Swin-PSA-SegNet: A Hybrid Framework Combining Swin Transformer, Polarized Self-Attention, and SegNext for Robust Medical Image Segmentation

Nasir Muhammad, Baoshan Sun

Abstract—**This paper presents a novel framework integrating the Swin Transformer, Polarized Self-Attention (PSA), and SegNext to advance medical image segmentation. The Swin Transformer leverages hierarchical self-attention to effectively model both global and local semantic relationship, addressing the challenges of long-range dependencies in medical imagin- g. PSA further advance the framework by preserving high resolution in both spatial and channel dimensions, o- ptimizing pixel level feature for segmentation tasks. SegNext contributes lightweight architectural elements, ensuring computational efficiency without sacrificing accuracy. The proposed approach is comprehensively evaluated on Synapse dataset, achieving state-of-the-art performance in multi-organ tasks. This unified framework sets a new standard for precision and efficiency in medical image segmentation.**

Index Terms—**Medical Image Segmentation, Semantic Segmentation, Swin Transformer, SegNext, Polarized Self Attention**

I. Introduction

Medical image segmentation is a fundamental aspect of contemporary healthcare systems, enabling accurate and efficient identification of anatomical structures. This process plays a crucial role in clinical applications such as disease diagnosis, surgical guidance, and treatment planning. Traditional manual segmentation, while effective in certain cases, is time-intensive, subject to variability among experts, and impractical for large-scale or time-sensitive analyses. These limitations underscore the growing need for automated segmentation methods powered by advanced computational techniques [1-3].

Deep learning has significantly transformed the field of medical image analysis, particularly through the introduction of convolutional neural networks (CNNs). Models like U-Net and its derivatives have become widely adopted for segmentation tasks due to their encoder-decoder architectures that efficiently extract and reconstruct spatial features. However, CNNs are inherently limited by their localized processing, which restricts their ability to model long-range dependencies and global semantic contexts. This limitation is particularly problematic for tasks that require an understanding of complex or variable anatomical structures [4] [5].

Transformers, which are built upon self-attention mechanisms, have gained significant attention for their ability to overcome the limitations of CNNs. Originally developed for natural language processing, transformers efficiently capture global feature relationships through self-attention, making them highly effective for vision tasks. The Swin Transformer, a vision-specific adaptation, integrates hierarchical representation learning with computational efficiency through its shifted window mechanism. This design is particularly well-suited for medical image segmentation, where capturing both global and local feature interactions is critical [4][6].

Attention mechanisms, forming the foundation of transformers, have evolved to include advanced techniques such as Polarized Self-Attention (PSA). PSA enhances segmentation accuracy by maintaining high-resolution feature representations, enabling precise pixel-level delineation of complex structures like organs and lesions. Additionally, lightweight architectures like SegNext complement these advancements by addressing the computational demands of segmentation models, ensuring efficiency and viability in resource-constrained clinical environments [7][8][2].

Despite the advancements offered by these methods, an integrated approach that combines the strengths of hierarchical transformers, advanced attention mechanisms, and lightweight designs remains underexplored. To address this gap, we propose a unified framework that leverages the Swin Transformer, PSA, and SegNext for enhanced medical image segmentation. This framework is designed to:

1. Utilize the Swin Transformer for effective hierarchical and global feature extraction [4] [6].
2. Enhance resolution and accuracy with PSA's pixel-level attention mechanisms [7].
3. Achieve computational efficiency through the lightweight architecture of SegNext [2] [8].

To validate the proposed approach, we conduct extensive experiments on Synapse dataset (for multi-organ segmentation). This dataset encompasses a range of challenging use cases, including segmentation of complex organ geometries and variations in patient anatomy. Our results demonstrate that the proposed method achieves superior segmentation accuracy and robustness while maintaining computational efficiency, establishing it as a leading approach in medical image analysis [3] [9].

II. ReLATED WORK

Advancements in medical image segmentation have been driven by innovations across convolutional architectures, attention mechanisms, and transformers. This section reviews these approaches in detail and discusses their contributions and limitations.

Manuscript received January 27, 2025

Nasir Muhammad, School of Computer Science and Technology, Tiangong University, Tianjin, China

Baoshan Sun, School of Computer Science and Technology, Tiangong University, Tianjin, China

A. CNN-Based Architectures

Convolutional neural networks (CNNs) are foundational to modern medical image segmentation. Early work such as U-Net [10] introduced an encoder-decoder structure with skip connections, enabling the recovery of spatial information lost during down sampling. This innovation proved highly effective for biomedical imaging tasks and inspired a range of subsequent enhancements. U-Net++ [5] extended the U-Net architecture by integrating nested skip pathways to improve feature refinement and fusion, which enhanced segmentation performance in complex anatomical scenarios.

For 3D imaging tasks, models like 3D U-Net [11] and V-Net [2] leveraged volumetric convolutions to process multi-dimensional data, making them particularly useful for medical images like MRI and CT scans. V-Net further addressed the class imbalance inherent in medical datasets by optimizing a Dice coefficient-based loss function, significantly improving segmentation accuracy in challenging cases. Despite these successes, CNNs remain constrained by their localized receptive fields, limiting their ability to model global relationships across an image.

To address these limitations, researchers explored hybrid approaches. For instance, ResU-Net [12] introduced residual connections to enhance feature propagation, while DenseU-Net incorporated densely connected layers to improve feature reuse. However, while these models advanced the state of the art, their reliance on convolutional operations still restricted their ability to capture long-range dependencies.

B. Attention Mechanisms

Attention mechanisms have emerged as a powerful tool for augmenting neural networks by allowing them to focus on relevant features. Squeeze-and-Excitation (SE) blocks [8] recalibrate channel-wise feature representations, improving model sensitivity to key image regions. Efficient Channel Attention (ECA) [13] further optimized this approach by introducing lightweight channel recalibration, reducing computational overhead without sacrificing accuracy.

For spatial attention, models like Attention U-Net [10] integrated attention gates to focus on target regions, such as organs or lesions, while suppressing irrelevant background areas. This approach enhanced segmentation precision without significantly increasing model complexity. Similarly, the Convolutional Block Attention Module (CBAM) [14] combined channel and spatial attention to further refine feature selection.

Polarized Self-Attention (PSA) [7] represents a more recent advancement in attention mechanisms, designed specifically for pixel-level tasks like segmentation. PSA preserves high resolution across both spatial and channel dimensions, making it ideal for tasks requiring fine-grained predictions. By maintaining this resolution, PSA has demonstrated significant improvements in segmentation accuracy for tasks involving complex or irregular anatomical structures.

C. Transformers in Medical Image Segmentation

Transformers have revolutionized computer vision by introducing self-attention mechanisms capable of capturing global dependencies. The Vision Transformer (ViT) [6] was one of the first models to adapt transformers for image processing, demonstrating their ability to achieve competitive performance on image classification tasks. However, ViT's reliance on large datasets and high computational costs limited its application in medical imaging, where data is often scarce.

To address these challenges, the Swin Transformer [4] introduced a hierarchical structure with shifted windows, enabling efficient computation while maintaining global context. This approach allowed for multi-scale feature learning, which is critical in medical segmentation tasks involving varying organ sizes and structures. Swin-Unet [4], a transformer-based U-shaped architecture, extended these principles by integrating Swin Transformer blocks with skip connections, achieving state-of-the-art results in tasks like multi-organ segmentation.

Other hybrid models, such as Medical Transformer [15] and TransUNet [16], combined the strengths of transformers and convolutional layers. These approaches utilized transformers for global feature extraction while retaining the convolutional encoder-decoder structures for local feature learning. While promising, these hybrid models often involve increased computational complexity, necessitating further optimization.

D. Lightweight Architectures

Efficiency has become a critical factor in medical segmentation, particularly for real-time applications. Lightweight models like SegNext [2] address this need by employing advanced feature extraction techniques with reduced computational demands. These architectures achieve competitive performance while maintaining efficiency, making them suitable for deployment in resource-constrained environments, such as portable imaging devices or edge computing systems.

Efforts to combine efficiency with accuracy include the MobileNet family [17] which employs depth-wise separable convolutions, and EfficientNet [18] which scales network dimensions using a compound scaling method. SegNext builds on these ideas by balancing architectural simplicity with the ability to capture fine-grained details, ensuring high-quality segmentation without excessive computational costs.

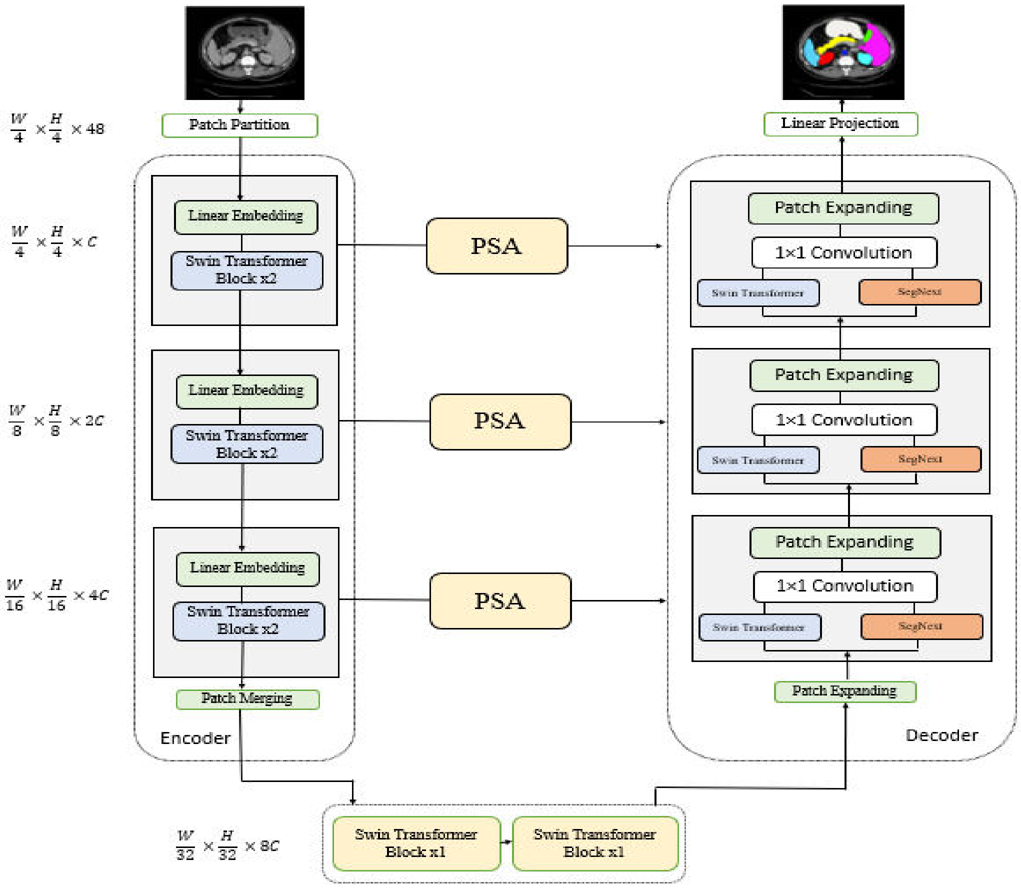

$\frac{W}{4}\times\frac{H}{4}\times 48$
$\frac{W}{4}\times\frac{H}{4}\times C$
$\frac{W}{8}\times\frac{H}{8}\times 2C$
$\frac{W}{16}\times\frac{H}{16}\times 4C$
$\frac{W}{32}\times\frac{H}{32}\times 8C$
Patch Partition
Linear Embedding
Swin Transformer Block x2
Patch Merging
Encoder
PSA
Linear Projection
Patch Expanding
1×1 Convolution
Swin Transformer
SegNext
Decoder
Swin Transformer Block x1

Fig.1 Overview

III. MATHODS

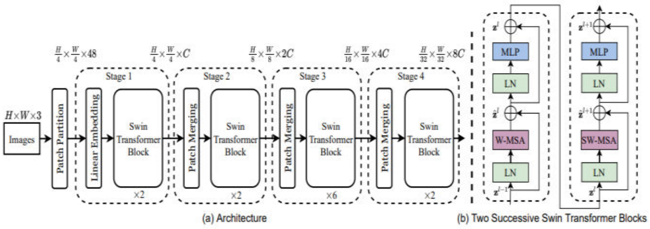
In this section, we outline the framework integrating the Swin Transformer, Polarized Self-Attention (PSA), and SegNext for medical image segmentation. The goal is to leverage the complementary strengths of these components—hierarchical feature extraction, high-resolution attention, and computational efficiency—to achieve precise and robust segmentation in challenging datasets.

A. Overview of the Framework

As shown in Fig.1, the proposed framework builds upon Swin-Unet, a transformer based model as the backbone. Swin Unet utilizes the Swin Transformer hierarchically to extract global-local features with PSA's ability to preserve fine-grained details at high resolution. SegNext serves as the backbone for efficient feature extraction, ensuring the model remains computationally lightweight. The integration of these components creates a unified architecture capable of addressing the unique challenges of medical imaging, including variability in organ shape and size, class imbalance, and the need for high accuracy in pixel-level segmentation.

B. Swin Transformer for Hierarchical Feature Learning

An overview of Swin Transformer is shown in Fig.2, the Swin Transformer is used as the encoder to extract multi-scale hierarchical features from the input image. It employs a shifted window-based self-attention mechanism, which computes attention within localized windows while maintaining connections between adjacent windows. This approach strikes an optimal balance between computational efficiency and the ability to capture long-range dependencies, which is crucial for tasks that require both global context and detailed local features. The hierarchical representation learning in the Swin Transformer allows the model to effectively extract features at multiple scales, making it well-suited for complex segmentation tasks. Additionally, the shifted window attention mechanism reduces the computational overhead typically associated with global self-attention methods, making the model more efficient while still maintaining high performance in capturing intricate relationships across the image.


$\frac{H}{4}\times\frac{W}{4}\times 48$
$\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{H}{32}\times\frac{W}{32}\times 8C$
$H\times W\times 3$
Images
Patch Partition
Stage 1
Linear Embedding
Swin Transformer Block
×2
Stage 2
Patch Merging
×2
Stage 3
×6
Stage 4
×2
z^l
MLP
LN
$\hat{z}^l$
W-MSA
z^{l-1}
z^{l+1}
$\hat{z}^{l+1}$
SW-MSA

(a) Architecture

(b) Two Successive Swin Transformer Blocks

Fig.2 (a) The architecture of a Swin Transformer (Swin-T) (b) two successive Swin Transformer Blocks. W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

C. Polarized Self-Attention for Fine-Grained Detail

Polarized Self-Attention (PSA) is incorporated into the decoder to refine the resolution of feature maps and improve the accuracy of boundary delineation. PSA achieves this b dividing the attention mechanism into two separate branches as shown in Fig.3: one that focuses on spatial features and another that emphasizes channel-wise relationships. The results from both branches are then combined to generate more refined feature maps. This dual attention mechanism helps preserve both spatial and channel-wise details, which is essential for achieving pixel-level precision in segmentation tasks. Moreover, PSA significantly enhances the accuracy of boundary detection, particularly for anatomical structures that have irregular shapes, ensuring that the model can accurately delineate complex and detailed regions in medical images.

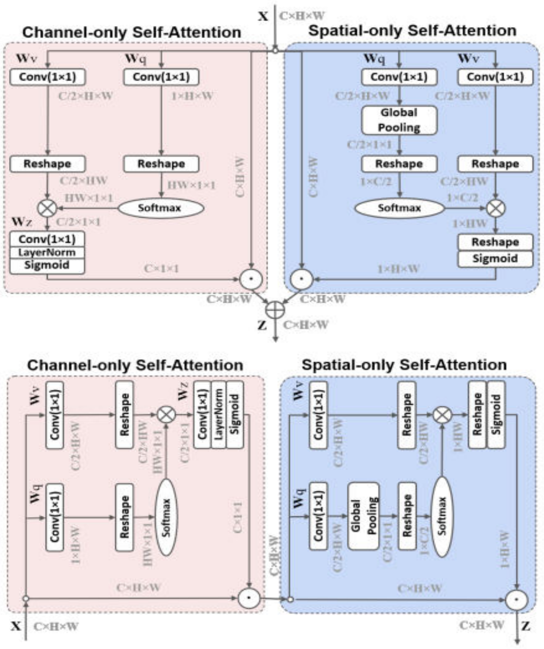


Fig.3 The Polarized Self-Attention (PSA) Structure

D. SegNext for Lightweight Feature Extraction

SegNext functions as the core architecture, offering a lightweight yet robust feature extraction backbone. By incorporating advanced convolutional designs alongside efficient attention mechanisms, SegNext ensures that the overall framework delivers high performance while minimizing computational demands. This makes it particularly suitable for real-time processing in clinical environments, where speed and efficiency are essential. Additionally, as shown in Fig.4, SegNext achieves competitive segmentation accuracy despite its reduced model complexity, allowing it to maintain strong performance while optimizing resource usage.

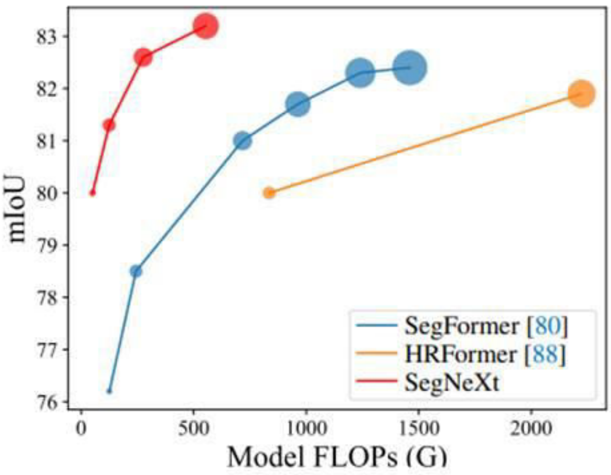


Fig.4 Performance-Computing curves on the Cityscapes validation sets.

E. Integration of Components

The integration of Swin Transformer, PSA, and SegNext follows a modular design that ensures both efficiency and high performance. The Swin Transformer is used as the encoder to process the input medical image, transforming it into a set of multi-scale features that capture both global and local contexts. In the next stage, PSA is applied in the decoder to refine these features, enhancing the resolution and ensuring that the segmentation outputs maintain high accuracy. Finally, SegNext serves as the lightweight backbone, providing efficient feature representation and minimizing computational overhead while still maintaining the robustness needed for accurate segmentation.

F. Loss Function

To address the class imbalance often observed in medical imaging datasets, we employ a hybrid loss function combining Dice Loss and Cross-Entropy Loss. Dice Loss ensures accurate segmentation of smaller regions, while Cross-Entropy Loss provides stable convergence.

Formulation:

$$\mathcal{L}_{total} = \lambda \cdot \mathcal{L}_{Dice} + (1-\lambda) \cdot \mathcal{L}_{CE} \ ,$$

where $\mathcal{L}_{Dice}$ is the Dice Loss, $\mathcal{L}_{CE}$ is the Cross-Entropy Loss, and λ = 0.6 balances their contributions. The Dice Loss is defined as:

$$LDice = 1 - \frac{2.\sum_{i=1}^{N} p_i \cdot g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i}$$

where p_i is the predicted probability for voxel i, g_i is the corresponding ground truth value (1 for foreground, 0 for background), and N is the total number of voxels. Dice Loss focuses on maximizing the overlap between the predicted and ground truth segmentation.

The Cross-Entropy **Loss** is defined as:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}[g_i\, log(p_i) + (1-g_i)\, log(1-p_i)]$$

where p_i and g_i retain their definitions. Cross-Entropy Loss penalizes incorrect predictions, encouraging the model to assign high confidence to the correct class.

G. Datasets and Evaluation Metrics

The framework is evaluated using the Synapse multi-organ segmentation dataset. The Synapse dataset tests the framework's ability to handle multi-class organ segmentation, challenging the model to manage inter-class variability. To assess the performance of the framework, several evaluation metrics are used, including the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and Mean Surface Distance (MSD), which provide comprehensive insights into the segmentation accuracy, spatial overlap, and boundary precision.

IV. EXPERIMENTS

A. Dataset

Experiments were conducted using the Synapse multi-organ segmentation dataset, which consists of 30 abdominal CT scans, comprising a total of 3,779 axial abdominal clinical images. Following the methodology established in Swin-Unet, the dataset was randomly divided into 18 scans for training and 12 scans for testing. The evaluation focused on segmenting eight specific abdominal organs: the aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach. Metrics used for evaluation included the average Dice Similarity Coefficient (DSC) to measure segmentation accuracy and the average Hausdorff Distance (HD) to assess boundary precision. Data augmentation techniques, such as random rotations, flipping, and elastic deformations, were applied during training to enhance the robustness of the model.

B. Results

The proposed framework was compared to state-of-the-art convolution-based and transformer-based segmentation models. Table 1 summarizes the performance of our model alongside existing methods, highlighting its superior performance in terms of segmentation accuracy and boundary precision.

Among convolution-based methods, U-Net achieved a DSC of 76.85% and an HD of 39.7, while Att-UNet improved on these results with a DSC of 77.77% and an HD of 36.02. However, these methods showed limitations in accurately segmenting complex organ boundaries. V-Net, another convolutional model, achieved a DSC of 68.81%, reflecting its weaker ability to handle inter-organ variability.

Transformer-based methods, such as ViT and R50 ViT, demonstrated moderate segmentation accuracy, achieving DSC scores of 67.86% and 71.29%, respectively. TransUnet improved on this performance with a DSC of 77.48% but struggled with boundary precision, as reflected in its HD value of 31.69. Swin-Unet, one of the strongest transformer-based methods, achieved a DSC of 79.13% and an HD of 21.55, highlighting its improved segmentation capabilities.

The proposed framework, combining Swin Transformer, PSA, and SegNext based on Swin-Unet outperformed all baseline models. It achieved the highest DSC score of 80.47% and the lowest HD value of 16.71, demonstrating its ability to achieve accurate segmentation while maintaining precise boundary delineation. For example, it achieved a DSC of 94.27% for the liver, 84.98% for the left kidney, and 70.72% for the gallbladder, consistently surpassing other methods. Fig.5, illustrates the superior classification accuracy of our approach, along with its enhanced capability to mitigate overfitting and boundary segmentation challenges, demonstrating substantial improvements over prior architectures.

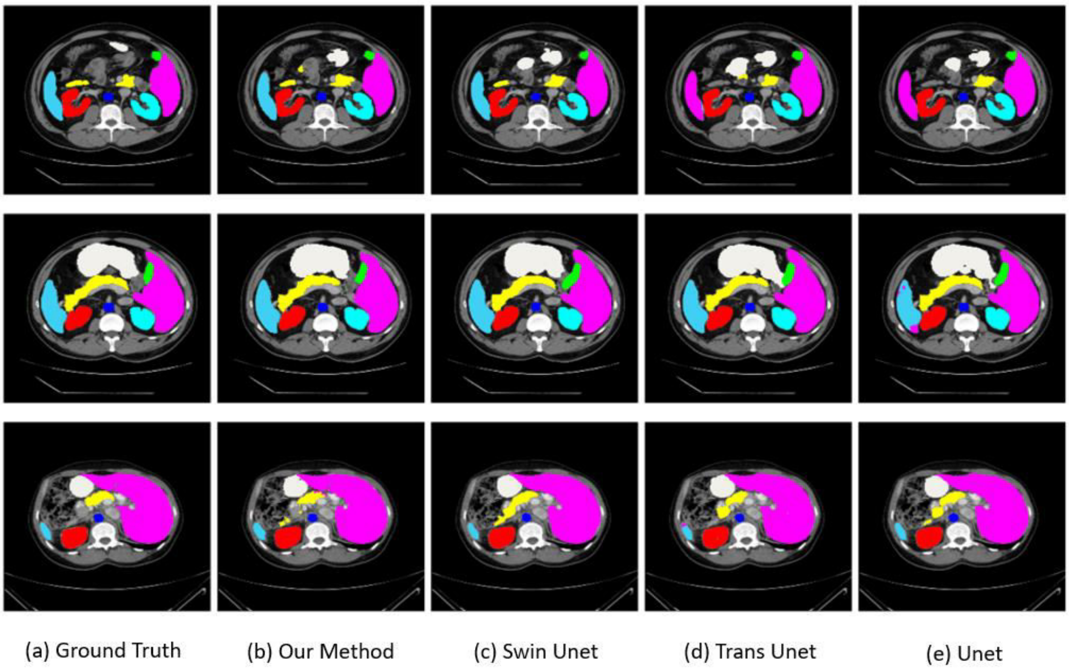


Fig.5 Result of our method on Synapse Dataset

Table 1 Result of our method on Synapse Dataset

Technology	Method	DSC	HD	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
Convolution	U-net	76.85	39.7	89.07	69.72	77.77	68.6	93.43	53.98	86.67	75.58
	V-net	68.81	-	75.34	51.87	77.1	80.75	87.84	40.05	80.56	56.98
	Att-UNet	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.3	75.75
Transformer	ViT	67.86	36.11	70.19	45.1	74.7	67.4	91.32	42	81.75	70.44
	R50 ViT	71.29	32.87	73.73	55.13	75.8	72.2	91.51	45.99	81.99	73.95
	TransUnet	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
	Swin Unet	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.6
Swin-PSA-SegNet	OurMethod	80.47	16.71	86.65	70.72	84.98	80.15	94.27	57.22	91.9	77.88

C. Ablation Study

To analyze the contribution of each component in the proposed framework, an ablation study was conducted by systematically adding PSA and SegNext to the baseline Swin-Unet model. The results, presented in Table 2, highlight the complementary benefits of each component.

The baseline Swin-Unet achieved a DSC of 79.13%. Adding PSA improved the model's segmentation accuracy to 79.5%, demonstrating the importance of advanced attention mechanisms for preserving spatial and channel-level details. Similarly, incorporating SegNext enhanced the DSC to 80.05%, reflecting the impact of efficient feature extraction. Combining both PSA and SegNext with Swin-Unet further improved the DSC to 80.47%, confirming that the integration of these components yields the best performance.

The ablation study illustrates that each module—PSA and SegNext—plays a critical role in enhancing segmentation accuracy and boundary delineation. Their synergy within the proposed framework ensures optimal performance, outperforming standalone models.

Table 2 Result of our method on Synapse Dataset

Method	Dice (%)
Swin Unet	79.13
Swin Unet + PSA	79.5
Swin Unet + SegNext	80.05
SwinUnet + PSA + SegNext	80.47

V. CONCLUSION

This paper proposed a novel framework integrating the Swin Transformer, Polarized Self-Attention (PSA), and SegNext for medical image segmentation. The framework demonstrated state-of-the-art performance on the Synapse dataset, achieving a Dice Similarity Coefficient (DSC) of 80.47% and a Hausdorff Distance (HD) of 16.71, outperforming both convolution-based and transformer based models.

The Swin Transformer effectively captured multi-scale features, PSA enhanced spatial and channel-wise details, and SegNext provided computational efficiency, making the framework both accurate and resource-friendly. The ablation study confirmed the complementary roles of these components, showcasing their combined impact on segmentation accuracy and boundary precision. The framework's robustness across diverse organ shapes and sizes, coupled with its computational efficiency, makes it suitable for real-world clinical applications. Future work will focus on optimizing inference speed, reducing memory usage, and extending the framework to other datasets and 3D segmentation tasks.

In summary, the proposed framework advances the field of medical image segmentation, offering a powerful and adaptable solution for clinical and research applications.

REFERENCES

[1] Oktay, O., et al., Attention u-net: Learning where to look for the pancreas. arXiv. arXiv preprint arXiv:1804.03999, 2018. **10**.

[2] Milletari, F., N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. in 2016 fourth international conference on 3D vision (3DV). 2016. Ieee.

[3] Antonelli, M., et al., The medical segmentation decathlon. Nature communications, 2022. **13**(1): p. 4128.

[4] Cao, H., et al. Swin-unet: Unet-like pure transformer for medical image segmentation. in European conference on computer vision. 2022. Springer.

[5] Zhou, Z., et al. 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, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. 2018. Springer.

[6] Alexey, D., An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929, 2020.

[7] Liu, H., et al., Polarized self-attention: Towards high-quality pixel-wise regression. arXiv preprint arXiv:2107.00782, 2021.

[8] Hu, J., L. Shen, and G. Sun. Squeeze-and-excitation networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[9] Bernard, O., et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging, 2018. **37**(11): p. 2514-2525.

[10] Ronneberger, O., P. Fischer, and T. Brox. 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. 2015. Springer.

[11] Çiçek, Ö., et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation. in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International

Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. 2016. Springer.

[12] Zhang, Z., Q. Liu, and Y. Wang, Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters, 2018. **15**(5): p. 749-753.

[13] Wang, Q., et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[14] Woo, S., et al. Cbam: Convolutional block attention module. in Proceedings of the European conference on computer vision (ECCV). 2018.

[15] Valanarasu, J.M.J., et al. Medical transformer: Gated axial-attention 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. 2021. Springer.

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

[17] Howard, A.G., Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[18] Tan, M. and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. in International conference on machine learning. 2019. PMLR.