HMDA: A Hybrid Model with Multi-scale Deformable Attention for Medical Image Segmentation

Mengmeng Wu, Tiantian Liu, Xin Dai, Chuyang Ye, Jinglong Wu, Shintaro Funahashi and Tianyi Yan

***Abstract*— Transformers have been applied to medical image segmentation tasks owing to their excellent long-range modeling capability, compensating for the failure of Convolutional Neural Networks (CNNs) to extract global features. However, the standardized self-attention modules in Transformers, characterized by a uniform and inflexible pattern of attention distribution, frequently lead to unnecessary computational redundancy with high-dimensional data, consequently impeding the model's capacity for precise concentration on salient image regions. Additionally, achieving effective explicit interaction between the spatially detailed features captured by CNNs and the long-range contextual features provided by Transformers remains challenging. In this architecture, we propose a Hybrid Transformer and CNN architecture with Multi-scale Deformable Attention(HMDA), designed to address the aforementioned issues effectively. Specifically, we introduce a Multi-scale Spatially Adaptive Deformable Attention (MSADA) mechanism, which attends to a small set of key sampling points around a reference within the multi-scale features, to achieve better performance. In addition, we propose the Cross Attention Bridge (CAB) module, which integrates multi-scale transformer and local features through channel-wise cross attention enriching feature synthesis. HMDA is validated on multiple datasets, and the results demonstrate the effectiveness of our approach, which achieves competitive results compared to the previous methods.**

This work was supported by Key-Area Research and Development Program of Guangdong Province (2023B0303030002); the STI 2030-Major Projects (grant number 2022ZD0208500); the National Natural Science Foundation of China (grant numbers U20A20191, 62336002, 82071912, 62373056, 62406025); the China Postdoctoral Science Foundation (grant numbers 2023TQ0027, 2024M754099); the Beijing Municipal Natural Science Foundation (grant number 7242273); the Beijing Natural Science Foundation (grant numbers IS23114, IS23115); and the Fundamental Research Funds for the Central Universities (grant number 2022CX11008). (Corresponding authors: Tianyi Yan and Tiantian Liu.)

Mengmeng Wu is with the School of Life, Beijing Institute of Technology, Beijing 100081, China. (e-mail:mengmengwubit@163.com).

Tiantian Liu, Xin Dai, Jinglong Wu, and Tianyi Yan are with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China. (e-mail:tiantian2bit@bit.edu.cn, 3120231412@bit.edu.cn, wujl@bit.edu.cn, yantianyi@bit.edu.cn)

Shintaro Funahashi is with the Advanced Research Institute for Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China; Department of Cognitive and Behavioral Sciences, Graduate School of Human and Environmental Science, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan; Kokoro Research Center, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan. (e-mail:funahashi@bit.edu.cn).

Chuyang Ye is with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China. (e-mail:chuyang.ye@bit.edu.cn).

***Index Terms*— medical image segmentation, hybrid model, multi-scale deformable attention, cross attention bridge**

I. Introduction

MEDICAL image segmentation is crucial for assisting physicians in disease assessment and optimizing preventive and control measures [1]. This technology enables the depiction of anatomical structures and organs with specific shapes, varying appearances, or different pathology levels on a pixel/voxel-level basis. Earlier segmentation algorithms heavily incorporated geometric shape priors and similar techniques [2]. However, these methods exhibit low segmentation quality. Advances in deep learning and computer vision techniques have facilitated medical image segmentation methods with remarkable progress [3], achieving more accurate segmentation results.

A convolutional neural network (CNN) is a typical deep learning method used for medical image segmentation [4], [5], [6], [7]. The CNN approaches the medical image segmentation task by treating it as a pixel/voxel-wise classification problem. In particular, U-Net [8], a famous encoder-decoder network architecture, has particularly excelled at performing medical image segmentation. Due to its effective utilization of skip connections and multi-scale feature extraction, it has arguably become the dominant choice for medical image segmentation [9], [3], [10], [11], [12]. Further iterations of U-Net architecture have been developed to enhance its design efficacy. For example, Ibtehaz et al. [13] introduced the MultiResUNet network by reconfiguring the skip connections in the conventional U-Net; Huang et al. proposed UNet++ [14] to optimize encoder-decoder connections and maximize the utilization of information between high-level and low-level layers. Although CNNs deliver remarkable results, their reliance on the limited receptive field and the inherent inductive bias can hinder their ability to effectively capture the extensive long-range dependencies present in data [26].

Recently, the Transformer architecture, which has achieved success in natural language processing and computer vision [17], [18], [19], [20], [22], [24], [25], has been explored for use in medical image segmentation tasks. The Transformer architecture employs a self-attention mechanism to model the nonlocal interactions between distant image regions and thus enable the capture of long-range dependencies. In medical

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3469230

image segmentation, researchers integrate Transformers with CNN networks to enhance segmentation performance. For example, TransUNet [26] is a pioneering hybrid model that employs a Transformer to encode the high-level features extracted by CNNs, which leverages this encoded information to refine the segmentation process. This pioneering integration has paved the way for methodologies like TransCeption [55] and nnFormer [27], which stack convolutional and Transformer layers in sequence. Simultaneously, other approaches such as UNETR [57] and SwinUNETR [28] have innovatively replaced CNNs with Transformers in critical components like the encoder or decoder. These advancements leverage the Transformers' prowess in capturing global information, thereby enhancing the feature representation capabilities of CNNs. However, these models, which utilize methods such as window-based attention and factorized attention mechanisms to decrease the computational complexity, still fail to effectively resolve computational redundancy issues, particularly the ability to quickly and accurately focus on salient regions within an image. As the paper [21] points out, multi-head self-attention not only incurs high memory usage and computational costs but also risks being influenced by irrelevant parts outside the region of interest, which can lead to computational redundancy. Furthermore, the integration of Transformer modules with CNNs, either in series or through simple skip connections, lacks the direct interaction necessary to fully leverage the complementarity of both architectures [31].

In this work, we develop a Hybrid Transformer and CNN architecture with Multi-scale Deformable Attention (HMDA), which employs a classic encoder-decoder framework to synergistically employ convolutional layers and the deformable-attention mechanism. This structure captures local and global image features across multiple scales, a strategy proven effective in various visual tasks [33], [34]. In the field of image processing, deformable convolution offers an efficient mechanism for attending to sparse spatial locations, as demonstrated by [38]. We specifically introduce a Multi-scale Spatial Adaptive Deformable Attention (MSADA) mechanism, integrating deformable operation capabilities with the Transformer's advanced relational modeling techniques, which reduces the redundancy of the attention process. Furthermore, we propose a Cross Attention Bridge (CAB) module that explicitly models the channel interdependencies, which effectively facilitates information complementarity between Transformers and CNNs. This module incorporates a feature recalibration mechanism, through which the network learns to selectively emphasize informative features using global channel information while suppressing less useful features [39]. The main contributions of our work are summarized as follows:

- We propose a novel hybrid model architecture, HMDA, which features a Multi-scale Spatial Adaptive Deformable Attention (MSADA) mechanism. Compared to other hybrid approaches, the MSADA mechanism reduces computational redundancy by using a deformable attention mechanism that selectively targets a small set of key sampling points across multi-scale input features.
- We introduce a Cross Attention Bridge called (CAB) module, which enhances the quality of the network's generated representations by explicitly modeling the channel interdependencies between the network's local and global features. This enhancement aims to provide the decoder with more detailed and pertinent information, thereby better restoring the fine details of the image.
- Our method achieves competitive results with those of the previous methods, as demonstrated through an extensive evaluation conducted on three public benchmark datasets, showcasing its effectiveness and robustness.

The remainder of this paper is organized as follows. In Section II, we review the related work on medical image segmentation. Section III provides a detailed presentation of the proposed method. In Section IV, we validate the proposed method through various experiments. In Section V, we draw the conclusions of this paper.

II. Related Work

A. CNN-Based Medical Image Segmentation

With its extensibility and symmetry, the U-Net framework [8] offers a wide range of design possibilities and has become a cornerstone in the medical image segmentation field due to its performance and elegant structure. Building upon the success of U-Net, researchers have proposed various extensions and modifications to address specific challenges in medical image segmentation. For instance, Chen et al. proposed a novel CNN architecture called Dense-Res-Inception Net (DRINet) [9], which addresses the problem by which standard convolution is insufficiently sensitive to intensity, location, shape, and size differences. The nnUNet algorithm [56] optimizes the processing pipeline based on input data and algorithmic design considerations without modifying the model architecture. Ibtehaz et al. [13] introduced the MultiResUNet network, which incorporates parallel pathways at different levels of the network to handle multi-scale feature maps. Studies [36] and [37] have explored the significance of data deformation augmentation techniques for boosting the accuracy of deep learning models in segmenting medical images.

To expand the receptive field and enhance the network's capability to capture long-range dependencies, several innovative approaches have been proposed. Wang et al. [35] proposed a multi-scale context-aware attention model for medical image segmentation. Oktay et al. [16] proposed an attention gate model by invoking an attention mechanism for medical image segmentation. Li et al. [39] proposed squeeze-and-excitation networks (SE-Net) to introduce channel attention to the image analysis field. Chen et al. [40] proposed the FED-Net model using SE blocks to implement feature channel attention. Wang et al. [42] introduced a novel non-local U-Net architecture to address the limitations associated with local convolution. Recently, the ConvNeXt [43] framework has sought to revitalize the traditional Convolutional Neural Network (ConvNet) design by integrating architectural elements inspired by Transformer models. MedNeXt [44] enhances the ConvNeXt architecture by introducing a Transformer-inspired large kernel segmentation, optimizing performance in medical imaging analysis.

B. Transformer-Based Medical Image Segmentation

Compared with CNNs, Transformers have recently demonstrated significant potential in medical imaging applications due to their larger effective receptive fields and superior understanding of contextual information. Medical image segmentation algorithms based on Transformer architecture can be categorized into two types: pure Transformer-based methods and hybrid architectures. Several pure Transformer models are available. Swin-Unet [30] is a U-Net-like architecture that leverages the Swin Transformer by utilizing hierarchical representations and self-attention mechanisms. Several models, such as the DS-TransUNet [48] and TransDeepLab [49], are pure transformer-based architectures founded on the Swin Transformer. In [59], a gated axial-attention mechanism that captures long-range spatial dependencies was utilized. There are also pure transformer-based methods like MISS-Former [47] and DAE-Former, which leverage the power of transformer architectures to address various challenges in the field of medical image segmentation. The Transformer reliance on large-scale datasets indicates a lower inductive bias for capturing local visual cues states [29]. However, in the medical imaging domain, there is a scarcity of data compared to the vast repositories available for natural images. To address these challenges, researchers have turned to hybrid architectures that combine the strengths of the Transformer with CNN methods to enhance the model's performance and generalization capabilities under conditions of limited data.

TransUNet [26] revolutionizes the domain of medical image segmentation by employing Transformers on CNN-derived low-resolution feature maps, thereby pioneering the incorporation of Transformers into this field. TransFuse [31] fuses information acquired from local and holistic contexts through a Bi-Fusion mechanism. HiFormer [32] utilizes a Swin Transformer and a CNN-based encoder to create two multi-scale feature representations, integrating a dual-layer fusion (DLF) module into the encoder-decoder skip connection. UNETR [57] architecture leverages pure Transformers as encoders to learn sequential representations of the input, effectively capturing global multi-scale information. SwinUNETR [28] fuses a Swin Transformer encoder with a U-shaped CNN decoder, creating a synergistic architecture for medical image segmentation. TransCeption [55] is designed with a hierarchical structure that integrates an inception-style CNN module in series with a Multi Branch (MB) Transformer Block that utilizes factorized attention mechanisms. The nnFormer [27] combines the interleaved convolutions and self-attention, further enhancing its capabilities with a dual focus on local and global volume-based self-attention to effectively learn volumetric representations. Zhu et al. [51] employed multi-head self-attention and sparse dynamic adaptive fusion in a 3D extended shifted window strategy. Zhu et al. [52] incorporated a Swin Transformer-based semantic segmentation module for deep feature extraction and a CNN-based edge detection module to accentuate the edge features.

However, few studies have focused on effectively resolving computational redundancy issues, particularly the ability to quickly and accurately focus on salient regions within an image [25]. Moreover, addressing how to harness the complementary advantages of local pattern recognition and overall contextual understanding is crucial for advancing medical image segmentation [29]. The proposed MSADA method alleviates computational redundancy by integrating deformable operation with the Transformer's relational modeling capabilities. At the same time, the Cross Attention Bridge (CAB) explicitly models the channel interdependencies between the convolutional and attentional features of the network, to provide the decoder with more detailed and pertinent information, thereby better restoring the fine details of the image.

III. Methodology

In this section, we introduce the architecture of HMDA, which consists of three components: a convolutional encoder with a Two-Branch Residual (TB-Residual) module, a Transformer encoder with a Multi-scale Spatial Adaptive Deformable Attention (MSADA) module, and a decoder with Cross Attention Bridging (CAB) module, as illustrated in Fig 1. The TB-Residual module enables comprehensive hierarchical feature representations to be obtained for local feature extraction purposes. Fine-grained information at different scales is captured and utilized by the MSADA module to incorporate contextual understanding at the global level. Finally, the first three layers of the features extraction incorporate the CAB module, harnessing the channel-wise feature between CNN and Transformer, thereby enhancing channel guidance and information filtration for CNN features. This approach is particularly crucial for preserving high-resolution details that could otherwise be compromised by the global feature extraction typically associated with Transformer architectures. The fourth layer retains Transformer features, ensuring that our model captures the broader contextual information necessary for precise segmentation, similar to the approach used in TransUNet [26].

A. Encoder

1) TB-Residual Enhanced Local Encoder: To explore the impact of convolutional design on the hybrid model, we utilize a TB-Residual convolutional block that aggregates transformations with the same topology to extract local spatial features. As demonstrated in the literature [53], incorporating transformations with identical topologies in an aggregate manner is proven to be more effective than simply increasing the depth or width of a network architecture. This principle also simplifies the selection of hyperparameters, such as the number of filters and stride sizes, by reducing the range of free choices available. An image I with a size of $C \times H \times W$, where H and W are the spatial dimensions of the 2D image features, and C is the number of channel dimensions, is initially fed into a 3×3 convolution (Conv) with a stride of 1, followed by Batch Normalization (BN) and a Rectified Linear Unit (ReLU) activation function, which is denoted as the "Conv-BN-ReLU" unit. This process is repeated twice, and a downsampling operation is performed to produce an encoded feature map $I' \in \mathbb{R}^{C' \times H/2 \times W/2}$. Next, the feature map I' is fed into four encoder stages with the TB-Residual module, which aggregates transformations with two same topologies. As these feature maps progress through these stages, their spatial resolutions are successively downsampled to $1/4, 1/8, 1/16$ and $1/32$ of the original, as shown in Fig 1.

The TB-Residual block consists of two same topologies, as shown in Fig 2, each of which comprises four fundamental Conv-BN-ReLU units. Given an input I' to the single-path network with dimensions of $C' \times H/2 \times W/2$, the input is processed through the stacking Conv-BN-ReLU units. The resulting output is then concatenated with the original input along the channel dimension with dimensions of $2C' \times H \times W$. In the l-th stage of the process, the computation is defined by the following equation:

$$x^l = Add(Cat[x^{\mathrm{l}-1}, x^{\mathrm{o}_1}]), (Cat[x^{\mathrm{l}-1}, x^{\mathrm{o}_2}]) \tag{1}$$

where Cat denotes the concatenation of the feature maps from $x^{\mathrm{l}-1}$ and its corresponding outputs along the channel dimension. x^{o_1} and x^{o_2} represent the outputs from the two distinct branches of the l-th block. This concatenated result is then added together to yield the final feature map for stage l, which is formulated with dimensions $x^l \in \mathbb{R}^{2C'_l \times H_l \times W_l}$.

2) MSADA Enhanced Global Encoder: The Transformer attention established between each region and all other locations within the input can result in an over-generalization of information. To address this, we proposed a Multi-scale Spatial Adaptive Deformable Attention (MSADA) module, focusing on a small set of critical sampling points around a reference. This module is agnostic to the spatial size of the feature maps. The strategy involves allocating a fixed set of keys for each query, ensuring a targeted concentration on the salient image regions, as shown in Fig 3.

Within the Transformer architecture, known for its sequence-to-sequence information processing capabilities, the input features derived from the CNN-encoder represented as $\{x^l\}_{l=1}^{L}$ with L layers, undergo a transformation to be flattened into a sequence $\{f^l\}_{l=1}^{L}$. After that, the dimensions of the input features transition from $C_l \times H_l \times W_l$ to $C_l \times (H_l \times W_l)$, resulting in the loss of spatial information that is imported for image segmentation. To address this issue, we employ sine and cosine functions with different frequencies [15] to embed the positional information $\{p^l\}_{l=1}^{L}$ into the flattened features f^l, where pos is the position and i is the dimension:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i}/d_{model}}\right) \tag{2}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i}/d_{model}}\right) \tag{3}$$

Furthermore, to discern the specific feature level of each spatial pixel, we incorporate a normalized scale embedding, denoted as $\{s^l\}_{l=1}^{L}$. This embedding is randomly initialized and is trained concurrently with the network. Distinct from the approach in [25], our method employs normalized scale embedding designed to align with the positional encoding's cyclical properties, ensuring a balanced model response to diverse positional cues.

The enhanced feature representation is achieved by augmenting each feature map $\{f^l\}_{l=1}^{L}$ with positional and scale-specific embeddings, denoted as $\{F^l\}_{l=1}^{L} = \{f^l\}_{l=1}^{L} + \{p^l\}_{l=1}^{L} + \{s^l\}_{l=1}^{L}$. After a 1×1 convolutional layer, the multi-scale features are transformed to possess a consistent channel depth of $C = 96$. This operation facilitates the aggregation of the multi-scale input features into a singular representation, denoted as $F = concatenate(F^1, F^2 \ldots, F^L)$, with dimensions of $C \times (H_1 W_1 + \ldots + H_L W_L)$. This aggregation results in a high-dimensional representation that poses challenges for computing self-attention, as the dimensions are too large, and the model may struggle to focus on the target regions.

To address this, we initially draw inspiration from [45] using the spatial reduction (SR) module to reduce the feature dimensions. The SR mechanism in stage l is articulated by the following formulation: $SR(F^l) = \mathrm{Reshape}(F^l, R_l) W_S$, where $F^l \in \mathbb{R}^{C \times H_l \times W_l}$ represents an input sequence and R_l denotes reduction ratio of the attention layer l. The operation $\mathrm{Reshape}(F^l, R_l)$ reshapes the input sequence F^l into a new dimension of $C \times (H_l W_l / R_l)$. W_S is a set of learnable weights corresponding to a 1×1 convolutional operation. Subsequently, we introduce a deformable attention mechanism that adaptively allocates focus to a fixed small set of key sampling points around a reference point within multi-scale features [25].

Given an multi-scale input feature maps $\{F^l\}_{l=1}^{L}$, where $F^l \in \mathbb{R}^{C \times H_l \times W_l}$ and l indexes the input feature level. Let $q \in \Omega_q$ denote a query element with representation feature F_q, where Ω_q defines the set of all query elements. The index k refers to the key elements within the sampled set, while K, being much smaller than the dimensions $H_l \times W_l$, represents the uniform total key count across all scales. The multi-scale deformable attention module of the m-th head is computed as follows:

$$f_m(F_q, \hat{r}_{lq}, \{F^l\}_{l=1}^{L}) = \sum_{l=1}^{L} \sum_{k=1}^{K} A_{lqk} \cdot W F^l(\hat{r}_{lq} + s_{lqk}) \tag{4}$$

where $\hat{r_{lq}} \in [0,1]^2$ denotes normalized 2-dimensional coordinates of reference point for each query element q within the scale l, and different scales of feature maps have their respective reference points. The $s_{lqk} \in \mathbb{R}^2$ indexes the sampling offset of the k-th sampling point at the l-th feature level, which are learned from the linear projection operator over the query features F_q. Given that $\hat{r}_{lq} + s_{lqk}$ is fractional, we employ bilinear interpolation for sampling the spatial locations within the representation features F^l. $W \in \mathbb{R}^{C \times C_v}$ denotes learnable weights with $C_v = C/M$, where M is the total number of attention heads. Thus, $WF^l(\hat{r}_{lq} + s_{lqk})$ represents the expression of the sampled value features. The attention weight A_{lqk} is calculated via the linear projection operator of MK channels over the query features F_q, which are then input into a softmax function. The result is a normalized distribution expressed as $\sum\limits_{l=1}^{L} \sum\limits_{k=1}^{K} A_{lqk} = 1$.

Then, an m-head deformable attention mechanism is formulated as $\theta(f_1, f_2, \cdots, f_M)$, where θ is a linear projection layer, to increase the capacity of the model for capturing

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3469230

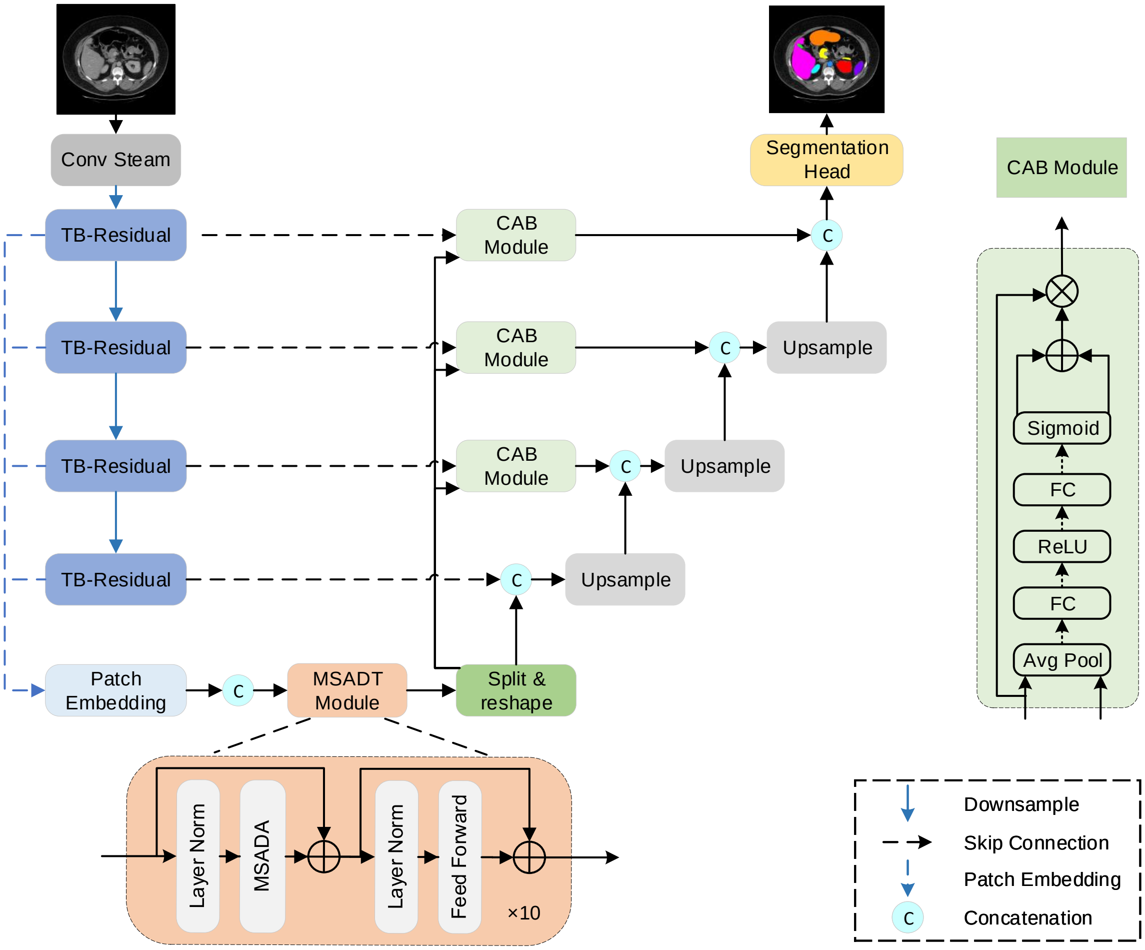


Fig. 1. Visualization of the overall pipeline of the proposed approach. The encoder consists of four stages, each of which contains a TB-Residual encoder that is responsible for extracting multi-scale feature maps. The flattened multi-scale feature maps are then sequentially processed by a Multi-scale Spatial Adaptive Deformable Transformer (MSADT) encoder with MSADA. It captures long-range dependencies through its generated features, which are integrated with the CNN-encoder features using the CAB module.

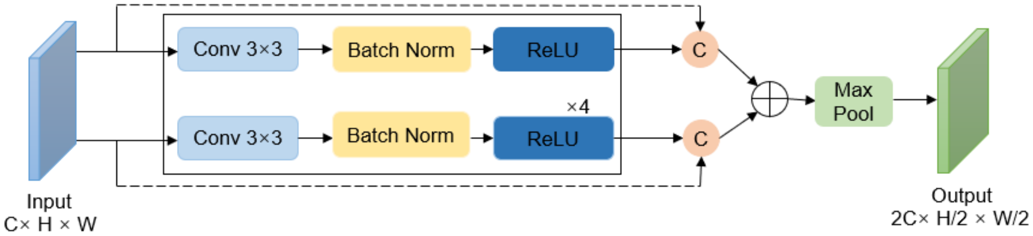


Fig. 2. The proposed TB-Residual encoder module consists of two single-path networks.

diverse patterns and improve its flexibility for capturing complex relationships. This mechanism effectively replaces the traditional self-attention layer in the Transformer architecture.

B. CAB: Cross Attention Bridge Module

The CAB module effectively integrates channel-wise features from both CNN and Transformer, enhancing channel guidance and information filtration for the CNN features, as illustrated in the right of Fig 1. This enhancement aims to provide the decoder with more detailed and pertinent information, thereby better restoring the fine details of the image. Specifically, the CAB module includes a feature recalibration mechanism. This mechanism enables the network to dynamically emphasize features that carry significant information, leveraging global and local channel insights, while concurrently suppressing features that are less relevant [39].

In our model, for the i-th level CNN output $L_i \in \mathbb{R}^{C\times H\times W}$ and the i-th Transformer output $G_i \in \mathbb{R}^{C\times H\times W}$, we employ Global Average Pooling (GAP) to squeeze global spatial

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3469230

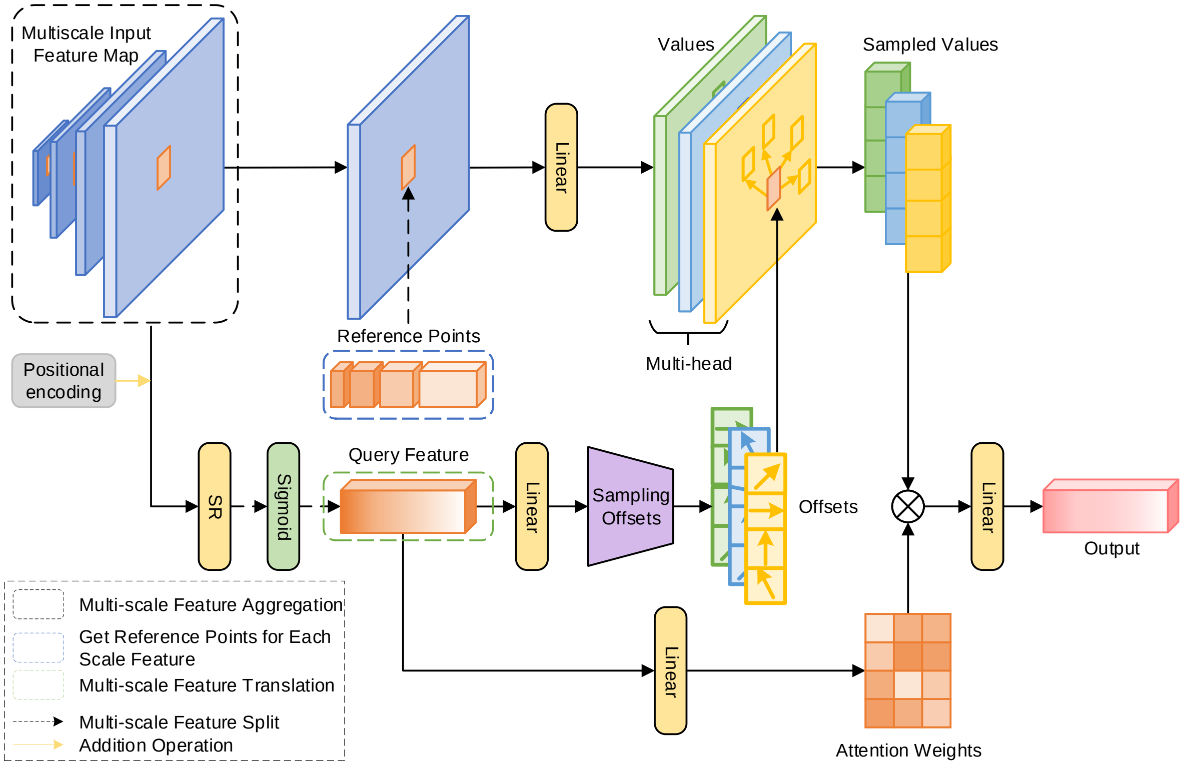


Fig. 3. The proposed Multi-scale Spatial Adaptive Deformable Attention (MSADA) module begins by receiving input from the CNN module, which consists of features at four different scales. These features are flattened into sequences and concatenated together. A spatial reduction transformation is applied, and the offset and attention weights are calculated. Subsequently, the sampled values are multiplied by the attention weights.

information into a channel descriptor with its k-th channel:

$$Z(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_k(i,j) \quad (5)$$

By separate aggregation of the two types of features along the spatial dimension, we derive $z_l = Z(L_i) \in \mathbb{R}^{C\times1\times1}$ for the CNN feature and $z_g = Z(G_i) \in \mathbb{R}^{C\times1\times1}$ for the Transformer feature. After the aggregation phase, an excitation operation followed, which is characterized by a simple self-gating mechanism designed for the acquisition of channel-specific weights The formulation of this mechanism is articulated as follows:

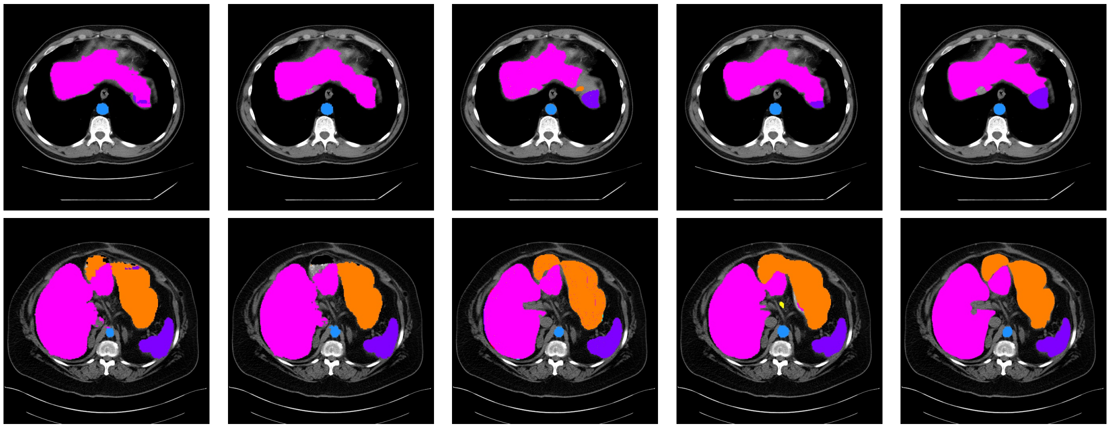
$$S(Z, W) = \varPhi(W_2(\varphi(W_1 Z))) \quad (6)$$

where φ refers to the ReLU function, $W_1 \in \mathbb{R}^{(C/r)\times1\times1}$ and $W_2 \in \mathbb{R}^{C\times1\times1}$ being weights of two Linear layers to capture the channel-wise dependencies. The Sigmoid function $\varPhi$ normalizes the output to a range between 0 and 1, reflecting the significance of each channel. The resulting attention weights, $S(z_l)$ and $S(z_g)$, both in $\mathbb{R}^{C\times1\times1}$, are then combined and applied to rescale L_i, yielding the final output: $\widetilde{X_k} = (S(z_l) + S(z_g)) \cdot L_i$.

IV. Experiments and Settings

A. Datasets

Our experimental analysis encompasses three datasets: the Synapse Multi-Organ Segmentation dataset (Synapse), the Automated Cardiac Diagnosis Challenge dataset (ACDC), and the SegTHOR dataset. The Multi-Organ Synapse dataset [46] is a collection of 30 abdominal CT scans within 8 abdominal organs (the aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, and stomach), comprising a total of 3779 axial contrast-enhanced clinical CT images. Each volumetric sample within the dataset consists of 85 to 198 slices. For training and testing purposes, the dataset is distributed according to the literature [26], consisting of 18 cases for training and 12 cases for validation. The ACDC dataset [58] consists of 100 MRI scans obtained from diverse patients. Each scan is meticulously labeled with respect to three distinct cardiac organs: the left ventricle (LV), right ventricle (RV), and myocardium (MYO). In accordance with the established methodologies, 70 cases from the dataset are allocated for training purposes, whereas 10 cases are allocated for validation. The remaining 20 cases are reserved for testing the performance of the compared algorithms. The SegTHOR dataset [65], comprising 40 CT scans acquired from different patients, includes meticulous labels for four cardiac organs: the esophagus, heart, trachea, and aorta. The spatial resolution of each scan is 512 × 512, with image spacings ranging from (0.9,0.9,2) mm to (1.37,1.37,2.5) mm. Each volume sample includes 147 to 284 slices, totaling 7,420 slices. For training, 20 cases are utilized, while 4 cases are reserved for validation. The remaining 16 cases are dedicated to testing the performance of the evaluated algorithms.



(a) MISSFormer (b) TranCeption (c) nnUNet (d) HMDA_{LW} (e) GroundTruth

Fig. 4. Shown in (a) to (e), the qualitative results on the Synapse dataset illustrate the segmentation outcomes for TranCeption, MISSFormer, nnUNet, HMDA_{LW}, and the ground truth, representing various anatomical structures: blue for the aorta, green for the gallbladder, red for the left kidney, cyan for the right kidney, pink for the liver, yellow for the pancreas, purple for the spleen, and orange for the stomach.

TABLE I
COMPARISON OF DIFFERENT METHODS ON THE SYNAPSE DATASET AND THE AVERAGE RESULTS

Methods	DSC % ↑	HD↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
R50 U-Net [26]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50 Att-UNet [26]	75.57	36.97	83.67	63.91	79.20	72.71	93.56	49.37	87.19	74.95
TransUNet [26]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [30]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
HiFormer [32]	80.39	14.70	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
MISSFormer [47]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
TransCeption [55]	82.24	20.89	87.60	71.82	86.23	80.29	95.01	65.27	91.68	80.02
nnUNet [56]	83.18	**9.46**	87.06	63.25	81.93	78.22	**97.28**	**76.14**	**93.76**	**87.79**
CTC-Net [63]	78.41	-	86.46	63.53	83.71	80.79	93.78	59.73	86.87	72.39
MCRformer [62]	80.46	20.79	85.78	67.55	85.83	80.29	94.34	59.34	90.85	79.73
HMDA	**83.77**	20.39	88.95	**71.82**	86.35	82.78	95.64	71.99	91.46	81.14
HMDA_{LW}	83.21	11.69	**89.38**	70.90	**87.02**	**84.29**	95.32	68.76	91.23	78.79

B. Implementation details

The proposed HMDA model is implemented using PyTorch on a single NVIDIA V100 GPU equipped with 32 GB of memory. This model undergoes training from scratch without pretraining. During preprocessing, we incorporate various data augmentation strategies, including random flipping, rotation, contrast transformation, and gamma noise. All medical images undergo intensity normalization, adjusting the pixel values to a uniform scale ranging from 0 to 1. Such techniques help enhance the ability of each model to handle diverse data variations. The input resolution for all datasets is set to 224 × 224. For network optimization, we use an SGD optimizer with a weight decay rate of 0.0001 and a momentum parameter of 0.9. Furthermore, we implement a poly learning rate policy. Specifically, the initial learning rate and batch size for Synapse are set to 0.05 and 6, for SegTHOR to 0.01 and 6, and for ACDC to 0.01 and 8, respectively. The number of epochs for the ACDC, SegTHOR, and Synapse datasets are 600, 1000, and 800, respectively. Our model employs both focal loss and Dice loss; each assigned a weight of 0.5. This combined approach aims to optimize the model's performance by addressing the class imbalance typically encountered in medical image segmentation through focal loss, while the Dice loss encourages the model to focus on the regions of interest. Finally, the model's performance is evaluated using the Dice score and Hausdorff Distance (HD).

Consistent with the methodologie [26], each 3D volume undergoes inference slice-by-slice, with each 2D slice's prediction being successively aggregated to assemble the complete 3D model for subsequent evaluation.

C. Comparison with the State-of-the-Art Methods

We compare our proposed method with the state-of-the-art methods on the aforementioned Synapse, ACDC, and SegTHOR datasets. This comparison encompasses various methods, including pure CNNs, Transformer-based approaches, and hybrid methods. Notably, our comparison is conducted without utilizing pre-training. By conducting such an evaluation, we can gain insights into the performance and effectiveness of our approach compared to those of the existing methods. In addition to the MSDA foundation, we applied two distinct transformations to the multi-scale features: one utilizes

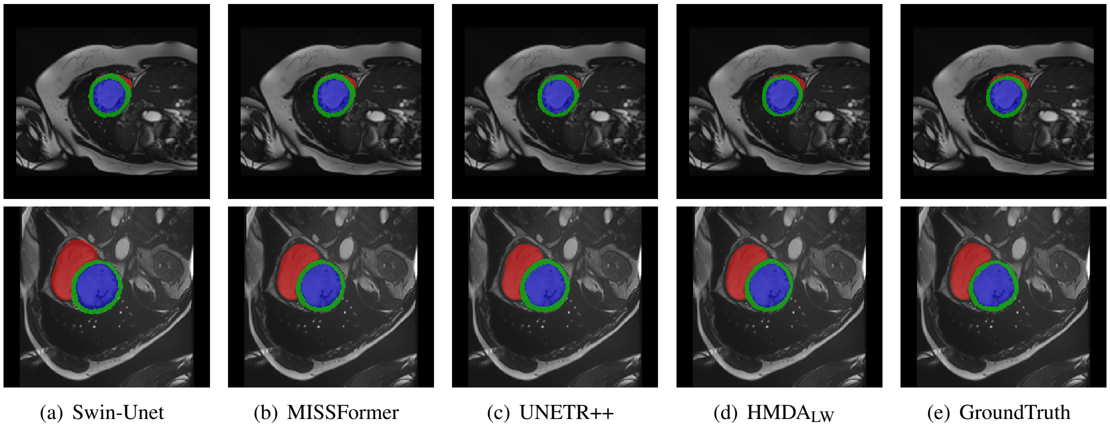


Fig. 5. The qualitative results obtained on the ACDC dataset (shown in (a) to (e)) illustrate the segmentation outcomes produced by Swin-Unet, MISSFormer, UNETR++, and our HMDA_{LW} method, and the ground truth, representing various anatomical structures: red for the RV, green for the Myo, and blue for the LV.

the spatial reduction (SR) module [45] introduced in the methodology section to reduce dimensions, which is labeled as HMDA_{LW}. Specifically, for the largest two dimensions of the multi-scale features, we applied a reduction ratio of 2, while maintaining the dimensions of the other features unchanged. The other employs a direct linear transformation, which is simply referred to as HMDA.

1) Synapse Dataset: In the comparative experimental, our method is benchmarked against performed well approaches, including nnUNet and TransCeption. Our model outperforms others in Dice with 83.77% and exhibits a relatively low HD with 20.39%. Compared to the best hybrid model (TransCeption), HDMA attains better performance, with a 1.53% improvement in Dice. Compared to the best CNN model (nnUNet), our method achieves better performance, with a 0.59% improvement in the Dice, as shown in Table I. It's significant to highlight that our method, trained from scratch, surpasses all prior architectures. This demonstrates the positive impact of our proposed approach on enhancing segmentation performance. Based on the experimental results, the HMDA_{LW} model, which incorporates the Spatial Reduction (SR) module, experiences a slight decrease in Dice score compared to the HMDA model. However, it exhibits a smaller HD with 11.69, indicating the model maintains better spatial accuracy and alignment with the ground truth segmentation boundaries. The qualitative results of the different models, including MISSFormer, TranCeption, nnUNet, and our HMDA_{LW} approach, are shown in Fig 4. It can be observed that our method performs well on organs with relatively regular shapes, while it has limitations when dealing with organs that have complex boundaries. Our method tends to identify more of the relevant areas of complex organs than other methods, such as the liver in the first row. In the second row, nnUNet demonstrates better boundary delineation for the stomach than our algorithm.

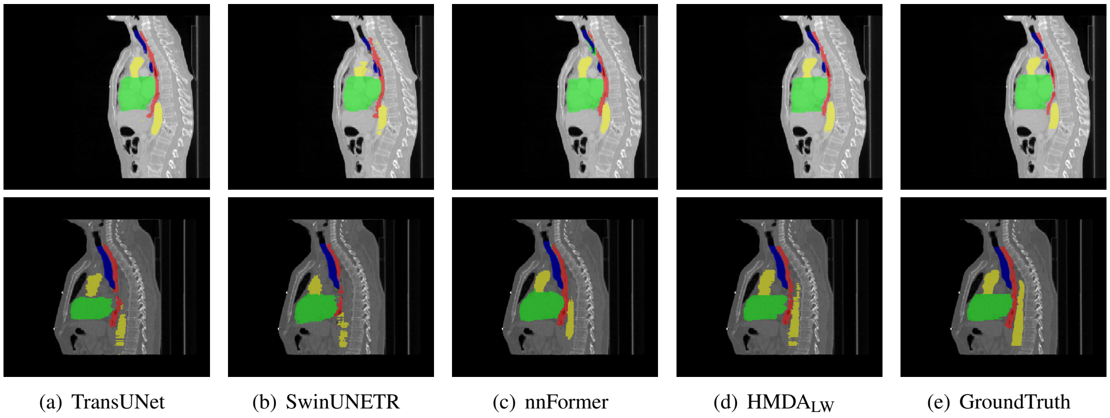
2) ACDC Dataset: In our comparative experiments, our method competes with performing well algorithms like nnUNet, nnFormer, and UNETR++. On the ACDC dataset, our method exhibits superior performance in terms of segmenting the challenging myocardium (MYO) organ, exhibiting a noticeable improvement of Dice 11.22% over the R50 Att-UNet, as shown in Table II. Our algorithm achieves the highest Dice value of 92.22%, outperforming all other methods. The HMDA model incorporates deep supervision on this MRI dataset, where the target regions specifically exhibit lower contrast. This is the same technique employed by nnFormer. Based on the experimental results, the HMDA_{LW} model, which incorporates the Spatial Reduction (SR) module, experiences a slight decrease in Dice score compared to the HMDA model. Fig 5. shows the qualitative results of the different models, including Swin-Unet, MISSFormer, UNETR++, and ours. It is evident that our method is capable of identifying a greater extent of the region of interest, such as the right ventricle (RV), compared to the other methods. In the ACDC dataset, the left ventricular myocardium has a relatively regular shape, making it more recognizable by models.

TABLE II
COMPARISON OF DIFFERENT METHODS ON ACDC DATASET AND THE AVERAGE RESULTS

Methods	DSC %↑	RV	Myo	LV
R50 U-Net [26]	87.55	87.10	80.63	94.92
R50 Att-UNet [26]	86.75	87.58	79.20	93.47
TransUNet [26]	89.71	88.86	84.53	95.73
Swin-Unet [30]	90.00	88.55	85.62	95.83
MISSFormer [47]	91.19	89.85	88.38	95.34
nnUNet [56]	91.61	90.24	89.24	95.36
MCRformer [62]	90.83	88.52	88.56	95.41
TC-CoNet [60]	91.58	90.27	88.98	95.47
nnFormer [27]	92.06	90.94	89.58	95.65
UNETR++ [61]	91.68	89.29	89.96	95.78
MedNeXt [44]	89.67	85.53	88.68	94.81
HMDA	**92.22**	**90.42**	**90.42**	95.82
HMDA_{LW}	92.02	89.95	90.10	**96.01**

3) SegTHOR Dataset: In the comparative experiments, our method is evaluated alongside performing well algorithms, including nnFormer and SwinUNETR. In Table III, HMDA improves the Dice coefficient by 4.44% over that of nnUNet,

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3469230



(a) TransUNet (b) SwinUNETR (c) nnFormer (d) HMDA$_{\mathrm{LW}}$ (e) GroundTruth

Fig. 6. The qualitative results obtained on the SegTHOR dataset (shown in (a) to (e)) illustrate the segmentation outcomes produced by TransUNet, SwinUNETR, nnFormer, our HMDA$_{\mathrm{LW}}$ method, and the ground truth, representing various anatomical structures: red for the Esophagus, green for the Heart, blue for the Trachea, yellow for the Aorta.

TABLE III
COMPARISON OF DIFFERENT METHODS ON THE SEGTHOR DATASET AND THE AVERAGE RESULTS

Methods	DSC%↑	HD↓	Esophagus	Heart	Trachea	Aorta
R50 U-Net [26]	77.81 ± 12.4	7.72 ± 3.58	56.76	88.37	83.59	82.54
TransUNet [26]	80.11 ± 10.3	6.35 ± 26.4	62.76	89.24	85.73	82.73
Swin-Unet [30]	70.58 ± 16.1	10.9 ± 5.41	44.88	86.75	81.25	69.45
nnUNet [56]	78.30 ± 10.4	17.9 ± 8.08	66.10	92.71	71.06	83.32
CTC-Net [63]	72.99 ± 11.9	20.1 ± 94.9	50.54	86.57	80.58	74.26
VTUNet [64]	68.55 ± 12.7	11.9 ± 13.9	53.31	82.43	58.74	79.73
SwinUNETR [28]	82.47 ± 9.27	33.5 ± 26.1	70.02	**93.70**	77.52	88.63
nnFormer [27]	83.71 ± 8.76	**6.57** ± 17.0	**70.21**	93.15	82.20	**89.26**
HMDA	82.74 ± 9.31	9.49 ± 3.26	67.01	91.13	85.39	87.41
HMDA$_{\mathrm{LW}}$	**83.76** ± 9.05	7.02 ± 1.31	68.59	92.51	**86.60**	87.33

and the Hausdorff distance is reduced by 8.39. In comparison to the Transformer-based architecture of TransUNet, our method achieves a 2.63% higher Dice coefficient. Although TransUNet has a slightly lower HD by 3.14, our method's HD variance of 3.26 is significantly lower than TransUNet's variance of 26.4. The nnFormer method demonstrates the best performance among hybrid transformer-based models, matching our lightweight model (HMDA$_{\mathrm{LW}}$) in terms of the Dice coefficient. To provide a more comprehensive evaluation of our approach, we compared the variances, revealing that our HD variance of 1.31 is significantly lower than nnFormer's 17.0. This indicates that our model is more consistent regarding the maximum distance between the segmentation boundaries and the ground truth.

Additionally, we conducted a computational complexity analysis on the SegTHOR dataset, comparing our approach with several high-performing models such as TransUNet, nnFormer, and SwinUNETR. It can be observed from Table IV that the overall performance of our lightweight model (HMDA$_{\mathrm{LW}}$) surpasses that of previous models (HMDA). HMDA$_{\mathrm{LW}}$ boasts a significantly lower parameter count of $41.88M$, compared to nnFormer's $149.6M$. This suggests a more efficient use of parameters in our model, which is a desirable trait in terms of model optimization and generalization. Our model exhibits a lower memory requirement, demanding a mere 7.56 GB of memory resources, which is substantially lower than the 17.0 GB consumption of the nnFormer. This achievement underscores our model's exceptional memory efficiency. HMDA$_{\mathrm{LW}}$ testing time is 21.3 seconds, and nnFormer is 20.8 seconds; the difference in speed is relatively minimal. When comparing computational performance measured in GFLOPs, our model achieves a throughput of 276 GFLOPs, whereas their model operates at 210 GFLOPs. It remains within a comparable range when compared to SwinUNETR's 444 GFLOPs.

The qualitative results of different models, including TransUNet, SwinUNETR, nnFormer, and HMDA$_{\mathrm{LW}}$, are shown in Fig 6. HMDA$_{\mathrm{LW}}$ demonstrates a closer approximation to the ground truth on the trachea in the first row of visualizations compared to other methods. The nnUNet appears to misclassify the heart as the target organ trachea. While our method exhibits slightly less distinct boundaries on the aorta than nnFormer, it is still superior to the other two.

D. Ablation Studies

In this subsection, We conduct ablation studies to assess the efficacy of the key components in our architecture. These assessments are conducted on the previously mentioned Multi-Organ Segmentation (Synapse) datasets. The ablation studies were all conducted on the basis of the MSDA framework,

TABLE IV
EVALUATION OF THE COMPUTATIONAL COMPLEXITY AMONG DIFFERENT METHODS

Models	GFLOPs	Param(M)	Memory(GB)	Test(s)	DSC%
TransUNet [26]	**148**	93.23	**4.77**	**19.5**	80.11
SwinUNETR [28]	444	61.99	15.7	25.3	82.47
nnFormer [27]	210	149.6	17.0	20.8	83.71
HMDA_{LW}	276	**41.88**	7.56	21.3	**83.76**
HDMA	504	123.99	25.6	25.2	82.74

TABLE V
EVALUATION OF THE TB-RESIDUAL ENCODER ON SYNAPSE DATASET

CNN Models	Bridge	DSC% ↑	HD ↓
Residual-Net	-	82.23	22.73
Residual-Net	✓	81.85	23.16
SI-Residual	✓	81.64	23.80
TB-Residual	-	82.39	22.49
TB-Residual	✓	**83.59**	**15.47**

disregarding the impact of multi-scale transformations on the experiments, which encompass the previously mentioned Spatial Reduction (SR) and linear transformations.

1) Impact of CNN Encoder Architecture on the Hybrid Model: In this subsection, we evaluate the performance of different CNN architectures within our hybrid model to determine their individual performance. Our evaluation encompasses the ResNet-50 network (Residual-Net), the SIngle-branch residual connections (SI-Residual) network, and our novel TB-Residual network, each integrated with a Transformer module utilizing MSDA.

Table V displays the segmentation results, where the Residual-Net and SI-Residual models achieve Dice scores of 81.85% and 81.64%, respectively, under the condition of having bridge connections. The slight reduction in Dice scores for the SI-Residual model is attributed to its simplified design, which uses a uniform number of Bottleneck blocks across stages, thus foregoing the hyperparameter present in the Residual-Net architecture. Nonetheless, the TB-Residual delivers a higher Dice score of 83.59%, reflecting a 1.74% increase in performance over the Residual-Net model. This performance underscores the efficacy of integrating identical topologies in an aggregated fashion within our hybrid model. Without bridge connections between the CNN and Transformer components, the TB-Residual model exhibits a Dice score of 82.39%, slightly surpassing the Residual-Net model, which records a Dice score of 82.23%. This comparison indicates that even without the bridge connections, the TB-Residual architecture maintains a marginal advantage in performance.

2) Comparative Evaluation of the Multi-scale Spatial Deformable Attention Encoder: In this section, we aim to assess the efficacy of the Multi-scale Spatial Deformable Attention (MSDA) by conducting a series of experiments. For comparative analysis, we have included control networks: the multi-head self-attention (MHSA), a deformable attention method from the literature [25], the Multi-scale Spatially Deformable Attention (MSDA) without the incorporation of multi-scale feature projection. Additionally, our ablation experiments are dedicated to fine-tuning the deformable attention hyperparameters within the MSDA module, with the goal of achieving optimal performance using the most efficient set of parameters. We explored three distinct parameter combinations for the MSDA module, as detailed in the middle rows of Table VI. The set involving parameters 96, 768, and 10 yielded the best performance, which corresponds to the dimensions of the self-attention mechanism, the hidden layer size of the multi-layer perceptron (MLP), and the depth of the Transformer layer, respectively.

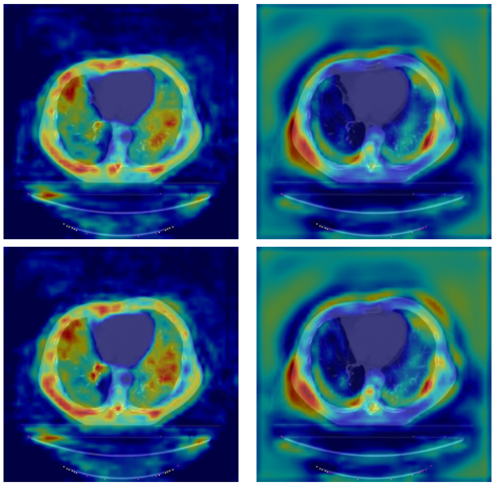
In the evaluation of optimal parameters, we compared MHSA, the deformable attention from [25], and our MSDA method. Our MSDA showed a 1.01% improvement in Dice score over MHSA, signifying superior segmentation accuracy. While the [25] method is configured with these specific parameters, it exhibits comparable performance to the MHSA method. This further substantiates the efficacy of our proposed MSDA algorithm, confirming its potency in enhancing model performance.

TABLE VI
EVALUATION OF THE MSADA ENCODER ON SYNAPSE DATASET

Methods	Parameters	DSC% ↑
TB-Residual + MHSA	$(96, 768, 10)$	82.58
TB-Residual + [25]	$(96, 768, 10)$	82.65
TB-Residual + MSDA	$(192, 768, 12)$	83.53
TB-Residual + MSDA	$(192, 768, 10)$	83.25
TB-Residual + MSDA	$(96, 768, 10)$	**83.59**

In a quantitative analysis comparing the multi-head self-attention mechanism to the Class Activation Mapping (CAM) heatmaps featured in the final feature extraction layer of our method, we've demonstrated that our approach excels at promptly identifying regions of interest. Conversely, the multi-head self-attention mechanism, despite its merits, allocates considerable attention to extraneous background areas, which can result in a less focused detection of the pertinent features. It has been observed that our model's focus has a minor component on the scanning bed, but it is negligible compared to the multi-head self-attention.

3) Assessment of the Cross Attention Bridge Connection: Considering the multitude of ways to integrate CNN and Transformer blocks, we conduct several experiments to verify the necessity of the coupling method within our architecture, presented in Table VII. Initially, we establish a baseline



(a) Deformable Attention (b) Self-Attention

Fig. 7. A comparative visualization on the Synapse dataset illustrating the proposed deformable attention module on the left and the multi-head self-attention mechanism on the right.

configuration, termed 'w/o-CAB,' in which the Transformer is directly linked to the decoder via skip connection rather than utilizing a direct bridge with the CNN. Subsequently, we compare an explicit bridging method from the [54] for connecting the two features. Furthermore, we explore configurations for connecting CNN and Transformer blocks across different numbers of stages. Specifically, 'CAB-stage4' utilized a bridging approach in all four layers, and 'CAB-stage3' implemented the bridge in the first three layers of the feature extraction but omitted it in the fourth layer.

TABLE VII
EVALUATION OF THE CAB BRIDGE CONNECTION MODULE ON SYNAPSE DATASET

Methods	DSC % ↑	HD ↓
w/o-CAB	82.39	22.49
[54]-stage3C	82.10	26.48
CAB-stage4	82.89	19.49
CAB-stage3	83.22	17.53
CAB-stage3C	**83.59**	**15.47**

As detailed in Table VII, the CAB-stage4 model, which employs explicit bridging methods, demonstrates a better performance over the w/o-CAB model, with a 0.5% increase in Dice score and a reduction of 3 in Hausdorff distance (HD). Specifically, we explored the performance when not using the CAB module in the fourth layer of the feature extraction but instead employing two alternative approaches. The method CAB-stage3, utilizing only Transformer features in the fourth layer, demonstrates a 0.33%increase in Dice value and a 1.96% decrease in HD compared to CAB-stage4. Building on this, CAB-stage3C is designed to concatenate CNN and Transformer features in the fourth layer, leading to a further improvement with a 0.37% increase in Dice value and a 2.06% decrease in HD. In literature [54]-stage3C, it enhances the transformer features over CNN features in the first three layers of the feature extraction, which leads to suboptimal experimental outcomes. The experimental outcomes demonstrate that a three-stage cross-attention model configuration, coupled with the concentration of local and global features in the fourth layer, optimally enhances segmentation performance. We note that our implementation utilizes a straightforward channel attention design, which serves as a solid foundation. Future work could explore more sophisticated attention mechanisms to further enhance model capabilities.

V. Conclusion and Discussion

We propose a Hybrid Transformer and CNN architecture with Multi-scale Deformable Attention (HMDA) for medical image segmentation tasks. Our approach incorporates a Multi-scale Spatially Adaptive Deformable Attention (MSADA) mechanism, which facilitates the fusion and coordination of information across different resolutions, enabling the integration of low-resolution and high-resolution features. Additionally, by introducing sparse spatial sampling, the model dynamically selects and focuses only on a small subset of crucial positions. Our approach addresses the challenges associated with multi-head self-attention, which not only leads to high memory usage and computational costs but also has the potential to be affected by irrelevant areas outside the region of interest. We propose a Cross-Attention Bridge (CAB) module that enhances the quality of the network's generated representations by explicitly modeling the channel interdependencies between the network's local and global features. This enhancement aims to provide the decoder with more detailed and pertinent information, thereby better restoring the fine details of the image. In the training process of the experiment, the model is randomly initialized and trained from scratch without using pre-training. Our method achieves promising segmentation results in medical image segmentation tasks compared to existing methods. However, the segmentation results obtained by the proposed algorithm reveal certain challenges, with discontinuities observed for ROIs that are more difficult to segment. Introducing contextual information and a priori knowledge, such as the geometric features and spatial relationships of anatomical structures, could also help to improve the continuity and accuracy of the output segmentation results. In addition, we explore the impacts of convolutional modules on the model architecture, which can be further explored when designing such modules. Lastly, further exploration should be conducted on lightweight models as well.

References

[1] Feng, Jiang, Aleksei, Grigorev, Seungmin, Rho, Zhihong, Tian, Yun-Sheng, and Fu, "Medical image semantic segmentation based on deep learning," *Neural Computing and Applications*, 2018.

[2] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. V. D. Kouwe, R. Killiany, D. Kennedy, and S. a. Klaveness, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain,"*Neuron* vol. 33, no. 3, pp. 0–355, 2002.

[3] Feng, Jiang, Aleksei, Grigorev, Seungmin, Rho, Zhihong, Tian, Yun-Sheng, and Fu, "Medical image semantic segmentation based on deep learning," *Neural Computing and Applications*, 2018.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[5] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018

[6] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021

[7] Z. Zhu, Z. Wang, G. Qi, N. Mazur, P. Yang, and Y. Liu, "Brain tumor segmentation in mri with multi-modality spatial information enhancement and boundary shape correction," *Pattern Recognition*, vol. 153, 2024.

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

[9] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Drinet for medical image segmentation," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2018.

[10] L. C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," *IEEE*, 2016.

[11] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *Computer Science*, no. 4, pp. 357–361, 2014.

[12] G. Papandreou, L. C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *IEEE International Conference on Computer Vision*, 2016.

[13] N. Ibtehaz and M. S. Rahman, "Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, 2019.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support*. Sep;11045:3-11. 2018.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. *Adv. Neural Inform. Process. Syst.* (2017) 2, 3, 5

[16] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. Mcdonagh, N. Y. Hammerla, and B. Kainz, "Attention u-net: Learning where to look for the pancreas", *arXiv preprint arXiv:1804.03999* 2018.

[17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, and D. Amodei, "Language models are few-shot learners", *arXiv preprint ArXiv:2005.14165*, 2020.

[18] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision Transformers," *Proceedings of the IEEE/CVF international conference on computer vision*, pp.22–31, 2021.

[19] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin Transformer: A general vision Transformer backbone with cross-shaped windows," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.12124–12134, 2021.

[20] L. Gao, J. Zhang, C. Yang, and Y. Zhou, "Cas-vswin Transformer: A variant swin Transformer for surface-defect detection," *Computers in Industry*, no. 140-, p. 140, 2022.

[21] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision Transformer with Deformable Attention," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision Transformer using shifted windows,"*Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012-10022, 2021.

[24] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision Transformer,"*arXiv preprint arXiv:2103.11886* 2021.

[25] X. Zhu, W. Su, L. Lu, B. Li, and J. Dai, "Deformable detr: Deformable Transformers for end-to-end object detection,"*arXiv preprint arXiv:2010.04159* 2020.

[26] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306.*

[27] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nn-Former: Interleaved transformer for volumetric segmentation," 2021, *arXiv:2109.03201.*

[28] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," *In: International MICCAI Brainlesion Workshop. pp. 272–284*. Springer (2022)

[29] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," 2021, *arXiv:2112.13492*

[30] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure Transformer for medical image segmentation," *arXiv:2105.05537*, 2021.

[31] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing Transformers and cnns for medical image segmentation," *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, pp. 14–24. 2021.

[32] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Hiformer: Hierarchical multi-scale representations using Transformers for medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 6202–6212.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[34] W. Wang et al., "CrossFormer: A versatile vision transformer hinging on cross-scale attention," 2021, *arXiv:2108.00154.*

[35] M. S. Alam, D. Wang, Q. Liao, and A. Sowmya, "A multi-scale context aware attention model for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 3731–3739, 2023.

[36] X. Liang, N. Li, Z. Zhang, J. Xiong, and Y. Xie, "Incorporating the hybrid deformable model for improving the performance of abdominal ct segmentation via multi-scale feature fusion network," *Medical Image Analysis*, vol. 73, no. 8, p. 102156, 2021.

[37] W. He, C. Zhang, J. Dai, L. Liu, T. Wang, X. Liu, Y. Jiang, N. Li, J. Xiong, and L. Wang, "A statistical deformation model-based data augmentation method for volumetric medical image segmentation," *Medical image analysis*, p. 91, 2024.

[38] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 764–773 (2017)

[39] J. Hu, L. Shen, G. Sun, and S. Albanie, "Squeeze-and-excitation networks," in *IEEE transactions on pattern analysis and machine intelligence.*, 2017.

[40] X. Chen, R. Zhang, and P. Yan, "Feature fusion encoder decoder network for automatic liver lesion segmentation," *arXiv:1903.11834* 2019.

[41] S. Chen, Y. Zou, and P. X. Liu, "Iba-u-net: Attentive bconvlstm u-net with redesigned inception for medical image segmentation," *Computers in Biology and Medicine*, vol. 135, no. 4, p. 104551, 2021.

[42] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local u-nets for biomedical image segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 6315–6322, 2020.

[43] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[44] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F.Jaeger, and K. H. Maier-Hein, "Mednext: transformer-driven scaling of convnets for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 405–415.

[45] W. Wang, E. Xie, X. Li, D. P. Fan, and L. Shao, "Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions," *arXiv:2102.12122*, 2021.

[46] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *arXiv:Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, vol. 5, 2015, p. 12.

[47] X. Huang, Z. Deng, D. Li, and X. Yuan, "Missformer: An effective medical image segmentation Transformer," *IEEE Trans Med Imaging*. pp:1484-1494, 2023.

[48] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[49] R. Azad, M. Heidari, M. Shariatnia, E. K. Aghdam, S. Karimijafarbigloo, E. Adeli, and D. Merhof, "Transdeeplab: Convolution-free transformer-

based deeplab v3+ for medical image segmentation," in Predictive Intelligence in Medicine. *Springer Nature Switzerland*, 2022, pp. 91–102.

[50] R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, and D. Merhof, "Dae-former: Dual attention-guided efficient transformer for medical image segmentation," *arXiv preprint* arXiv:2212.13504, 2022.

[51] Z. Zhu, M. Sun, G. Qi, Y. Li, X. Gao, and Y. Liu, "Sparse dynamic volume transunet with multi-level edge fusion for brain tumor segmentation," *Computers in Biology and Medicine*, vol. 172, 2024.

[52] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, and Y. Liu, "Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri," *Information Fusion*, 2023.

[53] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," CVPR. 2017. pp. 1492–1500, 2017.

[54] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer," *arXiv: 2109.04335*, 2021.

[55] R. Azad, Y. Jia, E. Khodapanah Aghdam, J. Cohen-Adad, and D. Merhof, "Enhancing medical image segmentation with transception: A multi-scale feature fusion approach," *arXiv e-prints*, 2023.

[56] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, and S. Wirkert, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[57] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, January 2022

[58] O. Bernard et al., "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?," in *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514-2525, Nov. 2018, doi: 10.1109/TMI.2018.2837502.

[59] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical Transformer: Gated axial-attention for medical image segmentation," *arXiv:2102.10662*, 2021

[60] 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.

[61] A. M. Shaker, M. Maaz, H. Rasheed, S. Khan, M. H. Yang, F. S. Khan. "UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation". *IEEE Trans Med Imaging*. 2024, May 9.

[62] J. Li, N. Chen, H. Zhou, T. Lai, H. Dong, C. Feng, L. Wei, MCRformer: Morphological Constraint Reticular Transformer for 3D Medical Image Segmentation, *Expert Syst. Appl*. (2023) 120877.

[63] F. Yuan, Z. Zhang, Z. Fang, "An effective CNN and Transformer complementary network for medical image segmentation", *Pattern Recognition 136* (2023) 109228.

[64] H. Peiris, M. Hayat, Z. Chen, G. Egan, M. Harandi, "A Robust Volumetric Transformer for Accurate 3D Tumor Segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2022.

[65] Lambert, Zoé, et al. "Segthor: Segmentation of thoracic organs at risk in ct images." *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. 2020.