OXFORD CDE

Journal of Computational Design and Engineering, 2022, 9(2), 616–632

https://doi.org/10.1093/jcde/qwac018
Journal homepage: www.jcde.org

RESEARCH ARTICLE

SwinE-Net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer

Kyeong-Beom Park and Jae Yeol Lee*

Department of Industrial Engineering, Chonnam National University, 77, Yongbong-ro, Buk-gu, Gwangju 61186, South Korea

*Corresponding author. E-mail: jaeyeol@chonnam.ac.kr

Abstract

Prevention of colorectal cancer (CRC) by inspecting and removing colorectal polyps has become a global health priority because CRC is one of the most frequent cancers in the world. Although recent U-Net-based convolutional neural networks (CNNs) with deep feature representation and skip connections have shown to segment polyps effectively, U-Net-based approaches still have limitations in modeling explicit global contexts, due to the intrinsic nature locality of convolutional operations. To overcome these problems, this study proposes a novel deep learning model, SwinE-Net, for polyp segmentation that effectively combines a CNN-based EfficientNet and Vision Transformer (ViT)-based Swin Ttransformer. The main challenge is to conduct accurate and robust medical segmentation in maintaining global semantics without sacrificing low-level features of CNNs through Swin Transformer. First, the multidilation convolutional block generates refined feature maps to enhance feature discriminability for multilevel feature maps extracted from CNN and ViT. Then, the multifeature aggregation block creates intermediate side outputs from the refined polyp features for efficient training. Finally, the attentive deconvolutional network-based decoder upsamples the refined and combined feature maps to accurately segment colorectal polyps. We compared the proposed approach with previous state-of-the-art methods by evaluating various metrics using five public datasets (Kvasir, ClinicDB, ColonDB, ETIS, and EndoScene). The comparative evaluation, in particular, proved that the proposed approach showed much better performance in the unseen dataset, which shows the generalization and scalability in conducting polyp segmentation. Furthermore, an ablation study was performed to prove the novelty and advantage of the proposed network. The proposed approach outperformed previous studies.

Keywords: polyp segmentation; convolutional neural networks; multidilation convolutional block; multifeature aggregation block; Swin Transformer; Vision Transformer

1. Introduction

Colorectal cancer (CRC) is the third most frequent cancer after breast and lung cancers (Ferlay, 2010). Prevention of CRC by inspecting and removing colorectal adenomas (i.e. paraneoplastic lesions) has become a worldwide priority concerning global healthcare. In 2020, over 104 000 cases and 53 200 deaths were estimated in the United States due to CRC (Siegel *et al.*, 2020). Colorectal polyps can be detected by colonoscopic inspection, performed by clinical endoscopists using video screening and diagnosis. However, roughly 8–37% of polyps are missed during

Received: 8 October 2021; **Revised:** 15 January 2022; **Accepted:** 23 January 2022

the colonoscopic examination due to mistakes and lack of experience (Van Rijn *et al.*, 2006).

Furthermore, the shape and complexity of polyps (e.g. tiny size, complicated shape, blurred boundary due to colors and textures, etc.) make the inspection more difficult. In this respect, computer-assisted polyp segmentation systems can reduce the endoscopic miss rates of colorectal polyps and assist clinical endoscopists in localizing complex colon polyps (Hassan *et al.*, 2020). These systems can allow endoscopists to focus their attention on the polyps shown on the screen, decreasing the likelihood of overlooked polyps (Mori & Kudo, 2018).

Deep learning-based convolutional neural networks (CNNs) have recently shown breakthrough in image segmentation tasks. In particular, fully convolutional networks (FCNs) significantly contributed to the research of semantic segmentation by conducting feature extraction through CNN-based downsampling and generating pixel-wise binary masks through image interpolation-based upsampling (Long *et al.*, 2015). U-Net also showed outstanding performance in semantic segmentation for biomedical applications (Ronneberger *et al.*, 2015). U-Net consists of a path for capturing context and a symmetric path that enables feature extraction and semantic mask generation. Recently, ResUNet++ has shown good performance in semantic segmentation using the neural network with residual blocks, atrous spatial pyramidal pooling, and attention blocks (Jha *et al.*, 2019). Although U-Net-based CNNs have shown very effective performance, it is still very challenging to segment various types of polyps accurately and robustly because the polyp's texture, color, size, and shape are different. In addition, it is very ambiguous to distinguish between the colorectal polyp and background, such as mucosa. Although recent U-Net-based CNNs have been proposed in medical segmentation, they have still difficulty in modeling and extracting global-level semantic features because the coverage of the receptive field of the CNN model is limited and narrow (Brandao *et al.*, 2017; Sun *et al.*, 2019; Fan *et al.*, 2020a; Guo *et al.*, 2020; Park *et al.*, 2020; Huang *et al.*, 2021; Tomar *et al.*, 2021a).

Recently, the Vision Transformer (ViT) was proposed to overcome the limitation of CNNs. ViT splits an input image into patches and conducts linear embeddings of these patches (Dosovitskiy *et al.*, 2020). Furthermore, a hierarchical transformer called Swin Transformer was proposed to extend ViT by extracting cascaded features like CNN (Liu *et al.*, 2021). Although good results are obtained using ViT, which converts the application in natural language processing to image processing, it still derived lower performance than CNNs because it requires very large datasets and spends much time for training models (Touvron *et al.*, 2021). In particular, ViT treats the input as 1D sequences and exclusively focuses on modeling the global context at all stages, which results in low-resolution features that lack detailed localization information. Furthermore, this information cannot be effectively recovered by direct upsampling to the full resolution, therefore leading to a coarse segmentation outcome (Chen *et al.*, 2021).

This study proposes a novel polyp segmentation method, SwinE-Net, that effectively combines the CNN-based EfficientNet and ViT-based Swin Transformer models by utilizing patch-level (low-level) spatial features and global-level (image-level) semantic features together. Thus, SwinE-Net takes full advantage of both models that complement each other. The main contribution is to effectively improve segmentation robustness and accuracy in maintaining global semantics without sacrificing low-level features of the CNN by combining Swin Transformer. First, the multidilation convolutional block refines multilevel feature maps extracted from EfficientNet and Swin Transformer separately by considering the relationship between receptive fields' eccentricity and size and enhancing the feature robustness and discriminability. Then, the multifeature aggregation block aggregates the refined feature maps to generate side outputs corresponding to the intermediate polyp masks for EfficientNet and Swin Transformer, respectively. Finally, the attentive deconvolutional network-based decoder upsamples the combined and refined feature maps through attention and inception module-based blocks, generating the final polyp mask more accurately and robustly. The segmentation loss function consists of the binary cross-entropy (BCE) loss and the intersection over union (IoU) loss for the global and local contexts. As shown in Fig. 1, the proposed SwinE-Net (Swin Transformer + EfficientNet) generates robust and accurate polyp segmentations, almost identical to the ground truth. To verify the proposed SwinE-Net, we conducted comparative evaluations with previous state-of-the-art (SOTA) approaches using five public datasets: Kvasir (Jha *et al.*, 2020), ClinicDB (Bernal *et al.*, 2015), ColonDB (Tajbakhsh *et al.*, 2015), ETIS (Silva *et al.*, 2014), and EndoScene (Vázquez *et al.*, 2017). As suggested by previous studies (Fan *et al.*, 2020a; Huang *et al.*, 2021), we trained only two seen datasets (Kvasir and ClinicDB). Then, the other three unseen datasets (ColonDB, ETIS, and EndoScene) were evaluated based on the result of the seen datasets to prove the robustness of the proposed approach. A statistical evaluation was also conducted to assess the robustness of the proposed approach. An ablation study was performed to verify the effectiveness of the proposed network using the Kvasir EndoScene datasets. Furthermore, we have conducted an additional comparative evaluation using the COVID-19 CT Segmentation Dataset (2020) to verify the extensibility and generality of the proposed method.

The proposed SwinE-Net has the following main contributions:

1) SwinE-Net is a novel deep learning model for polyp segmentation that effectively combines the CNN-based EfficientNet and the ViT-based Swin Transformer by applying multidilation convolution, multifeature aggregation, and attentive deconvolution. Thus, SwinE-Net can keep global-level semantic contexts without sacrificing low-level contexts.
2) The visualization of multilevel feature maps through the deep learning modules of SwinE-Net confirms the maintenance and refinement of global and local contexts for accurate segmentation.
3) The attentive deconvolutional network-based decoder conducts polyp segmentation more accurately and robustly by applying attention and inception block-based multiscale upsampling.
4) The comparative analysis proves that SwinE-Net outperforms previous SOTAs in five public datasets. In particular, it verified much better performance in the unseen dataset, which shows its generalization and scalability in conducting polyp segmentation.

The paper is organized as follows: In Section 2, we describe the related work. Section 3 presents the proposed SwinE-Net for polyp segmentation. Section 4 presents the comparative evaluation with previous studies and the ablation study. Comparative evaluation is conducted based on five public datasets for polyp segmentation, including two seen datasets and three unseen datasets. Finally, Section 5 concludes the paper with some future works.

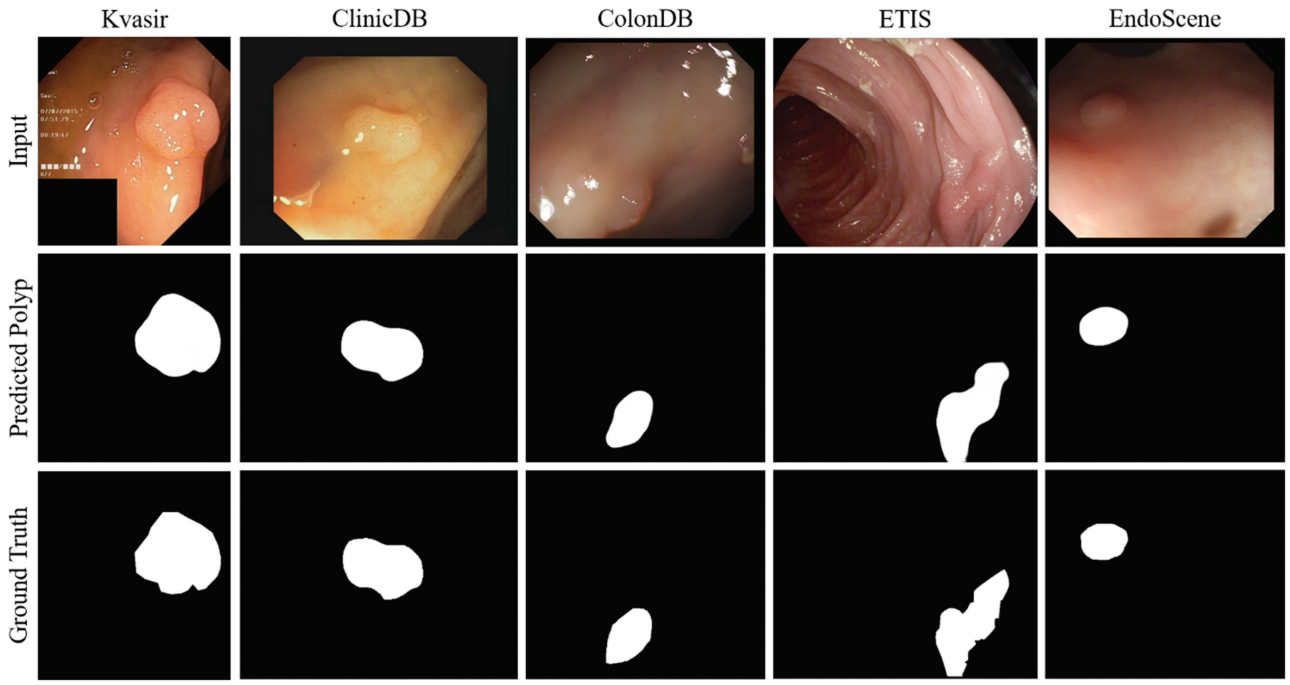

Figure 1: Examples of the colorectal polyp segmentation using the proposed SwinE-Net.

2. Related Work

We analyse previous studies on medical image segmentation. In particular, we focus on U-Net-like architectures because they showed good performance on pixel-wise semantic and medical segmentation. Recently, extended U-Net architectures using feature aggregation and attention modules derived better performance than the vanilla U-Net. In particular, some studies using ViT were conducted to complement the problems of CNN-based methods that are difficult to extract global features. This section reviews previous studies, including CNN-based, extended U-Net-based, and ViT-based medical image segmentation.

2.1. CNN-based medical image segmentation

Some studies conducted medical image segmentation using fully convolutional neural networks (FCN). Brandao *et al.* (2017) utilized FCN to identify and segment polyps in colonoscopy images. Their approach converted three established networks into a fully convolution network and fine-tuned their learned representations to the polyp segmentation task. Guo *et al.* (2020) proposed a polyp segmentation algorithm based on FCNs. They performed several experiments by examining design parameters, data augmentations, network configurations, and polyp characteristics. Vania *et al.* (2019) developed an automatic method to segment the spine from CT images. They combined CNN and FCN and utilized class redundancy as a soft constraint. Vania and Lee (2021) also proposed a multistage optimization mask RCNN to automatically segment the functional abnormalities in the intervertebral disc for treating lower back pain.

In particular, U-Net-based medical image segmentation was conducted by many previous studies. U-Net has an encoder–decoder network consisting of symmetric convolution layers. Sun *et al.* (2019) proposed a deep learning framework for colorectal polyp segmentation, consisting of an encoder to extract multiscale semantic features and a decoder to expand the semantic features to a polyp segmentation mask. Tomar *et al.* (2021a) proposed a dual decoder attention network for automatic polyp segmentation (DDANet). Their method followed an encoder–decoder mechanism and incorporated a single encoder that was shared by two parallel decoders. The first decoder acted as a segmentation network and the second decoder acted as an auto-encoder network.

Park *et al.* (2020) proposed M-GAN for retinal blood vessel segmentation using stacked U-Net and generative adversarial network (GAN). Safarov and Whangbo (2021) proposed an adaptive densely connected U-Net for poly segmentation. Their method applied multiple dilated convolutions with different atrous rates to observe a large field of view and an attention mechanism to reduce noise and inappropriate information for polyp segmentation. Tomar *et al.* (2021b) proposed a feedback attention network that unified the previous epoch mask with the feature map of the current training epoch. Although most CNN-based approaches could perform well in segmentation tasks, they still have difficulty dealing with complex polyp segmentation with the ambiguity between the background and polyp regions.

2.2. Advanced U-Net-based medical image segmentation

Dense U-Net-based approaches for medical image segmentation were also conducted for medical segmentation, which used dilated convolution or dense connection of multilevel features. Nguyen *et al.* (2020) proposed a multimodel deep encoder–decoder network called MED-Net. The method could capture multilevel contextual information by extracting discriminative features at different fields of view and multiple image scales. Fan *et al.* (2020c) proposed a multiscale attention network by utilizing a self-attention mechanism to adaptively integrate local features with their global dependences. Mahmud *et al.* (2021) designed depth dilated inception blocks, deep fusion skip module, and deep reconstruction module for automatic polyp segmen-

tation. Seo *et al.* (2019) proposed a method for segmenting livers and liver tumors using modified U-Net, which added a residual path with deconvolution and activation operations to the skip connection of the U-Net to avoid duplication of low-resolution information of features.

Instead of the complicated decoder, some studies conducted medical image segmentation using a receptive field block and a cascaded partial decoder, which effectively integrated multilevel features of the encoder to generate more accurate segmentation. Fan *et al.* (2020a) proposed a parallel reverse attention network (PraNet) for polyp segmentation. Using multi feature aggregation, they first aggregated the features in high-level layers, which generated a global map for polyp segmentation. Also, they mined the boundary cues using the reverse attention (RA) module, which established the relationship between areas and boundary cues. Fan *et al.* (2020b) proposed Inf-Net for COVID-19 lung infection segmentation, where the implicit RA and explicit edge attention were applied to model boundaries and enhance feature representations. Huang *et al.* (2021) proposed HarDNet-MSEG for polyp segmentation, which used a low-memory traffic CNN with a dense architecture. However, most previous studies generated a lot of blurs and false positives or false negatives in the segmented outputs. Therefore, it is challenging to extract global features from the whole point of view, but it is still difficult to extract them using only CNN-based approaches that mainly focus on local spatial features.

2.3. Transformer-based image processing

Some studies conducted transformer-based classification and segmentation in various fields. Dosovitskiy *et al.* (2020) showed that a transformer applied to the sequence of image patches could perform image classification tasks effectively without dependence on CNNs. The proposed ViT attained excellent results compared to the SOTA convolutional networks. Xie *et al.* (2021) proposed a transformer-based segmentation pipeline termed Trans2Seg for transparent object segmentation. In contrast to CNN's local receptive field, it provided the global receptive field, which proved advantageous over CNN architectures. Mao *et al.* (2021) researched applying the transformer networks for salient object detection. They investigated the contributions of two strategies to provide more substantial spatial supervision through the transformer layers within their framework, namely deep supervision and difficulty-aware learning. Liu *et al.* (2021) proposed a new ViT, Swin Transformer, with a hierarchical transformer whose representation was computed with shifted windows. This hierarchical architecture had linear computational complexity for the image size. Swin Transformer has shown better performance than ViT.

Chen *et al.* (2021) proposed TransUNet, which employed a fused CNN-transformer model to support the detailed spatial information of CNN features and the global contextual information of transformers for medical image segmentation. Cao *et al.* (2021) proposed Swin-UNet, a U-Net-like transformer-based segmentation. The tokenized image patches were fed into the Swin-UNet with skip connections using the hierarchical Swin Transformer. Although ViT-based approaches showed promising outcomes in some applications such as image classification compared with CNNs, CNN-based segmentation approaches still have many advantages over ViT-based segmentation approaches, such as detailed and low-level features. Thus, it is necessary to propose a new method that takes full advantage of both CNN and Swin Transformer, training low-level spatial features and global semantic features.

3. SwinE-Net for Polyp Segmentation

SwinE-Net segments colorectal polyps more accurately and robustly through multidilation convolution (MDC)- and multifeature aggregation (MFA)-based feature refinement and attentive deconvolution-based upsampling, which takes advantage of both CNN and Swin Transformer. The architecture of the proposed SwinE-Net is shown in Fig. 2. First, feature maps are extracted independently from the CNN-based EfficientNet and the ViT-based Swin Transformer. These extracted multilevel features are then refined through the multidilation convolutional block (Liu & Huang, 2018). Second, the refined features are aggregated to generate the side output through the multifeature aggregation block (Wu *et al.*, 2019) corresponding to the intermediate polyp segmentation map. Third, the decoder based on an attentive deconvolutional network upsamples the features generated by the MDC and generates the final polyp segmentation through the attention module and multilevel Inception-ResNet (Szegedy *et al.*, 2017) blocks. The loss function consists of the BCE loss and the IoU losses. The proposed approach simultaneously trains the side outputs and the final segmentation output to optimize segmentation performance.

3.1. EfficientNet and Swin Transformer for extracting multilevel features

EfficientNet plays the main role in extracting low-level spatial features in SwinE-Net. As shown in Fig. 3a, the performance of EfficientNet has been improved using a compound scaling method that uniformly scales all dimensions of depth (the number of layers), width (the number of channels), and resolution (size of the input image) with a simple yet highly effective compound coefficient (Tan & Le, 2019). This network achieved better accuracy and efficiency than previous ConvNets on ImageNet.

Swin Transformer is used for extracting global contexts, a hierarchical ViT whose network is computed with shifted windows. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to nonoverlapping local windows while also allowing for cross-window connection. This hierarchical architecture has linear computational complexity for the image size (Liu *et al.*, 2021). As shown in Fig. 3b, the input image is passed through the patch partition layer to separate the patch with a specific size, and then the feature is extracted through the linear embedding layer and the Swin Transformer block. The second stage passes through the patch merging layer and performs a downsampling process that merges adjacent 2×2 patches into one patch. As the network deepens, it can extract hierarchical representations like CNN such as VGG and ResNet. The Swin Transformer block is shown in Fig. 3b. Instead of multihead self-attention (MSA), Swin Transformer used window-multihead self-attention (W-MSA) and shifted window-multihead self-attention (SW-MSA) followed by a multilayer perceptron (MLP) with Gaussian error linear unit nonlinearity in between. In particular, a LayerNorm (LN) layer is introduced before each MSA module and MLP, and a residual connection is conducted after each module (Liu *et al.*, 2021). Swin Transformer block is represented as follows:

$$\hat{f}^{l} = \text{W\_MSA}\left(\text{LN}\left(f^{l-1}\right)\right) + f^{l-1}, \tag{1}$$

$$f^{l} = \text{MLP}\left(\text{LN}\left(\hat{f}^{l}\right)\right) + \hat{f}^{l}, \tag{2}$$

$$\hat{f}^{l+1} = \text{SW\_MSA}\left(\text{LN}\left(f^{l}\right)\right) + f^{l}, \tag{3}$$

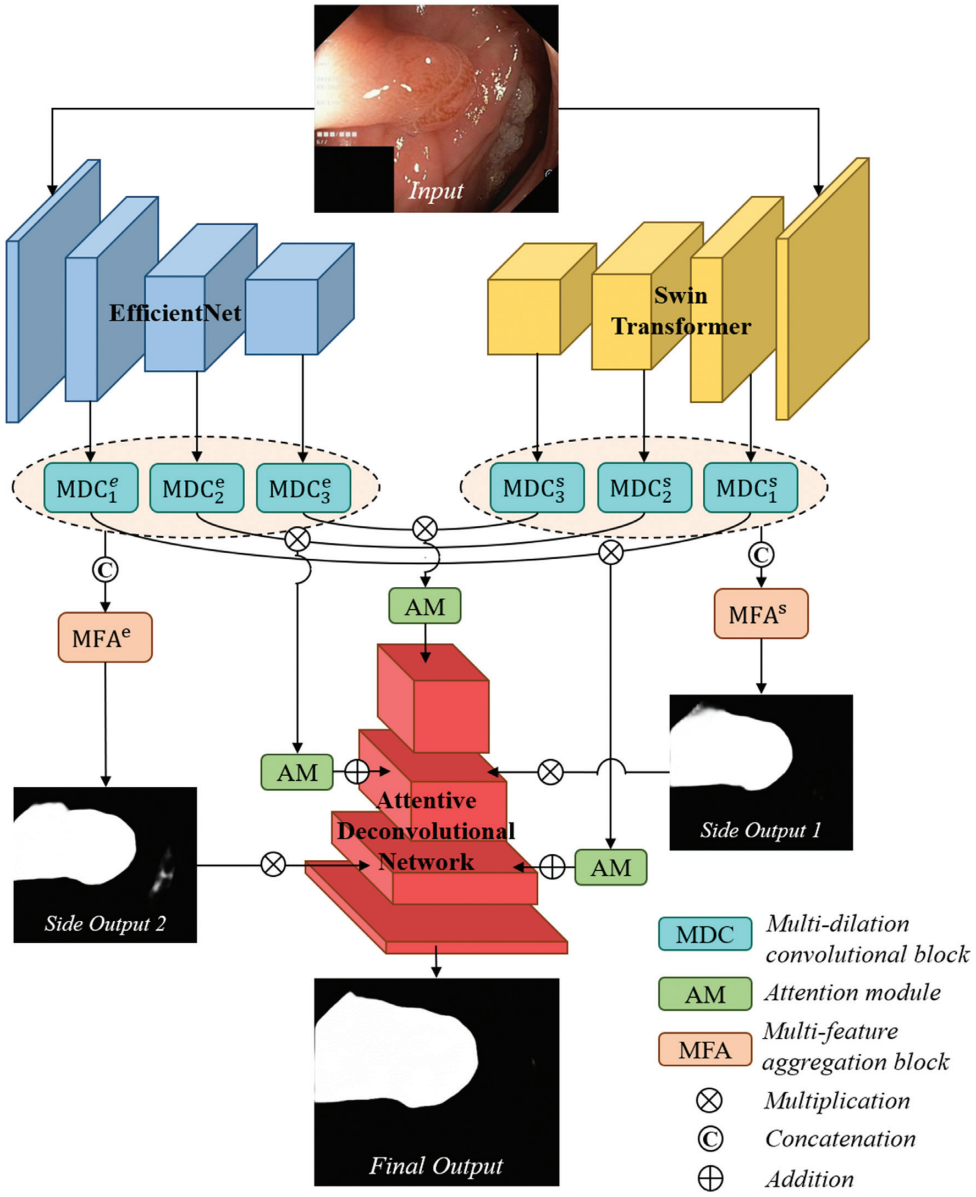

Figure 2: Architecture of the proposed SwinE-Net.

$$f^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{f}^{l+1}\right)\right) + \hat{f}^{l+1}, \tag{4}$$

where $\hat{f}^l$ and f^l represent the output features of the WMSA module and the MLP module of the l-th block, respectively. MLP represents MLP and LN represents layer normalization.

3.2. Multidilation convolutional and multifeature aggregation-based refinement and intermediate mask segmentation

The multidilation convolutional and multifeature aggregation blocks are applied to multilevel features extracted from EfficientNet and Swin ransformer to conduct feature enhancement and generate side outputs, respectively. The multidilation convolutional block refines multilevel features to keep the relationship between the size and eccentricity of receptive fields (Liu & Huang, 2018). The multidilation convolutional block aggregates the multilevel features to generate side outputs. The side output is used for generating intermediate polyp segmentation and for training segmentation masks.

As shown in Fig. 4a, the multidilation convolutional block trains convolution blocks with different sizes of receptive fields for multilevel features. It can effectively enhance the deep features of lightweight networks (Fan *et al.*, 2020a; Huang *et al.*, 2021). The segmentation accuracy can be improved by using different sizes of kernels and applying different sizes of dilated convolutions that can effectively store wider context information.

The multifeature aggregation block merges three refined features through the multidilation convolutional block by upsampling according to the size of the feature and performing element-wise multiplication and concatenated convolution (Fig. 5b). In particular, the multifeature aggregation block merges patch-level spatial features and global-level semantic features, which generates initial segmentation maps of EfficientNet and

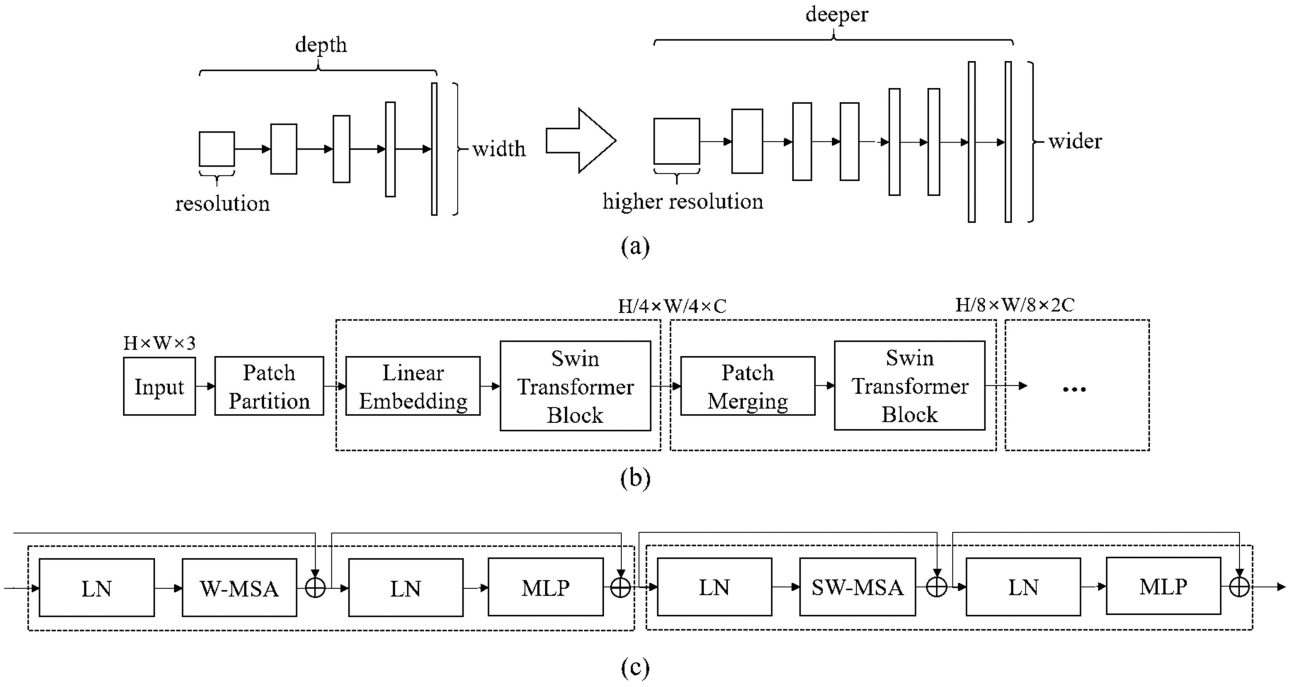

Figure 3: CNN and ViT models: (a) EfficientNet, (b) Swin Transformer, and (c) Swin Transformer block.

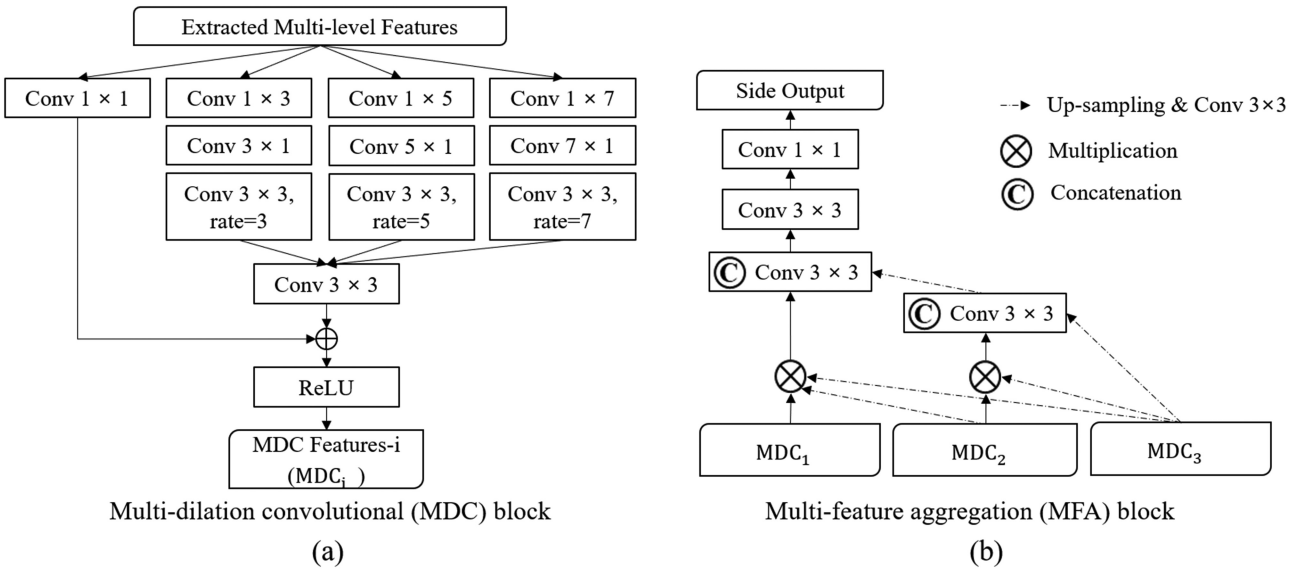

Figure 4: Multidilation convolutional block and multifeature aggregation block.

Swin Transformer. Therefore, it is possible to conduct polyp segmentation more accurately and robustly by considering spatial and global semantic features together.

To check whether feature maps extracted from Swin Transformer maintain more global information and those from the CNN-based network keep more low-level information, we visualized the feature maps of both models. Feature maps at the low level (MDC_1^f) and the high level (MDC_3^f) were extracted and visualized, as shown in Fig. 5. The visualization shows that the low-level features of CNN are trained mainly based on edges, and then the high-level features are pooled and refined based on the extracted low-level features, which increases the range of the receptive field and gradually finds the overall polyp area. However, the features are separated from each other according to the distance between pixels. In contrast, the low-level features and high-level features of Swin Transformer are extracted from the overall point of view, confirming that the features are related to each other, and the receptive range is wide. Thus, Swin Transformer generates less noise to find the global features effectively compared to the CNN features with more noise in the area away from the polyp area. Nevertheless, low-level features of Swin Transformer are less accurate than those of CNN.

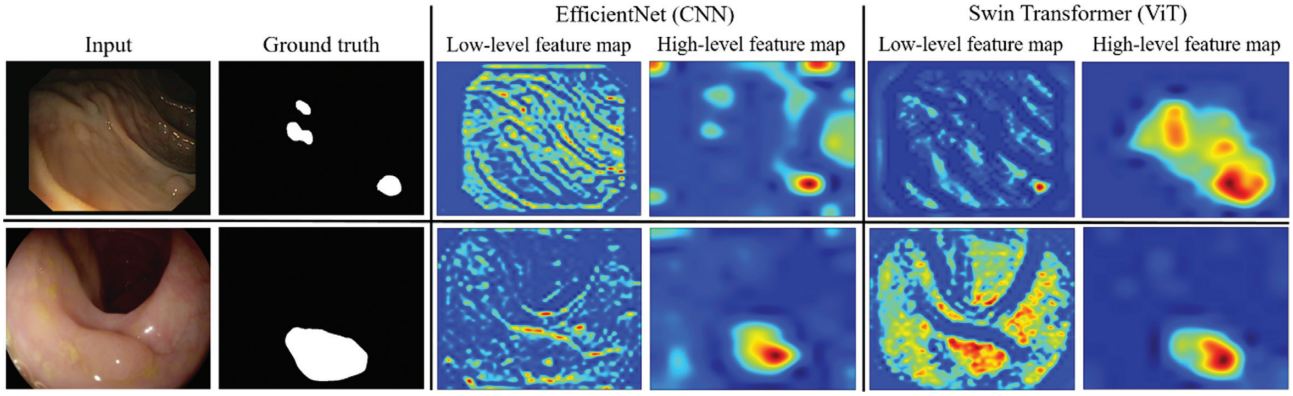

Figure 5: Examples of visualizing low-level and high-level feature maps of EfficientNet and Swin Transformer.

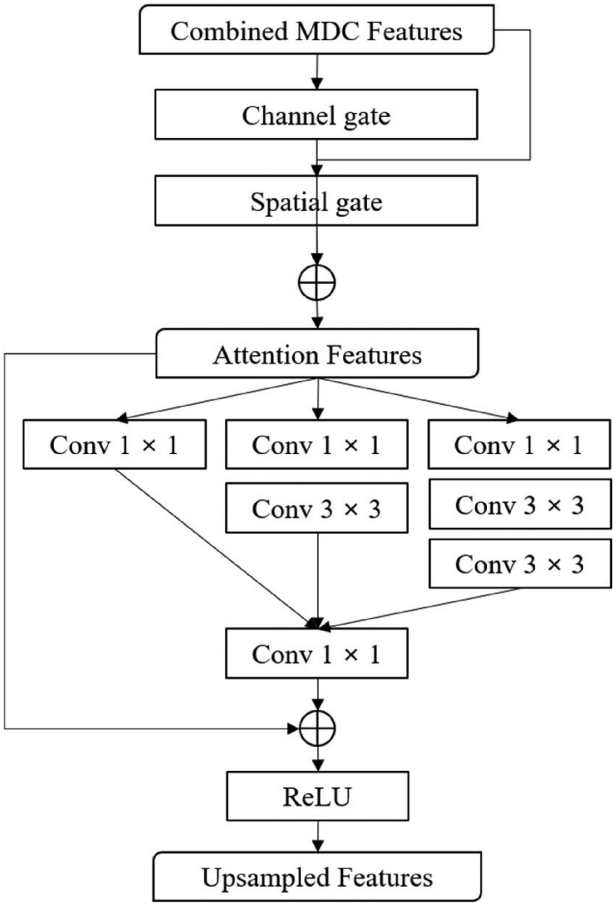

Figure 6: Decoder based on the attentive deconvolutional network.

Since SwinE-Net combines these features and upsamples them through attentive deconvolution, SwinE-Net can conduct more accurate and robust segmentation.

3.3. Attentive deconvolutional network-based decoder

The attentive deconvolution network-based decoder is applied to merge refined multilevel feature maps that independently pass through the multidilation convolutional block of EfficientNet and Swin Transformer and upsample them to generate the final polyp segmentation, as shown in Fig. 6. Using an attention module, the decoder conducts feature refinement of the merged feature maps. We use convolutional block attention module (CBAM; Woo *et al.*, 2018) as the attention module to improve segmentation performance. CBAM can sequentially infer attention maps along two separate dimensions, namely channel and spatial dimensions. The attention features are multiplied by the input features for adaptive feature refinement. Thus, the attention module can effectively extract the refined features and perform noise reduction of irrelevant clutters. Finally, the decoder generates the final polyp areas using Inception-ResNet blocks (Szegedy *et al.*, 2017) to effectively learn deeper networks and apply transposed convolution that performs upsampling.

The attention module diminishes noise (red circles) and makes polyp areas more highlighted (green circles), as shown in Fig. 7. CBAM is defined as follows:

$$f' = A_c\ (f) \otimes f, \quad f'' = A_s\ (f') \otimes f' \tag{5}$$

$$A_c\ (f) = \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}\left(f\right)\right) + \mathrm{MLP}\left(\mathrm{MaxPool}\left(f\right)\right)\right) \tag{6}$$

$$A_s\ (f') = \sigma\left(\mathrm{Conv}_{7\times7}\left(\mathrm{AvgPool}\left(f\right); \mathrm{MaxPool}\left(f\right)\right)\right), \tag{7}$$

where A_c is the channel attention function and A_s is the spatial attention function (Woo *et al.*, 2018).

Furthermore, as shown in Fig. 2, the final output through Inception-ResNet-based upsampling is more accurate and less noisy than the side outputs generated from Swin Transformer and CNN. Therefore, more sophisticated segmentation can be conducted through step-by-step adjustment and deconvolution using the Inception blocks, unlike previous studies that conduct simple upsampling or interpolation.

3.4. Segmentation loss function

The segmentation loss function of SwinE-Net consists of the weighted BCE loss and weighted IoU loss (Fan *et al.*, 2020a; Wei *et al.*, 2020). SwinE-Net trains both the final segmentation and side outputs to improve segmentation performance at spatial and semantic levels. The segmentation loss function is defined as follows.

$$L_{\mathrm{total}} = \alpha_f\ L_{\mathrm{seg}}\left(P_f, G\right) + \alpha_e L_{\mathrm{seg}}\left(P_e, G\right) + \alpha_s L_{\mathrm{seg}}\left(P_s, G\right) \tag{8}$$

$$L_{\mathrm{seg}} = L_{\mathrm{wbce}} + L_{\mathrm{wiou}} \tag{9}$$

L_{total} is composed of the segmentation loss of the final output P_f generated from the Inception decoder, that of the side output P_e generated from EfficientNet, and that of the side output P_s generated from Swin Transformer. α_f, α_e, and α_s represent the

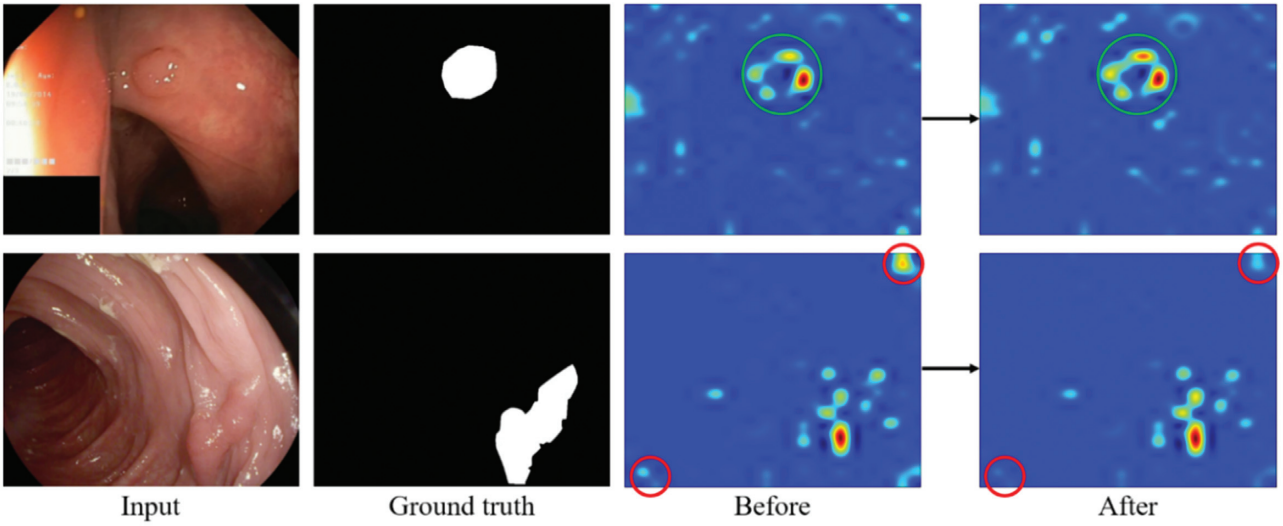

Figure 7: Examples of feature visualization before and after applying the attention module. The green circle of the feature map shows feature refinement, and the red circles show noise reduction.

weights of the final output and side outputs, which are set to 0.5, 0.3, 0.2, respectively. L_{seg} combines the weighted BCE loss and the weighted IoU loss. The weighted BCE and IoU loss are defined as follows.

$$L_{wbce} = \frac{\sum_{(x,y)} [G(x,y)\log(P(x,y)) + (1 - G(x,y))\log(1 - P(x,y))] \times \omega(x,y)}{\sum_{(x,y)} \omega(x,y)} \tag{10}$$

$$L_{wiou} = 1 - \frac{\sum_{(x,y)} P(x,y)\, G(x,y) \times \omega(x,y)}{\sum_{(x,y)} [P(x,y) + G(x,y) - P(x,y)\, G(x,y)] \times \omega(x,y)} \tag{11}$$

$$\omega(x,y) = 1 + \gamma \left| \text{avg\_pool}(G(x,y)) - G(x,y) \right| \tag{12}$$

The BCE loss is mainly used for binary semantic segmentation to reduce the pixel-wise difference between the predicted polyp area and the ground truth. It performs weighted learning by applying $\omega(x, y)$, which concedes different loss weights to each image. The IoU loss is proposed for semantic segmentation to reduce image-wise difference and improve the regional consistency and boundary response. Like the BCE loss, the IoU loss performs the weighted learning by $\omega(x, y)$. $\omega(x, y)$ is calculated by average pooling using the ground truth and finding the difference. It can extract more attention using weighted loss to hard pixels such as edge or hole. In this study, γ was set to 5, and the average pooling parameters were set to the kernel size 31, stride 1, and padding 15. Therefore, effective learning can be performed by calculating loss weights for each input through the weighted value $\omega(x, y)$.

4. Experiments

We conducted comparative evaluation to confirm the effectiveness and robustness of SwinE-Net for polyp segmentation using challenging public datasets.

4.1. Datasets

Five public datasets of colorectal polyps were used for the comparative evaluation: Kvasir (Jha *et al.*, 2020), ClinicDB (Bernal *et al.*, 2015), ColonDB (Tajbakhsh *et al.*, 2015), ETIS (Silva *et al.*, 2014), and EndoScene (Vázquez *et al.*, 2017). The same experimental setting for training and testing was used as described in related studies (Fan *et al.*, 2020a; Huang *et al.*, 2021). The training dataset contains 900 images of Kvasir and 550 images of ClinicDB, a total of 1450 images. Also, the test dataset contains 100 images of Kvasir, 62 images of ClinicDB, 380 images of ColonDB, 196 images of ETIS, and 60 images of EndoScene. The dataset includes the pixel-wise segmented ground truth annotation for each image. In particular, we trained only two seen datasets (Kvasir and ClinicDB), and the other three unseen datasets (ColonDB, ETIS, and EndoScene) were used for evaluation only, as suggested by Fan *et al.* (2020a).

For training and testing, each image was resized into 384 × 384. The main reason for resizing the image with 384 × 384 is to utilize the pretrained weight of Swin Transformer since Swin Transformer can only support the image with 224 × 224 or 383 × 384 using pretrained weights due to its generic network characteristics. All training datasets were augmented using rotation 90° intervals, flipping, scaling, and shearing. Instead of generating augmentation data in advance, we have applied random augmentation when the data are loaded for training and testing.

4.2. Implementation details

The proposed SwinE-Net was implemented using Pytorch (2016). We used Nvidia RTX 3090 GPU with 24GB for GPU-accelerated training. For training and testing, the input size was resized into 384 × 384 in all five public datasets. We used bilinear interpolation to convert the resized image to the original endoscopic image for the comparative testing. The training follows an end-to-end process. We set the batch size to 12 and the entire learning epoch to 200. In addition, we used the Adam optimizer and set the learning rate to 1e-4, beta1 to 0.9, beta2 to 0.999, and epsilon to 1e-08.

Table 1: Comparative evaluation of SwinE-Net with SOTA methods using the seen datasets (Kvasir and ClinicDB).

Metric/SOTAs	mDice	mIoU	FM	SM	EM	MAE↓
			Kvasir dataset			
ResUNet (Zhang *et al.*, 2018)	0.791	N/A	N/A	N/A	N/A	N/A
ResUNet++ (Jha *et al.*, 2019)	0.813	0.793	N/A	N/A	N/A	N/A
U-Net (Ronneberger *et al.*, 2015)	0.818	0.746	0.794	0.858	0.893	0.055
U-Net++ (Zhou *et al.*, 2018)	0.821	0.743	0.808	0.862	0.910	0.048
PraNet (Fan *et al.*, 2020a)	0.898	0.840	0.885	0.915	0.948	0.030
HarDNet-MSEG (Huang *et al.*, 2021)	0.912	0.857	0.903	0.923	0.958	0.025
TransUNet (Chen *et al.*, 2021)	0.913	0.857	N/A	N/A	N/A	N/A
TransFuse (Zhang *et al.*, 2021)	0.918	0.868	N/A	N/A	N/A	N/A
Swin-Up (Brandao *et al.*, 2017; Liu *et al.*, 2021)	0.883	0.823	0.876	0.917	0.956	0.033
SwinE-Net	**0.920**	**0.870**	**0.913**	**0.926**	**0.963**	**0.024**
			ClinicDB dataset			
ResUNet (Zhang *et al.*, 2018)	0.779	N/A	N/A	N/A	N/A	N/A
ResUNet++ (Jha *et al.*, 2019)	0.796	0.796	N/A	N/A	N/A	N/A
U-Net (Ronneberger *et al.*, 2015)	0.823	0.755	0.811	0.889	0.954	0.019
U-Net++ (Zhou *et al.*, 2018)	0.794	0.729	0.785	0.873	0.931	0.022
PraNet (Fan *et al.*, 2020a)	0.899	0.849	0.896	0.936	0.979	0.009
HardNet-MSEG (Huang *et al.*, 2021)	0.932	0.882	N/A	N/A	N/A	N/A
TransUNet (Chen *et al.*, 2021)	0.935	0.887	N/A	N/A	N/A	N/A
TransFuse (Zhang *et al.*, 2021)	0.918	0.868	N/A	N/A	N/A	N/A
Swin-Up (Brandao *et al.*, 2017; Liu *et al.*, 2021)	0.884	0.820	0.876	0.938	0.971	0.012
SwinE-Net	**0.938**	**0.892**	**0.936**	**0.950**	**0.989**	**0.006**

mDice: mean dice coefficient, mIoU: mean intersection over union, FM: mean weighted F-measure, SM: structure measure, EM: max enhanced-alignment measure, MAE: mean absolute error, ↓: the smaller value implies better performance.

4.3. Evaluation metrics

Various metrics were evaluated to compare the segmentation performance of SwinE-Net with previous studies, including mean IoU (mIoU), mean dice coefficient (mDice), weighted F-measure (FM), structure measure (SM), max enhanced-alignment measure (EM), and mean absolute error (MAE).

The mDice, FM (Margolin *et al.*, 2014), and mIoU metrics are calculated based on the 256 pairs of precision and recall between the binary mask of the predicted map and the ground truth, where the predicted map is converted into a total of 256 binary masks concerning thresholds changing from 0 to 255. Precision, recall, Dice, FM, and IoU are as defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

$$\text{Dice} = \frac{2\times\text{TP}}{2\times\text{TP}+\text{FP}+\text{FN}} \tag{14}$$

$$\text{weighted F-measure} = \frac{(1+\beta^2)\times\text{Precision}\times\text{Recall}}{\beta^2\times\text{Precision}+\text{Recall}} \tag{15}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}}. \tag{16}$$

TN, TP, and FN are true negative, true positive, and false negative, respectively. β^2 is set to 0.3.

SM (Fan *et al.*, 2017) is an evaluation metric that simultaneously calculates the object- and region-aware structural similarity between the predicted polyp mask and ground truth. EM (Fan *et al.*, 2018) is an evaluation metric for binary foreground map segmentation, consisting of a single term to account for pixel and image-level properties. MAE is an evaluation metric that calculates the average absolute per-pixel difference between the predicted polyp segmentation and the ground truth. MAE is described as follows.

$$\text{MAE} = \frac{1}{N}\sum_{(x,y)}\left|P(x,y)-G(x,y)\right| \tag{17}$$

N indicates the number of pixels in the image, and $P(x,\ y)$ and $G(x,\ y)$ represent the pixel value of the predicted polyp segmentation and that of the ground truth at the location $(x,\ y)$, respectively.

4.4. Comparative evaluation

Comparative evaluations with previous SOTA methods were conducted based on public polyp datasets to verify the effectiveness and robustness of the proposed SwinE-Net. First, we conducted an experiment using seen datasets consisting of Kvasir and ClinicDB. Experimental results demonstrate that the proposed approach achieves the best performance with the segmentation accuracy of 92.0% and 93.8% (mDice↑) and 87.0% and 89.2% (mIoU↓) for Kvasir and ClinicDB, respectively, as shown in Table 1. For example, the mDice metric was higher, about 10%, than traditional U-Net-based CNN approaches, such as U-Net (Ronneberger *et al.*, 2015), U-Net++ (Zhou *et al.*, 2018), ResUNet (Zhang *et al.*, 2018), and ResUNet++ (Jha *et al.*, 2019). Also, it was higher, about 1–3%, than advanced U-Net-based approaches, such as PraNet (Fan *et al.*, 2020a) and HarDNet-MSEG (Huang *et al.*, 2021).

SwinE-Net also outperformed TransUNet (Chen *et al.*, 2021) and Transfuse (Zhang *et al.*, 2021) that used ViT and CNN-based methods. It is important to note that these methods only used the last-level features rather than multilevel features of SwinE-Net that considers both low-level and high-level contexts. The Swin-Up method (Brandao *et al.*, 2017; Liu *et al.*, 2021) extracts features from only Swin Transformer and con-

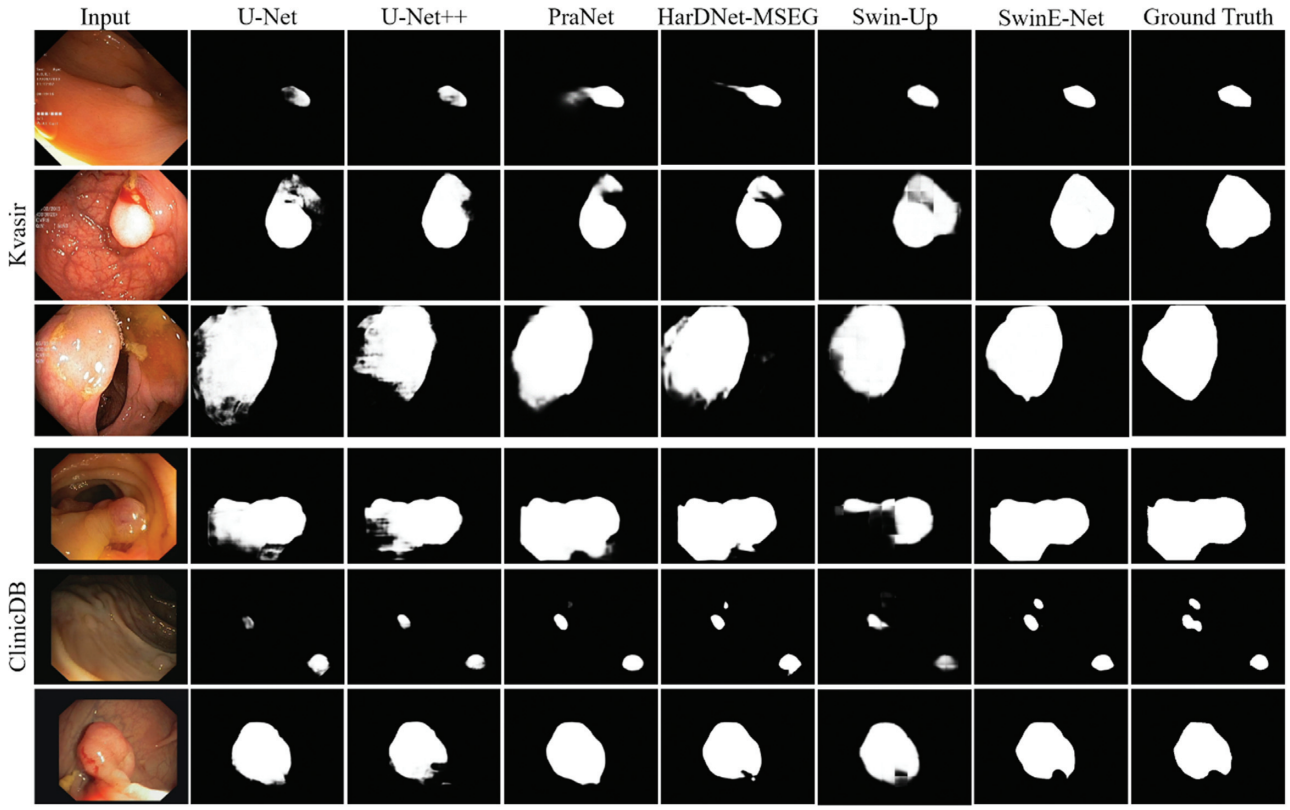

Figure 8: Qualitative evaluation of polyp segmentation results in the seen datasets.

ducts polyp segmentation through image upsampling. However, it yielded worse results. On the other hand, SwinE-Net confirms its performance and robustness by applying the multilevel features of Swin Transformer and CNN, MDC and multifeature aggregation blocks, and attentive deconvolution encoder. As a result, the proposed approach showed the best performance on various metrics, such as mDice, mIoU, FM, SM, EM, and MAE.

The qualitative analysis of the polyp segmentation using the seen datasets is shown in Fig. 8. The proposed approach can reduce false negatives and false positives by distinguishing the ambiguous boundaries between the background mucosa and polyp. For example, previous approaches generated many false-positive errors, such as blur and noise, for the test data in the first and third rows of the Kvasir dataset (Fig. 8). However, the proposed SwinE-Net verified accurate segmentation without such blur and noise. Also, for the test data in the second row of the Kvasir dataset and the test data in the first and second rows of the ClinicDB dataset, previous approaches could not generate polyp areas properly because the boundary between the background mucosa and polyp is very ambiguous. On the other hand, the proposed SwinE-Net could generate polyp areas more accurately and robustly despite ambiguous boundaries.

Second, we conducted another comparative evaluation using the unseen datasets, such as ColonDB, ETIS, and EndoScene. The proposed approach achieved the best performance in all the metrics for the unseen dataset compared with previous SOTA methods using the same weights trained with the seen datasets, as shown in Table 2. Notably, the proposed approach derived much better results with a large margin in the unseen datasets than with advanced U-Net approaches and traditional U-Net-based approaches. For example, the proposed approach showed 0.7% higher performance for the mDice metric than HarDNet-MSEG on the seen datasets, whereas it showed 5.8% higher performance for the mDice metric on the unseen datasets.

SwinE-Net also outperformed previous studies using ViT for the unseen dataset, as shown in Table 2. Through the transfer learning of Swin Transformer and EfficientNet whose weights were pretrained based on the image classification dataset, SwinE-Net can be easily applied to the unseen dataset in real situations, which verifies the generalization and robustness for polyp segmentation. Furthermore, SwinE-Net can be used in real medical environments since it runs at 18.5 frames per second.

The qualitative analysis of the polyp segmentation using the unseen datasets is shown in Fig. 9. The analysis confirms that the proposed approach can reduce false-negative and false-positive errors simultaneously compared with the previous approaches. While previous approaches generated a lot of blurs, the proposed approach segmented polyp areas as accurately as the ground truth without blur, as shown in the first and second rows of the ColonDB test data in Fig. 9. In addition, for the first rows of the ColonDB and ETIS test data, previous approaches could not segment polyp areas properly because the textures of mucosa and polyps are very similar. Still, the proposed approach segmented accurate polyp areas by detecting boundaries between mucosa and polyps.

Besides, k-fold cross-validation ($k = 5$) was conducted for statistical analysis of SwinE-Net. The total training data set with 1450 images is split into five sets (each split set has 290 images). Each set is used for validation, and the remaining four sets with 1160 images are used for training. Table 3 shows the results of cross-validation. Similar to the results of SwinE-Net

Table 2: Comparative evaluation of SwinE-Net with SOTA methods using the unseen datasets (ColonDB, ETIS, and EndoScene).

Metric/SOTAs	mDice	mIoU	FM	SM	EM	MAE↓
			ColonDB dataset			
U-Net (Ronneberger *et al.*, 2015)	0.512	0.444	0.498	0.712	0.776	0.061
U-Net++ (Zhou *et al.*, 2018)	0.483	0.410	0.467	0.691	0.760	0.064
PraNet (Fan *et al.*, 2020a)	0.709	0.640	0.696	0.819	0.869	0.045
HarDNet-MSEG (Huang *et al.*, 2021)	0.731	0.660	N/A	N/A	N/A	N/A
TransUNet (Chen *et al.*, 2021)	0.781	0.699	N/A	N/A	N/A	N/A
TransFuse (Zhang *et al.*, 2021)	0.773	0.696	N/A	N/A	N/A	N/A
Swin-Up (Brandao *et al.*, 2017; Liu *et al.*, 2021)	0.735	0.651	0.723	0.850	0.894	0.034
SwinE-Net	**0.804**	**0.725**	**0.787**	**0.869**	**0.910**	**0.028**
			ETIS dataset			
U-Net (Ronneberger *et al.*, 2015)	0.398	0.335	0.366	0.684	0.740	0.036
U-Net++ (Zhou *et al.*, 2018)	0.401	0.344	0.390	0.683	0.776	0.035
PraNet (Fan *et al.*, 2020a)	0.628	0.567	0.600	0.794	0.841	0.031
HarDNet-MSEG (Huang *et al.*, 2021)	0.677	0.613	N/A	N/A	N/A	N/A
TransUNet (Chen *et al.*, 2021)	0.731	0.660	N/A	N/A	N/A	N/A
TransFuse (Zhang *et al.*, 2021)	0.733	0.659	N/A	N/A	N/A	N/A
Swin-Up (Brandao *et al.*, 2017; Liu *et al.*, 2021)	0.661	0.581	0.619	0.831	0.863	0.019
SwinE-Net	**0.758**	**0.687**	**0.726**	**0.864**	**0.902**	**0.012**
			EndoScene dataset			
U-Net (Ronneberger *et al.*, 2015)	0.710	0.627	0.684	0.843	0.876	0.022
U-Net++ (Zhou *et al.*, 2018)	0.707	0.624	0.687	0.839	0.898	0.018
PraNet (Fan *et al.*, 2020a)	0.871	0.797	0.843	0.925	0.972	0.010
HarDNet-MSEG (Huang *et al.*, 2021)	0.887	0.821	N/A	N/A	N/A	N/A
TransUNet (Chen *et al.*, 2021)	0.893	0.824	N/A	N/A	N/A	N/A
Transfuse (Zhang *et al.*, 2021)	0.902	0.833	N/A	N/A	N/A	N/A
Swin-Up (Brandao *et al.*, 2017; Liu *et al.*, 2021)	0.874	0.799	0.847	0.942	0.976	0.008
SwinE-Net	**0.906**	**0.842**	**0.888**	**0.942**	**0.983**	**0.005**

in Tables 1 and 2 trained with a total of 1450 training images, excellent results were derived. Besides, standard deviations were very small, which confirms the robustness and stability of SwinE-Net.

We have also conducted fivefold cross-validation (80% training and 20% validation) for the previous studies, as shown in Table 4. Furthermore, the *t*-test was conducted to verify the statistical difference between the proposed approach and other methods. Table 4 shows that there is a significant statistical difference ($p < 0.001$), which proves that SwinE-Net outperforms previous studies. In conclusion, the comparative evaluation and statistical validation have verified the superiority of SwinE-Net to previous studies.

4.5. Ablation study

An ablation study was performed to prove the effectiveness of the proposed approach's architectural modules and loss functions. Table 5 shows the results of the ablation study using Kvasir and EndoScene datasets. The proposed segmentation loss function that combines the BCE loss and IoU loss achieved better performance than using only the pixel-wise BCE loss or the image-wise IoU loss. The original SwinE-Net is compared with methods using only EfficientNet, only Swin Transformer, without EfficientNet, without Swin Transformer, ResNet and DenseNet instead of EfficientNet, without the attentive deconvolutional network-based decoder, and without side output training (Fig. 10). The results verify that SwinE-Net with all the proposed networks and blocks shows optimal performance on segmentation.

We extracted feature maps from ResNet and DenseNet in addition to EfficientNet (see Fig. 5). Figure 11 visualizes low-level and high-level feature maps of ResNet and DenseNet. We have found that ResNet (He *et al.*, 2016) and DenseNet (Huang *et al.*, 2017) can also extract low-level and high-level features similar to EfficientNet. However, both focused more on low-level features than Swin Transformer. Although CNNs can extract high-level feature maps, we have found that Swin Transformer can extract high-level features more effectively and consistently. However, low-level features of Swin Transformer are less accurate than those of CNNs. It is important to note that EfficientNet with Swin Transformer showed better performance than ResNet and DenseNet with Swin Transformer, as shown in Table 5. Since SwinE-Net can take full advantage of the Swin Transformer and CNN, it can conduct more accurate and robust segmentation.

4.6. Additional medical segmentation experiment with COVID-19 dataset

It is essential to guarantee the generality of the deep learning model that can be applied to different datasets. We used the proposed hybrid approach to perform a lung infection segmentation experiment using the COVID-19 dataset (2020) to confirm the extensibility of the proposed approach. The training set contains

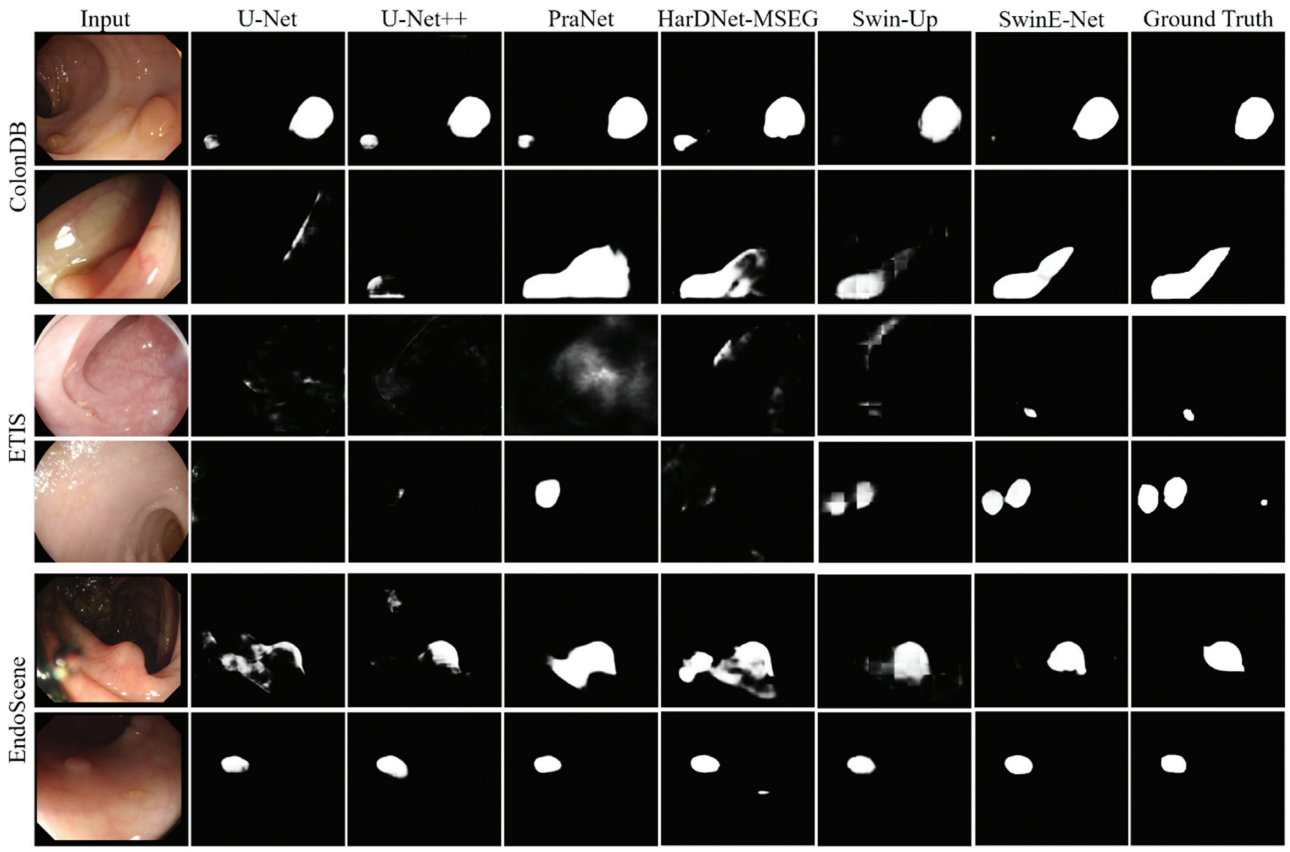

Figure 9: Qualitative evaluation of polyp segmentation results in the unseen datasets.

Table 3: Result of fivefold cross validation (80% training and 20% validation) using the seen and unseen datasets.

Metric		mDice	mIoU	FM	SM	EM	MAE↓
Seen dataset	Kvasir	0.914 ± 0.003	0.864 ± 0.003	**0.910 ± 0.002**	**0.923 ± 0.001**	**0.963 ± 0.002**	**0.025 ± 0.001**
	ClinicDB	0.926 ± 0.006	0.879 ± 0.006	**0.923 ± 0.006**	**0.945 ± 0.004**	**0.981 ± 0.006**	**0.007 ± 0.002**
Unseen dataset	ColonDB	**0.797 ± 0.004**	**0.716 ± 0.005**	0.777 ± 0.005	0.866 ± 0.004	0.912 ± 0.002	0.030 ± 0.001
	ETIS	**0.762 ± 0.011**	**0.683 ± 0.011**	0.718 ± 0.013	0.864 ± 0.005	0.908 ± 0.007	0.014 ± 0.001
	EndoScene	0.897 ± 0.004	0.831 ± 0.005	**0.876 ± 0.005**	**0.938 ± 0.004**	**0.982 ± 0.003**	**0.006 ± 0.001**

50 CT images, and the test set includes 48 CT images, as suggested by Fan *et al.* (2020b). It also includes the pixel-wise segmented ground truth annotation of each image. Each image is resized into 384 × 384 for training and testing since there are various sizes of CT images in the COVID-19 segmentation dataset.

Table 6 shows the comparative evaluation of the COVID-19 lung infection segmentation. The proposed SwinE-Net showed the best performance on various metrics compared to the previous studies, such as U-Net++ and Inf-Net. SwinE-Net derived excellent results on the CT image corresponding to the grayscale and the endoscopy image corresponding to the RGB. The result verifies the generality and robustness of the proposed approach for versatile medical image segmentation.

The qualitative analysis of the COVID-19 CT segmentation dataset is shown in Fig. 12. The proposed SwinE-Net generated high-quality segmentation results by reducing noise and blur by increasing true-positive rates and decreasing false-negative rates than previous approaches. Many blurs were still generated in most examples for the advanced U-Net-based approaches because they generated the final output that passed the last layer using a simple interpolation-based resizing method. On the other hand, the proposed SwinE-Net generated more accurate and robust polyp areas than previous approaches by performing cascaded weight refinement using the Inception decoder that includes transposed convolution and inception blocks.

5. Conclusion

CNN-based approaches have some limitations for modeling wide relations because of the intrinsic locality of convolutions. Transformer-based approaches are robust at modeling global semantics, but they have problems describing detailed information. This study proposed a novel colorectal polyp segmentation approach using the hybrid deep neural network, which utilizes low-level spatial features extracted from CNN and global-level semantic features extracted from ViT. Besides, multilevel features extracted from EfficientNet and Swin Transformer are re-

Table 4: Statistical evaluation using the *t*-test between SwinE-Net and other studies for the seen and unseen datasets ($^{*}p < 0.05$, $^{**}p < 0.01$).

Metric/SOTAs	mDice	mIoU	FM	SM	EM	MAE↓
			Seen dataset (Kvasir)			
U-Net (Ronneberger *et al.*, 2015)	0.739 ± 0.011**	0.630 ± 0.010**	0.706 ± 0.004**	0.801 ± 0.006**	0.861 ± 0.004**	0.071 ± 0.001**
PraNet (Fan *et al.*, 2020a)	0.769 ± 0.024**	0.679 ± 0.025**	0.727 ± 0.032**	0.828 ± 0.020**	0.875 ± 0.018**	0.069 ± 0.015**
HardNet-MSEG (Huang *et al.*, 2021)	0.855 ± 0.007**	0.785 ± 0.007**	0.835 ± 0.009**	0.883 ± 0.004**	0.920 ± 0.006**	0.043 ± 0.003**
TransFuse (Zhang *et al.*, 2021)	0.881 ± 0.007**	0.820 ± 0.009**	0.864 ± 0.012**	0.909 ± 0.005**	0.948 ± 0.004**	0.033 ± 0.002**
SwinE-Net	**0.914 ± 0.003**	**0.864 ± 0.003**	**0.910 ± 0.002**	**0.923 ± 0.001**	**0.963 ± 0.002**	**0.025 ± 0.001**
			Unseen dataset (EndoScene)			
U-Net (Ronneberger *et al.*, 2015)	0.590 ± 0.041**	0.487 ± 0.041**	0.546 ± 0.039**	0.762 ± 0.024**	0.807 ± 0.018**	0.024 ± 0.004**
PraNet (Fan *et al.*, 2020a)	0.583 ± 0.076**	0.479 ± 0.063**	0.504 ± 0.058**	0.766 ± 0.038**	0.839 ± 0.020**	0.036 ± 0.008**
HardNet-MSEG (Huang *et al.*, 2021)	0.805 ± 0.025**	0.715 ± 0.023**	0.759 ± 0.025**	0.878 ± 0.014**	0.933 ± 0.027*	0.014 ± 0.002**
TransFuse (Zhang *et al.*, 2021)	0.856 ± 0.010**	0.780 ± 0.013**	0.823 ± 0.013**	0.921 ± 0.006**	0.963 ± 0.004**	0.012 ± 0.002*
SwinE-Net	**0.897 ± 0.004**	**0.831 ± 0.005**	**0.876 ± 0.005**	**0.938 ± 0.004**	**0.982 ± 0.003**	**0.006 ± 0.001**

Table 5: Comparison of different loss and network module configurations of the proposed approach using the Kvasir dataset (seen dataset) and the EndoScene dataset (unseen dataset).

Dataset	Category	Configuration	mDice	mIoU	FM
Kvasir (seen)	Loss	Only BCE	0.893	0.835	0.886
		Only IoU	0.909	0.855	0.897
		BCE + IoU	**0.920**	**0.870**	**0.913**
	Modules	Only EfficientNet	0.906	0.853	0.900
		Only Swin Transformer	0.899	0.843	0.890
		Without EfficientNet	0.897	0.842	0.891
		Without Swin Transformer	0.904	0.848	0.893
		ResNet + Swin Transformer	0.906	0.854	0.901
		DenseNet + Swin Transformer	0.907	0.855	0.903
		Without attentive deconvolutional decoder	0.908	0.855	0.902
		Without side output training	0.904	0.852	0.901
		SwinE-Net	**0.920**	**0.870**	**0.913**
EndoScene (unseen)	Loss	Only BCE	0.872	0.796	0.838
		Only IoU	0.891	0.822	0.867
		BCE + IoU	**0.906**	**0.842**	**0.888**
	Modules	Only EfficientNet	0.876	0.805	0.849
		Only Swin Transformer	0.848	0.768	0.810
		Without EfficientNet	0.875	0.799	0.845
		Without Swin Transformer	0.893	0.823	0.867
		ResNet + Swin Transformer	0.894	0.826	0.870
		DenseNet + Swin Transformer	0.898	0.831	0.876
		Without attentive deconvolutional decoder	0.899	0.832	0.876
		Without side output training	0.889	0.821	0.866
		SwinE-Net	**0.906**	**0.842**	**0.888**

fined and enhanced through the multidilation convolutional and multifeature aggregation blocks. The attentive deconvolutional network-based decoder finalizes the polyp segmentation by fusing and upsampling refined features trained from the two networks. The side outputs from EfficientNet and Swin Transformer are also trained to optimize the loss function. The segmentation loss function combines IoU and BCE losses to reduce pixel-wise and image-wise differences and improve regional consistency and boundary response.

A comprehensive evaluation was conducted using five public datasets on colorectal poly segmentation to verify the novelty and advantages of the proposed SwinE-Net. The proposed

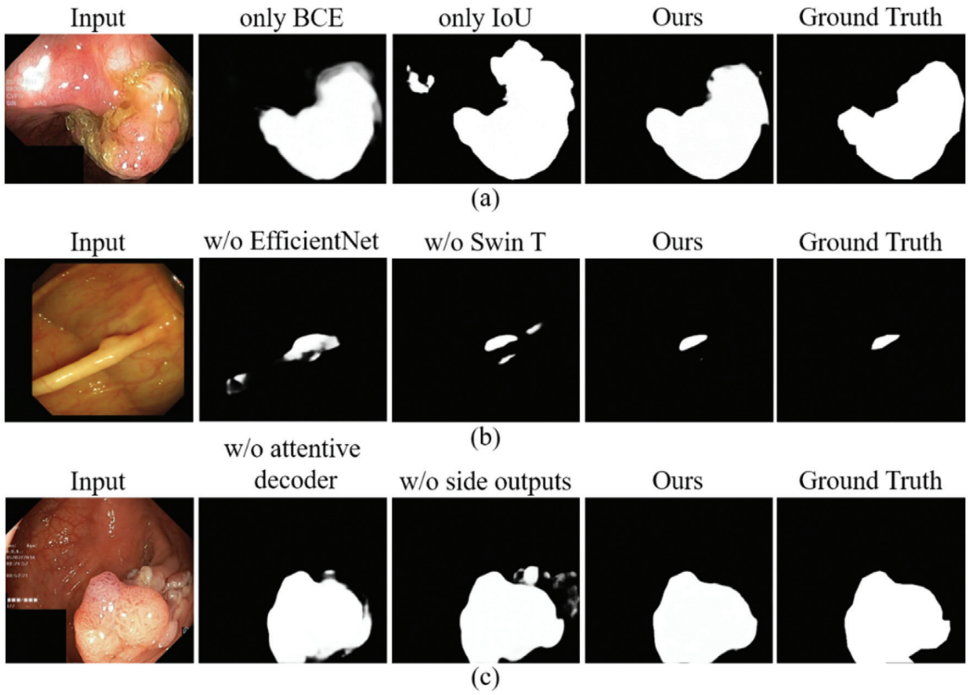

Figure 10: Qualitative results of ablation studies on polyp segmentation in Kvasir (a, c) and EndoScene (b) datasets: (a) comparative results based on the loss function, (b) comparative results based on the encoder architecture, and (d) comparative results based on the decoder architecture.

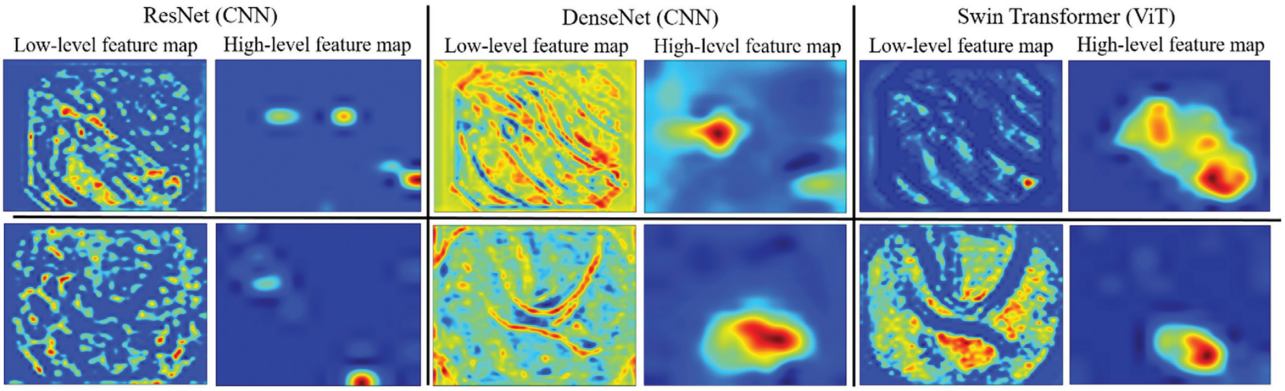

Figure 11: Examples of visualizing low-level and high-level feature maps of ResNet and DenseNet compared with Swin Transformer.

Table 6: Comparative evaluation of SwinE-Net using the COVID-19 CT segmentation dataset.

Metric/SOTAs	mDice	mIoU	FM	SM	EM	MAE↓
U-Net (Ronneberger *et al.*, 2015)	0.439	0.309	0.352	0.622	0.802	0.186
U-Net++ (Zhou *et al.*, 2018)	0.581	0.444	0.499	0.722	0.893	0.120
Inf-Net (Fan *et al.*, 2020b)	0.739	0.601	0.724	0.800	0.934	0.064
SwinE-Net	**0.759**	**0.628**	**0.757**	**0.801**	**0.938**	**0.056**

SwinE-Net achieved the best performance on all evaluation metrics than the other SOTA methods according to the comparative experiments. In particular, it showed much better performance for the unseen datasets that were not used for the training dataset. Furthermore, we conducted an additional comparative evaluation for COVID-19 CT segmentation. Even though the dataset consists of gray CT images, the proposed SwinE-Net derived the best result in all metrics with a large margin compared with previous approaches. Comprehensive experiment results verify that the proposed approach is more robust than the other

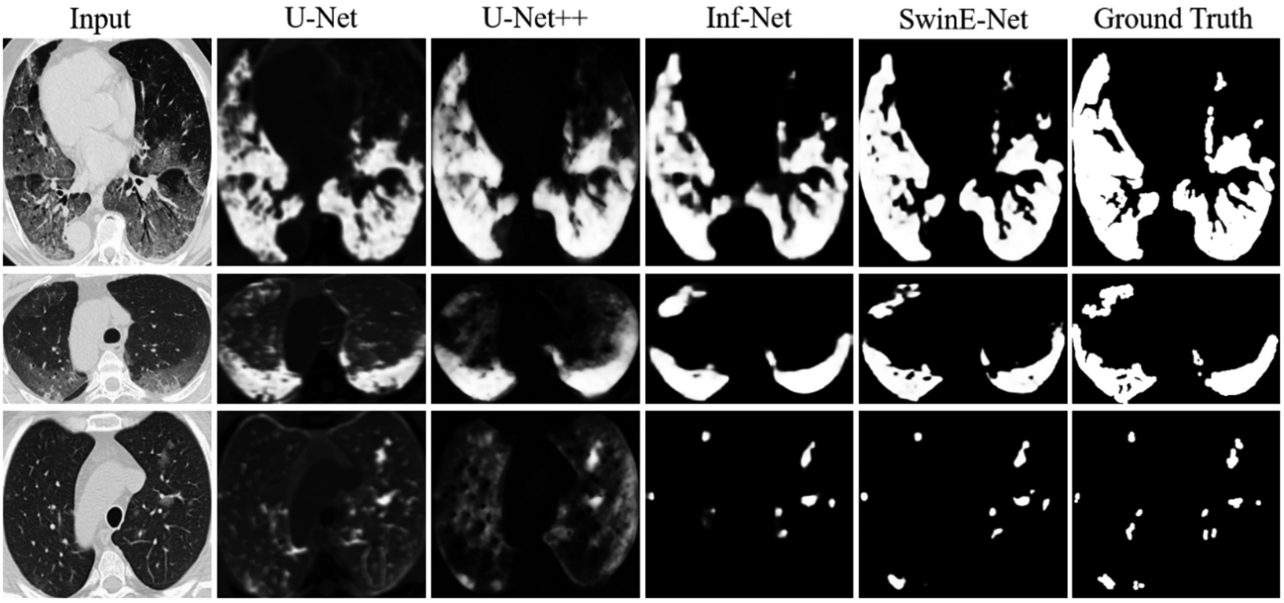

Figure 12: Infected region segmentation in the COVID-19 CT segmentation dataset.

in polyp segmentation tasks, proving generality and extensibility for various and different endoscopic data.

For future works, we will apply the proposed approach to different medical image segmentation applications by refining the network architecture and loss function for optimizing the relationship between CNN and transformer. We will apply the extended SwinE-Net to conduct 3D volume segmentation of other datasets, such as liver and Synapse multi-organ CT dataset.

Acknowledgment

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2019R1I1A3A01059082), and the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare (HI19C0642).

Conflict of interest statement

None declared.

References

Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., & Vilariño, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics, 43*, 99–111.

Brandao, P., Mazomenos, E., Ciuti, G., Caliò, R., Bianchi, F., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., & Stoyanov, D. (2017). Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Proceedings of the SPIE Medical Imaging*(Vol. *10134*).

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like pure transformer for medical image segmentation. *preprint arXiv:2105.05537*.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation.

COVID-19 CT Segmentation Dataset. (2020). *https://medicalsegmentation.com/covid19/*, accessed on Aug. 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *preprint arXiv:2010.11929*.

Fan, D. P., Cheng, M. M., Liu, Y., Li, T., & Borji, A. (2017). Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*(pp. 4548–4557).

Fan, D. P., Gong, C., Cao, Y., Ren, B., Cheng, M. M., & Borji, A. (2018). Enhanced-alignment measure for binary foreground map evaluation. *preprint arXiv:1805.10421*.

Fan, D. P., Ji, G. P., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. (2020a). Pranet: Parallel reverse attention network for polyp segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*(pp. 263–273).

Fan, D. P., Zhou, T., Ji, G. P., Zhou, Y., Chen, G., Fu, H., Shen, J., & Shao, L. (2020b). Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE Transactions on Medical Imaging, 39*(8), 2626–2637.

Fan, T., Wang, G., Li, Y., & Wang, H. (2020c). MA-Net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access, 8*, 179656–179665.

Ferlay, J. (2010). *GLOBOCAN 2008 v1. 2, Cancer incidence and mortality world-wide: IARC cancer base no. 10*. *https://gco.iarc.fr/*.

Guo, Y., Bernal, J., & Matuszewski, J. B. (2020). Polyp segmentation with fully convolutional deep neural networks—extended evaluation study. *Journal of Imaging, 6*(7), 69.

Hassan, C., Wallace, M. B., Sharma, P., Maselli, R., Craviotto, V., Spadaccini, M., & Repici, A. (2020). New artificial intelligence system: First validation study versus experienced endoscopists for colorectal polyp detection. *Gut, 69*(5), 799–800.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*(pp. 770–778).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*(pp. 4700–4708).

Huang, C. H., Wu, H. Y., & Lin, Y. L. (2021). HarDNet-MSEG: A simple encoder–decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. *preprint arXiv:2101.07172*.

Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., De Lange, T., Halvorsen, P., & Johansen, H. D. (2019). ResUNet++: An advanced architecture for medical image segmentation. In *Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM)*(pp. 225–2255).

Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., & Johansen, H. D. (2020). Kvasir-SEG: A segmented polyp dataset. In *International Conference on Multimedia Modeling (MMM)*(pp. 451–462).

Liu, S., & Huang, D. (2018). Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*(pp. 385–400).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using shifted windows. *preprint arXiv:2103.14030*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*(pp. 3431–3440).

Mahmud, T., Paul, B., & Fattah, S. A. (2021). PolypSegNet: A modified encoder–decoder architecture for automated polyp segmentation from colonoscopy images. *Computers in Biology and Medicine, 128*, 104119.

Mao, Y., Zhang, J., Wan, Z., Dai, Y., Li, A., Lv, Y., Tian, X., Fan, D. P., & Barnes, N. (2021). Transformer transforms salient object detection and camouflaged object detection. *preprint arXiv:2104.10127*.

Margolin, R., Zelnik-Manor, L., & Tal, A. (2014). How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255).

Mori, Y., & Kudo, S. E. (2018). Detecting colorectal polyps via machine learning. *Nature Biomedical Engineering*, *2*(10), 713–714.

Nguyen, N. Q., Vo, D. M., & Lee, S. W. (2020). Contour-aware polyp segmentation in colonoscopy images using detailed upsampling encoder–decoder networks. *IEEE Access, 8*, 99495–99508.

Park, K. B., Choi, S. H., & Lee, J. Y. (2020). M-GAN: Retinal blood vessel segmentation by balancing losses through stacked deep fully convolutional networks. *IEEE Access, 8*, 146308–146322.

Pytorch. (2016). https://pytorch.org/, accessed on Aug. 2021.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*(pp. 234–241).

Safarov, S., & Whangbo, T. K. (2021). A-DenseUNet: Adaptive densely connected UNet for polyp segmentation in colonoscopy images with atrous convolution. *Sensors, 21*(4), 1441.

Seo, H., Huang, C., Bassenne, M., Xiao, R., & Xing, L. (2019). Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. *IEEE Transactions on Medical Imaging*, *39*(5), 1316–1325.

Siegel, R. L., Miller, K. D., Goding, S. A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A., & Jemal, A. (2020). Colorectal cancer statistics. *A Cancer Journal for Clinicians*, *70*(3), 145–164.

Silva, J., Histace, A., Romain, O., Dray, X., & Granado, B. (2014). Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery, 9*(2), 283–293.

Sun, X., Zhang, P., Wang, D., Cao, Y., & Liu, B. (2019). Colorectal polyp segmentation by U-net with dilation convolution. In *Proceedings of the IEEE International Conference on Machine Learning And Applications (ICMLA)*(pp. 851–858).

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*(pp. 4278–4284).

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*(pp. 6105–6114).

Tajbakhsh, N., Gurudu, S. R., & Liang, J. (2015). Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging, 35*(2), 630–644.

Tomar, N. K., Jha, D., Ali, S., Johansen, H. D., Johansen, D., Riegler, M. A., & Halvorsen, P. (2021a). DDANet: Dual decoder attention network for automatic polyp segmentation. In *Proceedings of the International Conference on Pattern Recognition (ICCV)*(pp. 307–314).

Tomar, N. K., Jha, D., Riegler, M. A., Johansen, H. D., Johansen, D., Rittscher, J., Halvorsen, P., & Ali, S. (2021b). FANet: A feedback attention network for improved biomedical image segmentation. *preprint arXiv:2103.17235*.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*(pp. 10347–10357).

Vania, M., & Lee, D. (2021). Intervertebral disc instance segmentation using a multistage optimization mask-RCNN (MOM-RCNN). *Journal of Computational Design and Engineering, 8*(4), 1023–1036.

Vania, M., Mureja, D., & Lee, D. (2019). Automatic spine segmentation from CT images using convolutional neural network via redundant generation of class labels. *Journal of Computational Design and Engineering, 6*(2), 224–232.

Van Rijn, J. C., Reitsma, J. B., Stoker, J., Bossuyt, P. M., Van Deventer, S. J., & Dekker, E. (2006). Polyp miss rate determined by tandem colonoscopy: A systematic review. *The American Journal of Gastroenterology, 101*(2), 343–350.

Vázquez, D., Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., López, A. M., Romero, A., Drozdzal, M., & Courville, A. (2017). A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, *2017*, 4037190.

Wei, J., Wang, S., & Huang, Q. (2020). F^3Net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*(Vol. 34, No. 07, pp. 12321–12328).

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*(pp. 3–19).

Wu, Z., Su, L., & Huang, Q. (2019). Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(pp. 3907–3916).

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., & Luo, P. (2021). Segmenting transparent object in the wild with transformer. *preprint arXiv:2101.08461*.

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 749–753.

Zhang, Y., Liu, H., & Hu, Q. (2021). TransFuse: Fusing transformers and CNNs for medical image segmentation. *preprint arXiv:2102.08005*.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Proceedings of Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA)*(pp. 3–11).