Biomedical Signal Processing and Control 102 (2025) 107339

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

PSVT: Pyramid Shifted Window based Vision Transformer for cardiac image segmentation

Xingyu Zhang [a,b], Jiacheng Liu [a], Xiaoli Xian [c], Bo Chen [a], Dong Li [a], Fei Yang [a,*], Lei Zhang [d]

[a] School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, 264209, China
[b] School of Computer Science and Technology, Shandong University, Qingdao, 266237, China
[c] Department of Geriatrics, Weihai Municipal Hospital, Weihai, 264200, China
[d] Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland, Baltimore, MD 21201, USA

ARTICLE INFO

Keywords:
Cardiac image segmentation
Pyramid Swin Attention
Vision Transformer

ABSTRACT

Automating cardiac image analysis poses formidable challenges due to deformations during respiratory and cardiac cycles and the unique features of imaging modalities like CT and MRI. Transformer-based methods integrated with U-Net often miss essential local spatial details and incur high computational costs. To overcome these hurdles, we present the Pyramid Shifted-window-based Vision Transformer (PSVT), an innovative backbone for cardiac segmentation. Our model is meticulously designed to maintain a robust representation of both global and local features while operating on lower-dimensional inputs. By leveraging Continuous Position Bias in the Attention mechanism, we seamlessly integrate the attention module of Swin-Transformer-v2 with CNNs, yielding enhanced segmentation performance and computational efficiency. Additionally, we amplify receptive field understanding by incorporating Depth-Wise Convolution into the Feed-Forward module. Inspired by various adaptations of U-Net, we have restructured the Patch Merging and Patch Expanding modules, utilizing transposed convolution techniques to elegantly fuse multi-scale features into the classification head. Our experimental results compellingly demonstrate the superiority of PSVT, outperforming state-of-the-art cardiac segmentation models across three small-sample cardiac datasets: ACDC, MMWHS-CT, and LASC-2013. Our code is available in this URL.

1. Introduction

Cardiovascular diseases remain a significant global concern, with the World Health Organization reporting a staggering 19 million annual deaths worldwide [1]. Accurate segmentation of cardiac images is crucial for advancing clinical research and diagnosis, as it facilitates informed treatment decisions and minimizes misdiagnoses. Recently, the emergence of deep learning has revolutionized cardiac image segmentation, with applications spanning across various imaging modalities such as MRI, CT, and ultrasound [2]. For example, instance segmentation of diseased areas within different organs enhances surgeons' diagnostic capabilities through detailed visualizations. Moreover, deep neural networks support automated and efficient multi-class segmentation of cardiac images, underscoring the crucial role of deep learning in elevating the accuracy and efficiency of cardiac image segmentation [3].

Despite its potential, automating cardiac image analysis poses substantial challenges due to the inherent deformations during the respiratory and cardiac cycles. Meanwhile, different imaging modalities, such as CT and MRI, exhibit unique characteristics in brightness and intensity. Some efforts are made to supplement medical image datasets [4,5], but most of current cardiac segmentation datasets are still small-sample due to privacy and safety concerns. To overcome these difficulties, some unsupervised approaches [6–8] utilize cross-modal approaches to generate pseudo-labels, enriching datasets and improving model performance. Additionally, numerous supervised works leverage carefully designed neural networks for medical image segmentation, either in 2D [9–13] or 3D [14–19] medical images. Considering the heavy GPU memory demands of 3D segmentation, our research primarily focuses on the development of efficient 2D networks.

In 2D cardiac image segmentation, two mainstream architectures dominate: deep CNNs and Transformers. On the one hand, U-Net [20], often paired with CNNs backbone like ResNet [21], has become a fundamental model, delivering remarkable results in medical imaging tasks. On the other hand, the introduction of Transformers [22] has proposed new possibilities by capturing long-range dependencies

* Corresponding author.
E-mail addresses: 202315188@mail.sdu.edu.cn (X. Zhang), 202200800027@mail.sdu.edu.cn (J. Liu), xiaoli_xian@163.com (X. Xian), 202137544@mail.sdu.edu.cn (B. Chen), dongli@sdu.edu.cn (D. Li), feiyang@sdu.edu.cn (F. Yang), LeiZhang@som.umaryland.edu (L. Zhang).

https://doi.org/10.1016/j.bspc.2024.107339
Received 7 August 2024; Received in revised form 5 November 2024; Accepted 8 December 2024
Available online 9 January 2025

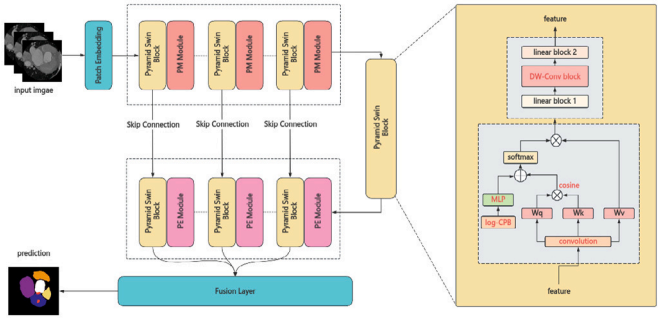

Fig. 1. Overview of Pyramid Shifted-window-based Vision Transformer (PSVT), the left shows the basic input flow and the right demonstrates Pyramid Swin Block (PSB), where improved part are colored with Red.

in cardiac images, with advanced models like ViT [23] and Swin-Transformer [24]. To leverage the strengths of both architectures, recent models such as Swin-Unet [9], PVT [10], and Swin UNETR [17] have attempted to integrate CNNs with Transformers. As one of the early efforts in this area, Swin-Unet combines Swin-Transformer with a linear-based up-sampling and down-sampling mechanism to enhance segmentation performance. However, it falls short by not fully utilizing CNNs' strengths in localized feature extraction and computational efficiency, resulting in suboptimal performance. More recent works, such as PVT [10] and nnFormer [14], integrate CNNs with ViT-based architectures to improve feature extraction. However, the computational burden of ViT remains significant, as the complexity of its attention mechanism grows quadratically with respect to the number of tokens.

To address these challenges, we introduce the Pyramid Sliding-window-based Vision Transformer (PSVT), a novel U-Net-based architecture that integrates traditional CNNs and Depth-Wise Convolution (DW-Conv) into swin-attention modules, as shown in Fig. 1. Our approach combines the strengths of CNNs and Transformers to enhance image comprehension while maintaining a low GPU memory footprint. Specifically, we propose Pyramid Swin Block (PSB), which employs the Swin-attention from Swin-Transformer-v2 [25] as the foundation attention mechanism, incorporating CNNs and DW-Conv to improve the model's ability to capture both global and local features. This combination allows us to significantly reduce computational complexity and memory usage when processing large feature maps. We also explore various paradigms for integrating CNNs and Transformers, prioritizing CNNs for up-sampling and down-sampling operations while using the attention mechanism to extract global features. We build a Patch Expanding module that utilizes various types of CNNs combined with multiple sampling methods to achieve the nonlinear representation of feature maps. Additionally, we introduce a multi-scale feature selection head to further refine the model's accuracy.

The contributions of our work are as follows:

1. We propose the PSVT model, which is a competitive backbone for cardiac image segmentation and surpasses a great number of previous models in different datasets, including ACDC, MMWHS-CT and LASC-2013. This achievement demonstrates the effectiveness and superiority of PSVT in cardiac image segmentation tasks.

2. We present the Pyramid Swin Block (PSB) as a highly effective and efficient feature extraction component. By integrating the attention mechanism from Swin-Transformer-v2 with convolutional blocks, and incorporating a DW-Conv layer within the Feed-Forward network, we enhance the model's ability to capture both local and global patch features comprehensively.

3. We investigate various forms of CNNs integration in the up-sampling and down-sampling processes, enhancing the Patch Merging and Patch Expanding modules in Swin-Unet with different CNNs and activation functions. This design provides a new perspective and solution for the field of cardiac image segmentation.

2. Related work

2.1. Methods based on CNNs

CNNs are dominating networks for processing computer vision tasks. Compared with models based on contours and traditional machine learning algorithms, CNNs demonstrate the ability that convolutional expressions can capture local spatial information in image pixels well. Previous studies demonstrate how to build a robust encoder architecture [21,26–28]. Researchers continuously optimize various aspects of the overall networks to enhance representation abilities. For example, ResNet combats gradient vanishing by designing skip connections, and others have extracted image-localized features by designing multiple forms of convolutional kernels like Atrous Deconvolution or Dilated Convolution.

CNNs are also widely used in medical image segmentation, where an extra decoder is required. U-Net has become the mainstream model for medical image segmentation due to the U-shaped design's simplicity and high performance. Therefore, its variants continue to emerge, such as [15,29,30]. They enrich local spatial features and further enhance segmentation performance by optimizing the network depth, the connectivity of skip connections, and the automatic search for the optimal parameters to obtain best results. Such architectural designs have also been applied to the 3D field, such as 3D-Unet [16] and V-net [31]. Meanwhile, a variety of up-sampling designs enrich feature representation of medical images, such as FCN [32], ASPP [33]. Recent works focus on the convolution kernel itself, such as the Depth-Wise Convolution kernel in PVT-v2 [11]. Although the attention mechanism has become mainstream in computer vision, the instability of training and large memory usage still make CNNs irreplaceable. Therefore, exploring favorable CNNs architectures and convolution kernel designs is still an significant research direction. At the same time, the introduction of the attention mechanism also provides new ideas and methods for optimizing CNNs.

2.2. Vision transformer and its variants

Inspired by the excellent performance of attention mechanisms in NLP, researchers work on transferring Transformers, especially self-attention, to computer vision (CV). There are various Transformers in CV, such as [10,23,24,34–36]. ViT, the foundation of these works, combines self-attention mechanism into visual models with the help of vision patches and achieves an impressive trade-off between speed and accuracy in many visual tasks. Compared with CNNs-based networks, the main disadvantages of ViT are that it requires large datasets when training, the expressive ability and computational memory cost of attention mechanisms are also in competition. Deit [37] proposes training strategies to alleviate these problems during training. Recently, Swin-Transformer have made improvements by implementing hierarchical scaling of attention with sliding windows, which enriches the expression of position relationships between different pixels. With the help of sliding windows, Swin-Transformer achieves linear computational complexity to image size. But the problem of huge computational memory usage when expanding the sliding windows in transformer still remain.

2.3. Combination of transformers and CNNs

UNet-based networks require the combined efforts of Transformer and CNNs, such as [9,12,38–41]. In previous works, Transformer is often used as a backbone and directly becomes a part of the encoder, such as ViT and Swin-Transformer. Their main work is to propose visual attention mechanism modules for feature extraction during down-sampling, and finally use a reasonable decoder head, such as FCN [32], for up-sampling as the output of image segmentation. However, in the U-Net network, the encoder and decoder modules often contain the

same size of feature extraction modules, which means that not only the use of Transformer-block in the backbone needs to be explored, but also the network structure during up-sampling and down-sampling require more designs.

Swin-Unet [9] mainly addresses the problem that CNNs cannot effectively learn global information and perform long-range semantic information interaction during image feature extraction. It is proposed to fully use Transformer and combine it with the U-Net architecture. In this paper, the authors choose Swin-Transformer-Block as the main feature extraction module for encoder and decoder modules, and use Linear layers between each module to learn the implicit expression of spatial image features during dimension transformation. PVT [10] focuses on reducing memory usage while maintaining high performance with Transformer. PVT uses a specific combination of CNNs modules for down-sampling, which allows the network to form a pyramid-style feature map during feature extraction. At the same time, PVT-v2 [11] proposes the linear complexity attention layer to reduce the computational complexity of PVT, making the entire attention module faster in the computation process of larger images. They also add DW-Conv to the MLP layer, which enriches the feature representation.

In contrast to the aforementioned approaches, Our work focuses on combining various forms of CNNs with attention mechanisms, leveraging the joint optimization of Swin-Transformer-v2 and DW-Conv to achieve visual feature perception at different granularities. We investigate the application of CNNs within the processes of up-sampling and down-sampling, striking a balance between computational demands and GPU memory consumption.

3. Method

3.1. Architecture overview

Our primary objective is to propose a backbone that seamlessly integrates the global and local feature extraction capabilities of Transformers and CNNs. In this study, we utilize Depth-Wise Convolution (DW-Conv) alongside Continuous Position Bias (CPB) to enhance the model's expressive power while ensuring improved training stability. Furthermore, we introduce the Pyramid Swin Block (PSB) to effectively capture features at multiple scales in cardiac images. Grounded in the U-Net framework, we explore straightforward yet effective Patch-Merging (PM) and Patch-Expanding (PE) modules, culminating in the overall architecture of our model, as depicted in Fig. 1.

3.2. Embedding

To input regular images as embedding vectors into the Transformer, we adopt the similar processing method as ViT. However, traditional position encoding such as absolute and relative position encoding fixes the position information directly at the patch embedding layer without caring about the variable information between image pixels. Therefore, we employ a learnable positional encoding bias, i.e., Continuous Position Bias demonstrated in Section 3.3, to enhance the positional feature representation in cardiac images. The patch embedding part is showncite Fig. 2.

Given a dataset $\{X, Y\}$, where X comprises the input images and Y consists of the corresponding semantic or binary segmentation labels, each image $x_i \in \mathbb{R}^{H\times W\times C}$ has spatial dimensions H and W, with $C = 3$ channels. Initially, the input image x_i is processed through the patch embedding module in RGB format. The image pixels are then segmented into non-overlapping tokens. These tokens are analogous to the basic inputs—words or subwords—formed by mapping a word through an embedding layer in NLP.

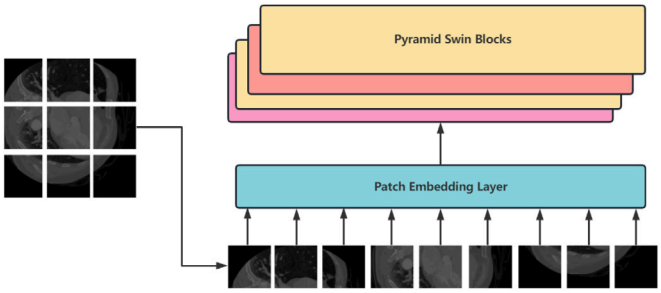

Fig. 2. Patch embedding layer in PSVT.

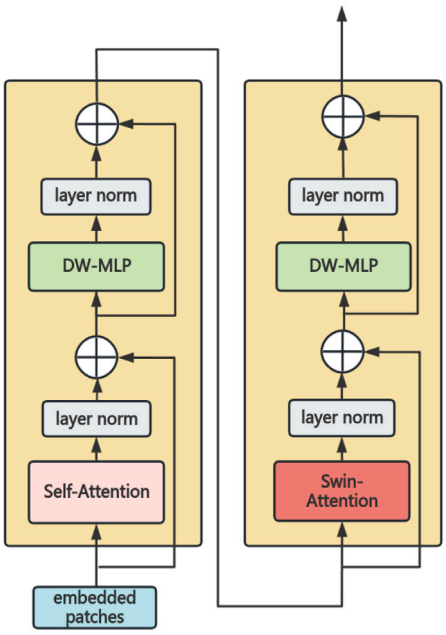

Fig. 3. Workflow of Pyramid Swin Block, include Multi-Head Self-Attention (left) and Pyramid Swin Attention (right).

3.3. Pyramid swin block

To seamlessly integrate CNNs and Swin-based attention mechanisms at the pixel level while enhancing the ability to capture both local and global representations, we propose the Pyramid Swin Block (PSB) consisting of two sub-blocks. The first sub-block is identical to the attention module in ViT, while the second sub-block applies Pyramid Swin Attention (PSA), an elaborately designed Swin Attention that incorporates a DW-Conv Block. In this process, we utilize a log-spaced Continuous Position Bias (CPB) method in Swin-transformer-v2 [25] to transfer spatial features between the underlying patches and to add learned spatial offsets. With the designed bias generation method, we also incorporate the Convolution Block prior to deriving queries, keys, and values in the attention mechanism for more fine-grained feature transformation. As shown in Fig. 3, input patches first undergo multi-head self-attention (MHSA) to perceive global features, after which they enter PSA for spatial feature extraction within the shifted window.

In more detail, MHSA is utilized in odd blocks, while attention based on the shifted window is employed in even blocks. The essence of shifted window attention aligns with the self-attention mechanism, but it employs CPB approach during the computation of queries, keys, and values. In attention layers, a larger attention window typically results in quadratic computational complexity and GPU memory footprint. Therefore, we employ PSA to reduce the size of the attention window,

typically setting the window size to 8. In self-attention mechanism, convolutional blocks are employed to query, key, and value, achieving channel-wise fusion perception with a constant number of channels. The ultimate attention formulations, as illustrated in Eq. (1), (2), and (3), effectively alleviate the computational burden by adjusting the input window size while simultaneously enhancing the model's ability to grasp spatial correlations among patches. For shifted-window attention, only corresponding positions can participate in the computation due to the bias generation mechanism of Swin. To address this, we leverage Swin's approach to bias generation, which compares mask positions, ensuring that the optimized shifted-window attention accurately encodes positional information. The impact of model parameters and FLOPs on computational efficiency is analyzed in Section 4.

$$Q = W_q Conv(q), K = W_k Conv(k), V = W_v Conv(v), \tag{1}$$

$$Attention(Q_i, K_i, V_i) = softmax(\frac{\cos(Q_i K_i^T)}{\tau} + B)V_i, \tag{2}$$

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W, \tag{3}$$

where Q, K, and V are input features, W is the feature transformation matrix, τ is a learnable parameters initialized in attention module, and B is the CPB with log scale, which also contains learnable blocks.

With respect to CPB, we primarily adopt the bias generation method from Swin-Transformer-v2 and integrate it into the PSB, shown in Fig. 1. First, we generate the relative position matrix for each head based on the patch resolution size of the query. Let M denote the resolution size of each head, with a range of values from $[-M+1, M-1]$. The relative position bias is parameterized as a bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M-1)}$, and the corresponding position is selected according to the position of the element to obtain the initial relative position matrix. To prevent the generated $\hat{B}$ from becoming excessively large, we employ a logarithmic function to scale the coordinates. This method requires smaller extrapolation ratios than using the original linearly spaced coordinates when transferring the relative position deviation across window resolutions. Subsequently, we use a small learnable meta-network to generate bias values for arbitrary relative coordinates, allowing for natural application in fine-tuning tasks where the window size varies. The equations are as follows.

$$\begin{aligned} \widehat{\Delta x} &= sign(x) \cdot \log(1 + |\Delta x|), \\ \widehat{\Delta y} &= sign(y) \cdot \log(1 + |\Delta y|). \end{aligned} \tag{4}$$

$$\begin{aligned} \hat{B}(\widehat{\Delta x}, \widehat{\Delta y}) &= \mathcal{G}(\widehat{\Delta x}, \widehat{\Delta y}), \\ B(x_i, y_i) &= \hat{B}(x_i, y_i), \end{aligned} \tag{5}$$

where $\Delta x, \Delta y, \widehat{\Delta x}, \widehat{\Delta y}$ are the linear-scaled and log-spaced coordinates, respectively. The formed coordinates are entered into the meta-network $\mathcal{G}$ to learn adaptive relative coordinates, after that we get the relative position bias in coordinates based on the corresponding x_i, y_i.

Regarding the PSB, each stage comprises two sub-blocks. The first sub-block performs self-attention transformation, followed by Layer Normalization and a forward network using Depth-Wise MLP (DW-MLP) for nonlinear feature transformation. After an additional Layer Normalization, the output is added to the original input to create a skip connection, facilitating the backward propagation of gradients. This result is then fed into the second sub-block for attention calculation based on the pyramid-shifted window. During this process, the intermediate variable undergoes regular patch transportation in Swin attention, resolution reduction, and the addition of CPB. These steps enable PSA to focus on the relationships between patches and establish the appropriate attention regions. The two processes alternate to perform the attention calculation for even blocks, as illustrated in Eq. (6).

$$\begin{aligned} z_a^l &= LN(MultiHead(z_m^{l-1})) + z_m^{l-1}, \\ z_m^l &= LN(DW - MLP(z_a^l)) + z_a^l, \\ z_a^{l+1} &= LN(SWMultiHead(z_m^l)) + z_m^l, \\ z_m^{l+1} &= LN(DW - MLP(z_a^{l+1})) + z_a^{l+1}, \end{aligned} \tag{6}$$

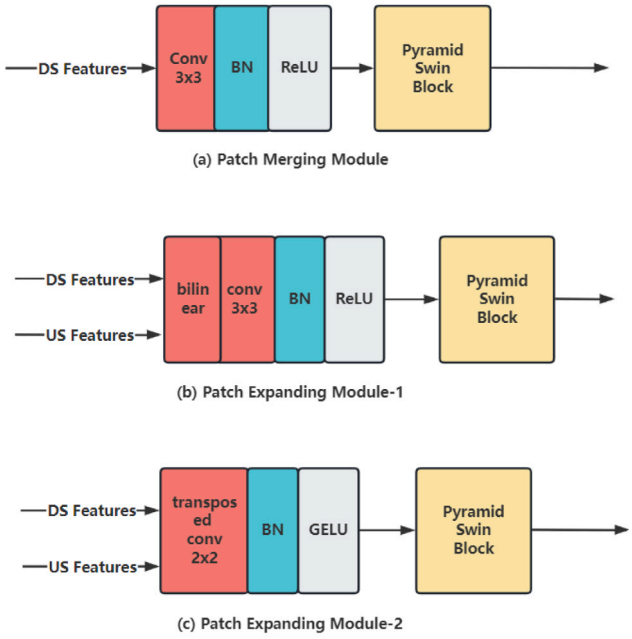

Fig. 4. Patching merging and patch expanding module.

where Z_m represents the final output from the DW-MLP layer, Z_a represents the intermediate output after MultiHead Attention, LN represents Layer Norm, and SWMultihead represents attention mechanism based on the shifted window in PSA.

3.4. Patch merging and patch expanding

To better emphasize positional relationships between patches and fully harness CNNs' strength in capturing local spatial details, we deliberately choose not to adopt conventional down-sampling methods based on linear transformations or hierarchical extraction, as implemented in Swin-Unet [9]. Instead, we introduced a simple yet carefully designed Patch Merging (PM) module. This module slides intermediate patches and computes features from adjacent positions, as illustrated in Fig. 4(a). Specifically, we draw inspiration from the down-sampling design in nnFormer [14] and opt to full convolutional networks for dimension transformation, which is called PM module. The advantage of PM module lies in its ability to utilize the complete spatial location information to improve the receptive field without adding too much computational burden. Since the input and output dimensions of the attention module are exactly the same, we often consider the PM module and PSB as a whole stage. As the resolution of the image is continuously halved and the number of channels is continuously increased, we obtain hierarchical feature maps like ResNet, U-Net and other networks, which can be conveniently used as additional inputs for feature merging during the up-sampling process by skip connections in the U-Net architecture.

For the up-sampling mechanism of patches, we adopt tailored strategies in our Patch Expanding (PE) module depending on the specific requirements of the scenario. Swin-Unet and nnFormer employ different approaches. Assuming a feature map dimension extracted by a transformer block, Swin-Unet, on the one hand, increases the number of channels using a linear layer, then divides these channels into four equal parts, thereby increasing the image size while reducing the number of channels. This can be expressed as $x_i \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C} \longrightarrow x_i \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C}$. On the other hand, nnFormer directly employs a transposed convolutional network to enhance image resolution. For the ACDC dataset, which is a 3-class segmentation problem, we utilize

patch expanding mode 1, as illustrated in Fig. 4(b). We employ bilinear interpolation for patch expansion while using a convolutional network (kernel size $k = 3$, stride $s = 2$, padding $p = 1$) to reduce the feature dimensions corresponding to the encoder part. For the MMWHS-2017 dataset, which involves a 7-class classification problem, we utilize patch expanding mode 2, as shown in Fig. 4(c). Since bilinear interpolation limits feature representation, we employ transposed convolution with GELU activation [42]. The convolutional kernel has a size of 2 and a stride of 2. Meanwhile, a subsequent convolutional network is used to further reduce the number of channels back to the original amount, which is then input into the Pyramid Swin Block.

In the final stage of the PE module, we primarily use a combination of patch expanding and convolution to focus on spatial features. We then combine the output features from each stage to form a hierarchical feature map. Finally, the original patch is restored through a similar image transformation approach, supplemented by a classification head, allowing for pixel-wise classification of each pixel to obtain the final segmentation result.

4. Analysis of parameters and FLOPs

Let us analyze how different Transformer models, particularly their attention mechanisms, influence computational complexity in terms of parameter counts and FLOPs. We assume the input image has dimensions H, W, and C, with a batch size and the number of heads in the attention mechanism both equal to 1.

4.1. Standard vit

For the Vision Transformer (ViT), the image is divided into patches of size 16×16, resulting in a patch representation of size $L = \frac{H}{16} \times \frac{W}{16}$.

Parameter Count: In ViT, the majority of parameters are attributed to the attention (Attn) and feed-forward network (FFN). For the attention mechanism, the query, key, and value undergo transformations via linear layers, resulting in intermediate variables Q, K, and V. Assuming the linear layers maintain their dimensions, the required parameters are $3N^2$. The attention calculation proceeds with $P = QK^T$, followed by masking, softmax, and $O = PV$, which does not introduce additional parameters. The normalization of O involves parameters N. Thus, the parameter count for attention is $3N^2 + N$ with a complexity of $O(N^2)$. For the FFN, with two linear layers expanding the dimensionality by a factor of 4, the parameter count is $4N^2 + N$. Therefore, the complexity of ViT is $O(N^2)$.

FLOPs: In attention, the query, key, and value each have shapes of $(1, L, N)$. After passing through linear layers, the required FLOPs for Q, K, and V is $3LN^2$. The FLOPs for calculating $P = QK^T$ is L^2N, followed by masking and softmax which require L^2. For the calculation of $O = PV$, the FLOPs is L^2N, and the normalization contributes LN. Therefore, the overall FLOPs for attention is approximately $3LN^2 + 2L^2N + L^2 + LN$. Notably, if $L \gg N$, the complexity becomes $O(L^2)$; if $N \gg L$, the complexity is $O(N^2)$. As the image resolution increases, L grows quadratically, leading to a significant increase in GPU memory requirements, which limits the balance between image resolution and GPU resources. For the FFN, the FLOPs required for the two linear layers is $2LN^2$, and the activation function and normalization require LN. Therefore, the total FLOPs is $8LN^2+LN$. If $L \gg N$, the complexity is $O(L)$; if $N \gg L$, the complexity is $O(N^2)$.

4.2. PSVT

In the Pyramid Shifted-window-based Vision Transformer (PSVT), we utilize the Pyramid Swin Block (PSB) as the core feature extraction block, which consists of the Pyramid Swin Attention (PSA) and an improved Feed-Forward Network. The PSA employs the Swin Transformer-v2 as the base model while integrating CNNs as a local feature optimizer. The FFN incorporates Depth-Wise Convolution (DW-Conv) for depth-wise feature optimization, resulting in different parameter counts and FLOPs compared to ViT.

Table 1
Comparison of Parameter Count and FLOPs for ViT and PSVT.

Model	Params	FLOPs	Complexity
ViT	$3N^2 + N$ (Attn)	$3LN^2 + 2L^2N + L^2 + LN$ (Attn)	$O(N^2)$ (Params)
	$4N^2 + N$ (FFN)	$8LN^2 + LN$ (FFN)	$O(L^2)$ (FLOPs)
PSVT (ours)	$12N^2 + N + C_1$ (Attn)	$9LN^2 + 2LNL_1 + L_1^2 + LN$ (Attn)	$O(N^2)$ (Params)
	$8N^2 + 13N$ (FFN)	$12LN$ (FFN)	$O(LL_1)$ (FLOPs)

Parameter Count: Similar to ViT, the parameter count for attention in PSVT primarily depends on N. Besides the parameters discussed in ViT, attention also introduces preceding CNNs (kernel size = 3) for local feature extraction, which maintains the feature dimension N, contributing $9N^2$ to the parameter count. The Swin Transformer-v2 introduces Continuous Positional Bias (CPB) that utilizes a meta-network for dynamic position encoding, transforming the 2D coordinates into a 512-dimensional space before mapping back. The fixed parameters required for this step is $C_1 = 2\times2\times512$. Thus, the total parameter count for attention is $12N^2+N+C_1$ with a complexity of $O(N^2)$. For the FFN, we add DW-Conv (kernel size = 3) between the two linear layers, which requires $12N$. Therefore, the total parameter count is $8N^2 + 13N$ with a complexity of $O(N^2)$.

FLOPs: A key distinction between Swin and ViT is the introduction of the window size concept, which alters the dimension L. Assuming the input dimensions are $(1, L, N)$ or $(1, H, W, N)$, the initial convolution (kernel size = 3) corresponds to FLOPs of $9N^2 \times H \times W = 9N^2L$. The additional FLOPs for CPB are $1 \times 2 \times 512 + 1 \times 512 \times 2$. Assuming a window size of 8, the transformed features change from $(1, H, W, N)$ to $(\frac{H}{ws} \times \frac{W}{ws}, ws \times ws, N)$, which we denote as (B_1, L_1, N). Consequently, the calculation of $P = QK^T$ requires FLOPs of $B_1L_1NL_1 = LNL_1$, followed by the calculation of the mask and softmax requiring L_1^2. The calculation of $O = PV$ requires FLOPs of $B_1L_1L_1N = LL_1N$, and the normalization contributes $B_1L_1N = LN$. Thus, the FLOPS of attention is $9N^2L + 2LNL_1 + L_1^2 + LN$. While the FLOPs formula for PSVT resembles that of ViT, the reduction in L to $L1$ (typically a smaller value) significantly decreases overall computations and GPU memory footprint. For the FFN, in addition to the FLOPs required for the two linear layers, we include the contribution from DW-Conv, resulting in an additional $12B_1L_1N = 12LN$.

Overall, we present the comparative results in Table 1. The results indicate that PSVT reduces the computational complexity from $O(L^2)$ to $O(L_1L)$ while maintaining the same parameter count as ViT. This reduction in complexity is a key factor contributing to the overall decrease in GPU memory footprint and computational load.

5. Experiment

5.1. ACDC

5.1.1. Dataset

ACDC [43] consists of cardiac data from 100 patients, including 20 healthy patients, 20 patients with a history of myocardial infarction, 20 patients with dilated cardiomyopathy, 20 patients with hypertrophic obstructive cardiomyopathy, and 20 patients with right ventricular abnormalities. The data are obtained through MRI scans and provides ground truth labels for the cardiac, including the left ventricle (LV), right ventricle (RV), and myocardium (Myo), as shown in Fig. 5.

To alleviate the problem of small sample sizes in medical data and make full use of cardiac data, this paper performs data augmentation during training, including image scaling, random cropping, random rotation, and photometric distortion. The optimizer used in this paper is SGD, with an initial learning rate of 1e-3. Using a polynomial learning rate decay, the learning rate gradually decreases to 1e-4. The paper trains a total of 25,000 steps with a batch size of 16 on a 3090 nvidia

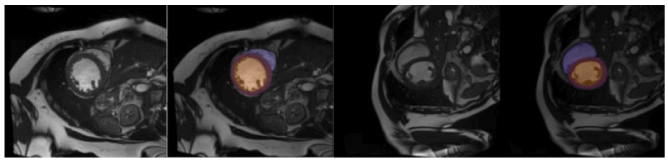

Fig. 5. Two representative images and labels in ACDC Dataset.

GPU provided in the laboratory and saved the final weights. This paper uses Dice and IoU as evaluation metrics, with Eq. (7) and (8).

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{7}$$

$$IoU = \frac{|X \cap Y|}{|X \cup Y|}, \tag{8}$$

where X denotes the region of predicted pixels for each class and Y denotes the ground truth of pixels respectively.

The cross-entropy loss function is used to guide the model to perform correct classification, but for image semantic segmentation problems, it cannot well represent the continuity of the same semantics. Therefore, we also introduced the Dice loss function, which aims to make the parts with the same semantics more continuous while clearly separating the boundaries of different semantics. In this process, the paper defines the following loss function, as shown in Eq. (9).

$$\begin{aligned} L_{class} &= -\sum_{i=0}^{C-1} w_i y_i \log(p_i), \\ L_{Dice} &= \sum_{i=0}^{C-1} 1 - Dice_i, \\ L &= \lambda_c \cdot L_{class} + \lambda_d \cdot L_{Dice}, \end{aligned} \tag{9}$$

where p_i and y_i denote the predicted probability and the true category, respectively, w_i denotes the weight of the category, $Dice_i$ denote the Dice loss of the category, and λ_c, λ_d denote the weights of cross-entropy and Dice loss, respectively.

5.1.2. Results

Table 2 and 3 present a comparative analysis of advanced deep learning models on the ACDC dataset, focusing on the segmentation of key cardiac structures: left ventricle (LV), myocardium (Myo), and right ventricle (RV). We compare PSVT with recent mainstream models in both 2D and 3D (marked with *). PSVT-transpose-8x model demonstrates outstanding 2D cardiac image segmentation performance, achieving an average Dice Similarity Coefficient (DSC) of 92.08% and an Intersection over Union (IoU) of 84.23%, closely competing with top 3D models. We also present the number of parameters and GFLOPs for each model in Table 7.

In LV and RV segmentation, PSVT-transpose-8x demonstrates a clear advantage over other 2D models, achieving up to 2.99% higher DSC and 1.06% higher IoU compared to models like SSTrans-Net. These results are also comparable to 3D models like neU-Net, which have slightly higher scores but require significantly more GPU memory footprint (161.28M and 129.83G). For Myo segmentation, PSVT-transpose-8x attains a DSC of 88.53% and IoU of 75.53%. This positions it favorably against models like SSTrans-Net (DSC 90.31%) and Swin-Unet (DSC 86.99%). Compared to 3D models like D-LKA Net and TC-CoNet, our PSVT-transpose-8x offers a more resource-efficient alternative, maintaining high accuracy while consuming less memory. This makes it a practical choice for deployment in resource-limited scenarios.

It is worth clarifying that the 8x and 32x parametric quantities of PSVT are not incrementally related; instead, by modifying the attention window in PSA, they only have a difference in computational volume, i.e., GPU memory footprint. Although the performance of the 32x model appears more favorable compared to the 8x models, subsequent results reveal that the 8x model exhibits superior performance in certain scenarios. We attribute this phenomenon to a similar effect observed with different-sized convolutional kernels on feature-awareness ability. And the results compared with other models can be found in Fig. 9.

Table 2
mDice(%) on the ACDC dataset (see [9,11,12,14,38,44–48]).

Model	RV	Myo	LV	mDice↑
PVT-v2-b2 [11]	87.59	84.44	93.13	88.39
Swin-Unet-pretrain [9]	86.59	81.45	92.92	86.99
TransUNet [12]	88.86	84.54	**95.73**	89.71
SSTrans-Net [44]	**89.02**	**87.51**	94.40	**90.31**
UNETR* [38]	85.29	86.52	94.02	88.61
D-LKA Net* [45]	91.35	88.04	94.68	91.36
MOSformer* [46]	89.43	89.09	95.77	91.43
TC-CoNet* [47]	90.27	88.98	95.47	91.58
nnFormer* [14]	90.04	89.58	95.65	92.06
neU-Net* [48]	**90.75**	**89.91**	**95.66**	**92.11**
PSVT-8x(ours)	86.47	84.52	93.1	88.04
PSVT-32x(ours)	91.01	87.79	94.76	91.19
PSVT-transpose-8x(ours)	**92.01**	**88.53**	**95.46**	**92.08**

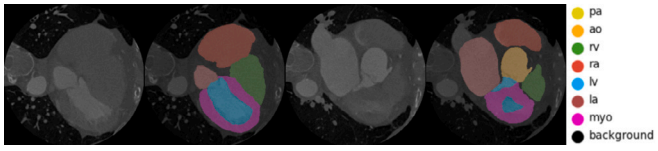

Fig. 6. Image and Label in MMWHS-2017.

5.2. MMWHS-2017

5.2.1. Dataset

The second dataset is an open-source dataset provided by the MWHS-2017 [49] Whole Heart Segmentation Challenge. To verify the robustness of the model on different medical scan formats, we mainly use CT scan images from 20 patients in the dataset, called MMWHS-CT. The dataset provides 7 labels, including the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium (Myo), pulmonary artery (PA), and aorta (AO), as shown in Fig. 6.

In addition to standard data augmentation for the ACDC dataset, we apply intensified augmentation specifically to core cardiac slices to address the increase in segmentation categories from 3 to 7. In the training set, since most of the cardiac data are background class, with a segmentation effect of up to 99% in the experiment, we perform data augmentation on other meaningful categories in the images. In this process, we use vertical random flipping, random rotation of 90 degrees, CLAHE (limited contrast adaptive histogram equalization), random brightness changes, and the addition of random noise, as shown in Fig. 7. During the training process, we use Adam as the optimizer with an initial learning rate of 1e-4, then it gradually decreases to 1e-5 through polynomial learning rate decay. During training, there are 25,000 steps with a batch size of 4 and linear warm-up in the first 1500 steps, where the learning rate increased from 1e-6 to 1e-5. Compared with the ACDC experiment, the loss function used in this paper do not undergo any essential changes, as shown in Eq. (9). The cross-entropy loss function is balanced based on the number of categories to obtain better results.

5.2.2. Results

Table 4 and Table 5 showcase the impressive capabilities of our PSVT-transpose-8x model on the MMWHS-2017 dataset, particularly in the segmentation of vital cardiac structures. We compare PSVT

Table 3
mIoU(%) on the ACDC dataset (see [11,12,14,24,38,44–48]).

Model	RV	Myo	LV	mIoU↑
PVT-v2-b2 [11]	77.92	73.07	87.14	79.2
Swin-Unet-pretrain [24]	76.35	68.71	86.78	76.98
TransUNet [12]	79.95	73.22	**91.81**	81.34
SSTrans-Net [44]	**80.21**	**77.79**	89.39	**82.33**
UNETR* [38]	74.35	76.24	88.71	79.55
D-LKA Net* [45]	84.08	78.64	89.9	84.09
MOSformer* [46]	80.88	80.33	91.88	84.21
TC-CoNet* [47]	82.27	80.15	91.33	84.47
nnFormer* [14]	81.88	81.13	91.66	85.29
neU-Net* [48]	**83.07**	**81.67**	**91.68**	**85.37**
PSVT-8x(ours)	76.16	73.19	87.09	78.64
PSVT-32x(ours)	83.5	78.24	90.04	83.81
PSVT-transpose-8x(ours)	**85.2**	**79.42**	**91.31**	**85.32**

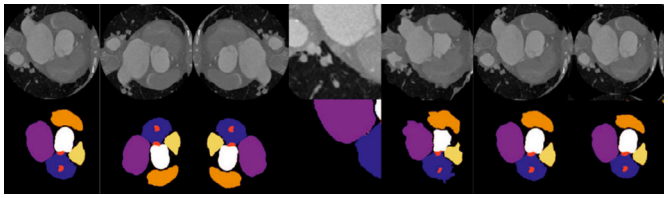

Fig. 7. Augmentation in MMWHS-2017.

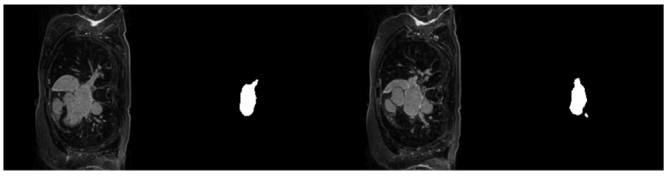

Fig. 8. Image and label in LASC-2013.

with recent mainstream models in both 2D and 3D (marked with *). With an average DSC of 91.17%, PSVT-transpose-8x emerges as a frontrunner in 2D and 3D segmentation models like UNETR and CoTr. PSVT effectively balance high performance and computational efficiency (46.38M and 15.24G shown in Table 7). Compared with 2D models, PSVT-transpose-8x outperforms the baseline GUT model up to 1.2% DSC, as well as other competitive models such as Swin-Unet and U-Net, showcasing its potential for practical deployment in resource-constrained environments.

PSVT-transpose-8x achieves impressive DSCs of 94.67% for the left ventricle (LV) and 89.94% for the right ventricle (RV). Moreover, the model also performs admirably in segmenting the aorta (AO) and pulmonary artery (PA), with DSCs of 96.00% and 85.04%, respectively. These results position PSVT-transpose-8x favorably against both traditional convolutional and recent transformer-based models, which often struggle with precise delineation in these regions. In challenging regions like the myocardium (Myo), PSVT-transpose-8x achieves a DSC of 88.46%. While this performance is competitive, it does fall short when compared to some 3D models like UNETR and CoTr, which achieves a Myo DSC of 95.0% and 94.4%. This gap highlights a limitation of our model in the context of more complex structures, suggesting that while our 2D architecture performs well, it may struggle in capturing the intricate spatial relationships inherent in 3D data.

Overall, the PSVT-transpose-8x model not only excels in segmentation accuracy but also offers significant advantages in computational efficiency. This combination positions it as an optimal choice for medical imaging applications, showcasing the potential of thoughtful model design to yield outstanding results while minimizing resource requirements. The results compared with other models can be found in Fig. 10.

5.3. LASC-2013

5.3.1. Dataset

In order to measure the model's performance on different tasks, we also test PSVT on instance segmentation. The Left Atrial Segmentation Challenge (LASC) [55] was held in 2013 at STACOM'13 and MICCAI'13, including 30 CT and 30 MRI data, with 10 training data and 20 testing data for both CT and MRI. LASC'13 was part of the STACOM'13 workshop, held in conjunction with MICCAI'13. Seven international research groups, comprising 11 algorithms, participated in the challenge. We follow the division rules of the challenge itself, using 10 MRI samples as the training set and the remaining 20 as the test set, images and labels are shown in Fig. 8.

Considering the simplicity of instance segmentation, we only use the default data augmentation method in the MMSegmentation framework, and unify the data to a size of 256×256 for network model training and testing. During the training process, we use AdamW as the optimizer, and used 1e-4 as the initial learning rate. We dynamically adjust the learning rate through a polynomial schedule. Meanwhile, we use the actual labeled exact slices for training as additional data, other than that we do no other data augmentation.

5.3.2. Results

We also demonstrate the results on the LASC-2013 dataset, where all models are compared under the same training environment and settings. Detailed results are presented in Table 6. Our findings indicate that the improved PSVT achieved competitive results (DSC 91.08% and IoU 83.62%) compared to most network models. With the same training setup, we outperform others on both evaluation metrics, despite not surpassing U-Net in accuracy. We attribute this to the nature of the LAseg-2013 instance segmentation problem; the relatively simple U-Net architecture tends to perform well with fewer datasets due to its capability to handle a wide range of truth values. Furthermore, accuracy alone is not a perfect measure for medical image segmentation, which allows our model to remain competitive.

5.4. Ablation

In this section, we first compare the FLOPs and parameter sizes of various models, as illustrated in Table 7. Based on empirical data, we observe that an increase in model parameters corresponds to a certain degree of improvement in performance. For instance, both PSVT-8x and PSVT-32x outperformed Swin-Unet and PVT on the ACDC dataset. However, while increasing the parameter size, we must also implement specific enhancements to the model itself. Notably, PSVT achieved

Table 4
mDice(%) on the MMWHS-CT dataset (see [9,11,20,38,50–54]).

Model	Myo	LA	LV	RA	RV	AO	PA	mDice↑
CUHK1 [50]	85.1	91.6	90.4	83.6	88.3	90.7	78.4	86.87
KTH [51]	85.6	93.0	92.3	87.1	85.7	89.4	83.5	88.09
GUT	88.1	92.9	91.8	88.8	90.9	93.3	84.0	89.97
PVT-v2-b3 [11]	76.46	82.14	86.97	74.24	74.89	85.85	75.51	79.43
V-net	79.1	85.3	81.3	**90.9**	81.6	76.3	71.7	80.89
CFUN	82.2	83.2	87.9	90.2	84.4	82.1	**94.0**	85.90
SEG-CNN	87.2	91.0	92.4	87.9	86.5	83.7	91.3	88.96
Swin-Unet [9]	85.62	-	**95.83**	-	88.55	-	-	90.00
U-net [20]	**88.47**	**91.76**	94.35	90.63	**89.61**	**97.08**	83.58	**90.78**
MAUNet* [52]	89.3	**91.0**	92.5	**92.8**	88.6	92.5	**86.6**	90.63
CoTr* [53]	94.4	89.6	93.4	89.8	90.8	95.3	84.5	91.1
Cui* [54]	–	–	–	–	–	–	–	91.1
UNETR* [38]	**95.0**	88.8	**93.9**	90.3	**91.9**	**96.4**	84.3	**91.5**
PSVT-32x(ours)	82.71	90.08	93.00	88.7	85.6	94.96	**93.27**	88.33
PSVT-transpose-8x(ours)	**88.46**	**92.25**	**94.67**	**91.14**	**89.94**	**96.00**	85,04	**91.17**

Table 5
mIoU(%) on the MMWHS-CT dataset (see [9,11,20,38,50–54]).

Model	Myo	LA	LV	RA	RV	AO	PA	mIoU↑
CUHK1 [50]	74.06	84.50	82.48	71.82	79.05	82.98	64.47	76.79
KTH [51]	74.83	86.92	85.70	77.15	74.98	80.83	71.67	78.72
GUT	78.73	86.74	84.84	79.86	83.32	87.44	72.41	81.77
PVT-v2-b3 [11]	61.89	69.69	76.94	59.03	59.86	75.21	60.66	65.88
V-net	65.43	74.37	68.49	**83.32**	68.92	61.68	55.88	67.91
CFUN	69.78	71.23	78.41	82.15	73.01	69.64	**88.68**	75.28
SEG-CNN	77.30	83.49	85.87	78.41	76.21	71.97	83.99	80.12
Swin-Unet [9]	74.86	–	**91.99**	–	79.45	–	–	81.82
U-net [20]	**79.32**	**84.77**	89.30	82.87	**81.18**	**94.33**	71.79	**83.12**
MAUNet* [52]	80.67	**83.49**	86.05	**86.57**	79.53	86.05	**76.37**	82.87
CoTr* [53]	89.39	81.16	87.62	81.49	83.15	91.02	73.16	83.65
Cui* [54]	–	–	–	–	–	–	–	84.33
UNETR* [38]	**90.48**	79.86	**88.50**	82.32	**85.01**	**93.05**	72.86	**84.33**
PSVT-32x(ours)	70.52	81.95	86.92	79.69	74.83	90.40	**87.39**	79.10
PSVT-transpose-8x(ours)	**79.25**	**85.64**	**89.9**	**83.72**	**81.71**	**92.31**	74.02	**83.79**

Table 6
Dice(%) and IoU(%) on the LAseg-2013 dataset.

Model	Dice↑	Acc↑	IoU↑
U-net	89.20	**95.91**	80.50
Swin-Unet-v2	89.14	95.1	80.40
PSVT-transpose-8x(ours)	**91.08**	95.22	**83.62**

Table 7
GFLOPs and Parameter.

Model	GFLOPs	Params(M)
Swin-Unet	5.9	27.14
PSVT-transpose-8x(ours)	**15.24**	**46.38**
PVT-v2-b2	19.91	42.56
PVT-v2-b3	23.57	62.44
TransUNet*	24.63	88.87
D-LKA Net*	42.41	39.95
nnFormer*	47.73	37.16
UNETR*	77.84	92.69
MOSformer*	100.96	77.09
neU-Net*	129.83	161.28
TC-CoNet*	145.16	86.79

better results than PVT-v2-b3 with a similar number of parameters. Our model has less than half the GFLOPs compared to nnFormer, while our performance on the ACDC dataset is better than nnFormer. Moreover, while models like D-LKA Net and nnFormer demonstrate significant parameter sizes and GFLOPs, they achieve performance gains that may not be proportionate to their computational cost. This reflects the differences in performance and accuracy preferences between 2D and 3D medical image segmentation models.

We also conduct partial ablation experiments on the PSVT network model. For the model without convolutional attention networks, we explore experimental combinations of the overall network structure. After combining convolution and attention, we perform experiments with the size of the attention window. The experimental results show that the addition of convolution greatly reduces the memory required for attention and improved the calculation speed. This allows us to use a larger attention window for calculation, thereby improving the model's segmentation performance, as shown in Table 8. We use ˆ and * to denote Patch Expanding modules 1 and 2, respectively.

We find that replacing linear scaling by up-sampled network structure (e.g., bilinear+conv), expanding the attention window by adding

Table 8
Ablation experiments.

Model	PE	FFN	PSA	RV	Myo	LV	mDice↑
No-upsample-8xˆ		✓		86.53	83.72	92.60	87.62
No-DW-8xˆ	✓			87.79	84.74	93.47	88.67
8xˆ	✓	✓		88.27	85.21	93.50	89.00
16xˆ	✓	✓		88.63	84.98	94.08	89.23
8x-pretrainˆ	✓	✓		90.34	86.57	94.51	90.47
8x*	✓	✓	✓	86.47	84.52	93.13	88.04
32x*	✓	✓	✓	**91.01**	**87.79**	**94.76**	**91.19**

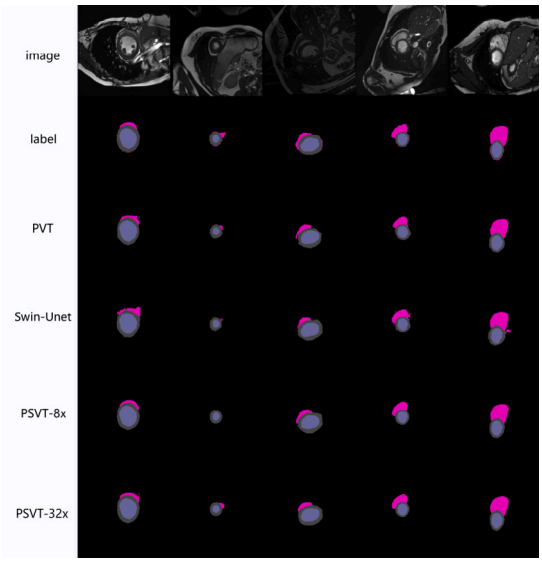

Fig. 9. Demonstration and comparison of the prediction results of different models for the acdc dataset.

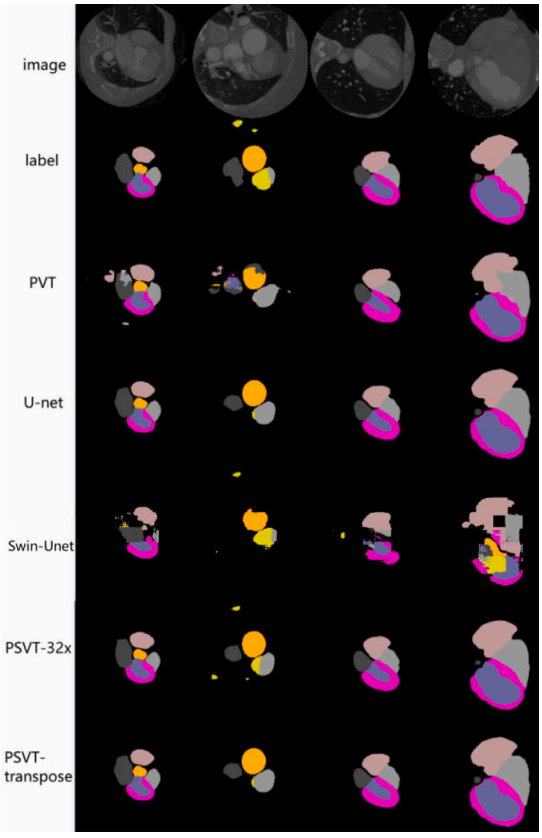

Fig. 10. Demonstration and comparison of the prediction results of different models for the MMWHS-CT dataset.

DW-Conv to the MLP layer of attention, and initializing attention parameters by using a pretrained model contribute to the segmentation performance. Additionally, combining convolution and attention can significantly improve model results while reducing training memory, even surpassing the results obtained by initializing with a pretrained model.

6. Discussion

Although PSVT demonstrates excellent performance compared to both 2D and 3D image segmentation models, we believe there is still room for improvement. In future work, we plan to incorporate self-supervised learning into the PSVT framework, adopting a contrastive learning approach. This method not only effectively addresses the challenges posed by the scarcity of medical training data but also significantly enhances performance on 3D segmentation tasks. Meanwhile, we aim to integrate advanced hierarchical attention mechanisms to more accurately capture the detailed information of target structures, thereby improving segmentation accuracy and boundary precision. Additionally, we plan to expand our research to include more classic 2D and 3D medical segmentation datasets to validate the generalization capabilities of our proposed methods.

7. Conclusion

In this paper, we introduce PSVT, a competitive backbone for cardiac image segmentation that explores the integration of CNNs and Transformer architectures. We also present the Pyramid Swin Block, which combines Depth-Wise Convolution and Swin Attention, effectively harnessing their strengths to capture both local and global spatial features. Our restructured Patch Merging and Patch Expanding modules, inspired by various U-Net adaptations, facilitate the seamless fusion of multi-scale features. Notably, our model outperforms 2D cardiac segmentation models on the ACDC, MMWHS-CT, and LASC-2013 datasets, and it stands out among mainstream 3D models due to its lower parameter counts and computational costs.

CRediT authorship contribution statement

Xingyu Zhang: Visualization, Validation, Software, Resources, Methodology, Investigation. **Jiacheng Liu:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xiaoli Xian:** Writing – review & editing, Supervision, Software, Resources, Funding acquisition, Conceptualization. **Bo Chen:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation. **Dong Li:** Supervision, Software, Resources, Formal analysis, Data curation. **Fei Yang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition. **Lei Zhang:** Supervision, Project administration, Investigation, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Fei Yang reports financial support was provided by the Natural Science Foundation of Shandong Province. Fei Yang reports financial support was provided by the National Natural Science Foundation of China. Fei Yang reports financial support was provided by Guangdong Basic and Applied Basic Research Foundation. Fei Yang reports equipment, drugs, or supplies was provided by NVIDIA Corp. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the Natural Science Foundation of Shandong Province, China (No. ZR2019MF011). This work was also supported by the National Natural Science Foundation of China (No. 62376136 and No. 62076149), Guangdong Basic and Applied Basic Research Foundation, China (No. 2024A1515011935) . We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V used for this research.

Data availability

I have disclosed the source of the code and dataset.

References

[1] S. Mendis, P. Puska, B.e. Norrving, W.H. Organization, et al., Global Atlas on Cardiovascular Disease Prevention and Control, World Health Organization, 2011.

[2] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, D. Rueckert, Deep learning for cardiac image segmentation: a review, Front. Cardiovascul. Med. 7 (2020) 25.

[3] J. El-Taraboulsi, C.P. Cabrera, C. Roney, N. Aung, Deep neural network architectures for cardiac image segmentation, Artif. Intell. Life Sci. 4 (2023) 100083.

[4] A. Zeng, C. Wu, G. Lin, W. Xie, J. Hong, M. Huang, J. Zhuang, S. Bi, D. Pan, N. Ullah, et al., Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images, Comput. Med. Imaging Graph. 109 (2023) 102287.

[5] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nature Commun. 15 (1) (2024) 654.

[6] S. Li, S. Zhao, Y. Zhang, J. Hong, W. Chen, Source-free unsupervised adaptive segmentation for knee joint MRI, Biomed. Signal Process. Control 92 (2024) 106028.

[7] J. Hong, Y.-D. Zhang, W. Chen, Source-free unsupervised domain adaptation for cross-modality abdominal multi-organ segmentation, Knowl.-Based Syst. 250 (2022) 109155.

[8] J. Hong, S.C.-H. Yu, W. Chen, Unsupervised domain adaptation for cross-modality liver segmentation via joint adversarial learning and self-learning, Appl. Soft Comput. 121 (2022) 108729.

[9] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.

[10] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.

[11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, Comput. Vis. Media 8 (3) (2022) 415–424.

[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[13] X. Zhou, D. Tan, Y. Su, C. Zheng, 2-D general network based on channel-space attention for medical image segmentation, in: 2023 4th International Conference on Computer, Big Data and Artificial Intelligence, ICCBD+AI, 2023, pp. 1–5, URL: https://api.semanticscholar.org/CorpusID:270395486.

[14] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, Nnformer: Interleaved transformer for volumetric segmentation, 2021, arXiv preprint arXiv:2109.03201.

[15] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P.F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, et al., nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018, arXiv preprint arXiv:1809.10486.

[16] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 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, Springer, 2016, pp. 424–432.

[17] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 272–284.

[18] X. Fu, Z. Sun, H. Tang, E.M. Zou, H. Huang, Y. Wang, L. Zhan, 3D bi-directional transformer U-Net for medical image segmentation, 5, 2023, URL: https://api.semanticscholar.org/CorpusID:255497673,

[19] J. Tian, B. Chen, Q. Li, R. Liu, Y. Feng, An efficient spatial modeling Conv-ViT using mask supervision for 3D medical image segmentation, in: 2024 International Joint Conference on Neural Networks, IJCNN, 2024, pp. 1–8, URL: https://api.semanticscholar.org/CorpusID:272538713.

[20] O. Ronneberger, P. Fischer, 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, Springer, 2015, pp. 234–241.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[25] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12009–12019.

[26] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[27] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, 2022, arXiv:2201.03545.

[28] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.S. Kweon, S. Xie, ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders, 2023, arXiv:2301.00808.

[29] Q. Jin, Z. Meng, C. Sun, H. Cui, R. Su, RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans, Front. Bioeng. Biotechnol. 8 (2020) 605132.

[30] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 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, Springer, 2018, pp. 3–11.

[31] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision, 3DV, Ieee, 2016, pp. 565–571.

[32] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[33] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[34] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: Revisiting the design of spatial attention in vision transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 9355–9366.

[35] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, Cswin transformer: A general vision transformer backbone with cross-shaped windows, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12124–12134.

[36] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, B. Fu, Shuffle transformer: Rethinking spatial shuffle for vision transformer, 2021, arXiv preprint arXiv:2106.03650.

[37] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[38] A. Hatamizadeh, D. Yang, H.R. Roth, D. Xu, UNETR: Transformers for 3D medical image segmentation, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2021, pp. 1748–1758, URL: https://api.semanticscholar.org/CorpusID:232290634.

[39] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, 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, Springer, 2021, pp. 36–46.

[40] Z. Yu, S. Han, Z. Song, 3D medical image segmentation based on multi-scale MPU-Net, 2023, arXiv:2307.05799.

[41] J. Pang, C. Jiang, Y. Chen, J. Chang, M. Feng, R. Wang, J. Yao, 3D shuffle-mixer: An efficient context-aware vision learner of transformer-MLP paradigm for dense prediction in medical volume, 2022, arXiv:2204.06779.

[42] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), 2016, arXiv preprint arXiv:1606.08415.

[43] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M.A.G. Ballester, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11) (2018) 2514–2525.

[44] L. Fu, Y. Chen, W. Ji, F. Yang, SSTrans-Net: Smart swin transformer network for medical image segmentation, Biomed. Signal Process. Control 91 (2024) 106071.

[45] R. Azad, L. Niggemeier, M. Hüttemann, A. Kazerouni, E.K. Aghdam, Y. Velichko, U. Bagci, D. Merhof, Beyond self-attention: Deformable large kernel attention for medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 1287–1297.

[46] D.-X. Huang, X.-H. Zhou, X.-L. Xie, S.-Q. Liu, Z.-Q. Feng, M.-J. Gui, H. Li, T.-Y. Xiang, X.-L. Liu, Z.-G. Hou, MOSformer: Momentum encoder-based inter-slice fusion transformer for medical image segmentation, 2024, arXiv preprint arXiv:2401.11856.

[47] Y. Chen, X. Lu, Q. Xie, Collaborative networks of transformers and convolutional neural networks are powerful and versatile learners for accurate 3D medical image segmentation, Comput. Biol. Med. 164 (2023) 107228.

[48] W. Yang, L. Xu, P. Wang, D. Geng, Y. Li, M. Xu, Z. Dong, More complex encoder is not all you need, 2023, arXiv preprint arXiv:2309.11139.

[49] X. Zhuang, J. Shen, Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI, Med. Image Anal. 31 (2016) 77–87.

[50] X. Yang, C. Bian, L. Yu, D. Ni, P.-A. Heng, 3D convolutional networks for fully automatic fine-grained whole heart partition, in: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8, Springer, 2018, pp. 181–189.

[51] C. Wang, Ö. Smedby, Automatic whole heart segmentation using deep learning and shape context, in: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8, Springer, 2018, pp. 242–249.

[52] Y. Ding, D. Mu, J. Zhang, Z. Qin, L. You, Z. Qin, Y. Guo, A cascaded framework with cross-modality transfer learning for whole heart segmentation, Pattern Recognit. 147 (2024) 110088.

[53] Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: Efficiently bridging cnn and transformer for 3D 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 III 24, Springer, 2021, pp. 171–180.

[54] H. Cui, Y. Wang, Y. Li, D. Xu, L. Jiang, Y. Xia, Y. Zhang, An improved combination of faster R-CNN and U-net network for accurate multi-modality whole heart segmentation, IEEE J. Biomed. Health Inform. 27 (7) (2023) 3408–3419.

[55] C. Tobon-Gomez, J. Peters, J. Weese, K. Pinto, R. Karim, T. Schaeffter, R. Razavi, K.S. Rhode, Left atrial segmentation challenge: a unified benchmarking framework, in: Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges: 4th International Workshop, STACOM 2013, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 26, 2013. Revised Selected Papers 4, Springer, 2014, pp. 1–13.