and Analysis Engineering Laboratory, Shenzhen, China, in part by the Shenzhen Science and Technology Program under Grant RCBS20200714114910193, and in part by Education Center of Experiments and Innovations at Harbin Institute of Technology, Shenzhen. The Associate Editor coordinating the review process was Rosenda Valdés Arencibia. *(Corresponding authors: Guangming Lu; Zheng Zhang.)*

Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, and Guangming Lu are with the Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: tianbaoge24@gmail.com; chenbingzhi.smile@gmail.com; jiayuxu1998@gmail.com; darrenzz219@gmail.com; luguangm@hit.edu.cn).

David Zhang is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518055, China (e-mail: davidzhang@cuhk.edu.cn).

Digital Object Identifier 10.1109/TIM.2022.3178991

I. Introduction

ACCURATE segmentation of lesions is useful for quantitative evaluation of disease prognosis and therapy efficacy. Measurements based on experts' segmentation might be highly accurate but generally costly and labor-intensive for standard clinical settings. In contrast, automated deployments reap the benefits of a reliable and repeatable process, which can further improve the accuracy and speed of identifying the segmenting objects of interest. Therefore, an automated medical image segmentation technique is highly demanded in clinical diagnosis and scientific research.

Over the past decade, convolutional neural networks (CNNs) have been widely used for various segmentation-related tasks and achieved encouraging performance on medical image segmentation. In particular, recent years have witnessed the remarkable success of fully convolutional networks (FCNs) [1], U-Net [2], and their variants. Both the architectures benefit from the encoder–decoder structure, in which the skip connections directly incorporate high-level semantic features provided by the encoder path with the fine-grained features from the decoder path. However, because there is an inevitable limitation of receptive field in convolution operation and because of the inherent inductive biases presented in the convolutional architectures, they fail to build long-range dependencies and global contexts in images, which impact further improvement of segmentation accuracy. Because of the variability from lesions in target images, these CNN-based methods cannot provide guidelines that are easily adapted to individual differences with the variations in sizes, shapes, and textures.

It should be noted that transformer [3] has sparked tremendous discussion in the computer vision (CV) community. Although transformer is originally designed for sequence-to-sequence modeling in natural language processing (NLP) models, now it is generally considered as more flexible alternatives to CNNs for processing various vision tasks. For example, vision transformers (ViTs) [4] achieve comparable performance on large-scale image classification using 2-D image patches with positional embedding as an input sequence and applying transformers with global self-attention to full-size images. Detection transformer (DETR) [5] is a fully end-to-end object detector based on transformer to explore the relationships of the objects and the global image context for object detection. Recent advances in accurate segmentation are also driven by the power of transformer. Inspired by ViT, TransUNet [6] further combines the advantages of transformer

and U-Net for medical image segmentation. Specifically, it is designed to adopt a transformer-based encoder to process sequences of image patches and a CNN-based decoder with skip connections to enable precise up-sampling feature recovery. However, the performance of TransUNet is still subjective to the limited number of annotated medical images. By introducing an additional gating mechanism in the self-attention module, medical transformer (MedT) [7] proposes a gated axial-attention model that extends the existing transformer architectures to match the dataset of any size.

The main core of the above transformer-based methods is to learn attentive interaction of different patch tokens by leveraging self-attention mechanisms in transformer to explore long-range dependencies. To some extent, the above works are encouraging but still suffer from limitations as follows.

1) These methods generally require dense prediction at the pixel level, which would suffer from tremendous computation requirements on high-resolution images. Note that the computational complexity of self-attention in standard transformer is quadratic to image size.
2) Meanwhile, the tokens of transformer used in most of the existing works are of a fixed scale, which is still difficult to deal with the diversity of pathological changes in challenging conditions.
3) Typically, an image is split into a sequence of nonoverlapping patches, which would be individually transformed into a vector embedding) in each stage. The existing approaches of patch division usually ignore pixel-level intrinsic structural information and local convex topology involved inside each patch, which make the model fail to preserve the local continuity around those patches.

To avoid quadratic complexity of traditional transformer, swin transformer [8] only computes self-attention within nonoverlapping local windows and leverages the shifted window partition to build connections among the windows of each preceding layer. Moreover, swin transformer uses patch merging layer for down-sampling to introduce the pyramid structure which is very important for dense prediction tasks. In this way, swin transformer is able to construct a hierarchical representation by starting from small-sized patches and gradually merging neighbor patches as the network gets deeper. It can be seen that the swin transformer can effectively incorporate inductive bias for spatial locality, as well as hierarchy and translation invariance. Therefore, it is of great significance to study swin transformer for developing reliable vision models on downstream tasks. To overcome the above challenges and take advantage of swin transformer, we propose a novel swin transformer-based encoder–decoder framework, dubbed DS-TransUNet, which is designed to incorporate the advantages of swin transformer to optimize the structure of the standard U-shaped architecture for automatic medical image segmentation.

The proposed DS-TransUNet is formulated by a novel dual-scale encoding mechanism that adopts dual-scale encoders based on hierarchical swin transformer to replace the traditional encoder structure for learning multiscale feature representations. Specifically, each medical image is sliced into nonoverlapping patches at large and small scales, respectively. By taking these two different scale patches as inputs, the proposed dual-scale encoder subnetworks can effectively extract the coarse and fine-grained feature representations of different semantic scales. Moreover, to form a comprehensive integration of these multiscale features obtained by the dual encoder, we develop a well-established transformer interactive fusion (TIF) module. Considering feature misalignment and semantic gap between multiscale feature representations, the TIF module performs traditional self-attention mechanism on them to yield unified feature representations. Such a fusion strategy can fully explore the global dependencies between multiscale features and make them complement each other. Consistent with the traditional U-shaped architectures, the extracted context features would be up-sampled by the designed decoder and fused with unified feature representations from encoder via skip connections. Unlike the most existing methods that are built with the CNN-based decoder, the swin transformer block is also introduced into the decoder of DS-TransUNet to further explore long-range contextual information during the up-sampling process, significantly improving the decoding capability. Benefitting from these improvements, the proposed DS-TransUNet can effectively improve the quality of medical image segmentation.

We mainly evaluate the effectiveness of our DS-TransUNet on multiple medical image datasets. Our experiments mainly involve four typical challenges in clinical diagnosis, including: 1) polyp segmentation on endoscopic images [9]–[11] [12], [13]; 2) skin lesion segmentation on dermatoscopic images [14], [15]; 3) gland segmentation on histology images [16]; and 4) nuclei segmentation on divergent images [17]. Our main contributions are summarized as follows.

1) This article proposes a novel deep swin-transformer-based framework called DS-TransUNet, which effectively incorporates the advantages of hierarchical swin transformer to enhance the functionality and flexibility of the traditional encoder–decoder architecture. The core idea of the proposed method is to combine the swin transformer with the U-shaped architecture for automatic medical image segmentation.
2) A well-designed dual-scale encoding mechanism is proposed to make full use of coarse–fine-tuning features from the encoders with different semantic scales to generate the discriminate feature representations, which can guarantee the semantic consistency between the coarse and fine features.
3) The proposed TIF module is developed to establish global dependencies between features of different scales through the self-attention mechanism, hence effectively fusing multiscale contexts and yielding high-quality semantic segmentation performance.
4) Extensive experiments across four typical tasks for medical image segmentation show that the proposed DS-TransUNet consistently outperforms previous state-of-the-art methods, which can demonstrate the effectiveness and superiority of our method.

The rest of this article is organized as follows. Section II reviews the related works of automatic medical image segmentation, and the description of our proposed DS-TransUNet is given in Section III. Next, the comprehensive experiments and visualization analyses are conducted in Section IV. Finally, Section V makes a conclusion of the whole work.

II. Related Work

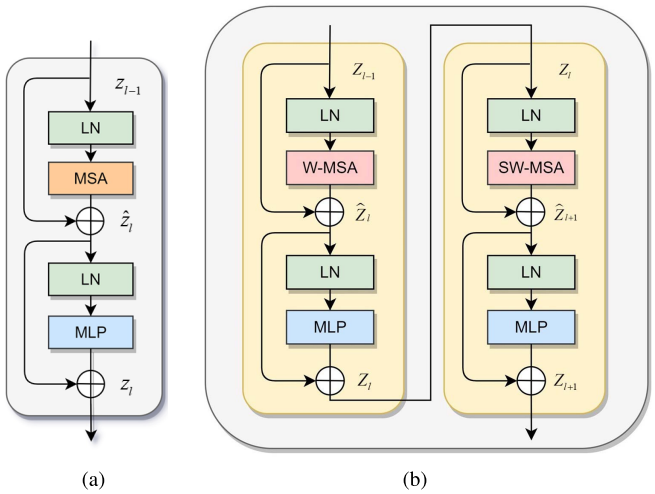
In this section, we first summarize the most typical CNN-based methods used in medical image segmentation, and then we make an overview of the recent related works about the applications of transformer used in CV, especially in the field of segmentation.

A. Medical Image Segmentation Based on CNNs

CNNs, especially the variants of FCN [1] and U-Net [2], have demonstrated superb performance in medical image segmentation. For example, UNet++ [18] designs a series of nested and dense skip connections to reduce the semantic gap between the encoder and the decoder. Attention U-Net [19] proposes a novel attention gate (AG) mechanism that enables the model to focus on targets of different sizes and suppress the irrelevant feature responses. Res-UNet [20] adds a weighted attention mechanism and ResNet-based [21] skip connection scheme to improve the performance of retinal vessel segmentation. R2U-Net [22] combines the strengths of residual networks and U-Net to achieve better feature representation. The parallel reverse attention network (PraNet) [23] is specifically designed for polyp segmentation through the parallel partial decoder (PPD) and reverse attention (RA) module. KiU-Net [24] proposes a novel architecture using both under-complete and over-complete features that makes an improvement in segmenting small anatomical structures. DoubleU-Net [25] is a strong baseline for medical image segmentation using two U-Net in sequence and adopting atrous spatial pyramid pooling (ASPP). FANet [26] unifies the previous epoch mask with the current epoch feature map during training. Note that all these methods are still based on CNNs, so they lack the ability to build long-range dependencies and global context connections. Although there are some works trying to model long-range dependencies for convolution such as [27], [28] [29], [30] [31], [32], they still encounter great limitations in modeling contextual dependencies.

B. Vision Transformer

1) Transformer for Various Vision Tasks: Inspired by the success of transformer [3] in various NLP tasks, more and more transformer-based methods appear in CV tasks because the ability of modeling long-range dependencies based on multihead self-attention (MSA) is also suitable for pixel-based image processing. Among the recent vision transformers, DEtection TRansformer (DETR) [5] uses an elegant design based on transformer to build the first fully end-to-end object detection model. ViT [4] is the first attempt that proves that pure transformer-based architecture can achieve SOTA performance on image recognition when pre-training on large



(a) (b)

Fig. 1. (a) Architecture of a standard transformer block [notation presented with (1)]. (b) Schematic of a swin transformer block [notation presented with (2) and (4)].

datasets such as ImageNet-22K and JFT-300M. DeiT [33] introduces data-efficient training strategies and knowledge distillation that allows ViT to perform well on smaller ImageNet-1K dataset. Swin transformer [8] has linear computational complexity through the proposed window-based MSA (W-MSA) and shifted window-based MSA (SW-MSA), and it achieves state-of-the-art performance in image recognition and dense prediction tasks such as object detection and semantic segmentation. Unlike most previous transformer-based models, swin transformer is flexible to be a general-purpose backbone network by introducing the hierarchical architecture for dense prediction.

2) Transformer for Segmentation: For semantic segmentation, SETR [34] treats semantic segmentation as a sequence-to-sequence prediction task using transformer as encoder, which proves that transformer-based methods can achieve SOTA performance in segmentation tasks. Segmenter [35] is a segmentation framework based purely on transformer, which uses ViT as the encoder and proposes a mask transformer decoder to generate class embeddings. SegFormer [36] is a simple yet powerful segmentation architecture, which consists of a transformer-based hierarchical encoder and a lightweight multilayer perceptron (MLP) decoder. In the field of medical image segmentation, TransUNet [6] is the first attempt to establish self-attention mechanisms by combining transformer with U-Net and proves that transformer can be used as powerful encoders for medical image segmentation. TransFuse [37] was proposed to improve efficiency for modeling global context by fusing transformers and CNNs in a parallel style. MCTrans [38] is a unified transformer network that incorporates rich contextual dependencies and semantic relationships for accurate biomedical segmentation. Furthermore, to train the model effectively on medical image datasets, MedT [7] introduces the gated axial-attention based on axial-deeplab [39] and a local–global training strategy (LoGo). However, all these transformer-based methods designed for medical image segmentation fail to establish pixel-level intrinsic structural features inside patch sequences, while only take

the advantages of transformer in the encoder. Unlike these approaches, we propose a UNet-like architecture which applies swin transformer block to both the encoder and decoder and uncovers the intrinsic structural features by introducing dual-scale encoding mechanism. It is our belief that a unified architecture across the encoder and decoder based on transformer can provide stronger performance in medical image segmentation.

III. Methodology

In this section, our proposed DS-TransUNet is introduced in detail and illustrated in Fig. 2. We first give a brief overview of DS-TransUNet. Then we introduce the standard transformer and swin transformer adopted in DS-TransUNet. Subsequently, we elaborate the encoder based on the swin transformer block. Moreover, we show that DS-TransUNet can benefit from the dual-scale encoding mechanism and describe how we can fuse multiscale feature representations by the TIF module as shown in Fig. 3. Finally, we present the proposed decoder scheme in detail.

A. Overview of DS-TransUNet

Fig. 2 illustrates a detailed pipeline of the proposed DS-TransUNet. In practice, given the input of medical image $I \in \mathbb{R}^{3\times H\times W}$, where $H \times W$ represents the spatial resolution of image instance. Unlike the existing encoder–decoder works, the proposed DS-TransUNet adopts a novel dual-scale encoding mechanism, in which two parallel encoders directly apply swin transformer for encoding coarse and fine feature representations from decomposed image patches of different semantic scales. Before the image I is fed into the dual encoders, it will be split and transformed into sequences of patch embeddings $F^0 \in \mathbb{R}^{(H/4\times W/4)\times C}$ and $G^0 \in \mathbb{R}^{(H/8\times W/8)\times c}$, respectively. Benefitting from the self-attention in transformer, a well-designed TIF module is embedded into each encoding stage to aggregate the learned coarse–fine-tuning features into the unified feature representations. Moreover, our DS-TransUNet introduces the swin transformer into the decoder blocks to further explore the long-range contextual information in the up-sampling process. To recover valuable spatial information, the up-sampling features in each decoder stage are combined into unified feature representations of the corresponding encoding stage using skip connections. Finally, our DS-TransUNet can accurately predict the corresponding pixel-wise semantic label maps with the size of $H \times W$.

B. Swin Transformer as Encoder

The traditional transformer encoder [3] is composed of a stack of N identical blocks. As shown in Fig. 1(a), each block consists of multihead self-attention (MSA) and multilayer perceptron (MLP). Besides, a LayerNorm (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module. Therefore, the output z^l of the lth layer in the transformer encoder can be expressed as

$$\begin{aligned}\hat{\mathbf{z}}^l &= \mathrm{MSA}\big(\mathrm{LN}\big(\mathbf{z}^{l-1}\big)\big) + \mathbf{z}^{l-1},\\ \mathbf{z}^l &= \mathrm{MLP}\big(\mathrm{LN}\big(\hat{\mathbf{z}}^l\big)\big) + \hat{\mathbf{z}}^l.\end{aligned} \tag{1}$$

In the standard MSA mechanism, every token needs to be computed based on its relationships to all other tokens, where the computational complexity is quadratic to the number of tokens, making it unacceptable for many dense prediction and high-resolution image tasks. For efficient modeling, swin transformer [8] proposes W-MSA and SW-MSA.

In W-MSA, the input is a sequence of patches, $z^{l-1} \in \mathbb{R}^{L\times D}$ with length L and dimension D. Specially, z^0 is equal to F^0 or G^0. The input feature will be divided into non-overlapping windows, and each window contains $M \times M$ patches (set to 7 by default). W-MSA will only conduct self-attention within local windows. As shown in Fig. 1, $\hat{z}^l$ and z^l, respectively, represent the outputs of W-MSA and MLP in the lth layer, which are computed by

$$\begin{aligned}\hat{\mathbf{z}}^l &= \text{W-MSA}\big(\mathrm{LN}\big(\mathbf{z}^{l-1}\big)\big) + \mathbf{z}^{l-1},\\ \mathbf{z}^l &= \mathrm{MLP}\big(\mathrm{LN}\big(\hat{\mathbf{z}}^l\big)\big) + \hat{\mathbf{z}}^l.\end{aligned} \tag{2}$$

The problem of W-MSA lies in the weak interaction between windows. To introduce cross-window interaction without additional computation, there is an SW-MSA module followed by the W-MSA structure.

The window configuration of SW-MSA is different from the previous W-MSA layer in which an efficient batch processing approach is developed by cyclic-shifting toward the upper left direction. After this shift, a batch window may be composed of multiple non-adjacent sub-windows in the feature map, while keeping the equal number of batch windows as a regular partitioning in W-MAS. While conducting self-attention within local windows in both W-MSA and SW-MSA, the relative position bias is included in computing similarity.

With such shifted window partitioning mechanism, the outputs of the SW-MSA and MLP module can be written as

$$\begin{aligned}\hat{\mathbf{z}}^{l+1} &= \text{SW-MSA}\big(\mathrm{LN}\big(\mathbf{z}^l\big)\big) + \mathbf{z}^l\\ \mathbf{z}^{l+1} &= \mathrm{MLP}\big(\mathrm{LN}\big(\hat{\mathbf{z}}^{l+1}\big)\big) + \hat{\mathbf{z}}^{l+1}\end{aligned} \tag{3}$$

where $\hat{z}^{l+1}$ and z^{l+1}, respectively, represent the outputs of SW-MSA and MLP in the $(l+1)$th layer, as illustrated in Fig. 1(b). The self-attention computed in W-MSA and SW-MSA can be written as follows:

$$\begin{aligned}Q = z^l\mathbf{W}_Q,\ K = z^l\mathbf{W}_K,\ V = z^l\mathbf{W}_V,\\ \text{Attention}\,(z_l) = \text{SoftMax}(QK^T/\sqrt{d} + B)V\end{aligned} \tag{4}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V \in \mathbb{R}^{D\times d}$ are the learnable parameters of three projection matrices. $Q, K, V \in \mathbb{R}^{L\times d}$ are the *query*, *key*, and *value* matrices; d represents the dimension of *query* or *key*. $B \in \mathrm{R}^{L\times L}$ denotes the relative position bias.

In the overall structure of our model, we refer to [2] using the U-shaped design with the encoder–decoder architecture. For the encoder in DS-TransUNet, we use swin transformer for feature extraction. As shown in Fig. 2, the input medical image will first be sliced into $H/s \times W/s$ nonoverlapping patches, where s is the patch size. Each patch is treated as a "token" and will be projected to C-dimensionality by a linear embedding layer. Since these patches are obtained by convolution operation, no additional position information is needed here. Subsequently, these patch tokens are formally fed into swin transformer that contains four stages. Each

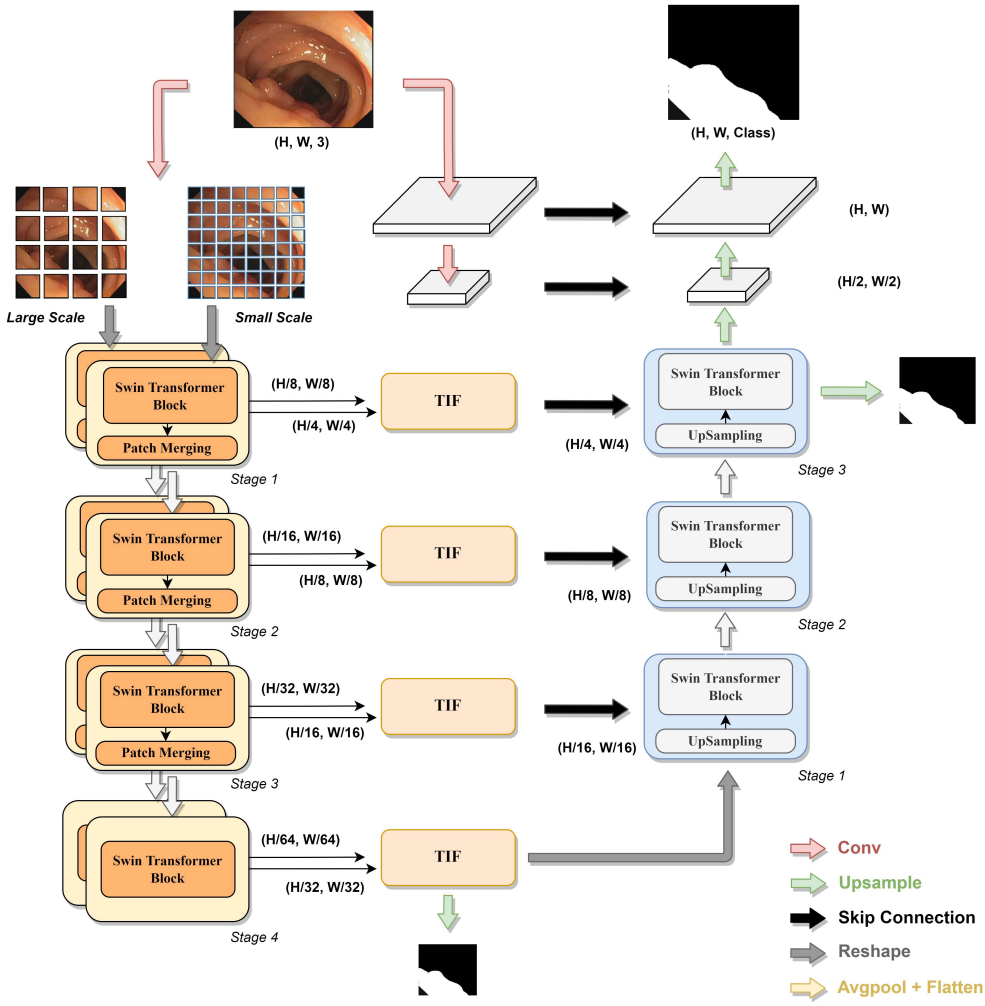


Fig. 2. Illustration of the proposed dual swin transformer U-Net (DS-TransUNet). Given an input medical image, we first split it into nonoverlapping patches at two scales and feed them into the two branches of encoder separately, and then the output feature representations of different scales will be fused by the TIF module. Finally, the fused features are restored to the same resolution as an input image after the up-sampling process based on the swin transformer block. As such, the final mask predictions are obtained.

stage holds a certain number of swin transformer blocks that include window multihead self-attention (W-MSA) and shifted window multihead self-attention (SW-MSA). To produce a hierarchical representation, the number of tokens will be reduced as the network gets deeper. In the first three stages, the inputted features will go through patch merging layer to reduce the feature resolution and increase dimension after swin transformer blocks' transformation. Specifically, the patch merging layer concatenates features of each group of 2×2 neighboring patches, and then applies a linear layer on the channel-dimensional concatenated features. This will reduce the number of tokens by $2\times$, perform $2\times$ downsampling of resolution, and increase the output dimension by $2\times$. So the output resolutions of four stages are $H/s \times W/s$, $H/2s \times W/2s$, $H/4s \times W/4s$, and $H/8s \times W/8s$, and the dimensions are C, $2C$, $4C$, and $8C$, respectively.

C. Dual-Scale Encoding Mechanism

Although the self-attention mechanism can effectively build long-range dependencies between patches, the patch division used in ViT [4] and swin transformer [8] treats an image as a sequence of non-overlapping patches, ignoring the pixel-level intrinsic structure features inside each patch, which will lead to loss of shallow features such as edges and lines information. Although the patch merging layer can introduce location information between patches implicitly, information loss is inevitable at pixel level within each patch. Moreover, Segmenter [35] can obtain better segmentation performance with fine-grained patch size without introducing any parameters. Taking these into account, we use dual-scale swin transformer for feature extraction in the encoder to explore multiscale contexts and preserve the local continuity, which also helps enhance the robustness of our model and improve the segmentation performance.

Patches of different scales can complement each other in feature extraction. Notably, larger scales are privileged to capture coarse-grained feature, while small patches tend to locate fine-grained feature. In [40], dual-scale transformer can alleviate the above problems to a certain extent and achieve better performance than ViT in image recognition.

Motivated by this, the proposed dual-scale encoding mechanism is designed to use dual-scale encoders based on swin transformer. Although patches of each scale will lose the local continuity to a certain extent, the fusion between dual-scale encoders can provide implicit overlap between patches of different scales, just like the overlapping patch embedding used in [36] and [41], which promotes the feature interaction within patches and preserves the local continuity around each patch. Moreover, such an encoding design allows the predictions between features of different scales to complement each other, ensuring consistency between pixel-level features of mask and prediction. Specifically, we use two independent branches with patch size of $s = 4$ (primary) and $s = 8$ (complementary) for feature extraction at different spatial levels. As a result, the output with sizes of $(H/4) \times (W/4) \times C$, $(H/8) \times (W/8) \times 2C$, $(H/16) \times (W/16) \times 3C$, and $(H/32) \times (W/32) \times 4C$ can be obtained from small-scale branch, while the output sizes of large-scale are $(H/8) \times (W/8) \times c$, $(H/16) \times (W/16) \times 2c$, $(H/32) \times (W/32) \times 3c$, and $(H/64) \times (W/64) \times 4c$, respectively. The effect of different patch size combinations will be discussed in Section IV-F.

D. Transformer Interactive Fusion Module

After obtaining the output features from the dual-scale encoder, the remaining problem is how to effectively fuse them to formulate multiscale feature representations' learning. A direct approach is to simply concatenate the multiscale features and then perform convolution operation. However, such a straightforward approach fails to capture the long-range dependencies and global context connection between features at different scales. Therefore, we propose a novel TIF module, which uses the MSA mechanism to enable efficient and effective interaction between multiscale features. In particular, we select the standard transformer block [3] instead of the swin transformer block in TIF, mainly because the latter essentially operates on rectangle-based feature map. But in the multiscale features' fusion module, we need to generate a token at specified size based on feature map of one branch, and then compute self-attention together with the token sequence reshaped by another branch. Moreover, we only need to perform monolayer self-attention operation twice at each stage, which means the computational complexity is acceptable.

As shown in Fig. 3, the proposed TIF can integrate features from two branches of different scales. Herein, we choose small-scale branch for specific analysis, and the same procedure is also applicable to large-scale branches.

To be specific, for outputs of two branches from the same stage $i\,\{i = 1, 2, 3, 4\}$ are denoted as $F^i = [f_1^i; f_2^i; \ldots; f_{h\times w}^i] \in \mathbb{R}^{(h\times w)\times(i\times C)}$ (primary branch) and $G^i = [g_1^i; g_2^i; \ldots; g_{h/2\times w/2}^i] \in \mathbb{R}^{(h/2\times w/2)\times(i\times c)}$ (complementary branch), where $h = H/2^{i+1}$ and $w = W/2^{i+1}$. Then, we obtain the transformed output of G^i by

$$\hat{\mathbf{g}}^i = \mathrm{LP}(\mathrm{Flatten}(\mathrm{Avgpool}(G^i))) \tag{5}$$

where $\hat{g}^i \in \mathbb{R}^{1\times(i\times C)}$ and $\mathrm{LP}(\cdot)$ stands for the linear projection. $\mathrm{Avgpool}(\cdot)$ means the average-pooling layer, followed by the flatten operation. The token $\hat{g}^i$ represents the global abstract information of G^i to interact with F^i at the pixel level. Specifically, F^i is concatenated with $\hat{g}^i$ into a sequence of $1 + h \times w$ tokens, which are fed into the transformer layer for computing global self-attention

$$\begin{aligned}\hat{F}^i &= \mathrm{Transformer}([\hat{g}^i; F^i]) \\ &= [\hat{f}_0^i; \hat{f}_1^i; \ldots; \hat{f}_{h\times w}^i] \in \mathbb{R}^{(1+h\times w)\times(i\times C)} \\ F_{\mathrm{out}}^i &= [\hat{f}_1^i; \hat{f}_2^i; \ldots; \hat{f}_{h\times w}^i] \in \mathbb{R}^{(h\times w)\times(i\times C)}\end{aligned} \tag{6}$$

where transformer$(\cdot)$ plays the same role as (1) and F_{out}^i as the final output of small-scale branch in TIF. This approach introduces connections between each token in $F^i = [f_1^i; f_2^i; \ldots; f_{h\times w}^i]$ and the whole G^i, so that fine-grained feature can also obtain coarse-grained information from the large-scale branch. More importantly, there are indirect interactions between features of different scales, which can effectively preserve the local continuity around the patches. After similar processing, we can obtain $G_{\mathrm{out}}^i \in \mathbb{R}^{(h/2\times w/2)\times(i\times c)}$ from the large-scale branch. Finally, the output feature representation can be acquired as follows:

$$Z_{\mathrm{out}}^i = \mathrm{Conv}_{3\times3}([\mathrm{Up}(G_{\mathrm{out}}^i); F_{\mathrm{out}}^i]) \in \mathbb{R}^{(h\times w)\times(i\times C)} \tag{7}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ is a 3×3 convolution layer and $\mathrm{Up}(\cdot)$ means the $2\times$ bilinear up-sampling process. The resulting features Z_{out}^i are passed to the decoder via skip connections. In this way, the TIF module can bring effective feature fusion of multiscale branches which can improve the segmentation performance. The impact of TIF compared with ordinary multiscale features' fusion based on CNN will be discussed in Section IV-D.

E. Decoder

As shown in Fig. 2, the decoder is mainly composed of three stages, which will be described below. Unlike the previous U-Net [2] and its variants, each stage of our model includes up-sampling (nearest upsampling), skip connection, and swin transformer block. Specifically, the output of stage 4 in the encoder is used as the initial input of the decoder. In each stage of the decoder, the input features are up-sampled by $2\times$, and then concatenated with the appropriate skip connection feature maps from the encoder in the corresponding stage. After that, the output is fed into the swin transformer block for self-attention computation. There are some advantages of such a design: 1) it allows us to make full use of the features from the encoder and up-sampling and 2) it can build long-range dependencies and obtain global context information during the up-sampling process to achieve better decoding performance. The impact of introducing swin transformer block in the decoder will be discussed in Section IV-D. Each stage in the decoder will increase the resolution of feature maps by $2\times$ and reduce the output dimension by $2\times$, so the output resolutions of these three stages are $(H/16) \times (W/16)$, $(H/8) \times (W/8)$, and $(H/4) \times (W/4)$; and the dimensions are C, $2C$, and $4C$, respectively. Finally, we down-sample the input image by cascading two blocks to get low-level features with resolution of $(H/2) \times (W/2)$ and $H \times W$, where each block consists a 3×3 convolutional layer, a group normalization layer, and a ReLU layer successively. All these output features will be

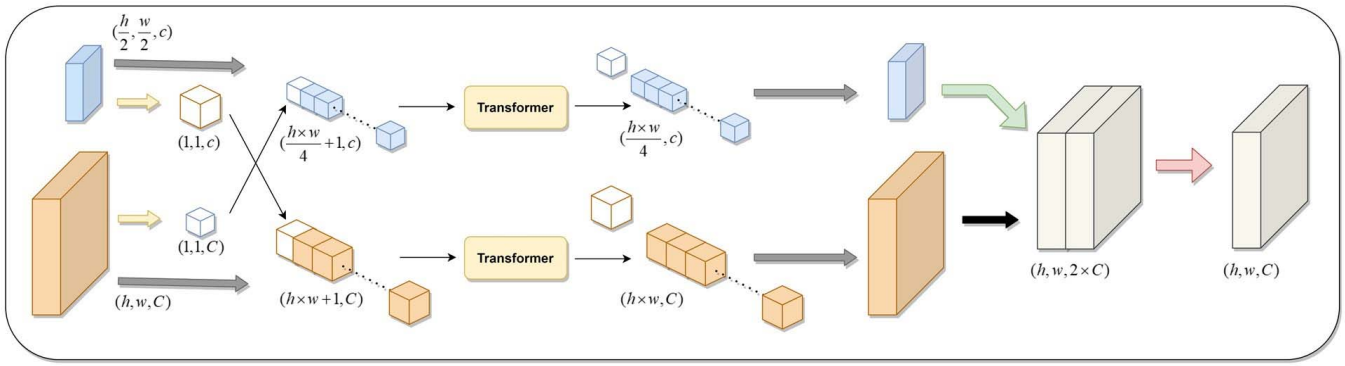


Fig. 3. Illustration of TIF module, which serves as the core component of DS-TransUNet in the multiscale features' fusion process. The annotations of arrows are presented in Fig. 2.

used to get the final segmentation mask with a $H \times W \times N_{\text{cls}}$ resolution, where N_{cls} denotes the number of categories.

F. Loss Function

In the training phase, the proposed DS-TransUNet is trained with the objective function in an end-to-end manner. The loss function is composed of weighted IoU loss $\mathcal{L}_{IoU}^{W}$ and binary cross-entropy loss $\mathcal{L}_{\text{BCE}}^{W}$. Inspired by [23], we find deep supervision helps the model training by additionally supervising the output *S2* of stage 4 in the encoder and *S3* of stage 1 in the decoder, which means the final loss function $\mathcal{L}_{\text{total}}$ can be written as

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \alpha\mathcal{L}(G,S_1)+\beta\mathcal{L}(G,S_2)+\gamma\,\mathcal{L}(G,S_3) \\ \mathcal{L} &= \mathcal{L}_{\text{IoU}}^{W}+\mathcal{L}_{\text{BCE}}^{W} \end{aligned} \tag{8}$$

where G is the ground truth in the training sample. In our experiments, the hyperparameters of $\alpha, \beta, and \gamma$ are empirically set to 0.6, 0.2, and 0.2, respectively.

IV. Experiments

In this section, we first evaluate the performances of our proposed DS-TransUNet on four challenging medical image segmentation tasks in comparison to the state-of-the-art methods. Moreover, we also perform ablation studies to analyze the effect of each component used in DS-TransUNet.

A. Datasets

In our experiments, we evaluate the validity and generalizability of the proposed DS-TransUNet for the polyp segmentation task, which mainly involves five public endoscopic image datasets, including the Kvasir [9], CVC-ClinicDB (ClinicDB) [13], CVC-ColonDB (ColonDB) [10], EndoScene [11], and ETIS [12] datasets. Consistent with the settings of prior works [25], [42], [23], we first perform the comparative experiments on two independent datasets, i.e., Kvasir (880 images for training and 220 images for testing) and CVC-ClinicDB (550 images for training while 62 for testing). To make a fair comparison, the input images from Kvasir and ClinicDB are resized into 512 × 512 and 384 × 384, respectively. Moreover, a cross-study evaluation is also conducted on all the datasets, in which the training set consists of 1450 images (Kvasir: 900 and ClinicDB: 550) and the testing sets contains 798 images (Kvasir: 100, ClinicDB: 62, ColonDB: 380, EndoScene: 60, and ETIS: 196). All the input images in cross-study are resized into 384 × 384 for training and testing.

1) *Kvasir:* As the most commonly used endoscopic image dataset in the field of polyp segmentation, the Kvasir dataset is collected by Vestre Viken Health Trust in Norway from inside the gastrointestinal (GI) tract. It mainly contains 1000 images related to endoscopic polyp removal, which can be used for computer-aided gastrointestinal lesion segmentation.
2) *CVC-ClinicDB:* The CVC-Clinic DB is published by MICCAI'2015 Automatic Polyp Detection in Colonoscopy Videos Sub Challenge. It contains 612 image frames extracted from 29 different colonoscopy sequences. Meanwhile, their pixel-wise ground truth is made up of different segmentation masks, each corresponding to the region covered by the polyp in the image.
3) *ColonDB:* The ColonDB dataset is a database of annotated video sequences from colonoscopy videos. It contains 15 short colonoscopy sequences, which come from 15 different studies. For the polyp segmentation task, only 300 frames are selected from all the sequences and provided with the corresponding high-quality annotations of the whole area covering the polyp region.
4) *EndoScene:* The EndoScene dataset is a combination of CVC-ClinicDB and CVC-300 that contains 912 images from 44 colonoscopy sequences that were acquired from 36 patients total. In our experiments, we only use the 60 samples in CVC-300 as the testing set since part of CVC-612 may not be complete in the training stage.
5) *ETIS:* The ETIS dataset is an early established dataset for early diagnosis of colorectal cancer. It contains 196 polyp images extracted from 34 colonoscopy videos. In contrast, this dataset is more complex and challenging, since the segmenting objects of polyp in ETIS generally have various sizes and shapes.

Furthermore, the proposed DS-TransUNet is also evaluated on three additional medical image segmentation tasks,

including: 1) skin lesion segmentation on ISIC 2018 dataset; 2) gland segmentation on gland segmentation (GLAS) dataset [16]; and 3) nuclei segmentation on 2018 Data Science Bowl (Bowl) dataset [17]. Following the existing works [24], [26], the input images from ISIC 2018 and Bowl are resized into 256×256 in a unified manner, while the input images from GLAS are resized into 128×128.

1) *ISIC 2018:* The ISIC 2018 dataset comes from ISIC-2018 challenge [14], [15] and is useful for skin lesion analysis. It includes 2596 images and their corresponding annotations. In this part, we perform experiments using fivefold cross-validation to show the effectiveness of our DS-TransUNet.
2) *GLAS:* The GLAS dataset [16] is collected from 2015 challenge on gland segmentation in histology images, which provides images of hematoxylin and eosin (H&E)-stained slides. It contains 165 images which are split into 85 images for training and 80 for testing according to [24].
3) *Bowl:* The dataset is collected from 2018 Data Science Bowl challenge [17] and is used to find the nuclei in divergent images, including 670 images in total. We use the same settings as [23], that is, 80% of dataset for training, 10% for validation, and 10% for testing.

B. Experimental Settings

1) Baselines: In addition to the vanilla U-Net [2], two broad approaches are involved in our comparative experiments as baselines, i.e., CNN-based approaches and transformer-based approaches.

1) *CNN-Based Methods:* Some advanced CNN-based models are introduced to compare with our proposed DS-TransUNet, including U-Net [2], Seg-Net [45], UNet++ [18], Attention U-Net [19], R2U-Net [22], BCDU-Net [46], PraNet [23], KiU-Net [24], DoubleU-Net [25], HarDNet-MSEG [44], and FANet [26].
2) *Transformer-Based Methods:* Moreover, several transformer-based models are considered as major contenders, including TransUNet [6], MedT [7], TransFuse [37], Swin-Unet [43], SegFormer [36], and MCTrans [38].

2) Implementation Details: Two variants are presented in this article, i.e., DS-TransUNet-B and DS-TransUNet-L, in which the primary encoders are built on Swin-Base and Swin-Large [8], respectively. Different from the primary encoders, their complementary encoders are initialized with pre-trained Swin-Tiny [8]. In our experiments, the proposed DS-TransUNet is trained with SGD [47] optimizer with a momentum of 0.9 and a weight decay of 0.0001. All the models are trained for 300 epochs with an initial learning rate is 0.01. Moreover, early stopping and cosine annealing schedule are also used in our experiments to adjust the learning rate. All the models are built using PyTorch [48] platform and trained on a NVIDIA RTX 3090 GPU. In our experiment, the results are given as the probability map directly output by the model, which can be binarized to get the final binary mask predictions for performance evaluation. Notably, we only apply the multiscale training strategy in all the experiments, instead of data augmentation.

3) Evaluation Metrics: To compare the existing state-of-the-art methods, we adopt the mean dice coefficient (mDice) (a.k.a. F1), mean intersection over union (mIoU), precision (Pre.), and recall (Rec.) as the key evaluation metrics to measure the extent of similarity between the predicted mask and ground truth.

C. Experimental Results

1) Results on Polyp Segmentation: In this section, the performance of the proposed DS-TransUNet for polyp segmentation is tested under different settings. To make a fair comparison, we first conduct experiments of polyp segmentation with two typical independent datasets, i.e., Kvasir and CVC-ClinicDB, respectively. Moreover, we further perform the cross-study based on all the five independent datasets to verify the effectiveness of the proposed DS-TransUNet. The comparative results with the state-of-the-art methods are presented in Table I, while the corresponding qualitative results are illustrated in Fig. 4.

Based on the above experimental results, we have the following observations.

1) Comparing with the vanilla U-Net, we can observe that various variants of FCN and U-Net, such as UNet++, Attention U-Net, and DoubleU-Net, have met with various degrees of success. For example, Attention U-Net and DoubleU-Net can improve the mDice score by 0.4% and 3.0% on the Kvasi dataset, respectively. Thus, it is of great value to further enhance the stability and plasticity of the standard encoder–decoder architecture.
2) In contrast, some transformer-based models, i.e., MCTrans and TransUNet, are clearly superior to the above variants with the guidance of standard transformer. In particular, TransUNet achieves encouraging mDice scores of 0.896 and 0.923 on the Kvasir and CVC-ClinicDB datasets, respectively. Due to the limitations of the standard transformer, the performance of these models is still below the level of advanced CNN-based methods, such as HarDNet-MSEG and FANet in Table I(a).
3) It can be seen that the proposed DS-TransUNet-L can achieve the highest scores on almost all evaluation metrics for independent datasets. Specifically, the mDice scores of our DS-TransUNet on the Kvasir and CVC-ClinicDB datasets are 0.913 and 0.942, respectively, which clearly outperform the previous state-of-the-art competitors, such as HarDNet-MSEG (0.904) and FANet (0.936). Meanwhile, Fig. 4(a) shows a more precise and fine segmentation output of the proposed network than the existing baselines. This improvement not only can demonstrate the effectiveness of the proposed DS-TransUNet but also indicates that swin transformer has tremendous potential to replace the traditional CNNs.
4) Moreover, as shown in Table II(b), the evaluation results of the proposed DS-TransUNet on cross datasets

TABLE I

QUANTITATIVE RESULTS ON POLYP SEGMENTATION TASK. RESULTS OF THE MODEL WITH "*" ARE REIMPLEMENTED BY THE RELEASED SOURCE CODES. "–" DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. (a) EXPERIMENTAL RESULTS OF POLYP SEGMENTATION ON KVASIR AND CVC-CLINICDB, RESPECTIVELY. (b) EXPERIMENTAL RESULTS BASED ON CROSS-STUDY OF POLYP SEGMENTATION TASK. RESULTS OF SETR-PUP ARE OBTAINED FROM [37]

(a)

Kvasir						CVC-ClinicDB					
Method	Year	mDice	mIoU	Rec.	Pre.	Method	Year	mDice	mIoU	Rec.	Pre.
U-Net* [2]	2015	0.783	0.684	0.808	0.828	U-Net* [2]	2015	0.872	0.804	0.868	0.917
UNet++* [18]	2018	0.784	0.678	0.817	0.820	UNet++* [18]	2018	0.881	0.819	0.910	0.885
Attention U-Net* [19]	2018	0.787	0.686	0.793	0.852	Attention U-Net* [19]	2018	0.890	0.827	0.887	0.909
DoubleU-Net [25]	2020	0.813	0.733	0.840	0.861	Swin-Unet* [43]	2021	0.906	0.849	0.918	0.907
MCTrans [38]	2021	0.862	-	-	-	SegFormer* [36]	2021	0.911	0.860	0.942	0.911
FANet [26]	2021	0.880	0.810	0.906	0.901	HarDNet-MSEG* [44]	2021	0.918	0.864	0.912	0.945
Swin-Unet* [43]	2021	0.890	0.825	0.906	0.906	MCTrans [38]	2021	0.923	-	-	-
TransUNet* [6]	2021	0.896	0.833	0.912	0.913	TransUNet* [6]	2021	0.923	0.869	0.942	0.917
HarDNet-MSEG [44]	2021	0.904	0.848	0.923	0.907	DoubleU-Net [25]	2020	0.924	0.861	0.846	**0.959**
SegFormer* [36]	2021	0.909	0.848	0.935	0.904	FANet [26]	2021	0.936	0.894	0.934	0.940
DS-TransUNet-B (ours)	-	0.911	0.856	0.935	0.914	DS-TransUNet-B (ours)	-	0.935	0.885	0.946	0.931
DS-TransUNet-L (ours)	-	**0.913**	**0.859**	**0.936**	**0.916**	DS-TransUNet-L (ours)	-	**0.942**	**0.894**	**0.950**	0.937

(b)

		Kvasir		ClinicDB		ColonDB		EndoScene		ETIS		Average	
Method	Year	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net [2]	2015	0.818	0.746	0.823	0.755	0.512	0.444	0.398	0.335	0.710	0.626	0.652	0.581
Attention U-Net* [19]	2018	0.814	0.730	0.850	0.789	0.561	0.484	0.773	0.682	0.371	0.305	0.674	0.598
U-Net++ [18]	2018	0.821	0.743	0.794	0.729	0.483	0.410	0.401	0.344	0.707	0.624	0.641	0.570
PraNet [23]	2020	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567	0.800	0.739
Swin-Unet* [43]	2021	0.896	0.835	0.899	0.836	0.759	0.666	0.850	0.764	0.681	0.586	0.817	0.737
SegFormer* [36]	2021	0.904	0.844	0.891	0.826	0.762	0.674	0.856	0.780	0.748	0.658	0.832	0.756
SETR-PUP [34]	2021	0.911	0.854	0.934	0.885	0.773	0.690	0.889	0.814	0.726	0.646	0.847	0.778
HarDNet-MSEG [44]	2021	0.912	0.857	0.932	0.882	0.731	0.660	0.887	0.821	0.677	0.613	0.828	0.767
TransUNet* [6]	2021	0.912	0.860	0.910	0.856	0.797	0.715	0.887	0.815	0.754	0.671	0.852	0.783
TransFuse-S [37]	2021	0.918	0.868	0.918	0.868	0.773	0.696	0.902	0.833	0.733	0.659	0.849	0.785
TransFuse-L [37]	2021	0.918	0.868	0.934	0.886	0.744	0.676	0.904	0.838	0.737	0.661	0.847	0.786
DS-TransUNet-B (ours)	-	0.934	0.888	**0.938**	**0.891**	0.798	0.717	0.882	0.810	**0.772**	**0.698**	0.865	0.801
DS-TransUNet-L (ours)	-	**0.935**	**0.889**	0.936	0.887	**0.798**	**0.722**	**0.911**	**0.846**	0.761	0.687	**0.868**	**0.806**

consistently outperform the previous competitors, which can effectively prove the generalization ability of our DS-TransUNet. Compared with TransFuse, our DS-TransUNet-L has improvements of 2.1% and 2.0% in terms of the averaged mDice and averaged mIoU scores, respectively. From Fig. 4(b), we can see that DS-TransUNet produces high-quality segmentation masks on cross-study of the polyp segmentation task.

5) Both DS-TransUNet-B and DS-TransUNet-L have promising capability of identifying the segmenting objects of interest from endoscopic images, but DS-TransUNet-L achieves a better performance than DS-TransUNet-B (0.868 versus 0.865). It is because larger model scale will bring better learning ability, but the computational cost is greatly increased at the same time. To sum up, these comparative results demonstrate the superiorities of the proposed DS-TransUNet for automated polyp segmentation.

2) Results on ISIC 2018 Dataset: To evaluate the effectiveness of the proposed DS-TransUNet, we conduct comparative experiments on the ISIC-2018 dataset for the task of skin lesion segmentation. The comparison results with the state-of-the-art methods are presented in Table II, while the corresponding quantitative results are illustrated in Fig. 5(a). From Table II, we have the following observations.

1) It can be seen that the attention-guided models, such as Attention R2U-Net (0.691) and FANet (0.873), achieve a better performance than the vanilla U-Net (0.674),

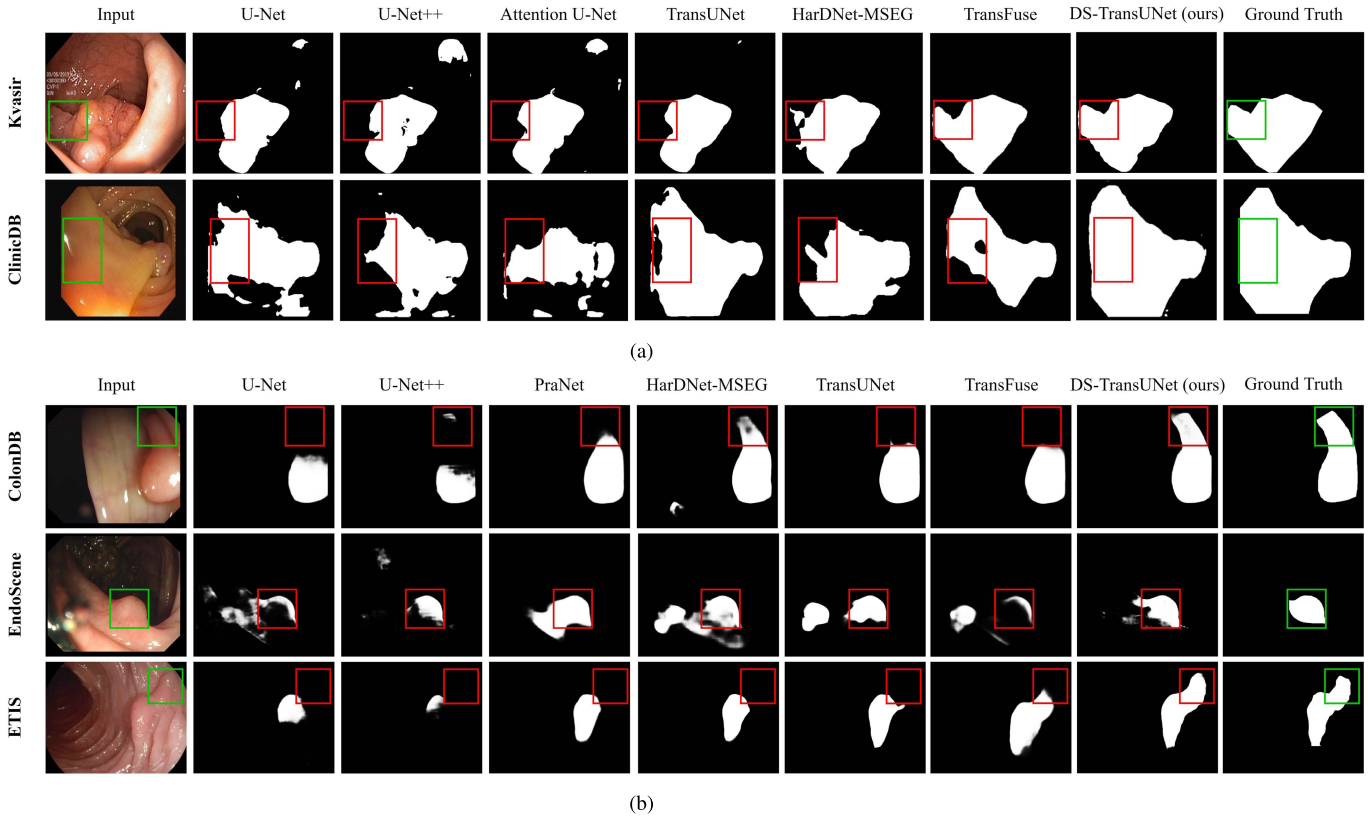


Fig. 4. Comparison of qualitative results between DS-TransUNet and the existing models on the polyp segmentation task. To better visualize the differences between segmentation predictions and ground truths, we highlight the key region with appropriate boxes. (a) Qualitative results of polyp segmentation on Kvasir and CVC-ClinicDB, respectively. (b) Qualitative results based on cross-study of polyp segmentation on ColonDB, EndoScene, and ETIS, respectively.

TABLE II

QUANTITATIVE RESULTS ON ISIC 2018 DATASET. RESULTS OF THE MODEL WITH "*" ARE REIMPLEMENTED BY THE RELEASED SOURCE CODES. "-" DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Method	Year	mDice	mIoU	Rec.	Pre.
U-Net [2]	2015	0.674	0.549	0.708	-
Attention U-Net [19]	2018	0.665	0.566	0.717	-
R2U-Net [22]	2018	0.679	0.581	0.792	-
Attention R2U-Net [22]	2018	0.691	0.592	0.726	-
BCDU-Net (d=3) [46]	2019	0.851	-	0.785	-
FANet [26]	2021	0.873	0.802	0.865	0.924
DoubleU-Net [25]	2020	0.896	0.821	0.878	**0.946**
Swin-Unet* [43]	2021	0.897	0.829	0.903	0.920
SegFormer* [36]	2021	0.902	0.836	0.911	0.921
MCTrans [38]	2021	0.904	-	-	
TransUNet* [6]	2021	0.906	0.841	0.913	0.923
DS-TransUNet-B (ours)	-	0.910	0.848	0.911	0.934
DS-TransUNet-L (ours)	-	**0.913**	**0.852**	**0.922**	0.927

TABLE III

QUANTITATIVE RESULTS ON THE GLAS DATASET. RESULTS OF THE MODEL WITH "*" ARE REIMPLEMENTED BY THE RELEASED SOURCE CODES. "–" DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Method	Year	mDice	mIoU	Rec.	Pre.
Seg-Net [45]	2017	0.786	0.660	-	-
U-Net [2]	2015	0.796	0.672	0.845	0.778
MedT [7]	2021	0.810	0.696	-	-
UNet++ [18]	2018	0.813	0.696	0.857	0.798
Attention UNet* [19]	2018	0.816	0.701	0.844	0.813
TransUNet* [6]	2021	0.818	0.704	0.871	0.795
KiU-Net [24]	2020	0.833	0.728	0.889	0.809
Swin-Unet* [43]	2021	0.867	0.773	0.890	0.861
DS-TransUNet-B (ours)	-	0.873	0.785	**0.895**	0.863
DS-TransUNet-L (ours)	-	**0.878**	**0.791**	0.888	**0.878**

which indicates that the traditional U-shaped model can be optimized by additional attention computation.

2) Moreover, some CNN-based methods take advantage of multiscale contexts to optimize the structure of U-Net, such as BCDU-Net (0.851) and DoubleU-Net (0.896), which can demonstrate the validity of multiscale context fusion.
3) In contrast, the transformer-based models, i.e., Swin-Unet (0.897), SegFormer (0.902), MCTrans (0.904), and TransUNet (0.906), outperform the above methods. It is verified that transformer enables the scheme to identify

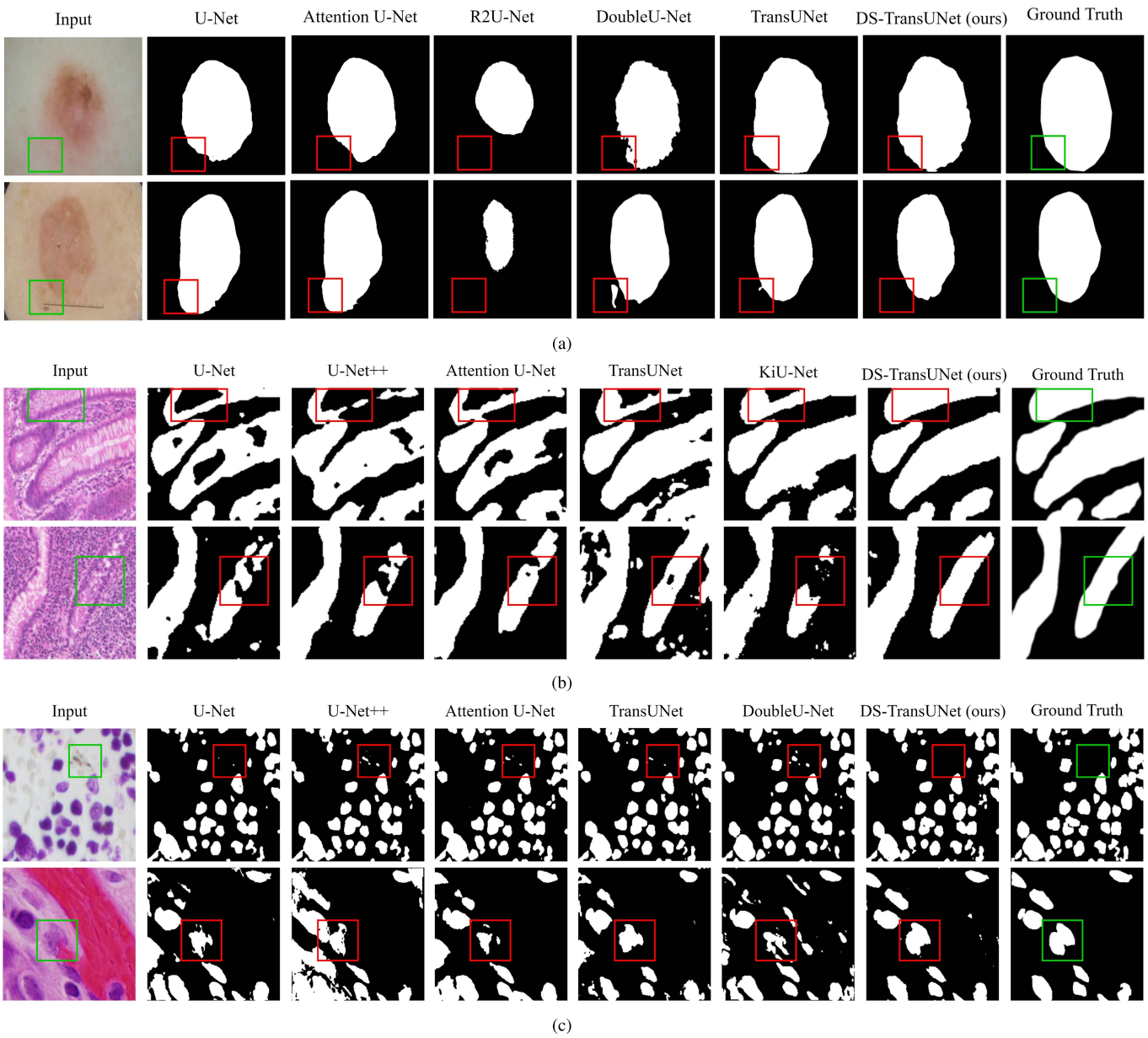


Fig. 5. Qualitative results of DS-TransUNet on three medical image segmentation tasks compared with other models. (a) ISIC 2018 dataset, (b) GLAS dataset, and (c) 2018 Data Science Bowl dataset, respectively. To better visualize the differences between segmentation predictions and ground truths, we highlight the key region with appropriate boxes.

the segmenting objects of interest to be implemented with a better operability, which offers a promising way to solve the defections in the conceptual design of the encoder–decoder architecture.

4) Benefitting from the self-attention computation in swin transformer and multiscale contexts provided by the dual-scale encoding mechanism, the proposed DS-TransUNet consistently outperforms these transformer-based competitors, improving the F1 scores from 0.906 to 0.913. As illustrated in Fig. 5(a), we can observe that our DS-TransUNet can effectively capture the boundaries of skin lesions and generate better segmentation prediction. Thus, these comparative results again demonstrate the powerful ability of our DS-TransUNet in skin lesion segmentation.

3) Results on GLAS Dataset: Moreover, the proposed DS-TransUNet is also evaluated on a small-scale dataset, i.e., GLAS, to automatically quantify the morphology of glands. The experimental results with the state-of-the-art methods are presented in Table III, and the corresponding quantitative results are illustrated in Fig. 5(b).

From Table III, we have the following observations.

1) It is obvious that the proposed DS-TransUNet still outperforms the previous baselines and yields the highest mDice and mIoU scores of 0.878 and 0.791, which can prove the effectiveness of DS-TransUNet.
2) Compared with the previous state-of-the-art baseline, i.e., KiU-Net, our DS-TransUNet easily surpasses KiU-Net by 4.5% and 6.3% in terms of mDice and mIoU scores, which effectively proves that our method can also

TABLE IV

QUANTITATIVE RESULTS ON THE 2018 DATA SCIENCE BOWL DATASET. RESULTS OF THE MODEL WITH "*" ARE REIMPLEMENTED BY THE RELEASED SOURCE CODES. "−" DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Method	Year	mDice	mIoU	Rec.	Pre.
U-Net [2]	2015	0.757	0.910	-	-
UNet++ [18]	2018	0.897	**0.926**	-	-
TransUNet* [6]	2021	0.907	0.836	0.923	0.897
Attention UNet [19]	2018	0.908	0.910	-	0.916
DoubleU-Net [25]	2020	0.913	0.841	0.641	**0.950**
Swin-Unet* [43]	2021	0.916	0.851	0.924	0.915
FANet [26]	2021	0.918	0.857	0.922	0.919
SegFormer* [36]	2021	0.919	0.855	0.931	0.912
DS-TransUNet-B (ours)	-	0.920	0.859	**0.943**	0.906
DS-TransUNet-L (ours)	-	**0.922**	0.861	0.938	0.912

produce high-quality segmentation performance with a small number of training samples.

3) In particular, the proposed DS-TransUNet is clearly superior to the recent transformer-based work, i.e., MedT (0.810), TransUNet (0.818), and Swin-Unet (0.867), which again demonstrate the advantages of dual-scale encoding mechanism and TIF module for gland segmentation.
4) Besides, we also present the visualization of generated mask images in Fig. 5(b), which demonstrates that our DS-TransUNet can bring excellent performance to distinguish the gland itself from the surrounding tissue.

4) Results on 2018 Data Science Bowl: Furthermore, we also evaluate the proposed DS-TransUNet on the 2018 Data Science Bowl dataset for the task of multiple nuclei segmentation. The comparison results with these state-of-the-art methods are presented in Table IV, and the corresponding quantitative results are illustrated in Fig. 5(c).

From Table IV, we have the following observations.

1) Consistent with other evaluation results, the proposed DS-TransUNet outperforms the existing baselines by exploring the self-attention computation and multiscale contexts.
2) Specifically, our DS-TransUNet achieves the highest scores 0.922 and 0.943 in terms of F1 and recall, respectively. Compared with previous advanced works, such as DoubleU-Net (0.913) and FANet (0.918), the F1 score of our DS-TransUNet achieves the improvements of 0.9% and 0.4%, respectively.
3) As illustrated in Fig. 5, we can see that our DS-TransUNet can concurrently predict the boundaries of dozens of cell nuclei much more accurately than the existing baselines. This phenomenon indicates that our DS-TransUNet also has a powerful ability of nuclei segmentation on divergent images.
4) Therefore, the above experimental results can further verify the generalization ability of DS-TransUNet for different tasks of medical image segmentation.

D. Ablation Study

In this section, we further conduct the ablation studies on the task of polyp segmentation to evaluate the ability of each component in the proposed DS-TransUNet. The experimental results are presented in Table V. Here, U-Net is considered as a vanilla baseline. "U w/ TE" denotes the U-shaped model with a standard transformer-based encoder. "U w/ SE" denotes the U-shaped model with the swin-transformer-based encoder, "U w/ SE + SD" represents the U-shaped model with both the swin-transformer-based encoder and decoder. "U w/ DSE + SD" is the U-shaped model with the proposed dual-swin-transformer-based encoder and swin-transformer-based decoder. "U w/ DSE + SD + TIF" is the full DS-TransUNet architecture, with the proposed dual-scale encoding component, swin-transformer-based decoder and TIF module.

From Table V, we have the following observations.

1) When we use transformer-based encoder to replace the traditional encoder, "U w/ TE" achieves a significant improvement of 19.5% and 20.5% in terms of the average mDice and mIou scores, respectively. Compared with vanilla U-Net, it can demonstrate that transformer is beneficial to encode contextual information.
2) In contrast, "U w/ SE" outperforms "U w/ TE," especially for the average mDice score by the margin of 0.6%, which can validate the superiorities of swin transformer in the encoder against the standard transformer.
3) Meanwhile, the proposed decoder can effectively model the long-range dependencies with swin transformer, so that "U w/ SE + SD" is clearly superior to "U w/ SE" (0.863 versus 0.853).
4) By adding the dual-scale encoding mechanism, "U w/ DSE + SD" can achieve an improvement of 0.2% and 0.3% in terms of the average mDice and mIoU scores, respectively, which shows that the use of additional encoding branch can generate discriminate feature representations, hence improving segmentation performance.
5) Although "U w/ DSE + SD" can yield a reliable performance, "U w/ DSE + SD + TIF" can further enhance the average mDice and mIoU scores from 0.865 to 0.868 and 0.801 to 0.806, respectively. Such an improvement proves unquestioningly that our proposed TIF module can guarantee the consistency between different features and improve segmentation performance.

Based on the above experimental results, it can be concluded that all the designed components play an indispensable role in the task of medical image segmentation.

E. Number of Parameters

To compare the model size and computational complexity, we further conduct experiments on the ClinicDB dataset. From Table VI, we have the following observations: 1) to enlarge the receptive field, the CNN-based methods usually require stacking sufficiently deep convolutional layers, which leads to high computational cost; 2) the self-attention mechanism requires more parameters than convolution operation, which makes transformer-based methods larger size; and

TABLE V

ABLATION STUDIES BASED ON CROSS-STUDY OF POLYP SEGMENTATION TASK. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

	Kvasir		ClinicDB		ColonDB		EndoScene		ETIS		Average	
Method	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net (baseline) [2]	0.818	0.746	0.823	0.755	0.512	0.444	0.398	0.335	0.710	0.626	0.652	0.581
U w/ TE (Transfuse) [37]	0.918	0.868	0.934	0.886	0.744	0.676	0.904	0.838	0.737	0.661	0.847	0.786
U w/ SE	0.926	0.876	0.923	0.875	0.791	0.709	0.889	0.816	0.734	0.650	0.853	0.785
U w/ SE + SD	0.929	0.879	0.929	0.880	0.795	0.717	0.904	0.836	0.759	0.677	0.863	0.798
U w/ DSE + SD	0.932	0.885	0.931	0.881	0.797	0.719	0.908	0.841	0.758	0.677	0.865	0.801
U w/ DSE + SD + TIF	**0.935**	**0.889**	**0.936**	**0.887**	**0.798**	**0.722**	**0.911**	**0.846**	**0.761**	**0.687**	**0.868**	**0.806**

TABLE VI

PERFORMANCE COMPARISON OF THE MODEL SIZE (PARAMS) AND THEORETICAL COMPUTATIONAL COMPLEXITY (FLOPS) BETWEEN DS-TRANSUNET AND OTHER LEADING METHODS ON THE CLINICDB DATASET. "*" MEANS DIRECTLY PERFORM 32× UP-SAMPLING ON THE OUTPUT OF SWIN-L

Method	Params	FLOPs	mDice	mIoU
U-Net [2]	24.56M	38.26G	0.872	0.804
U-Net++ [18]	25.09M	84.30G	0.881	0.819
Attention U-Net [19]	25.09M	40.07G	0.890	0.827
DoubleU-Net [25]	29.30M	107.9G	0.924	0.861
MCTrans [38]	23.79M	39.71G	0.923	-
SegFormer [36]	84.59M	19.06G	0.911	0.860
TransUNet [6]	105.28M	24.66G	0.923	0.869
TransFuse [37]	115.59M	38.73G	0.928	0.876
Swin-Unet [43]	149.22M	30.14G	0.906	0.849
Swin-L* [8]	198.63M	34.02G	0.912	0.849
DS-TransUNet-B (ours)	171.44M	30.97G	0.935	0.885
DS-TransUNet-L (ours)	287.75M	51.09G	0.942	0.894

TABLE VII

PERFORMANCE COMPARISON OF USING DIFFERENT COMBINATIONS OF PATCH SIZE ON THE KVASIR AND GLAS DATASETS

	Kvasir		GLAS		
Patch Size	mDice	mIoU	mDice	mIoU	FLOPs
(4, 8)	0.911	0.856	0.873	0.785	30.97G
(4, 4)	0.900	0.844	0.868	0.778	34.27G
(4, 12)	0.907	0.853	0.869	0.780	30.52G
(8, 12)	0.878	0.811	0.828	0.720	30.26G

3) although self-attention is introduced into TIF and decoder, DS-TransUNet can not only produce a good complexity parameter trade-off but also achieve the best segmentation performance.

F. Effect of Patch Size

To achieve a trade-off between accuracy and memory footprint, the transformer-based models usually divide the image into non-overlapping patches. A smaller path size is beneficial to a more detailed feature representation, but resulting in a long sequence that requires more compute time and memory footprint. To investigate the effect of patch size in our work, we conduct experiments on the Kvasir and GLAS datasets by testing the combinations of different patch sizes.

From Table VII, we have the following observations: 1) when both the encoding branches have the same patch size, i.e., the combinations of (4, 4), our DS-TransUNet struggles to have a satisfactory result due to the lack of complementary feature information; 2) to ensure the quality of the proposed dual-scale encoding mechanism, we further develop one of the two branches with larger patch sizes, i.e., the combinations of (4, 8). It can be seen that our DS-TransUNet can achieve the best performance with acceptable FLOPs; and 3) however, we observe a significant drop in performance when the patch sizes of the two branches are set to 4 and 8. It is because an oversize patch size of the encoder provides inadequate fine-grained features for medical image segmentation, causing the pixel-level accuracy to decrease. Based on the above experimental results, it can be concluded that the optimal combination of patch sizes used in the encoding branches is 4 and 8.

V. CONCLUSION

In this article, we presented the dual swin transformer U-Net (DS-TransUNet), a U-shaped encoder-decoder-based framework for improving the segmentation quality of biomedical image. Our DS-TransUNet was designed based on the hierarchical swin transformer. Besides the encoder, we also innovatively added the swin transformer block to the decoder, allowing to model global contexts throughout the network. Moreover, we introduced a novel dual-scale encoding mechanism in the encoder to extract multiscale feature representations. We further proposed a novel TIF module to build long-range dependencies between features of different scales through the self-attention mechanism, which could effectively fuse the multiscale feature representations from the encoder. Extensive experiments on four medical image segmentation tasks demonstrated that our DS-TransUNet significantly outperformed the previous state-of-the-art methods especially in polyp segmentation task. In the future, we will focus on designing more lightweight transformer-based models and efficiently learning pixel-level intrinsic structural features generated by patch division in vision transformers.

REFERENCES

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

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[3] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[4] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[6] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[7] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 36–46.

[8] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

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

[10] J. Bernal *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

[11] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2015.

[12] D. Vázquez *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 2017, pp. 1–9, Jul. 2017.

[13] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, 2014.

[14] N. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[15] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, Dec. 2018.

[16] K. Sirinukunwattana *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, Jan. 2017.

[17] J. C. Caicedo *et al.*, "Nucleus segmentation across imaging experiments: The 2018 data science bowl," *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[18] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

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

[20] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted Res-UNet for high-quality retina vessel segmentation," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Oct. 2018, pp. 327–331.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[22] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[23] D.-P. Fan *et al.*, "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2020, pp. 263–273.

[24] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "KiU-Net: Towards accurate segmentation of biomedical images using over-complete representations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2020, pp. 363–373.

[25] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[26] N. K. Tomar *et al.*, "FANet: A feedback attention network for improved biomedical image segmentation," 2021, *arXiv:2103.17235*.

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

[28] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[29] J. Li, H. Huo, C. Li, R. Wang, C. Sui, and Z. Liu, "Multigrained attention network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[30] J. Tang, B. Zou, C. Li, S. Feng, and H. Peng, "Plane-wave image reconstruction via generative adversarial network and attention mechanism," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[31] X. Zhou *et al.*, "Dense attention-guided cascaded network for salient object detection of strip steel surface defects," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[32] J. Huang, M. Tu, W. Yang, and W. Kang, "Joint attention network for finger vein authentication," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[34] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[35] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

[36] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.

[37] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 14–24.

[38] Y. Ji *et al.*, "Multi-compound transformer for accurate biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 326–336.

[39] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 108–126.

[40] C.-F.-R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 357–366.

[41] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[42] D. Jha *et al.*, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40496–40510, 2021.

[43] H. Cao *et al.*, "Swin-UNet: Unet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

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

[45] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.

[46] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with Densley connected convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 406–415.

[47] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[48] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.

Ailiang Lin received the B.S. degree in computer science and technology from South China Normal University, Guangzhou, China, in 2020. He is currently pursuing the M.S. degree with the School of Computer Technology, Harbin Institute of Technology, Shenzhen, China.

His current research interests include computerized medical diagnosis, pattern recognition, deep learning, and machine learning.

Bingzhi Chen received the B.S. degree in software engineering from South China Normal University, Guangzhou, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China.

His current research interests include computerized medical diagnosis, pattern recognition, deep learning, and machine learning.

Jiayu Xu received the B.S. degree in software engineering from South China Normal University, Guangzhou, China, in 2020. She is currently pursuing the M.S. degree with the School of Computer Technology, Harbin Institute of Technology, Shenzhen, China.

Her current research interests include pattern recognition, deep learning, and machine learning.

Zheng Zhang (Senior Member, IEEE) received the M.S. degree in computer science and the Ph.D. degree in computer applied technology from the Harbin Institute of Technology, Harbin, China, in 2014 and 2018, respectively.

He was a Post-Doctoral Research Fellow with The University of Queensland, Queensland, NSW, Australia; a Research Associate with The Hong Kong Polytechnic University, Hong Kong; and a Visiting Researcher with the National Laboratory of Pattern Recognition (NLPR), Chinese Academy of Sciences (CAS), Beijing, China. He is currently an Assistant Professor with the Harbin Institute of Technology, Shenzhen, China. He has authored or coauthored over 60 technical papers at prestigious international journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), the IEEE TRANSACTIONS ON CYBERNETICS (TCYB), the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), the European Conference on Computer Vision (ECCV), the Association for the Advancement of Artificial Intelligence (AAAI), the ACM International Conference on Multimedia (ACMM), the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), and the International Joint Conference on Artificial Intelligence (IJCAI). His current research interests include machine learning, computer vision, and multimedia analytics.

Dr. Zhang serves/served as a (lead) Guest Editor for *Information Processing and Management* journal and *Neurocomputing* journal, a Publication Chair for the 16th International Conference on Advanced Data Mining and Applications (ADMA 2020), and an SPC/PC Member for several top conferences.

Guangming Lu (Member, IEEE) received the B.S. degree in electrical engineering, the M.S. degree in control theory and control engineering, and the Ph.D. degree in computer science and engineering from the Harbin Institute of Technology (HIT), Shenzhen, Harbin, China, in 1998, 2000, and 2005, respectively.

He was a Post-Doctoral Fellow with Tsinghua University, Beijing, China, from 2005 to 2007. He is currently a Professor with the Bio-Computing Research Center, HIT. He has published over 120 technical papers at prestigious international journals and conferences, including IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), IEEE TRANSACTIONS ON CYBERNETICS (TCYB), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), the Association for the Advancement of Artificial Intelligence (AAAI), the ACM International Conference on Multimedia (ACMM), and the International Joint Conference on Artificial Intelligence (IJCAI). His current research interests include pattern recognition, image processing, and automated biometric technologies and applications.

David Zhang received the B.S. degree in computer science from Peking University, Beijing, China, in 1974, the M.Sc. and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Shenzhen, Harbin, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Post-Doctoral Fellow with Tsinghua University, Beijing, and then an Associate Professor with Academia Sinica, Beijing. He has been a Chair Professor with The Hong Kong Polytechnic University, Hong Kong. He is currently a Presidential Chair Professor with The Chinese University of Hong Kong, Shenzhen.

Dr. Zhang is a fellow of the Academy of Science of the Royal Society of Canada.