Heliyon 10 (2024) e26775

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

SCANeXt: Enhancing 3D medical image segmentation with dual attention network and depth-wise convolution

Yajun Liu [a], Zenghui Zhang [a], Jiang Yue [b,*], Weiwei Guo [c,*]

[a] Shanghai Key Laboratory of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, China
[b] Department of Endocrinology and Metabolism, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, China
[c] Center for Digital Innovation, Tongji University, China

ARTICLE INFO

Keywords:
3D medical image segmentation
Dual attention
Depth-wise convolution
Swin transformer
InceptionNeXt

ABSTRACT

Existing approaches to 3D medical image segmentation can be generally categorized into convolution-based or transformer-based methods. While convolutional neural networks (CNNs) demonstrate proficiency in extracting local features, they encounter challenges in capturing global representations. In contrast, the consecutive self-attention modules present in vision transformers excel at capturing long-range dependencies and achieving an expanded receptive field. In this paper, we propose a novel approach, termed **SCANeXt**, for 3D medical image segmentation. Our method combines the strengths of dual attention (**S**patial and **C**hannel **A**ttention) and Conv**NeXt** to enhance representation learning for 3D medical images. In particular, we propose a novel self-attention mechanism crafted to encompass spatial and channel relationships throughout the entire feature dimension. To further extract multiscale features, we introduce a depth-wise convolution block inspired by ConvNeXt after the dual attention block. Extensive evaluations on three benchmark datasets, namely Synapse, BraTS, and ACDC, demonstrate the effectiveness of our proposed method in terms of accuracy. Our SCANeXt model achieves a state-of-the-art result with a Dice Similarity Score of 95.18% on the ACDC dataset, significantly outperforming current methods.

1. Introduction

In recent years, vision transformers (ViTs) [1] have gradually surpassed and replaced Convolution Neural Network (CNN) and found wide applications in various downstream tasks of medical imaging, including segmentation [2–5], classification [6–9], restoration [10–13], synthesis [14–17], registration [18–21], and object detection in medical images [22,23]. In particular, significant progress has been observed in 3D medical image segmentation with the adoption of Vision Transformers (ViTs) [24–28]. Transformers have emerged as powerful tools for medical image segmentation, either as standalone techniques or as components of hybrid architectures. The self-attention mechanism, a core component of transformers, plays a crucial role in their success by enabling direct capture of long-range dependencies. Drawing on the efficacy of self-attention, scholars have suggested incorporating spatial-wise attention alongside channel-wise attention within transformer blocks to capture interactions [29–31]. These dual attention schemes based on self-attention have demonstrated improved performance across medical image segmentation tasks [32–36]. Taking inspira-

* Corresponding authors.
E-mail addresses: rjnfm3083@163.com (J. Yue), weiweiguo@tongji.edu.cn (W. Guo).

https://doi.org/10.1016/j.heliyon.2024.e26775
Received 5 January 2024; Received in revised form 19 February 2024; Accepted 20 February 2024

2405-8440/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

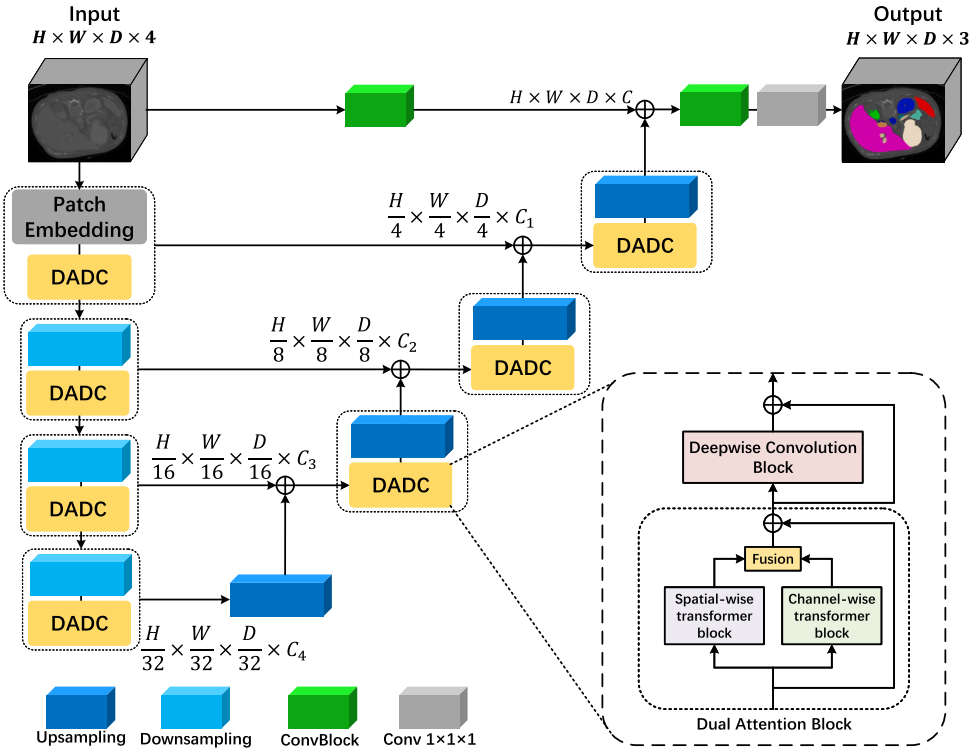

Fig. 1. Overview of our SCANeXt structure.

tion from the aforementioned research, we propose a novel transformer block that leverages spatial-wise and channel-wise attention mechanisms to enhance the capture of both spatial and channel information in 3D medical images.

Drawing insights from the successful experiences of both ViT and CNN, ConNeXt [37] is proposed as a pure convolutional network that surpasses sophisticated transformer-based models in performance and it also has widespread applications in 3D medical image segmentation [37–40]. However, ConNeXt faces challenges in terms of high computational FLOP demands associated with 3D depth-wise convolutions. Inspired by the recent InceptionNeXt [41], we further incorporate grouped depth-wise convolutions with different group convolution kernels in our novel architecture. Our approach allows extracting multiscale information while simultaneously reducing model computational complexity.

In this paper, we propose a synergistic hybrid DADC module combining spatial and channel-wise attention-based transformers with the multi-scale feature extraction capabilities of 3D depth-wise convolutions for processing 3D medical images. As a fundamental component, the DADC module is seamlessly incorporated into both the encoder and decoder sections within our SCANeXt network architecture. Our approach leverages the Swin Transformer-based [42] spatial-wise attention module to capture a broad range of spatial information, surpassing the capabilities of traditional ViTs-based spatial attention modules. Concurrently, the channel-wise attention transformer effectively captures channel information. To address the potential oversight of spatial details in the channel-wise transformer block, we introduce a gated-convolutional feed-forward network (GCFN) block. This block adeptly incorporates spatial information between layers within the channel-wise transformer, enhancing the overall depth of feature extraction. Our novel fusion mechanism seamlessly integrates these two distinct attention paradigms, forming a cohesive Dual Attention (DA) mechanism within the transformer framework in contrast to other methods that typically design spatial and channel attention separately. Complementing the DA transformer module, our Depthwise Convolution (DC) module divides the input into three distinct groups, each subjected to 3D grouped convolutions with kernel sizes of 3, 11, and 21, respectively. This stratification enables a comprehensive extraction of multi-scale features. Our evaluations of SCANeXt, conducted on three diverse public volumetric datasets, have consistently demonstrated its superiority in supervised segmentation tasks, significantly outperforming current state-of-the-art (SOTA) networks across all datasets.

We summarize our contributions as below:

- We introduce a hybrid hierarchical architecture tailored for 3D medical image segmentation, aiming to leverage the strengths of both transformers' attention mechanism and depth-wise convolution. This amalgamation capitalizes on the unique advantages of each technique, resulting in enhanced segmentation performance.
- We introduce a novel dual attention module that comprises a Swin Transformer-based spatial attention block and a channel attention block with the ability to capture local spatial information. This module effectively captures both short-range and long-range visual dependencies, enhancing the ability to understand complex image structures and relationships.

- We redesign a depth-wise convolution module based on InceptionNeXt for 3D medical image segmentation to extract deeper-level information from feature maps processed by the dual attention module, resulting in significantly improved 3D medical image segmentation performance.

2. Related work

Depthwise Convolution Based Segmentation Methods: Deep learning techniques have gained widespread adoption in medical image segmentation tasks [43–46], showcasing outstanding performance owing to their exceptional feature extraction capabilities. Numerous U-Net variants have been employed in medical image segmentation [47–51]. UNet++ and UNet3+ have improved upon the basic skip connection structure of UNet by integrating multi-scale skip connections and full-scale skip connections. This enhancement facilitates more effective feature extraction at various levels. Attention UNet [52] introduces attention gates to suppress irrelevant regions and focus on salient features. ResUNet [48] builds upon the UNet architecture by incorporating residual connections, while MultiResUNet [49] utilizes a multi-resolution approach to replace traditional convolutional layers. Among the notable UNet variants in recent years, nnUNet [53] stands out for its ability to autonomously adapt to data preprocessing and automatically select the optimal network architecture, eliminating the need for manual interventions. STNet [54] extends the nnUNet framework to construct a segmentation model with enhanced scalability and transferability.

Apart from these advancements in the traditional convolution-based UNet architecture, recent studies have started to reexamine the concept of depthwise convolution, tailoring its characteristics to enhance robust segmentation. Sharp U-Net [38] integrates innovative connections between the encoder and decoder subnetworks, which entail applying depthwise convolution to the encoder feature maps with a sharpening spatial filter. The $3D^2$U-Net [55] utilizes depthwise convolutions as domain adapters to extract domain-specific features within each channel.

Both studies show that utilizing depthwise convolution can improve volumetric problems. Only a tiny kernel size is utilized for depthwise convolution. Previous studies have explored the efficacy of large-kernel (LK) convolution in medical image segmentation. For instance, LKAU-Net [56] is the first use of LK and dilated depthwise convolution for brain tumor segmentation. Following this, ConvNeXt, a modified convolutional neural network model rooted in the standard ResNet, is introduced by Liu et al. [37]. It replaces the cutting-edge Swin transformer in multiple computer vision applications. Inspired by ConvNeXt, ASF-LKUNet [57] devises a residual block with enlarged kernels and incorporates large-kernel global response normalization (GRN) channel attention, leveraging depthwise convolution to concurrently capture both global and local features. Introducing volumetric depthwise convolution with substantial kernel dimensions (7×7×7), 3D-UXNet [58] emerges as the pioneer in simulating large receptive fields within the Swin transformer to generate self-attention. RepUX-Net [39] achieves better adaptation with extremely large kernel sizes (21×21×21) depthwise convolution to get a larger receptive field for volumetric segmentation. However, 3D-UXNet and RepUX-Net only use ConvNeXt blocks partially in a standard convolution encoder, limiting their possible benefits. Presented by Saika et al. [40], MedNeXt stands as the inaugural fully ConvNeXt 3D segmentation network. It facilitates scaling in width (more channels) and receptive field (larger kernels) at both standard and up/downsampling layers.

Transformer-based Segmentation Methods: To our best knowledge, TransUNet [2] is the first to use a hybrid CNN-Transformer encoder that combines spatial information at high resolution from CNN features with the global context extracted from Transformers in the U-shaped architecture, and the decoder maintains CNN-based upsampling structure. Meanwhile, SwinUNet [4] is the first pure Transformer-based U-shaped architecture. Self-attention mechanism in the transformer-based encoder captures local and global features after patch embedding. In the encoder, patch merging is applied for downsampling, while in the decoder, patch expanding is utilized to achieve upsampling.

SwinMM [59] employs a self-supervised learning approach, emphasizing the use of Swin Transformer for multiview information processing to enhance the performance of self-supervised learning in medical images. DS-TransUNet [60] adopts an effective dual-scale encoding mechanism that simulates non-local dependencies and multiscale contexts. It combines this mechanism with the Swin Transformer in both the encoder and decoder to enhance semantic segmentation quality. DS-TransUNet primarily focuses on improving segmentation performance through dual-scale encoding and a cross-fusion module. SwinPA-Net [61] introduces two modules, namely the Dense Multiplicative Connection (DMC) module and the Local Pyramid Attention (LPA) module, to aggregate multiscale contextual information in medical images. By combining these two modules with the Swin Transformer, it learns more powerful and robust features. ST-UNet [62] uses Swin Transformer as the encoder and CNNs as the decoder, introducing the Cross-Layer Feature Enhancement (CLFE) module in the skip connection part. This design aims to leverage low-level features comprehensively to enhance global features and reduce the semantic gap between the encoding and decoding stages. These Swin-based medical segmentation methods provide inspiration for subsequent Transformer-based medical image segmentation approaches. For instance, UNTER [25] and SwinUNETR [28] replace the hybrid CNN-Transformer architecture of the encoder in TransUNet with a cascade of several ViTs and Swin Transformer modules, respectively, and remain the decoder unchanged. DBT-UNETR [63] adopts the patch embedding and patch merging from SwinUNETR and replaces the encoder with a straightforward concatenation of the encoders from UNETR and SwinUNETR. nnFormer [26] builds upon the SwinUNet architecture by replacing the patch embedding and patch merging operations in the encoder with 3D convolutions, and the patch expanding operation in the decoder with 3D deconvolution. This enables the interleaving of convolutional and transformer-based blocks in the encoder and decoder, fully leveraging their respective advantages in feature extraction. TFCNs [64] introduce the Convolutional Linear Attention Block (CLAB), which encompasses two types of attention: spatial attention over the image's spatial extent and channel attention over CNN-style feature channels. The aforementioned methods simply rearrange the CNN and transformer modules within the encoder and decoder struc-

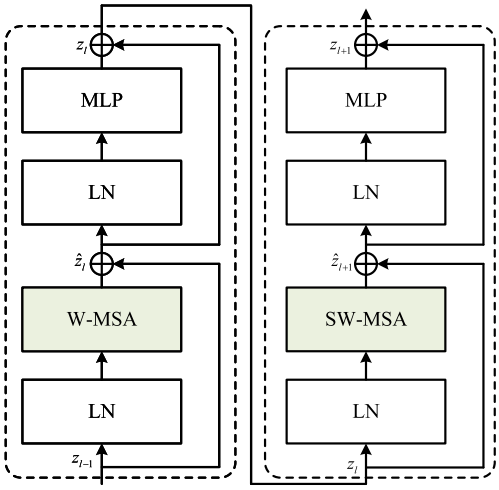

Fig.2(a) The spatial-wise transformer block

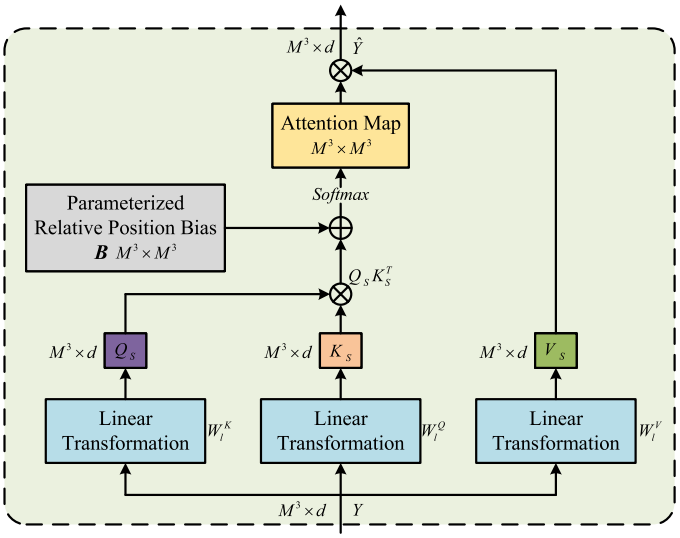

Fig.2(b) The window-based multi-head self-attention

⊕ : Element-wise Addition **⊗ : Matrix Multiplication**

Fig. 2. Components of the spatial-wise transformer.

tures of UNet, without introducing any novel modules, resulting in limited improvements in segmentation performance. Compared with SwinUNet, MISSFormer [65] introduces a redesigned transformer block in the network structure of remix-FFN for integrating global information and local content. Additionally, MISSFormer proposes a novel ReMixed transformer context bridge to replace the skip connection between the encoder and decoder. CoTr, as introduced in [24], presents a novel mechanism for self-attention, termed the deformable self-attention. This mechanism constitutes the efficient DeTrans module, tailored for processing feature maps that are both multi-scale and high-resolution. These two approaches, which involve the redesign or improvement of the attention module in transformers, have also been widely applied in subsequent methods. Take the recent two SOTA methods as examples, VTUNet [27] proposes parallel cross-attention and self-attention in the decoder to preserve global context.

The central aspect of UNETR++ [66] involves introducing a novel efficient paired attention (EPA) block. This block effectively captures spatial and channel-wise discriminative features by employing inter-dependent branches that incorporate spatial and channel attention mechanisms. However, the channel attention and spatial attention modules in EPA are independently designed and share the parameters during attention computations, without taking into account the interplay between the two attention modules. While we design a channel attention module to acquire channel information, we also consider the impact on local spatial information and use the designed gate structure to better capture local spatial information and obtain better feature extraction results. Further, we use the Swin transformer in the spatial attention module to obtain spatial information at different scales.

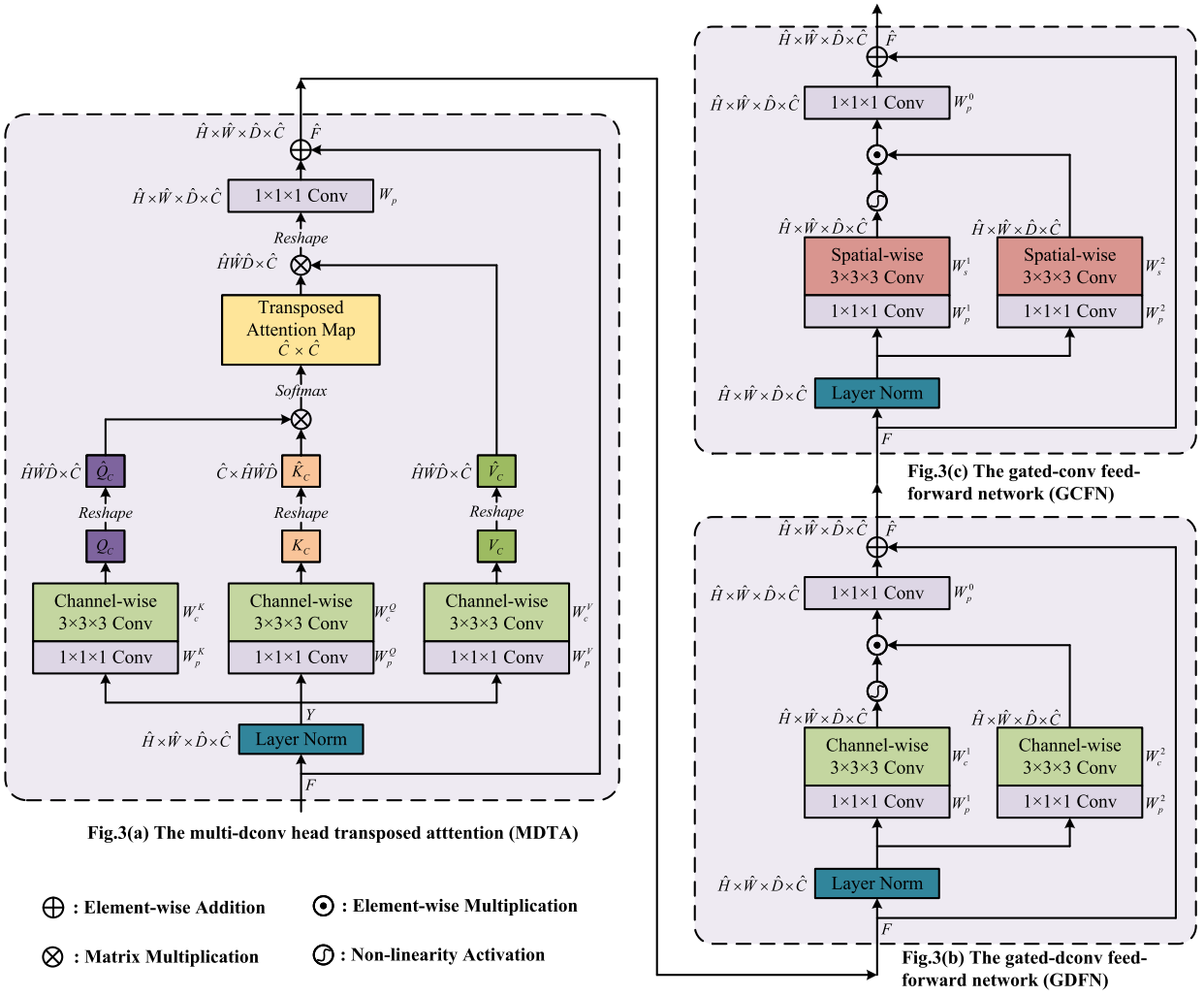

Fig. 3. Components of the channel-wise transformer.

3. Method

Introducing SCANeXt, a transformer backbone characterized by its cleanliness, efficiency, and effectiveness, integrating local fine-grained features with global representations. This section begins with an overview of the hierarchical layout of our model, succeeded by a comprehensive elucidation of both the dual attention module and the depthwise convolution module.

3.1. Overall architecture

Fig. 1 shows the SCANeXt architecture, which consists of a hierarchical encoder-decoder structure with skip connections between them, and convolutional blocks (ConvBlocks) for generating the prediction masks. The SCANeXt utilizes a hierarchical design that reduces feature resolution by a factor of two at each stage, efficiently capturing both local and global context information. Our SCANeXt framework distinguishes itself from the recently proposed TransUNet [60] and TFCNs [64]. TransUNet primarily focuses on integrating the U-Net architecture with Transformers to capture local features and global context information. In comparison, Our SCANeXt goes beyond by not only merging spatial and channel attention but also effectively extracting multi-scale features through depth convolution blocks. TFCNs introduce the Convolutional Linear Attention Block (CLAB) for spatial and channel attention while SCANeXt integrates transformer Attention Block with three components: Spatial Attention Block, Channel-wise Attention Block, and Spatial Channel Fusion Block. Inspired by the Swin Transformer, the Spatial Attention Block utilizes the W-MSA and SW-MSA structures to capture spatial features and the Channel-wise Attention Block comprises MDTA, GDFN, and GCFN, designed based on the MSA in Vision Transformer.

In our SCANeXt framework, the encoder is structured with four stages. Initially, during the first phase of patch embedding, the volumetric input undergoes a segmentation into 3D patches. This process is succeeded by the incorporation of our innovative dual

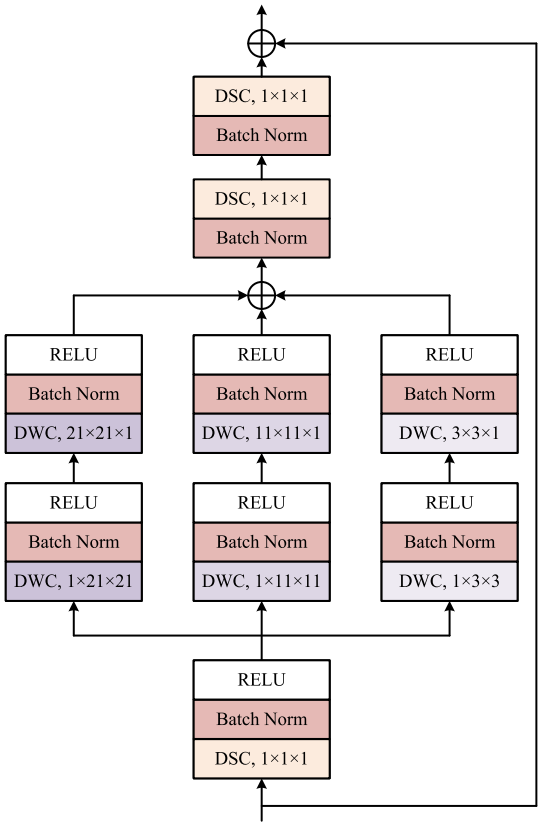

Fig. 4. Components of the depthwise convolution module.

attention and depthwise convolution (DADC) module. In the context of patch embedding, the division of each 3D input (volume) $\boldsymbol{x}_u \in \mathbb{R}^{H\times W\times D}$ takes place, resulting in non-overlapping patches $\boldsymbol{x}_u \in \mathbb{R}^{N\times(P_1,P_2,P_3)}$, where (P_1, P_2, P_3) signifies the resolution of each patch and $N = \frac{H}{P_1} \times \frac{W}{P_2} \times \frac{D}{P_3}$ represents the sequence length. The specified patch resolution is denoted as $(P_1, P_2, P_3) = (2, 2, 2)$. Following this, the patches undergo projection into C channel dimensions, generating feature maps sized $\frac{H}{P_1} \times \frac{W}{P_2} \times \frac{D}{P_3} \times C$. In the subsequent encoder stages, downsampling layers employ non-overlapping convolution to decrease the resolution by a factor of two. Subsequently, the DADC block is integrated into the process.

In the proposed SCANeXt framework, each block of the DADC module integrates modules for spatial attention and channel attention, followed by a depthwise convolution module. The dual attention module effectively captures enriched representations by incorporating information from both spatial and channel dimensions. Simultaneously, the depthwise convolution module adjusts to a broad receptive field, enhancing features through the expansion of independent channels. To establish connectivity between encoder and decoder stages, skip connections are employed. These connections merge outputs at varying resolutions, facilitating the retrieval of spatial information that may be lost the downsampling procedures. This, in turn, contributes to the generation of more accurate predictions. In a parallel to the encoder's configuration, the decoder is structured into four stages. Within each stage, an upsampling layer is incorporated, employing deconvolution to double the resolution of feature maps. Following the upsampling layer, the DADC block is applied, with the exception of the last decoder stage. The channel count reduces by half between successive decoder stages. The final decoder's outputs are fused with convolutional feature maps to not only restore spatial information but also enhance feature representation. The resultant output undergoes further processing through 3×3×3 and 1×1×1 convolutional blocks, culminating in voxel-wise final mask predictions. Subsequently, we provide a detailed exposition on our DADC block.

3.2. Dual attention block

This section introduces the Dual Attention Block (DAB), designed to comprehensively capture spatial and channel-wise dependencies. Illustrated within the dashed box in Fig. 1, the DAB is composed of three fundamental sub-blocks: the Spatial Attention Block (SAB), Channel-wise Attention Block (CAB), and Spatial Channel Fusion Block (S-CFB). Incorporating the DAB into the standard transformer allows the synergistic utilization of spatial and channel-wise attention, enhancing the exploration of visual information for volumetric medical image segmentation.

3.2.1. Spatial attention block (SAB)

Illustrated in Fig. 2(a), our approach involves the utilization of window-based multi-head self-attention (W-MSA) and shift window-based multi-head self-attention (SW-MSA) [42] for applying spatial attention to volumetric medical image features. The rationale behind opting for W-MSA and SW-MSA to capture spatial dependencies lies in the distinct characteristics compared to the multi-head self-attention (MSA) in [67], the shifted window mechanism of W-MSA and SW-MSA can extract feature representations at several spatial resolutions while reducing computational complexity. Additionally, the W-MSA and SW-MSA blocks need to be modified to suit 3D medical image data. The two consecutive Swin transformer blocks depicted in Fig. 2(a) can be described mathematically as follows:

$$\begin{aligned} \hat{z}_l &= \text{W-MSA}(\text{LN}(z_l)) + z_{l-1}, \\ z_l &= \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l, \\ \hat{z}_{l+1} &= \text{SW-MSA}(\text{LN}(z_l)) + z_l, \\ z_{l+1} &= \text{MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1}, \end{aligned} \tag{1}$$

Equation (1) defines the operations of window-based MSA (W-MSA) and shifted window-based MSA (SW-MSA). Here, z_{l-1} and z_l represent the inputs to W-MSA and SW-MSA, while $\hat{z}_l$ and $\hat{z}_{l+1}$ denote their respective outputs. The final output of spatial attention feature is denoted as z_{l+1}. To simplify the notation, we use U to represent the input feature map ($U = z_{l-1}$) and $\hat{U}$ for the output feature map ($\hat{U} = z_{l+1}$). Additionally, MLP and LN stand for multi-layer perception and layer normalization, respectively.

In Fig. 2(b), the diagram illustrates the window-based multi-head self-attention. Specifically, the calculation of self-attention can be expressed as:

$$\text{Attention}(Q_S, K_S, V_S) = \text{Softmax}\Big(\frac{Q_S K_S^{\top}}{\sqrt{d}} + B\Big)V_S. \tag{2}$$

In Equation (2), $B \in \mathbb{R}^{M^3 \times M^3}$ represents the relative position of tokens within each window. Here, d denotes the dimension of the query and key, and M^3 corresponds to the number of patches in the 3D image window. The matrices for query Q_S, key Q_S, and value V_S are computed as follows:

$$Q_S = W_l^Q Y, K_S = W_l^K Y, V_S = W_l^V Y. \tag{3}$$

In Equation (3), $W_l^Q, W_l^K, W_l^V \in \mathbb{R}^{M^3 \times M^3}$ represent the linear projection matrices, while $Y \in \mathbb{R}^{M^3 \times d}$ denotes the input after the LN layer.

3.2.2. Channel-wise attention block (CAB)

The channel-wise transformer comprises three sequential components illustrated in Fig. 3: MDTA (multi-Dconv head transposed attention), GDFN (gated-Dconv feed-forward network), and GCFN (gated-Conv feed-forward network). In the work by Zamir et al. [68], the MDTA module within Restormer is designed to perform self-attention along the channel, diverging from the spatial dimension. This configuration allows for global attention computation with linear computational complexity. GDFN, on the other hand, selectively transmits valuable information to subsequent layers while suppressing less informative features. While Restormer has demonstrated competitive performance in natural image enhancement, our experiments revealed a limitation in terms of losing local spatial information. In addition to the parallel transfer of local spatial information through spatial attention blocks, we introduce a channel-wise attention block in our approach. This block incorporates a gated-conv feed-forward network (GCFN) to enhance the capture of local spatial information, thereby improving organ segmentation performance. The subsequent sections delve into the specifics of MDTA, GDFN, and GCFN.

MDTA Fig. 3(a) illustrates the MDTA module's structure. Starting with an input feature map $F \in \mathbb{R}^{\hat{H} \times \hat{W} \times \hat{D} \times \hat{C}}$, we first apply a layer normalization module to obtain $Y \in \mathbb{R}^{\hat{H} \times \hat{W} \times \hat{D} \times \hat{C}}$. The matrices $Q_C, K_C, V_C \in \mathbb{R}^{\hat{H} \times \hat{W} \times \hat{D} \times \hat{C}}$ are then derived through a 1×1×1 pixel-wise convolution operation (for encoding channel-wise context) and a 3×3×3 channel-wise convolution operating (for aggregating pixel-wise cross-channel context). After reshaping operations, $\hat{Q}_c, \hat{K}_c^T, \hat{V}_c \in \mathbb{R}^{\hat{H}\hat{W}\hat{D} \times C}$ are obtained from Q_c, K_c and V_c, respectively. These matrices are utilized to generate a transposed-attention map $A \in \mathbb{R}^{C \times C}$, serving as an alternative to the extensive spatial attention map of size $\hat{H}\hat{W}\hat{D} \times \hat{H}\hat{W}\hat{D}$ [67], [69]. The output feature of the MDTA module, denoted as $\hat{F}$, can be expressed as:

$$\begin{aligned} \hat{F} &= W_p \cdot \text{Attention}(\hat{Q}_c, \hat{K}_c, \hat{V}_c) + F \\ \text{Attention}(\hat{Q}_c, \hat{K}_c, \hat{V}_c) &= \hat{V}_c \cdot \text{Softmax}(\hat{Q}_c^T \cdot \hat{K}_c / \alpha) \end{aligned} \tag{4}$$

In Equation (4), the parameter α is a learnable scaling factor that governs the magnitude of the dot product output $\hat{Q}_c^T \cdot \hat{K}_c$, and W_p represents a linear projection matrix. It is crucial to highlight that the result of the self-attention operation undergoes reshaping to align with the input size of F.

GDFN Introducing a gating mechanism [67] proves beneficial in selectively transmitting valuable information to subsequent layers within the network architecture. This gating network effectively suppresses less informative features. In the GDFN module, as illustrated in Fig. 3(b) and utilized in [68], the gating mechanism is computed as follows:

$$\begin{aligned} \hat{F} &= W_p^0 \cdot \text{Gating}(F) + F \\ \text{Gating}(F) &= \phi\left(W_c^1 W_p^1 \text{LN}(F)\right) \odot W_c^2 W_p^2 \text{LN}(F). \end{aligned} \tag{5}$$

In Equation (5), F and $\hat{F}$ represent the input and output features, respectively. The operator $\odot$ signifies element-wise multiplication, ϕ denotes the GELU non-linearity activation layer, and LN indicates layer normalization. The linear projection matrices are denoted as W_p^0, W_p^1, and W_p^2. Additionally, W_c^1 and W_c^2 correspond to 3×3×3 channel-wise convolutions.

GCFN To address the potential loss of spatial information by the MDTA and GDFN modules, which predominantly leverage channel information for 3D medical image segmentation, a gated-conv feed-forward network (GCFN) module (depicted in Fig. 3(c)) is incorporated into the channel-wise transformer block. Notably, the GCFN differs from the GDFN in that the 3×3×3 channel-wise convolutions (W_c^1 and W_c^2) in GDFN are replaced with 3×3×3 spatial-wise convolutions (W_s^1 and W_s^2). Our hypothesis posits that spatial-wise convolutions may more effectively harness spatial information within volumetric medical images, thereby enhancing the segmentation of regions of interest.

3.2.3. Spatial-channel fusion block (S-CFB)

To maintain model conciseness, we opt for the element-wise sum to amalgamate the two attention feature maps:

$$F_{dual} = \text{LN}\left(\rho(\hat{U}) + \rho(\hat{F})\right) \tag{6}$$

In Equation (6), the symbols $\hat{U}$ and $\hat{F}$ stand for the spatial and channel-wise attention feature maps, respectively. The operation $\rho(\cdot)$ signifies the dropout, with a set probability of 0.1 for zeroing an element. Additionally, LN$(\cdot)$ represents the layer normalization.

3.3. Depthwise convolution module

The 3DInceptionNeXt block is composed of 3×3×3 convolutional layers and multi-branch convolutional layers incorporating multi-scale depthwise separable kernels. In our experiments, we utilize kernel sizes of 3, 11, and 21 for these multi-scale depthwise separable kernels. Similar to ConvNeXt [37], the 3DInceptionNeXt block employs the inverted bottleneck design, where the channel size of the hidden 1×1×1 convolutional layer is four times wider than the input along the channel dimensions, as illustrated in Fig. 4.

Depthwise separable kernel To address the computational and memory inefficiency associated with using a large kernel in the convolutional layer, we adopt the depthwise convolution layer with separable kernels, following the kernel design principles introduced in ConvNeXt and InceptionNeXt [41]. In this section, we decompose the kernel size of $x \times x \times x$ into $1 \times x \times x$ and $x \times x \times 1$ kernels. This decomposition is illustrated for 21×21×21, 11×11×11, and 3×3×3 kernels in Fig. 4. Beyond the computational and memory savings, it has been demonstrated that separable kernels enable the model to independently extract temporal and frequency features, leading to improved accuracy in medical image segmentation.

Inverted bottleneck In contrast to the convolutional design of the inverted bottleneck in MobileNetV2 [70], our approach involves situating the multi-branch convolutional layers at the top, preceding the application of 1×1×1 convolution layers to perform the channel-wise expansion and squeeze operation. This design choice, akin to the ConvNeXt block [37], aids in mitigating the memory footprint and computational time associated with the utilization of large kernels.

Non-linear layers To augment the nonlinearity within the model, batch normalization and RELU activation layers follow each separable kernel, as illustrated in Fig. 4.

3.4. Loss function

In alignment with the baseline approach UNETR++, we concurrently employ both the Dice Loss and the Cross-Entropy Loss to optimize our network. The expression for our loss function is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{dice} + \mathcal{L}_{CE} \tag{7}$$

$$\mathcal{L}_{dice} = 1 - \frac{2}{C} \sum_{j=1}^{C} \frac{\sum_{i=1}^{N} P_{ij} Y_{ij}}{\sum_{i=1}^{N} P_{ij} + \sum_{i=1}^{N} Y_{ij}} \tag{8}$$

$$\mathcal{L}_{CE} = -\frac{1}{NC} \sum_{i=1}^{N} \sum_{j=1}^{C} (Y_{ij} \log P_{ij}) \tag{9}$$

Equation (7) represents the total loss, while Equation (8) and Equation (9) correspond to the Dice Loss and Cross-Entropy (CE) Loss, respectively. These equations involve parameters where N represents the total number of voxels, C signifies the number of target classes, P denotes the Softmax output of the prediction segmentation, and Y represents the ground truth.

Table 1
The Synapse dataset's segmentation accuracy is assessed for various approaches, with evaluation metrics expressed in DSC (in %) and HD95 (in mm). Eight organs are subject to segmentation: spleen (Spl), right kidney (RKid), left kidney (LKid), gallbladder (Gal), liver (Liv), stomach (Sto), aorta (Aro), and pancreas (Pan). The optimal performance is highlighted in bold, while the second-best is underlined.

Methods	Spl	RKid	LKid	Gal	Liv	Sto	Aor	Pan	Average	
									HD95 ↓	DSC ↑
U-Net [43]	86.67	68.60	77.77	69.72	93.43	75.58	89.07	53.98	-	76.85
TransUNet [2]	85.08	77.02	81.87	63.16	94.08	75.62	87.23	55.86	31.69	77.49
Swin-UNet [71]	90.66	79.61	83.28	66.53	94.29	76.60	85.47	56.58	21.55	79.13
UNETR [25]	85.00	84.52	85.60	56.30	94.57	70.46	89.80	60.47	18.59	78.35
MISSFormer [65]	91.92	82.00	85.21	68.65	94.41	80.81	86.99	65.67	18.20	81.96
nnUNet [53]	94.81	84.41	87.30	66.22	96.15	75.20	91.73	76.04	12.56	83.98
Swin-UNETR [28]	95.37	86.26	86.99	66.54	95.72	77.01	91.12	68.80	10.55	83.48
nnFormer [26]	90.51	86.25	86.57	70.17	96.84	**86.83**	92.04	**83.35**	10.63	86.57
3D-UXNet [58]	94.90	**92.46**	**94.98**	70.61	96.74	86.61	91.17	71.30	16.01	86.72
UNETR++ [66]	**95.77**	87.18	87.54	71.25	96.42	86.01	92.52	81.10	7.53	87.22
Ours	95.69	87.46	85.92	**73.66**	**96.98**	85.60	**92.96**	82.95	**7.47**	**89.67**

Table 2
The BraTS dataset's segmentation accuracy is assessed for various approaches, with evaluation metrics expressed in DSC (in %) and HD95 (in mm). Segmentation is performed on three recombined regions: tumor core (TC), whole tumor (WT), and enhancing tumor (ET).

Methods	WT		ET		TC		Average	
	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑
nnUNet [53]	3.64	**91.9**	4.06	81.0	4.95	85.6	4.60	86.2
UNETR [25]	8.27	78.9	9.35	58.5	8.85	76.1	8.82	71.1
nnFormer [26]	3.80	91.3	3.87	81.8	4.49	**86.0**	4.05	86.4
VT-UNet-S [27]	4.01	90.8	2.91	81.8	4.60	85.0	3.84	85.9
UNETR++ [66]	3.61	91.4	2.82	78.6	**3.82**	84.5	**3.42**	84.8
Ours	**3.35**	91.6	**2.64**	**83.6**	4.70	84.5	3.56	**86.6**

Table 3
The ACDC dataset's segmentation accuracy is assessed for various approaches. The evaluation metric utilized is DSC (in %). Segmentation is performed on three sub-structures, encompassing the right ventricle (RV), left ventricle (LV), and myocardium (MYO).

Methods	RV	Myo	LV	Average
TransUNet [2]	88.86	84.54	95.73	89.71
Swin-UNet [71]	88.55	85.62	95.83	90.00
UNETR [25]	88.49	82.04	91.62	87.38
MISSFormer [65]	86.36	85.75	91.59	87.90
nnUNet [53]	90.24	89.24	95.36	91.61
3D-UXNet [58]	77.97	81.84	92.41	84.07
nnFormer [26]	91.18	86.24	94.07	90.50
UNETR++ [66]	91.47	86.47	94.25	90.73
Ours	**96.51**	**92.19**	**96.84**	**95.18**

4. Experiments

4.1. Experimental setup

Our experimentation involves three benchmark datasets for medical image segmentation: Synapse [72], ACDC [73], and BraTS [74].

Datasets: Synapse is a multi-organ segmentation dataset, consisting of 30 abdominal CT scans in 8 abdominal organs (spleen, right kidney, left kidney, gallbladder, liver, stomach, aorta, and pancreas). We adhere to the dataset division strategy employed in TransUNet [2] and nnFormer [26]. Specifically, we use 18 cases to form the training set, of which 4 cases are chosen for the validation set, and the remaining 14 cases are used for testing.

ACDC comprises 100 scans from 100 patients, with target regions of interest (ROIs) including the left ventricle (LV), right ventricle (RV), and myocardium (MYO). Following the data split protocol outlined in UNETR++ [66], we allocate 70 cases for training, 10 for validation, and 20 for testing purposes.

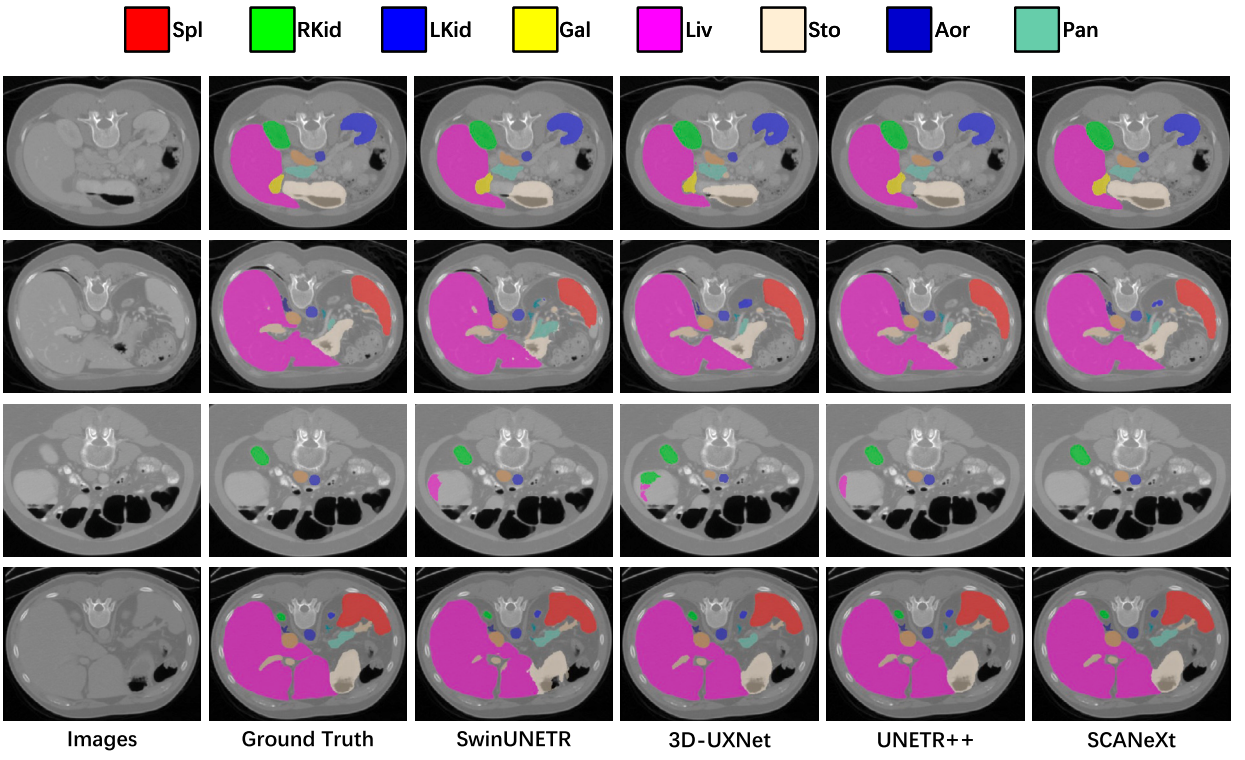

Fig. 5. Qualitative comparison of the segmentation performance for the Synapse dataset.

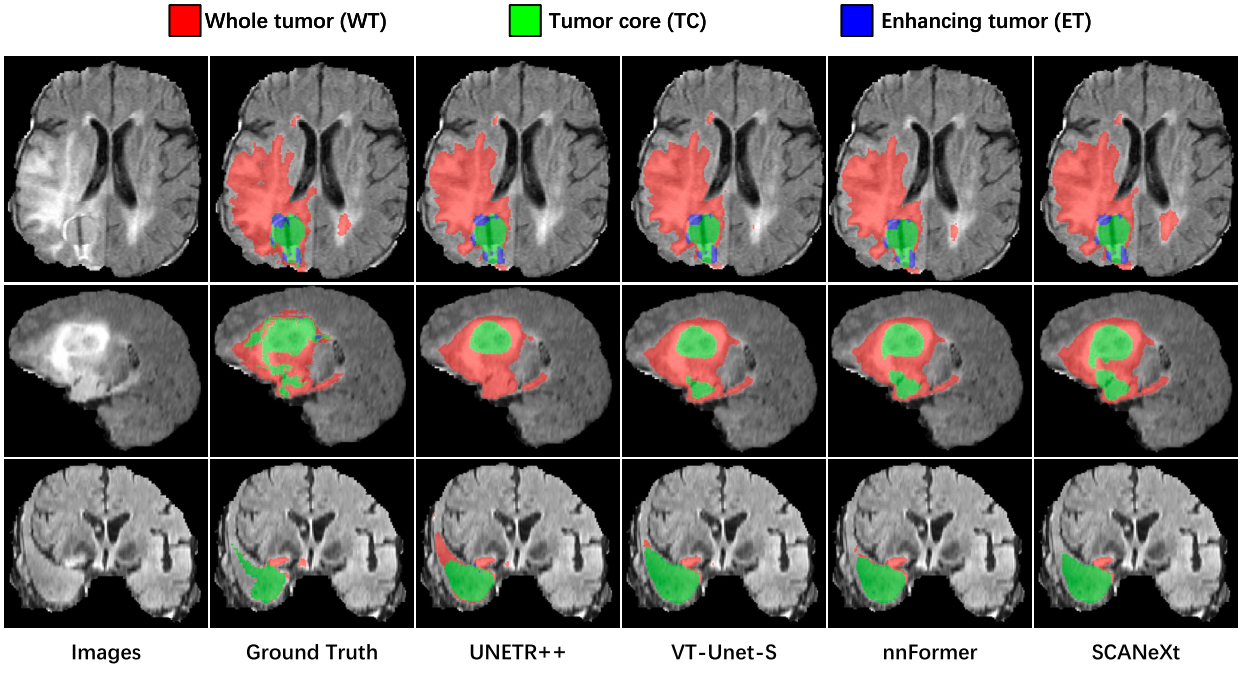

Fig. 6. Qualitative comparison of the segmentation performance for the BraTS dataset.

The Medical Segmentation Decathlon (MSD) BraTS dataset comprises 484 MRI scans with multi-modalities (FLAIR, T1w, T1-Gd, and T2w). The ground-truth segmentation labels cover peritumoral edema, GD-enhancing tumor, and the necrotic/non-enhancing tumor core. Evaluation of performance involves three combined regions: tumor core, whole tumor, and enhancing tumor. The dataset undergoes random partitioning into training (80%), validation (15%), and test (5%) sets.

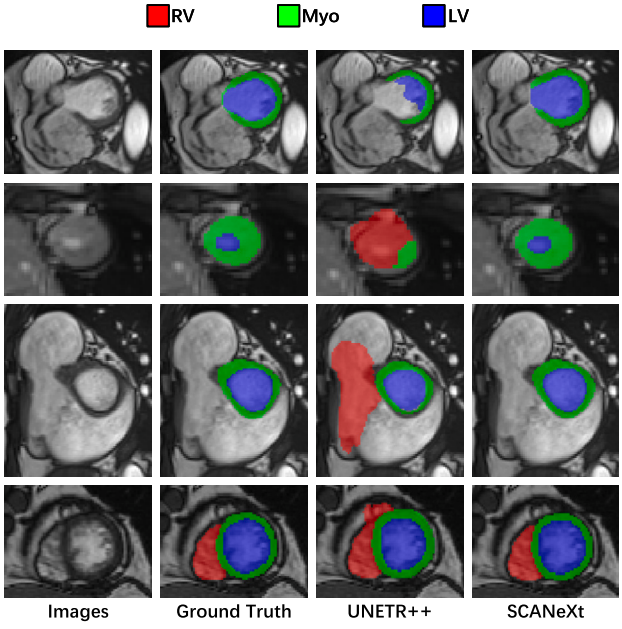

Fig. 7. Qualitative comparison of the segmentation performance for the ACDC dataset.

Evaluation Metrics: Model performance evaluation relies on two metrics: the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95). Typically, superior segmentation performance is reflected in a higher score for the region-based metric (DSC) and a lower score for the boundary-based metric (HD95). The calculation of similarity between the predicted segmentation P and the ground truth Y based on the region is given by Equation (10):

$$DSC(Y,P) = 2 * \frac{|Y * P|}{|Y| + |P|} \tag{10}$$

For boundary measurement, the calculation is articulated through Equations (11), (12), and (13):

$$HD_{95}(Y,P) = \max(d_{YP}, d_{PY}) \tag{11}$$

$$d_{YP} = \max_{p \in P} \left(\min_{y \in Y} \|y - p\| \right) \tag{12}$$

$$d_{PY} = \max_{y \in Y} \left(\min_{p \in P} \|p - y\| \right) \tag{13}$$

Equation (11) quantifies the unidirectional Hausdorff distance between Y and P, with $\max(\cdot)$ representing the calculation of the 95th percentage of the distance between the boundary points of Y and P.

Implementation Details: Our approach is implemented in PyTorch v1.10.1, utilizing the MONAI libraries [75]. To ensure a fair comparison with both the baseline UNETR++ and nnFormer, we adopt identical input sizes, pre-processing strategies, and training data amounts. The models are trained on a single Quadro RTX 6000 24 GB GPU. For the Synapse dataset, all models undergo training for 1000 epochs with input 3D patches sized $128 \times 128 \times 64$, employing a learning rate of 0.01 and weight decay of $3e^{-5}$. Concerning the ACDC and BRaTs datasets, we train all models at resolutions of $160 \times 160 \times 16$ and $128 \times 128 \times 128$, respectively. During training, all other hyper-parameters are kept consistent with those of nnFormer. Specifically, the input volume is segmented into non-overlapping patches, which are then utilized for learning segmentation maps via back-propagation. Additionally, the same data augmentations are applied for UNETR++, nnFormer, and our SCANeXt throughout the training process.

4.2. Comparison with state-of-the-art methods

The evaluation of each task against state-of-the-art methods encompasses both quantitative and qualitative aspects.

4.2.1. Experimental results on the Synapse dataset

As listed in Table 1, results from another ten approaches are presented, including three classical methods namely, U-Net, TransUNet, Swin-UNet, and seven state-of-the-art methods namely UNETR, MISSFormer, nnUNet, SwinUNETR, nnFormer, 3D-UXNet, and UNETR++. We reproduce the results of nnUNet, SwinUNETR, 3D-UXNet, and UNETR++, while the results of the remaining methods are directly selected from their original papers, all the methods follow the same date split as mentioned in Section 4.1. According

to Table 1, with the DSC of 89.67% and the HD95 of 7.47 mm, our method achieves the best performance in both evaluation metrics. Our method outperforms the second-best method UNETR++ by 2.45% in the DSC and by 0.06 mm in the HD95. Specifically, compared with nnFormer, 3D-UXNet, and UNETR++, SCANeXt achieves the highest DSC in the segmentation of the following three organs, which are gallbladder, liver, and aorta, and achieves the second highest DSC in the following three organs, which are spleen, right kidney, and pancreas.

In Fig. 5, a qualitative comparison of SCANeXt with three state-of-the-art methods for abdominal multi-organ segmentation is presented. SCANeXt demonstrates a substantial reduction in the number of incorrectly segmented pixels overall. In the first row, both SwinUNETR and UNETR++ fail to achieve complete segmentation of the Sto region. Although 3D-UXNet performs slightly better, it still falls short compared to SCANeXt in terms of achieving complete segmentation. The most significant difference becomes evident in the fourth row, where 3D-UXNet and UNETR++ also struggle to achieve complete segmentation of the Pan region, while SwinUNETR succeeds in fully segmenting Pan, but makes errors at the boundaries of Sto and Liver regions. In the second row, SwinUNETR incorrectly segments a region that does not belong to any label as Pan. On the other hand, 3D-UXNet and UNETR++ both misclassify this region as Aor, with 3D-UXNet additionally incorrectly segmenting part of the Sto region. Only SCANeXt exhibits no misclassification in this scenario. Similar to the second row, in the third row, SwinUNETR, 3D-UXNet, and UNETR++ mistakenly segment regions without labels as liver or Rkid. Overall, SCANeXt achieves the closest segmentation results to the ground truth. This could be attributed to the SCANeXt method's incorporation of spatial and channel attention mechanisms in the transformer, combined with the advantages of separable kernel convolutions in extracting contextual features.

4.2.2. Experimental results on the BraTS dataset

The experiment results on the BraTS dataset are shown in Table 2. SCANeXt surpasses other methods in segmenting enhancing tumor (ET) regions, yielding the highest DSC and the lowest HD95. It also performs competitively in whole tumor (WT) segmentation, with only a marginal 0.3% DSC difference from the classical CNN-based method nnUNet, and achieves the lowest HD95 for WT segmentation among all compared methods. Furthermore, SCANeXt's average performance, in terms of DSC, is the highest, and its HD95 is the second-lowest. It is particularly noteworthy that SCANeXt's performance exceeds that of the dual-attention-based method UNETR++ by 1.80% in DSC, which validates the effectiveness of SCANeXt's redesigned dual-attention mechanism over the EPA block in UNETR++. Our SCANeXt outperforms the SOTA UNETR++ and the secondary nnFormer on average. All of them are based on the transformer structure. In nnFormer, they employed the basic transformer block as the main building block for local 3D volume embedding with high computational complexity, while our proposed DADC decomposes global transformer attention into spatial-wise and channel-wise transformer attention with high computation efficiency.

Qualitative segmentation results on the BraTS dataset are depicted in Fig. 6. Specifically, the first to third rows illustrate the segmentation outcomes of MRI scans in the transverse, sagittal, and coronal planes. In the transverse plane, UNETR++, VT-UNet-S, and nnFormer exhibit limitations in achieving complete segmentation of the WT subregion. In the coronal plane, these three methods struggle to accurately distinguish between the TC and WT regions. Additionally, in the sagittal plane, normal tissue is misidentified as tumor regions by all three methods. Nevertheless, SCANeXt consistently produces favorable segmentation results in all cases, showcasing its potential for effectively delineating irregular and complex lesions.

4.2.3. Experimental results on the ACDC dataset

The experiment results on the ACDC dataset are presented in Table 3. Compared with other methods, SCANeXt achieves the highest DSC for all three sub-structures segmentation, including RV, Myo and LV, surpassing the second-best methods by 5.04%, 2.95%, 1,01%. With an average DSC of 95.18%, SCANeXt significantly outperforms other methods, achieving a remarkable 3.57% higher accuracy compared to the second-best performing method, nnUNet.

Fig. 7 illustrates qualitative comparisons between SCANeXt and UNETR++ on the ACDC dataset, focusing on four different cases. In the first row, UNETR++ exhibits under-segmentation of the left ventricle (LV) cavity, whereas SCANeXt accurately delineates all three categories. The second row presents a challenging sample with comparatively smaller sizes for all three heart segments. In this instance, UNETR++ mistakenly segments a significant portion of the LV and myocardium (Myo) regions as the right ventricle (RV), while SCANeXt provides a more accurate segmentation. In the third row, UNETR++ incorrectly segments a large normal region as RV, which is similar to the problem shown by UNETR++ in the case of the third row in Fig. 5. In the last row, UNETR++ over-segments the RV cavity, while SCANeXt delivers improved delineation, producing a segmentation that closely aligns with the ground truth. These qualitative examples show that SCANeXt successfully achieves precise segmentation of the three heart segments without either under-segmenting or over-segmenting. This underscores the superior capability of SCANeXt, which leverages a collaborative approach combining the dual-attention module and depthwise separable convolution to learn highly discriminative feature representations.

4.2.4. Comparison of the number of network parameters

During our comparative experiments on the Synapse dataset, we recorded the model complexities of five methods, denoted as TransUNet, UNETR, Swin-UNETR, nnFormer, and UNETR++. We measured the model complexities using two metrics, parameters, and FLOPs (floating-point operations). Combining the complexity measurements with the corresponding Dice Similarity Coefficient (DSC) obtained for each method, we created an intuitive visualization in Fig. 8, where the circle sizes represent the computational complexity of each method.

From Fig. 8, it can be observed that methods such as TransUNet and UNETR, which combine convolutional and attention mechanisms in transformer operators, exhibit lower DSC values and higher parameter counts. SwinUNETR, which replaces ViT in UNETR with Swin Transformer, enhances DSC while reducing the model's parameter count. However, due to the inclusion of 3D moving

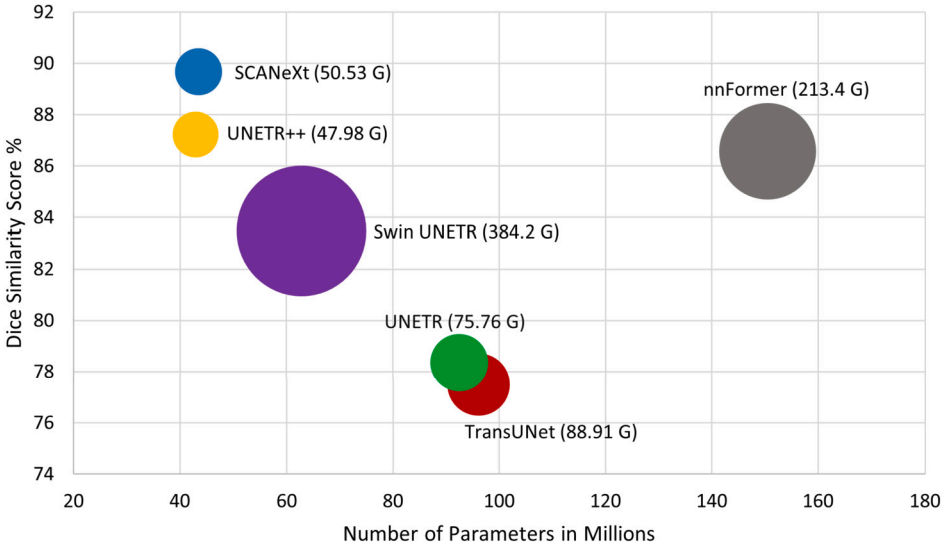

Fig. 8. The model size vs. DSC is shown in this plot. Circle size indicates computational complexity by FLOPs.

Table 4
Influence of DAM and DCM on the segmentation performance. w/o indicates that the module is not included.

Methods	RV	Myo	LV	Average
UNETR [25]	88.49	82.04	91.62	87.38
3D-UXNet [58]	77.97	81.84	92.41	84.07
UNETR++ [66]	91.47	86.47	94.25	90.73
SCANeXt w/o DAM	79.86	80.34	92.89	84.36
SCANeXt w/o DCM	92.77	88.24	94.02	91.68
SCANeXt	**96.51**	**92.19**	**96.84**	**95.18**

window operations, SwinUNETR requires 4 times more FLOPs than UNETR. On the other hand, nnFormer, a pure transformer-based method, reduces FLOPs by 44.5% compared to SwinUNETR while improving accuracy. However, its UNet structure with encoder and decoder constructed entirely using transformers leads to a model with more than twice the parameters of SwinUNETR. UNETR++, an improvement on UNETR with revised attention modules from ViT, significantly enhances DSC while reducing model parameters and required FLOPs.

Our proposed SCANeXt, building upon UNETR++, shows a 0.98% increase in model parameters and a 5.31% increase in FLOPs but achieves a 2.81% improvement in DSC. This further demonstrates the superiority of our newly proposed attention mechanism over UNETR++'s EPA and showcases the better performance achieved with the incorporation of DCM. Overall, our findings underscore the effectiveness of the proposed attention mechanism and its combination with DCM in SCANeXt, offering notable improvements over UNETR++ and contributing to enhanced segmentation results.

4.3. Ablation studies

We conducted ablation experiments on the ACDC segmentation dataset to investigate the impact of the DAM and DCM modules on SCANeXt's performance, with the results shown in Table 4.

When the DAM module was removed, leaving only the DCM module, SCANeXt transformed into a pure convolutional UNet segmentation network. On average, its performance improved by only 0.29% compared to the baseline 3D-UNXet based on depthwise convolution. The improvement was not significant, and in fact, its accuracy in segmenting Myocardium (Myo) was lower than 3D-UXNet. Both 3D-UNXet and SCANeXt with only DCM performed worse than the baseline UNETR. In contrast, retaining only the DAM module showed a performance improvement of 4.3% over UNETR and 0.95% over UNETR++, and outperformed both methods in segmenting all three subregions. When combining the DAM and DCM modules, SCANeXt's performance further improved by 3.5% compared to using only the DAM module. This suggests that the DCM module effectively extracted features from the feature maps obtained after applying the DAM module at multiple scales.

Overall, the results demonstrate that the DAM module plays a crucial role in enhancing segmentation performance, while the combination of DAM and DCM modules leads to the best overall performance improvement in SCANeXt.

4.4. Discussion

Through the comparative analysis of experimental results, SCANeXt demonstrates superior performance over state-of-the-art methods across datasets of varying difficulty, encompassing Synapse, BraTS, and ACDC. This compelling evidence underscores that the enhancement of the attention mechanism in the Transformer, coupled with its integration with depthwise convolution, represents a more advanced approach compared to existing methods. Ablation studies shed light on the pivotal role of two key modules, namely, DAM and DCM, indicating their potential to contribute to overall performance improvements. However, for a comprehensive model addressing medical image segmentation challenges, further validation across additional datasets is essential to ensure generalization, versatility, and robustness. Consequently, we anticipate that our research will attract future collaborations aimed at advancing algorithms for medical image segmentation.

5. Conclusion

This paper introduces SCANeXt as a versatile model designed for precise 3D medical image segmentation. We have re-designed a dual attention mechanism within the transformer framework. Notably, we introduced an innovative multi-Dconv head transposed attention to calculate attention along the channel dimension. In addition, we adapted the InceptionNeXt architecture to 3D medical images within the depthwise convolution module. Extensive experiments showcase SCANeXt's superiority over state-of-the-art approaches, particularly in terms of DSC and HD95 evaluation metrics. Through ablation experiments, we validated the proposed combination of Dual Attention-based Transformers and Depthwise Convolutions proves to be a superior approach in segmentation tasks. Our future research will predominantly concentrate on the development of more lightweight hybrid models based on ViT and CNNs. Additionally, we plan to explore segmentation methods that prioritize efficiency.

CRediT authorship contribution statement

Yajun Liu: Writing – original draft, Software, Methodology. **Zenghui Zhang:** Writing – review & editing, Supervision, Methodology, Investigation. **Jiang Yue:** Supervision, Resources. **Weiwei Guo:** Writing – review & editing, Writing – original draft, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

In our study, the datasets employed for experiments are readily accessible for download from public sources. Consequently, there was no need to create a dedicated database for storage. Readers can retrieve the pertinent datasets from the official website, adhering to the specified guidelines, to meet their specific requirements.

Acknowledgement

This work is supported in part by National Natural Science Foundation of China under Grant 62271311, 62071333, and 62201343.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., An image is worth 16×16 words: transformers for image recognition at scale, arXiv preprint, arXiv:2010.11929, 2020.
[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang Le Lu, Alan L. Yuille, Yuyin Zhou, Transunet: transformers make strong encoders for medical image segmentation, arXiv preprint, arXiv:2102.04306, 2021.
[3] Zhuangzhuang Zhang, Weixiong Zhang, Pyramid medical transformer for medical image segmentation, arXiv preprint, arXiv:2104.14702, 2021.
[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang, Swin-Unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.
[5] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, David Zhang, DS- TransUNet: dual swin transformer U-Net for medical image segmentation, IEEE Trans. Instrum. Meas. 71 (2022) 1–15.
[6] Yin Dai, Yifan Gao, Fayu Liu, Transmed: transformers advance multi-modal medical image classification, Diagnostics 11 (8) (2021) 1384.
[7] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, Kevin Smith, Is it time to replace CNNs with transformers for medical images?, arXiv preprint, arXiv:2108.09038, 2021.
[8] Chengeng Liu, Qingshan Yin, Automatic diagnosis of Covid-19 using a tailored transformer-like network, J. Phys. Conf. Ser. 2010 (2021) 012175, IOP Publishing.
[9] Xiaohong Gao, Yu Qian, Alice Gao, COVID-VIT: classification of Covid-19 from ct chest images based on vision transformer models, arXiv preprint, arXiv: 2107.01682, 2021.
[10] Zhicheng Zhang, Lequan Yu, Xiaokun Liang, Wei Zhao, Lei Xing, Transct: dual-path transformer for low dose computed tomography, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24, Springer, 2021, pp. 55–64.
[11] Yilmaz Korkmaz, Salman U.H. Dar, Mahmut Yurt, Muzaffer Özbey, Tolga Cukur, Unsupervised MRI reconstruction via zero-shot learned adversarial transformers, IEEE Trans. Med. Imaging 41 (7) (2022) 1747–1763.

[12] Yanmei Luo, Yan Wang, Chen Zu, Bo Zhan, Xi Wu, Jiliu Zhou, Dinggang Shen, Luping Zhou, 3D transformer-GAN for high-quality pet reconstruction, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24, Springer, 2021, pp. 276–285.
[13] Alper Güngör, Baris Askin, Damla Alptekin Soydan, Emine Ulku Saritas, Can Barış Top, Tolga Çukur, TranSMS: transformers for super-resolution calibration in magnetic particle imaging, IEEE Trans. Med. Imaging 41 (12) (2022) 3562–3574.
[14] Xuzhe Zhang, Xinzi He, Jia Guo, Nabil Ettehadi, Natalie Aw, David Semanek, Jonathan Posner, Andrew Laine, Yun Wang Ptnet, A high-resolution infant MRI synthesizer based on transformer, arXiv preprint, arXiv:2105.13993, 2021.
[15] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, Salah A. Baker, Vtgan: semi-supervised retinal image synthesis and disease prediction using vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3235–3245.
[16] Nicolae-Catalin Ristea, Andreea-Iuliana Miron, Olivian Savencu, Mariana-Iuliana Georgescu, Nicolae Verga, Fahad Shahbaz Khan, Radu Tudor Ionescu, Cytran: cycle-consistent transformers for non-contrast to contrast ct translation, arXiv preprint, arXiv:2110.06400, 2021.
[17] Onat Dalmaz, Mahmut Yurt, Tolga Çukur, Resvit: residual vision transformers for multimodal medical image synthesis, IEEE Trans. Med. Imaging 41 (10) (2022) 2598–2614.
[18] Junyu Chen, Yufan He, Eric C. Frey, Ye Li, Yong Du, Vit-v-net: vision transformer for unsupervised volumetric medical image registration, arXiv preprint, arXiv:2104.06468, 2021.
[19] Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, Yong Du, Transmorph: transformer for unsupervised medical image registration, Med. Image Anal. 82 (2022) 102615.
[20] Yungeng Zhang, Yuru Pei, Hongbin Zha, Learning dual transformer network for diffeomorphic registration, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24, Springer, 2021, pp. 129–138.
[21] Gijs van Tulder, Yao Tong, Elena Marchiori, Multi-view analysis of unregistered medical images using cross-view transformers, 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. 104–113.
[22] Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu, Sen Song, Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network, IEEE Trans. Neural Netw. Learn. Syst. 30 (11) (2019) 3484–3495.
[23] Shijie Liu, Hongyu Zhou, Xiaozhou Shi, Junwen Pan, Transformer for polyp detection, arXiv preprint, arXiv:2111.07918, 2021.
[24] Yutong Xie, Jianpeng Zhang, Chunhua Shen, Yong Xia, Cotr: efficiently bridging CNN and transformer for 3D medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 171–180.
[25] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, Daguang Xu, UNETR: transformers for 3D medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.
[26] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, Yizhou Yu, nnFormer: interleaved transformer for volumetric segmentation, arXiv preprint, arXiv:2109.03201, 2021.
[27] Himashi Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, Mehrtash Harandi, A robust volumetric transformer for accurate 3D tumor segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 162–172.
[28] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, Daguang Xu, Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images, in: International MICCAI Brainlesion Workshop, 2022.
[29] A. Burakhan Koyuncu, Han Gao, Atanas Boev, Georgii Gaikov, Elena Alshina, Eckehard Steinbach, Contextformer: a transformer with spatio-channel attention for context modeling in learned image compression, in: European Conference on Computer Vision, Springer, 2022, pp. 447–463.
[30] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, Lu Yuan Davit, Dual attention vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 74–92.
[31] Gaurav O. Gajbhiye, Abhijeet V. Nandedkar, Generating the captions for remote sensing images: a spatial-channel attention based memory-guided transformer approach, Eng. Appl. Artif. Intell. 114 (2022) 105076.
[32] Tongxue Zhou, Shan Zhu, Uncertainty quantification and attention-aware fusion guided multi-modal MR brain tumor segmentation, Comput. Biol. Med. (2023) 107142.
[33] Xin Hua, Zhijiang Du, Hongjian Yu, Jixin Ma, Fanjun Zheng, Cheng Zhang, Qiaohui Lu, Hui Zhao, Wsc-trans: a 3D network model for automatic multi-structural segmentation of temporal bone ct, arXiv preprint, arXiv:2211.07143, 2022.
[34] Zhengyong Huang, Sijuan Zou, Guoshuai Wang, Zixiang Chen, Hao Shen, Haiyan Wang, Na Zhang, Lu Zhang, Fan Yang, Haining Wang, et al., ISA-Net: improved spatial attention network for PET-CT tumor segmentation, Comput. Methods Programs Biomed. 226 (2022) 107129.
[35] Reza Azad, René Arimond, Ehsan Khodapanah Aghdam, Amirhosein Kazerouni, Dorit Merhof, Dae-former: dual attention-guided efficient transformer for medical image segmentation, arXiv preprint, arXiv:2212.13504, 2022.
[36] Gorkem Can Ates, Prasoon Mohan, Emrah Celik, Dual cross-attention for medical image segmentation, arXiv preprint, arXiv:2303.17696, 2023.
[37] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[38] Hasib Zunair, A. Ben Hamza, Sharp u-net: depthwise convolutional network for biomedical image segmentation, Comput. Biol. Med. 136 (2021) 104699.
[39] Ho Hin Lee, Quan Liu, Shunxing Bao, Qi Yang, Xin Yu, Leon Y. Cai, Thomas Li, Yuankai Huo, Xenofon Koutsoukos, Bennett A. Landman, Scaling up 3D kernels with Bayesian frequency re-parameterization for medical image segmentation, arXiv preprint, arXiv:2303.05785, 2023.
[40] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, Klaus Maier-Hein, Mednext: transformer-driven scaling of ConvNets for medical image segmentation, arXiv preprint, arXiv:2303.09975, 2023.
[41] Weihao Yu, Pan Zhou, Shuicheng Yan, Xinchao Wang, Inceptionnext: when inception meets ConvNeXt, arXiv preprint, arXiv:2303.16900, 2023.
[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[43] Olaf Ronneberger, Philipp Fischer, Thomas 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.
[44] Fausto Milletari, Nassir Navab, Seyed-Ahmad 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.
[45] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[46] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, Olaf 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.
[47] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, Pheng-Ann Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imaging 37 (12) (2018) 2663–2674.
[48] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, Chen Wu, Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data, ISPRS J. Photogramm. Remote Sens. 162 (2020) 94–114.

[49] Nabil Ibtehaz, M. Sohel Rahman, Multiresunet: rethinking the u-net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.
[50] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, Jianming Liang, Unet++: redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.
[51] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Jian Wu, Unet 3+: A Full-Scale Connected Unet for Medical Image Segmentation, 2020.
[52] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, et al., Attention U-Net: learning where to look for the pancreas, arXiv preprint, arXiv:1804.03999, 2018.
[53] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, Klaus H. Maier-Hein, nNU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[54] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al., Stu-net: scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training, arXiv preprint, arXiv:2304.06716, 2023.
[55] Chao Huang, Hu Han, Qingsong Yao, Shankuan Zhu, S. Kevin Zhou, 3D U^2-Net: a 3D universal U-Net for multi-domain medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 291–299.
[56] Hao Li, Yang Nan, Guang Yang, Lkau-net: 3D large-kernel attention-based u-net for automatic MRI brain tumor segmentation, in: Annual Conference on Medical Image Understanding and Analysis, Springer, 2022, pp. 313–327.
[57] Rongfang Wang, Zhaoshan Mu, Kai Wang, Hui Liu, Zhiguo Zhou, Shuiping Gou, Jing Wang, Licheng Jiao, ASF-LKUNet: adjacent-scale fusion U-Net with large-kernel for multi-organ segmentation, Available at SSRN 4592440.
[58] Ho Hin Lee, Shunxing Bao, Yuankai Huo, Bennett A. Landman, 3D UX-Net: a large kernel volumetric ConvNet modernizing hierarchical transformer for medical image segmentation, arXiv preprint, arXiv:2209.15076, 2022.
[59] Yiqing Wang, Zihan Li, Jieru Mei, Zihao Wei, Li Liu, Chen Wang, Shengtian Sang, Alan L. Yuille, Cihang Xie, Yuyin Zhou, SwinMM: masked multi-view with swin transformers for 3D medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 486–496.
[60] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, David Zhang, Ds-transunet: dual swin transformer u-net for medical image segmentation, IEEE Trans. Instrum. Meas. 71 (2022) 1–15.
[61] Hao Du, Jiazheng Wang, Min Liu, Yaonan Wang, Erik Meijering, Swinpa-net: swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[62] Jing Zhang, Qiuge Qin, Qi Ye, Tong Ruan, St-UNet: swin transformer boosted u-net with cross-layer feature enhancement for medical image segmentation, Comput. Biol. Med. 153 (2023) 106516.
[63] Haojie Tao, Keming Mao, Yuhai Zhao, DBT-UNETR: double branch transformer with cross fusion for 3D medical image segmentation, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2022, pp. 1213–1218.
[64] Zihan Li, Dihan Li, Cangbai Xu, Weice Wang, Qingqi Hong, Qingde Li, Jie Tian, Tfcns: a CNN-transformer hybrid network for medical image segmentation, in: International Conference on Artificial Neural Networks, Springer, 2022, pp. 781–792.
[65] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, Missformer: an effective medical image segmentation transformer, arXiv preprint, arXiv:2109.07162, 2021.
[66] Shaker Abdelrahman, Maaz Muhammad, Rasheed Hanoona, Khan Salman, Yang Ming-Hsuan, Khan Fahad Shahbaz, UNETR++: delving into efficient and accurate 3D medical image segmentation, arXiv preprint, arXiv:2212.04497, 2022.
[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[68] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, Restormer: efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.
[69] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words, arXiv preprint, arXiv:2010.11929, 2020.
[70] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[71] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang, Swin-UNet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision Workshops, 2022.
[72] Bennett Landman, Zhoubing Xu, J. Igelsias, Martin Styner, T. Langerak, Arno Klein, Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, 2015.
[73] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, Pierre-Marc Jodoin, 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.
[74] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M.S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, Koen Van Leemput, The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2015) 1993–2024.
[75] Project-MONAI, Medical open network for AI, https://github.com/Project-MONAI/MONAI, 2020.