PSwinUNet: Bridging Local and Global Contexts for Accurate Medical Image Segmentation with Semi-Supervised Learning

Zhixuan Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail:18535049831@163.com

Chentao Qian
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail:qct0331@gmail.com

Bailin Liu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail:498194312@qq.com

Yijian Zhang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail:2278662342@qq.com

Hongpei Zhang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail:2464094384@qq.com

***Abstract*—It's highly crucial to divide up medical photos correctly in order to make diagnoses and plan treatments. Convolutional Neural Networks (CNNs) are very good at picking up local information, but they have problems with long-range dependencies. On the other side, Vision Transformers (ViTs) are good at modeling global context, but they need a lot of computer power and labeled data. To get surrounding these difficulties, we establish PSwinUNet, a hybrid CNN-Transformer system based on a partially supervised learning the structure. Adding a SwinTransformer block to a U-shaped structure makes PSwinUNet better at learning internationally semantics and up-sampling. It also uses a polarized self-attention mechanism in skip connections to keep spatial information from getting lost when the image is downsampled. PSwinUNet does a better job than the best gets closer that are currently accessible when tested on the BUSI, DRIVE, and CVC-ClinicDB datasets. For instance, it earned Dice Similarity Coefficient (DSC) scores of 0.781, 0.896, and 0.960 based on the BUSI data set with 1/8, 1/2, and entire labeled information, respectively. These scores are substantially better than those of the old UNet and UNet++ models.**

Keywords-Machine Learning; PSwinUNet; CNN-Transformer; Semi-supervised learning

I. INTRODUCTION

Deep learning (DL) has done very well in computer vision [1] and is a key part of improving medical image analysis technology. Accurate segmentation is a key part of computer-aided diagnosis and surgical guidance because it shows the boundaries of lesions and anatomical structures. [2,3] However, manual segmentation across different imaging modalities is very difficult because it requires a lot of clinical knowledge, takes a lot of time, and often gives inconsistent results because different clinicians have different preferences. [4, 5] These problems show how badly we need smart automated segmentation methods right now.

The U-shaped design that Unet [7] displays well is presently the most popular one. It may employ convolutional and pooling methods to extract features in a hierarchical way while

DOI: 10.2478/ijanmc-2025-0024

keeping spatial information during up-sampling because of its encoder-decoder configuration with skip connections. [6-10] These CNN-based methods will always have problems, even though newer ones like UNet++[9], UNet3+[11], nnUNet [10], and Attention U-Net [12] have done better.This is a big problem with medical photographs because you need to see the complete picture to see the rich textures and hazy borders. [13, 14]

Recent improvements to Vision Transformers (ViT) make them even more intriguing because they use self-attention approaches to explain how things interact with one another on a global scale. [15-18]There are two main drawbacks with the current ViT-based medical segmentation models. First, they need large-scale annotated datasets that are difficult to find in medical areas. Second, pure Transformer topologies frequently don't have the fine-grained spatial precision that is needed for pixel-level segmentation tasks.

To deal with these problems, it is recommended to use PSwinUNet. This is a semi-supervised hybrid architecture with excellent design performance, and it works well together.Our method aims to achieve efficient long-range dependency modeling without putting too much of a strain on the computer, strong performance even when there aren't many annotations through semi-supervised learning, and better preservation of spatial information during down-sampling and up-sampling processes. This goal is achieved by adding a Swin-Transformer block at the bottom layer of the network and introducing a polarization self-attention mechanism in the skip connections. There are three things we have done:

The research presents a U-shaped hybrid convolutional neural network - the transformer network, with a Swin-Transformer block placed at the bottleneck layer. This approach makes it possible to get global contextual information while still being able to do feature up-sampling quickly. We develop a semi-supervised training framework that significantly reduces the need for fully annotated datasets. This solves the problem of not having enough data in medical imaging fields.

This study employed the "Polarized Self-Attention Skip Connection" (SCPSA) technique, which is capable of enhancing the interaction between different dimensions in the channel domain and the spatial domain. This reduces the loss of information that happens when you down-sample many times. Extensive tests on three public medical datasets (BUSI [22], DRIVE [23], and CVC-ClinicDB [24]) show that PSwinUNet has the best segmentation accuracy and is better at generalizing than other approaches.

II. Relate Work

A. Semi-Supervised in Medical Sementic Segmentation

Semi-supervised methods use supervised learning on a small number of labeled data and extract useful representations from a large set of unlabeled samples. Making reliable annotations is hard work and takes a long time for medical professionals. Also, manual annotations can cause differences in segmentation because the experts are subjective. There is typically a lot of unlabeled data in healthcare institutions, and it is important to make the most of it. More and more people are using semi-supervised learning, as seen by works like CA-Net [20] and SCP-Net [26]. This method has worked very well on a number of medical semantic segmentation tasks, showing how strong the model is and how well it can handle problems that come up when there isn't a lot of annotated data.

B. CNN Methods

Most of the early methods for medical picture segmentation were based on contour-based approaches and typical machine learning algorithms [27, 28]. Deep convolutional neural networks (CNNs) were a big development, and the publication of Unet [7] revolutionized how medical images are cut up. There are several versions of U-Net, like UNet++, UNet3+, nnUNet, and Acc-UNet, that have been built since the U-shaped design is easy to use and works well. These methods work well because they can find complex features and patterns, which has helped the field make a lot of progress.

C. Combining CNNs with Self-attention mechanisms and Transformer

Researchers have been working hard to incorporate the self-attention mechanism into CNNs so that they operate better [29]. It has been said that rethinking skip connections is very important [30]. A lot of people are using other ways, like SwinUNet [15], which exclusively uses Transformer designs. The Transformer architecture was made just for sequence-to-sequence prediction, which makes it easier to collect semantic data on both a local and global level. But if you don't pay enough attention to little details, it could be hard to localize. Some studies are now trying to merge CNNs and transformers, which goes against the concept that CNNs are the most important [16, 31-32].

D. Summary of the Main Design

The article is about PSwinUNet, a semi-supervised way of cutting up medical images that attempts to combine the best features. The main aspects of the design are:

Hybrid CNN-Transformer Architecture: We can use CNNs to gain a lot of low-level features and Transformers that are used to get dependency chains and global semantics by adding a Swin-Transformer block to the U-shaped architecture. This makes it easy to go back to the original resolution information when you up-sample.

Semi-Supervised Learning Framework: This framework learns from a lot of data that doesn't have labels and a few samples that do. This makes the model stronger.

Skip Connection with Polarized Self-Attention (SCPSA): This involves adding a Polarized Self-Attention the mechanism to skip connections to change how they work. This boosts interactions in both the channel and spatial domains, which makes up for the loss of spatial information that arises when you down-sample.

These design features make PSwinUNet work very well on publicly available medical imaging datasets such as BUSI, DRIVE, and CVC-ClinicDB. It indicates that it can not just break up visuals into components but also use what it learns in new scenarios. In a number of various situations with annotation data, experimental results show that PSwinUNet works better than other cutting-edge approaches. When there aren't many annotations, the speed benefits are much bigger.

III. METHOD

This section begins with an overview of the proposed PSwinUNet architecture. Subsequently, a summary of the polarized self-attention (PSA) mechanism is provided. Finally, the Swin-Transformer block and Skip Connection with Polarized Self-Attention (SCPSA) mechanisms are introduced.

A. Architecture Overview

The architecture of our proposed PSwinUNet is illustrated in Figure 1, which comprises an encoder, decoder, Swin-Transformer block. Fig. 1 illustrates the architecture of PSwinUNet, comprising an encoder, decoder, Swin-Transformer block, and SCPSA module.

During the encoder phase, we receive an input medical image with dimensions ranging from H to W to 3. Subsequently, the feature dimension of each patch is projected into a specified dimension (denoted as C) via patch embedding. The encoding component of the PSwinUNet utilizes a pre-trained VGG16 model as its feature extraction network. We have redesigned the encoder component of the U-shaped architecture to closely resemble the VGG network [34], proposing the use of VGG16 as the backbone for the encoder. This modification is aimed at expediting the training of PSwinUNet and achieving superior segmentation performance. The transformed patch tokens traverse successive layers of CNNs and undergo patch merging operations.

This study proposes a convolutional neural network - transformer architecture composed of Swin-Transformer blocks and multiple convolutional neural network layers. In the modeling process with Transformers, the absence of low-resolution features is observed. Directly using features encoded by Transformers and up-sampling is not effective in recovering information at the original resolution. Convolutional encoding can yield more abundant low-level features, providing an improved solution to address the challenge of insufficient low-resolution features generated by Transformer encoding. We will elaborate on each block in the following sections.

B. Polarized Self-Attention (PSA) Module

Polarized self-attention represents an attention mechanism characterized by minimal compression in both spatial and channel dimensions, thus enabling sustained high performance on fine-grained pixel-level tasks. In photography, horizontal random light reflections commonly occur. Motivated by the notion that allowing light orthogonal to the horizontal direction to pass through polarization filtering could potentially enhance photo contrast, Polarized self-attention was introduced.

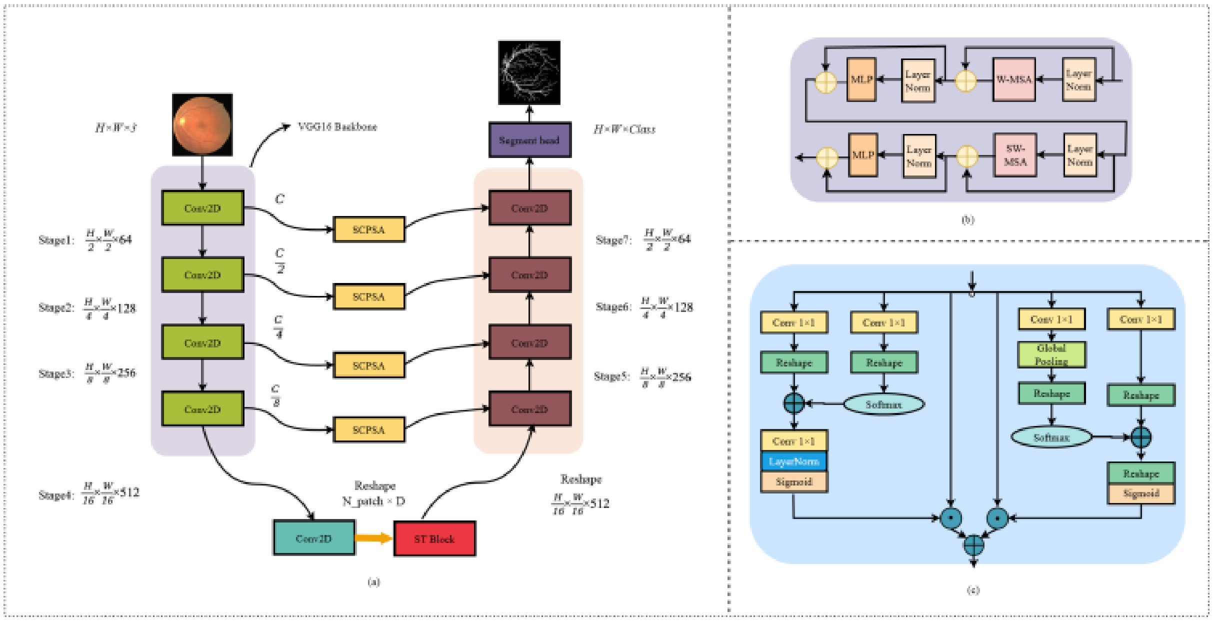

Figure 1. (a): We present the architecture of our PSwinUNet, a hybrid CNN-Transformer architecture. (b): TheSwin-Transformer block (c): The SCPSA module enhances cross-dimensional interactions from both channel and spatial aspects, compensating.

(1)This process involves folding features exclusively in one direction while simultaneously preserving high resolution.

(2)The comprehensive architecture of the paralleled layout within the PSA module is illustrated in Fig. 1(b).

C. Channel-only Branch1

The Channel-only branch (Ach(X) ∈ RCin×1×1) can be formulated as:

$$A^{ch}(X)=F_{SG}[W_{Z|\theta_1}(\sigma_1(W_V(x))\times F_{SM}(\sigma_2(W_q(x))))] \quad (1)$$

The operator FSM (·) corresponds to a softmax operation, and "×" denotes the matrix dot product operation defined as:

$$F_{SM}(X)=\sum_{j=1}^{N_p}\frac{e^{x_j}}{\sum_{m=1}^{N_p}e^{x_m}}x_j \qquad (1)$$

The number of internal channels shared between Wv, Wq, and Wz is C/2.

D. Spatial-only Branch

The spatial-only branch (Asp(X) ∈ R1×H×W) can be formulated as:

$$A^{sp}(X)=F_{SG}[\sigma_3(F_{SM}(\sigma_1(F_{GP}(W_q(X))))\times\sigma_2(W_v(X)))] \quad (3)$$

In this context, Wv and Wq represent standard 1×1 convolution layers. Additionally, δ1, δ2, and δ3 are three tensor reshape operators, and FSM(·) denotes a SoftMax operator.

The operator FGP(·) corresponds to a global pooling operation defined as:

$$F_{GP(X)}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X(:,i,j) \qquad (4)$$

Where × indicates the matrix dot-product operation.

E. Connection Modes

The outputs of the two branches can be arranged in a parallel layout

$$\begin{aligned} PSA_{\rho}(X) &= Z^{ch} + Z^{sp} \\ &= A^{\mathrm{ch}}(\mathrm{X}) \square^{\mathrm{ch}} X + A^{sp}(X) \square^{sp} X \end{aligned} \tag{2}$$

Or a sequential layout

$$\begin{aligned} PSA_{\rho}(X) &= Z^{sp}(Z^{ch}) \\ &= A_{sp}(\mathrm{A}_{\mathrm{ch}}(\mathrm{X}) \square_{ch} X) \square_{sp} A_{ch}(X) \square_{ch} X \end{aligned} \tag{6}$$

F. Double Swin-Transformer Block

The first one uses a common way for dividing a window that starts at the top-left pixel. The last module, on the other hand, uses a shifted windowing configuration, which is different from the usual partitioning mechanism used in the layer before it. This adjustment moves the windows by ([M2],[M2]) pixels, which is different from how they are usually split. Figure 2 shows you how to figure out how much self-attention there is in a window that has been relocated. Figure 2 shows how to get all the self-attention values you need in one forward pass. You only need to calculate four windows to do this. To find consecutive Swin-Transformer blocks using the shifted window partitioning approach, execute the following:

$$\hat{Z}^{l} = W - MSA(LN(Z^{l-1})) + Z^{l-1} \tag{3}$$

$$Z^{l} = MLP(LN(\hat{Z}^{l})) + \hat{Z}^{l} \tag{4}$$

$$\hat{Z}^{l+1} = SW - MSA(LN(Z^{l})) + Z^{l} \tag{5}$$

$$Z^{l+1} = MLP(LN(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \tag{6}$$

Where Zˆl and Zl denote the output features of the (S) W-MSA module and the MLP module for block l.In [15], a U-shaped encoding-decoding architecture is employed. This approach led to an absence of low-resolution features during Transformer modeling, and up-sampling might not effectively recover information at the original resolution. However, convolutional operations in the encoder are capable of providing rich low-level features.

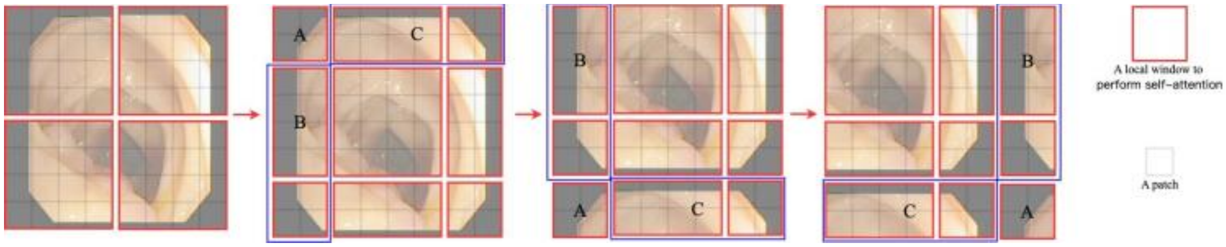

Figure 2. Lustration of an efficient batch computation approach for self-attention in shifted window configuration.

IV. EXPERIMENT

A. Datasets and Evaluation

This study tested the proposed model on three different medical image datasets: colonoscopy (colon) photos, digital retinal tissue images for vessel extraction (DRIVE) images. We use a random splitting method to separate the images in each dataset into labeled and unlabeled sets. We choose a random sample of photos (for example, 1/8, 1/4, 1/2, or the entire dataset) to be labeled information and leave the rest to be unlabeled data. We do this random selection for each dataset to ensure our results can be used elsewhere. We also think about category balance when we partition the data such that the labeled and unlabeled sets have a representative distribution of every class in the dataset.

Breast Ultrasound Image Dataset (BUSI): This dataset contains breast ultrasound images of 600 women aged 25 to 75 in 2018. There are 780 pictures in it, and each one measures approximately 500 pixels wide. There are three types of images: normal, benign, and malignant.

We made sure that all the photographs were the same size by resizing them to 512×512.

CVC-ClinicDB data set, There are 612 pictures in this dataset that were taken from colonoscopy videos. All of them have been reduced to 512× 512 pixels. There is a ground truth: a respirator in each photography frame that shows where the polyp is located.

Digital Retinal Images (Drive): This work uses the same dataset that was used to segment retinal vessels. There are 40 JPEG color fundus photos in this set, and 7 of them show signs of disease. At first, each image had a resolution of 584×565 pixels, but they were all shrunk to 512×512 pixels to make them all the same size.

Metrics for Evaluation: We use the average Dice-Similarity Coefficient (DSC) as a way to measure how well our models work. Dice loss is a useful tool for optimizing models since it works well with DSC, a widely used measure. There are four types of predictions: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The DSC is calculated like this:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \qquad (7)$$

B. Implementation Details

The Ubuntu 20.04 system was used for all of the experiments in this study. There are two parts to the instruction process. During fully supervised training, the model goes through a freezing phase and then an unfreezing phase. The first step of semi-supervised education is an unfreezing phase, which is followed by a freezing phase. In semi-supervised learning, the VGG16 backbone stays frozen for the first 50 epochs. Then, at epoch 100, the semi-supervised training starts. The goal of carefully freezing all VGG16 backbone parameters is to speed up the training process while using less GPU memory. Next, the VGG16 backbone parameters are unfrozen so that they can be changed. This method is used to make sure that training moves faster and uses less GPU memory in the beginning. After that, updating the parameters helps the model perform better.

C. Comparison with State-of-the-artss

This study conducted tests using different quantities of labeled data. specifically 1/8, 1/4, 1/2, and the full amount of labeled data. Then, PSwinUNet's segmentation performance was compared to that of the best segmentation models available. Figure 3 shows the segmentation results of our proposed PSwinUNet compared to existing state-of-the-art models, utilizing 50% labeled data. We did experiments on the BUSI and DRIVE datasets to make sure the results were accurate. Table 1 shows that PSwinUNet works best on the three datasets when there is 50% labeled data.

To test the semi-supervised technique on the BUSI dataset, we randomly choose 1/8, 1/4, and 1/2 of the photos as labeled data and the balance of the training images as unlabeled data. Under each semi-supervised and fully supervised setting, PSwinUNet gets DSC ratings of 0.781, 0.813, 0.896, and 0.960, which are higher than those of UNet, UNet++, and UNet3+. The Swin-Transformer module added to the bottom of the PSwinUNet network is what made this upgrade possible. This module quickly and easily captures global relationships in one step while also highlighting local connections between items, this improves the segmentation performance. We also compared PSwinUNet with swwin-unet and TransUNet. The CNN-Transformer architecture always makes things work better, with increases ranging from 0.027 to 0.052, no matter how much training data is used. Our method is better compared to the SwinUNet that uses only the Swin-Transformer block.

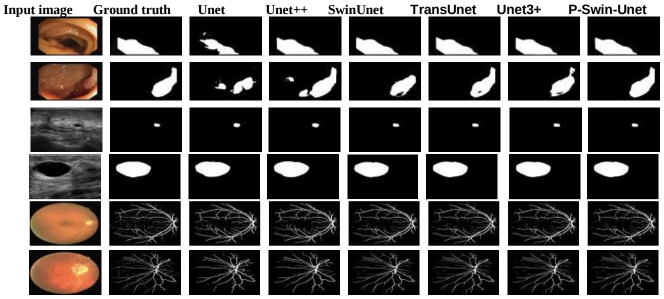

Figure 3. The visual segmentation results of various methods in the semi-supervised experiment with 1/2 labeled data amounts on the BUSI, DRIVE, and CVC-ClinicDB datasets are displayed in Fig. 3. Notably, our PSwinUNet demonstrates relatively superior visualizations compared to other methods.

It shows that all of the experimental results are superior. The PSwinUNet model makes segmentation work better by adding polarized attention to the skip connections. This model mixes deep and shallow semantic data. We also did the same tests again on the DRIVE and CVC-ClinicDB datasets, and in every case, our technique delivered the best results. Also, the best results occur when there is 1/8, 1/2, or all of the data.

In the end, PSwinUNet is always better than other networks like UNet, UNet++, TransUUNet, SwinUUNet, and UNet3+. The proposed PSwinUNet architecture shows that it performs effectively for the challenging task of images with only some supervision. PSwinUNet can work well with different datasets and amounts of labeled data. It could be valuable in the real world because it does well in many different scenarios. PSwinUNet is strong because it keeps growing stronger as additional data with tags becomes accessible. Other methods are also growing better, but PSwinUNet is the best since it keeps getting better over time.

D. Ablation Study

The next series of ablation testing will begin with the UNet. We look at the Dice values on the test set for the following setups to observe how each module changes the model as a whole: There are different ways to connect to PSA, with and without the SCPSA module and with and without Table 1. Table 1 displays the numerical results for DSC of various methods on the BUSI, DRIVE, and CVC-ClinicDB datasets using 1/8, 1/4, 1/2, and all of the labeled data.

As illustrated in Table 2, in our evaluation, we analyze each module’ s impact on the overall model by comparing Dice values on the test set for the following configurations: without (w/o) the SCPSA module, without (w/o) the Swin-Transformer block, and the complete PSwinUNet model.

E. Effect of PSA Connection Modes

In Table 2, we conduct experiments with five PSA connection modes on a resolution scale of 512×512 using the BUSI dataset. PSA comprises channel self-attention and spatial self-attention, with two types of connection modes: sequential layout and parallel layout. Therefore, we define five connection modes for PSA, including no attention, individual spatial branch, individual channel branch, parallel layout, and sequential connections. The results indicate that the parallel connection mode essentially achieves optimal performance across varying amounts of labeled training data. Our method, PSwinUNet, demonstrates improvements in DSC by 0.781, 0.896, and 0.960 when the labeled data amounts are 1/8, 1/2, and full, respectively. When the labeled data amount is 1/4, the difference between the parallel layout and series layout is only 0.009. To ensure a fair comparison with other methods and to consider segmentation performance, we conduct all experiments based on the default parallel connection mode.

TABLE I. THE QUANTITATIVE RESULTS FOR DSC OF VARIOUS METHODS ON 1/8, 1/4, 1/2, AND FULL LABELED DATA AMOUNTS ARE PRESENTED

Method	**Dataset**	**Labeled Data Amount**			
		1/8	***1/4***	***1/2***	***full***
UNet	***BUSI DRIVE***	***0.612 0.686***	***0.744 0.753***	***0.791 0.801***	***0.836 0.844***
	CVC-ClinicDB	***0.570***	***0.691***	***0.731***	***0.804***
UNet++	***BUSI DRIVE***	***0.597 0.691***	***0.783 0.744***	***0.832 0.795***	***0.893 0.852***
	CVC-ClinicDB	***0.602***	***0.686***	***0.751***	***0.804***
SwinUNet	***BUSI DRIVE***	***0.688 0.723***	***0.761 0.772***	***0.869 0.815***	***0.952 0.846***
	CVC-ClinicDB	***0.711***	***0.745***	***0.815***	***0.883***
TransUNet	***BUSI DRIVE***	***0.712 0.738***	***0.795 0.791***	***0.891 0.842***	***0.943 0.894***
	CVC-ClinicDB	***0.732***	***0.791***	***0.844***	***0.934***
UNet3+	***BUSI DRIVE***	***0.662 0.663***	***0.736 0.720***	***0.806 0.791***	***0.897 0.862***
	CVC-ClinicDB	***0.705***	***0.818***	***0.864***	***0.907***
PSwinUNet	***BUSI DRIVE***	***0.781 0.740***	***0.813 0.786***	***0.896 0.872***	***0.960 0.896***
	CVC-ClinicDB	***0.750***	***0.802***	***0.874***	***0.939***

TABLE II. ABLATION STUDY ON THE IMPACT OF PSA CONNECTION MODES ON BUSI DATASET.

Connection Modes	Data Amount			
	1/8	*1/4*	*1/2*	*full*
None	***0.692***	***0.733***	***0.791***	***0.855***
Channel-only branch	***0.702***	***0.782***	***0.858***	***0.891***
Spatial-only branch	***0.723***	***0.757***	***0.844***	***0.889***
parallel layout	***0.781***	***0.813***	***0.896***	***0.960***
sequential layout	***0.744***	***0.822***	***0.868***	***0.915***

F. Effect of Swin-Transformer Block

The proposed PSwinUNet requires the Swin-Transformer block to work. It has a hierarchical architecture with window partitions at different levels to make calculations easier. It also has shifted windows to make it easier for nearby windows to interact, which gives it global modeling capabilities. Table 3 shows the results of PSwinUNet on the BUSI dataset, both with and without the Swin-Transformer block. When you compare PSwinUNet with the Swin-Transformer to PSwinUNet with no Swin-Transformer, you can see a big difference in performance (from 0.017 to 0.119) with varying amounts of training data.

For example, adding the Swin-Transformer block to the DSC enhances it by 0.119, 0.017, and 0.051 for 1/8, 1/4, and 1/2 labeled data amounts, respectively. This shows that taking off the Swin-Transformer block, which is in charge of getting multi-scale characteristics, is not enough. It makes it such that features can't be extracted from the image's local space.

TABLE III. ABLATION STUDY ON THE IMPACT OF SWIN-TRANSFORMER BLOCK ON BUSI DATASET.

Method	Data Amount			
	1/8	*1/4*	*1/2*	*full*
Baseline	***0.612***	***0.744***	***0.791***	***0.836***
PSwinUNet(w/o)	***0.662***	***0.796***	***0.845***	***0.887***
PSwinUNet(Ours)	***0.781***	***0.813***	***0.896***	***0.960***

V. DISCUSSION

This study looks at what PSwinUNet, a new way to break apart medical images, can do. By integrating Swin-Transformer blocks with a Geographically Polarized Self-Attention (SCPSA) component, PSwinUNet has done very well on benchmark datasets like DRIVE, CVC-ClinicDB, and BUSI. In that order, the dice coefficients are 0.781, 0.896, and 0.960. Using both CNNs and transformers, PSwinUNet strikes an improved harmony between precision and rapidity than typical U-Net topologies. This makes it very useful in therapeutic settings where resources are limited. Also, PSwinUNet's self-attention mechanism just uses picture data, which makes it better than examples like LViT that need more than one form of input. The SCPSA module overcomes a big problem with regular attention approaches by adding polarization. This retains the precision of the space while also capturing long-range relationships. Tests have demonstrated that this architecture improves segmentation accuracy by 3 – 5% on challenging characteristics such as narrow vascular networks. In the context of semi-supervised learning, the study combines a patch-based contrastive loss function to create fake labels. This strategy not only cuts down on noise in unlabeled data, but it also makes sure that the meaning is the same in all local locations. This does rid of the challenge of class collision that happens a lot in traditional contrastive approaches. The model does so well at tasks like retinal vascular segmentation (with Dice scores > 0.896) that it could cut the period it takes to edit it manually by 60 – 70% in clinical situations. Some possible future research areas that could improve medical image segmentation for more accurate and useful applications in clinical settings include combining different types of data, dynamic share pruning, and federated learning.

VI. CONCLUSIONS

This study designed a well-structured semi-supervised U-shaped hybrid convolutional neural network - Transformer architecture, named PSwinUNet. This architecture interacts among remote semantic information. We also suggest a new module identified as the SCPSA module that combines channel and spatial features. Our technique can get the best results on the BUSI and DRIVE, as well as the CVC-ClinicDB datasets, according to experimental results. In general, PSwinUNet has demonstrated that it can learn important anatomical correlations from medical images. But this study showed that even while outcomes were outstanding with just a tiny amount of labeled data, they were still not good enough. Because of this, it is still uncommon to apply these findings in clinical settings. In the future, we will focus on how to make the architecture for medical picture segments more complex and stable, with a focus on using less labeled data.

REFERENCES

[1] Guan, Z. et al. Artificial intelligence in diabetes management: advancements, opportunities, and challenges. Cell Reports Medicine (2023).

[2] Shen, D, Wu, G. & Suk, H. -I. Deep learning in medical image analysis. Annu. Review biomedical engineering 19, 221-248 (2017).

[3] Hatamizadeh, A. et al. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 574-584 (2022).

[4] Chen, X. et al. Learning active contour models for medical image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 11632-11640 (2019).

[5] Chen, B., Liu, Y., Zhang, Z., Lu, G. & Kong, A. W. K. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. IEEE Transactions on Emerg. Top. Comput. Intell. (2023).

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

[7] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 234-241 (Springer, 2015).

[8] Ibtehaz, N. & Kihara, D. Acc-unet: A completely convolutional unet model for the 2020s. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 692-702 (Springer, 2023).

[9] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018

[10] 1Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. methods 18, 203-211 (2021).

[11] Huang, H. et al. Unet 3+: A full-scale connected unet for medical image segmentation. In ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), 1055-1059 (IEEE, 2020).

[12] Oktay, O. et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018).

[13] Nazir, A. et al. Off-enet: An optimally fused fully end-to-end network for automatic dense volumetric 3d intracranial blood vessels segmentation. IEEE Transactions on Image Process. 29, 7192-7202 (2020).

[14] Cao, H. et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision, 205-218 (Springer, 2022).

[15] 16. Chen, J. et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021).

[16] Zhou, H.-Y. et al. nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021).

[17] Heidari, M. et al. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 6202-6212 (2023).

[18] Wang, Y., Xiao, B., Bi, X., Li, W. & Gao, X. Mcf: Mutual correction framework for semi-supervised medical image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15651-15660 (2023).

[19] Zhao, C. et al. Context-aware network fusing transformer and v-net for semi-supervised segmentation of 3d left atrium. Expert. Syst. with Appl. 214, 119105 (2023).

[20] Bai, Y., Chen, D., Li, Q., Shen, W. & Wang, Y. Bidirectional copy-paste for semi-supervised medical image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11514-11524 (2023).

[21] Yap, M. H. et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal biomedical health informatics 22, 1218-1226 (2017).

[22] Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A. & Van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging 23, 501-509 (2004).

[23] Bernal, J. et al. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Medical imaging graphics 43, 99-111 (2015).

[24] Zhang, Z. et al. Self-aware and cross-sample prototypical learning for semi-supervised medical image segmentation. arXiv preprint arXiv:2305.16214 (2023).

[25] Tsai, A. et al. A shape-based approach to the segmentation of medical imagery using level sets. IEEE transactions on medical imaging 22, 137-154 (2003).

[26] Held, K. et al. Markov random field segmentation of brain mr images. IEEE transactions on medical imaging 16, 878-886 (1997).

[27] Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018).

[28] Peng, Y., Sonka, M. & Chen, D. Z. U-net v2: Rethinking the skip connections of u-net for medical image segmentation. arXiv preprint arXiv:2311.17791 (2023).

[29] Zhang, Y., Liu, H. & Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention - MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I 24, 14-24 (Springer, 2021).

[30] Wang, Z. et al. Smeswin unet: Merging cnn and transformer for medical image segmentation. In International Conference on Medical Image Computing

and Computer-Assisted Intervention, 517‐526 (Springer, 2022).

[31] Chen, Y. et al. Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 7759‐7767 (2024).

[32] Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016).