Bidirectional Dynamic Transformer for Medical Image Segmentation

Jing Tong*, Ling Ma*, Yanbiao Ji*, Shaokai Wu*, Yue Ding* and Hongtao Lu*
*Shanghai Jiao Tong University, Shanghai, China
Email: {tj_19_hf, maling920827, jiyanbiao, shaokai.wu, dingyue, htlu}@sjtu.edu.cn

***Abstract*—Medical image segmentation is a challenging task, where U-shaped networks are widely used. In order to better model long-range dependencies, many recent works integrate Transformer or state-space model (SSM) into U-Net and claim superior performance. However, such methods either suffer from semantic redundancy and heavy computation in self-attention, or require unnatural scanning routes in SSM. To validate their limitations, we first conduct an empirical study to show that existing Transformer- or SSM-based methods may actually fail to outperform vanilla U-Net under fair comparison settings, *i.e.*, with the same pre-trained encoder and proper ablations. Then, to mitigate the limitations of self-attention, we propose BIdirectional Dynamic TransFormer (BIDFormer) to model spatial relationships and semantic correlations separately. *(i) spatial relationships:* to efficiently model global context, we replace self-attention with *Bidirectional Linear Attention (BLA)* between image patches and learnable class embeddings; *(ii) semantic correlations:* to explicitly model the correlation among target classes, we utilize class embeddings as *Dynamic Low-rank Classifiers (DLC)*. Extensive experiments on public datasets demonstrate that BIDFormer can be easily integrated into existing models, leading to remarkable performance improvements. Code is available at https://github.com/JingTongsh/UN-Segmentation.**

***Index Terms*—computer vision, medical image analysis, semantic segmentation, transformer**

I. Introduction

Medical image segmentation is a challenging and important task, which enables precise delineation of anatomical structures or pathological regions. Convolutional Neural Networks (CNNs) like U-Net [1] have dominated the field with outstanding performance, but they frequently struggle to model global context [2], [3]. Consequently, abundant research [4], [5], [6], [7] has focused on integrating Transformer [8] or state-space model (SSM) Mamba [9], [10] modules into the U-Net framework, claiming better performance than vanilla U-Net.

Nevertheless, these Transformer- or Mamba-based methods still face notable drawbacks, *i.e.*, semantic redundancy and heavy computation in Transformers, and complex scanning techniques in Mamba. Self-attention can successfully enlarge receptive fields at the cost of **semantic redundancy** and **heavy computation**. Considering that many neighboring pixels belong to the same tissue or structure, they actually provide duplicated semantic information for the attention calculation [11], [12]. Mamba appears to be an efficient alternative to Transformer modules, but when applied to computer vision tasks, the core SSM operation inevitably requires **complex scanning techniques** [13] that disrupt spatial relationships among image patches [14], [15], leading to suboptimal performance in some scenarios [16].

In this paper, we first conduct an empirical study to show that Transformer- or Mamba-based methods are not necessarily superior to CNN-based methods like U-Net. Specifically, we identify two issues often overlooked in current comparisons: discrepancy in encoder capacity and lack of SSM ablation. While recent methods tend to employ stronger encoders, the baseline U-Net is implemented with ResNet [17] encoder, which is somewhat outdated. Hence, we ensure U-Net is equipped with a comparable strong encoder, leading to competitive performance against state-of-the-art (SOTA) methods. The commonly used Visual State Space (VSS) module consists of SSM operation within a gated CNN architecture [18]. In natural image classification, removing the SSM to retain a pure gated CNN structure has been shown to improve performance [16]. Therefore, We do the same in segmentation decoders, and extend the finding to medical image segmentation.

To this end, we propose **BI**directional **D**ynamic Trans**Former** (BIDFormer), a simple yet effective module that enhances CNNs with both long-range context modeling and deep semantic guidance. Within each layer, BIDFormer utilizes *Bidirectional Linear Attention (BLA)* between the image patches and learnable class embeddings, successfully capturing global context with linear complexity relative to the number of patches. Across layers, the class embeddings serve as *Dynamic Low-rank Classifiers (DLC)* that automatically adapt to the input and capture multi-scale semantic relationships via a Transformer decoder. Extensive experiments on public datasets demonstrate that BIDFormer can be easily integrated into existing models, leading to remarkable performance improvements.

The main contributions of this work are:

- We conduct an empirical study on U-shaped networks, and find that CNN methods can actually outperform Transformer- or Mamba-based methods.
- We propose BIDFormer, a plug-and-play module that effectively enhances CNNs with both long-range context modeling and in-depth semantic guidance.
- Extensive experiments on multiple medical image segmentation datasets demonstrate the superior performance of BIDFormer over state-of-the-art (SOTA) methods.

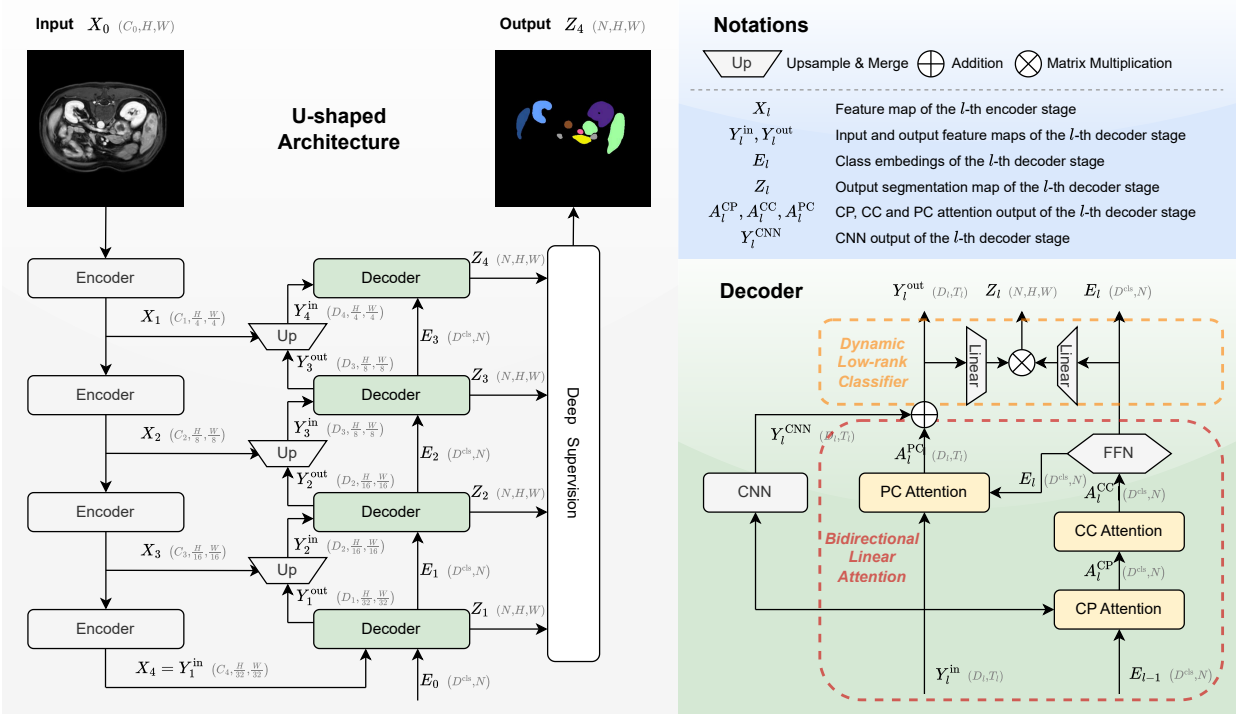


Fig. 1: The overall framework and components of the proposed BIDFormer model. *Left:* The model is composed of a pre-trained encoder providing multi-scale features and a BIDFormer decoder for pixel classification. Note that each decoder stage has an output for deep supervision. *Right:* Each BIDFormer stage encompasses 3 parts: a CNN module for local feature modeling, BLA for long-range context processing, and DLC for semantic guidance.

II. Related Work

A. Convolution-based Methods

Convolutional Neural Networks (CNNs) have laid the foundation for both natural and medical semantic segmentation. CNN encoders serve as strong backbones that can extract multi-scale features efficiently, but their limited receptive fields hinder the modeling of long-range dependencies [3]. CNN decoders up-sample the low-resolution encoder features and provide pixel-wise classification, with abundant research trying to enlarge the receptive field [19], [20] or leverage multi-scale features [21], [22].

B. Transformer-based Methods

There are mainly 3 ways to apply Transformers [8] in semantic segmentation. Self-attention encoders like ViT [23] and its multi-scale variants [24], [25] can be applied to replace [26], [27] or enhance [4] CNN encoders, trading low resolution for larger receptive fields and quadratic complexity for better capacity [28]. Self-attention decoders can capture long-range dependencies better than CNN decoders, but suffer from low resolution [29], restricted local windows [5], and quadratic complexity with respect to the number of input tokens. Cross-attention decoders [30], [31] extract class embeddings from low-resolution encoder features as semantic prototypes. Despite their success in natural images, these methods do not leverage the multi-scale encoder features, which are crucial for medical image segmentation [1].

C. Mamba-based Methods

State-space models including Mamba [9], [10] can efficiently scale up the context window through linear-time inference and parallelized training, thus considered as promising alternatives to Transformers in long sequence modeling. Although it is not trivial to apply such sequence models to vision tasks, which require carefully designed scanning techniques [14], several methods have adapted Mamba as image encoders [32], [18], [33] to efficiently enlarge the receptive fields. As for medical image segmentation, the application of Mamba has become a topic of interest [13]. To be specific, most methods try to adapt Visual State Space (VSS) modules from VMamba [18] into their encoders [34], [35] or decoders [7].

TABLE I: Comparison between prevalent standard self-attention with our BLA method. BLA is more efficient in all three aspects.

Prevalent Methods	Our Method	Explanation
Self-attention	Cross-attention	III-B1
Softmax attention	Linear attention	III-B2
Linear projection	Bottleneck MLP projection	III-B2

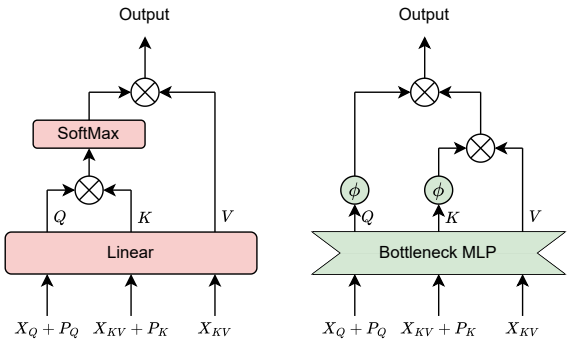


Fig. 2: Comparison between Standard softmax attention with Low-rank Linear Attention.

III. Method

In this section, we present our BIDFormer model for medical image segmentation, which enhances CNN-based methods with Bidirectional Linear Attention for global context modeling, and Dynamic Low-rank Classifier for semantic correlation modeling, as shown in Figure 1.

A. Overview

1) Image encoder: We apply a pre-trained image encoder to effectively extract generalizable multi-scale representations. First, the input image $\mathbf{X}_0$ is resized to height H_0 and width W_0, and the number of input channels C_0 equals 1 for grayscale images and 3 for RGB images. Then, the encoder outputs hierarchical features at 4 different resolutions, all utilized as input to the decoder:

$$\mathbf{X}_l = \texttt{Encoder}_l(\mathbf{X}_{l-1}) \in \mathbb{R}^{C_l \times H_l \times W_l}, \quad l = 1, 2, 3, 4, \tag{1}$$

where the spatial sizes $H_l \times W_l$ gradually decrease from $\frac{H}{4} \times \frac{W}{4}$ to $\frac{H}{32} \times \frac{W}{32}$, *i.e.*, $H_l = H_0 \cdot 2^{-l-1}, W_l = W_0 \cdot 2^{-l-1}$.

2) BIDFormer decoder: We focus on designing the segmentation decoder so as to better interpret the semantic relationships from multi-scale encoder features. As a plug-and-play module, BIDFormer can be easily integrated into CNN-based decoders, brining benefits from two aspects: *(i) Bidirectional Linear Attention* provides large receptive fields similar to self-attention, but endure less semantic redundancy and computational cost; *(ii) Dynamic Low-rank Classifier* provides deep supervision that not only adapts to the input automatically, but also pass on information across different layers through class embeddings.

Note that we draw inspiration from natural image segmentation methods like Segmenter [29] and Mask2Former [30], which extract class embeddings from low-resolution (16×16 or 7×7) features that do not fit with the U-Net architecture. The major difference is that we closely integrate the class embeddings into multi-scale U-Net decoders for both global context modeling and layer-wise deep supervision. Moreover, we implement BLA as an efficient substitution for standard self-attention, as summarized in Table I.

Overall, we introduce learnable class embeddings $\mathbf{E}_0 \in \mathbb{R}^{D^{\text{cls}} \times N}$ in correspondence with N semantic classes to be segmented, including the background. Formally, there are two inputs for the l-th decoder stage: class embeddings $\mathbf{E}_{l-1} \in \mathbb{R}^{D^{\text{cls}} \times N}$ as semantic prototypes, and image feature map $\mathbf{Y}_l^{\text{in}} \in \mathbb{R}^{D_l \times H_{5-l} \times W_{5-l}}$ from the previous stage:

$$\mathbf{Y}_l^{\text{in}} = \begin{cases} \mathbf{X}_4, & l = 1, \\ \texttt{Merge}(\mathbf{X}_{5-l}, \texttt{UpSample}(\mathbf{Y}_{l-1}^{\text{out}})), & l > 1, \end{cases} \tag{2}$$

where `Merge` is implemented as addition to reduce channels, and `UpSample` is implemented as patch expanding, unless configured otherwise. Then, the l-th decoder stage has 3 outputs: the updated class embeddings $\mathbf{E}_l \in \mathbb{R}^{D^{\text{cls}} \times N}$, image feature map $\mathbf{Y}_l^{\text{out}} \in \mathbb{R}^{D_l \times H_l \times W_l}$, and segmentation map $\mathbf{Z}_l \in \mathbb{R}^{N \times H_{5-l} \times W_{5-l}}$ for deep supervision

$$(\mathbf{E}_l, \mathbf{Y}_l^{\text{out}}, \mathbf{Z}_l) = \texttt{Decoder}_l(\mathbf{Y}_l^{\text{in}}, \mathbf{E}_{l-1}). \tag{3}$$

The inner details are explained in the following subsections.

B. Bidirectional Linear Attention

1) Bidirectional cross-attention: We find self-attention between all patch pairs unnecessary, and calculate attention in a bidirectional manner between patch tokens and class embeddings instead. For brevity, the patch tokens $\mathbf{Y}_l^{\text{in}}$ are flattened from $H_l \times W_l$ into a sequence of length $T_l = H_l W_l$. The overall time complexity is linear $O(T_l N D_l)$ as opposed to quadratic (inherent for self-attention) $O(T_l^2 D_l)$, since the number of classes N is much smaller than the number of patches T_l.

*a) Class-patch (*CP*) attention:* First, BLA aggregates information from patch tokens that are most informative for each class, so that all embeddings can get a global receptive field with complexity $O(NT_l D_l)$:

$$\begin{aligned} \mathbf{A}_l^{\text{CP}} &= \texttt{Attn}(\mathbf{E}_{l-1}, \mathbf{Y}_l^{\text{in}}, \mathbf{Y}_l^{\text{in}}), \\ \mathbf{E}_l^{\text{CP}} &= \texttt{Norm}(\mathbf{E}_{l-1} + \mathbf{A}_l^{\text{CP}}). \end{aligned} \tag{4}$$

*b) Class-class (*CC*) interaction:* Then, the class embeddings are further refined using self-attention with complexity $O(N^2 D_l)$, and a 2-layer fully-forward network (FFN) with complexity $O(N D_l^2)$:

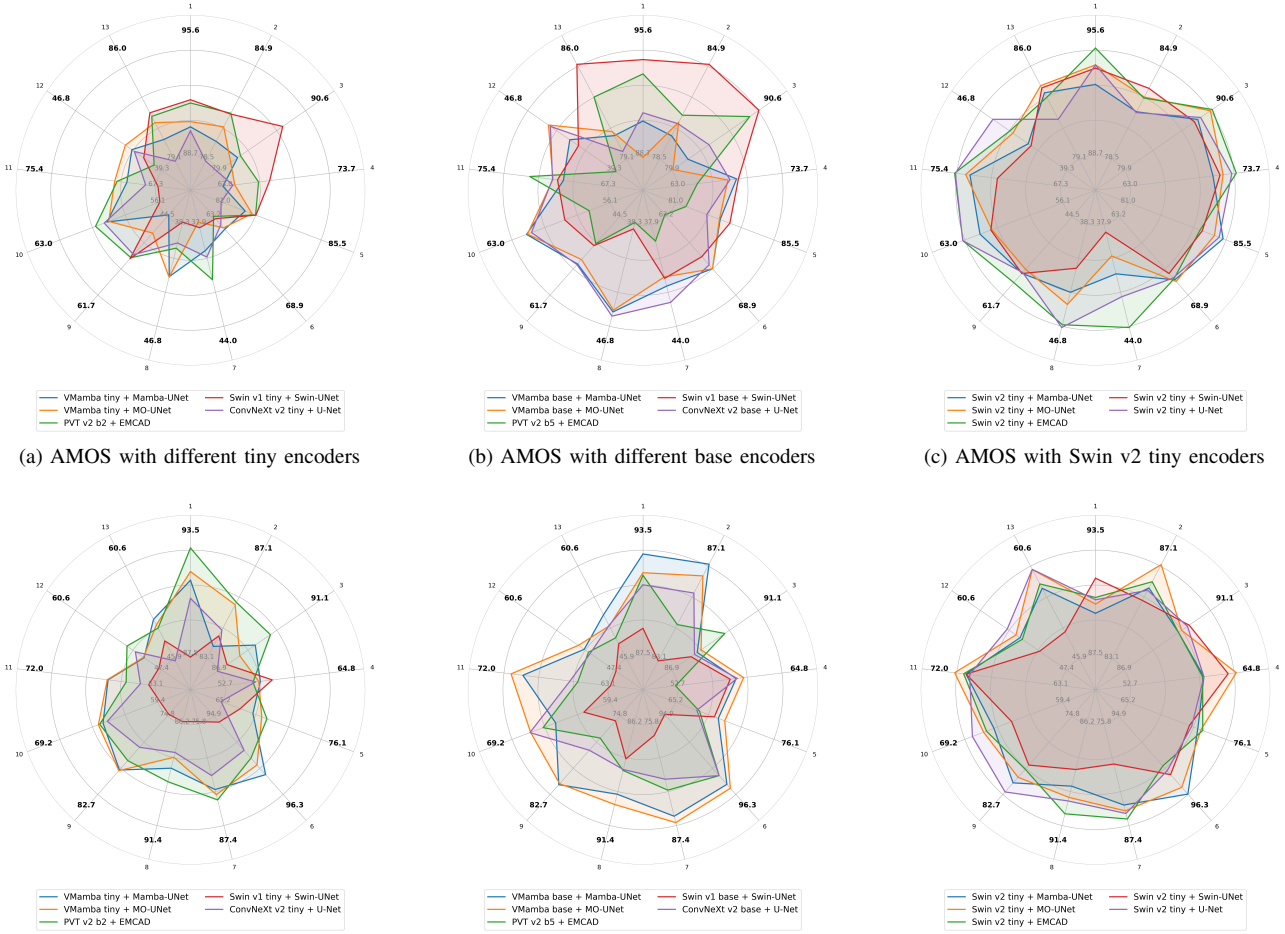
$$\begin{aligned} \mathbf{A}_l^{\text{self}} &= \texttt{Attn}(\mathbf{E}_l^{\text{CP}}, \mathbf{E}_l^{\text{CP}}, \mathbf{E}_l^{\text{CP}}), \\ \mathbf{E}_l^{\text{self}} &= \texttt{Norm}(\mathbf{E}_l^{\text{CP}} + \mathbf{A}_l^{\text{self}}), \\ \mathbf{E}_l &= \texttt{Norm}(\mathbf{E}_l^{\text{self}} + \texttt{FFN}(\mathbf{E}_l^{\text{self}})). \end{aligned} \tag{5}$$

*c) Patch-class (*PC*) attention:* Finally, BLA re-allocates semantic information from the class embeddings back to patch tokens:

$$\begin{aligned} \mathbf{A}_l^{\text{PC}} &= \texttt{Attn}(\mathbf{Y}_l^{\text{in}}, \mathbf{E}_l, \mathbf{E}_l), \\ \mathbf{Y}_l^{\text{out}} &= \texttt{CNN}(\mathbf{Y}_l^{\text{in}}) + \mathbf{A}_l^{\text{PC}}, \end{aligned} \tag{6}$$

where `CNN` can be any convolutional decoder module, and the time complexity for this step is $O(NT_l D_l)$.

When put together, CP, CC and PC attention can achieve global context modeling similar to self-attention. By compressing global information into a few class tokens, BLA endures less feature redundancy and computational cost.



(a) AMOS with different tiny encoders (b) AMOS with different base encoders (c) AMOS with Swin v2 tiny encoders

(d) BTCV with different tiny encoders (e) BTCV with different base encoders (f) BTCV with Swin v2 tiny encoders

Fig. 3: Dice score per class for the empirical study. *What the subfigures mean:* We compare several decoders (Mamba-UNet [36], EMCAD [37], Swin-UNet [5], U-Net [1], and MO-UNet by omitting the SSM in Mamba-UNet) across 6 different settings, and each subfigure represents a specific setting. Left column: different tiny encoders; Middle column: different base encoders; Right column: Swin v2 tiny encoder [25]. First row: AMOS dataset; Second row: BTCV dataset. *What the axes mean:* Each axis represents a class, and is scaled according to dice scores of all methods for the corresponding class in the dataset.

2) Low-rank linear attention: Instead of linear-projected softmax attention, we design low-rank linear attention with bottleneck MLP projection to further reduce computation cost, as shown in Figure 2. Denote the query input as $\mathbf{X}_Q \in \mathbb{R}^{D_Q \times T}$ and key-value input as $\mathbf{X}_K, \mathbf{X}_V \in \mathbb{R}^{D_{KV} \times S}$, and their position encodings as $\mathbf{P}_Q \in \mathbb{R}^{D_Q \times T}$ and $\mathbf{P}_K \in \mathbb{R}^{D_{KV} \times S}$. The projections for H-head attention with d-dim per head can be formulated as

$$\begin{aligned} \mathbf{Q} &= \mathbf{W}_Q^{\text{up}} \sigma(\mathbf{W}_Q^{\text{down}} \mathbf{X_Q} + \mathbf{P}_Q) \in \mathbb{R}^{hd \times T}, \\ \mathbf{K} &= \mathbf{W}_K^{\text{up}} \sigma(\mathbf{W}_K^{\text{down}} \mathbf{X_K} + \mathbf{P}_K) \in \mathbb{R}^{hd \times S}, \\ \mathbf{V} &= \mathbf{W}_V^{\text{up}} \sigma(\mathbf{W}_V^{\text{down}} \mathbf{X_V}) \in \mathbb{R}^{hd \times S}, \end{aligned} \tag{7}$$

where the non-linearity σ is GELU [38]. We apply the original kernel function for linear attention [39] $\phi(x) = \text{elu}(x) + 1$:

$$\texttt{Attn}(\mathbf{X_Q}, \mathbf{X_K}, \mathbf{X_V}) = \mathbf{W}^{\text{out}} \frac{\sum_{s=1}^{S} \mathbf{V}_s \phi(\mathbf{K}_s^\top) \phi(\mathbf{Q})}{\sum_{s=1}^{S} \phi(\mathbf{K}_s^\top) \phi(\mathbf{Q})}. \tag{8}$$

C. Dynamic Low-rank Classifiers

To provide semantic guidance before the final output, we apply deep supervision at each decoder stage. Instead of the commonly used linear classifiers which are independent across stages and kept static for all input images, we use the class embeddings $\mathbf{E}_l$ as dynamic classifiers, which automatically adapt to the current input and propagate semantic information across layers of different resolutions. Specifically, we linearly project both the class embeddings and image features to the same dimension D^{cls}, and compute the segmentation map $\mathbf{Z}_l \in \mathbb{R}^{N \times H_{5-l} \times W_{5-l}}$ with dot product followed by a softmax normalization:

$$\mathbf{Z}_l = \texttt{Softmax}((\mathbf{W}_l^{\text{cls}} \mathbf{E}_l)^\top \cdot \mathbf{W}_l^{\text{feat}} \mathbf{Y}_l^{\text{out}}), \tag{9}$$

where $\mathbf{W}_l^{\text{cls}} \in \mathbb{R}^{D^{\text{cls}} \times D^{\text{cls}}}$ and $\mathbf{W}_l^{\text{feat}} \in \mathbb{R}^{D^{\text{cls}} \times D_l}$ are learnable projection matrices.

TABLE II: Results comparing segmentation decoders using different image encoders. ResNet50 + U-Net is a common baseline. "TF": Transformer, "VM": VMamba. Mean ± std over 3 seeds. **Bold**: best in the group; underline: second best in the group; green background: better than U-Net in the same group; yellow background: worse than U-Net in the same group; red background: worse than ResNet50 + U-Net.

Encoder	Enc. Type	Decoder	Dec. Type	Skip Con.	Params (M)	FLOPs (G)	AMOS	BTCV
Group 0: ResNet 50 + U-Net as baseline								
ResNet 50	CNN	U-Net	CNN	concat	40.17	7.62	63.44±0.30	73.77±0.50
Group 1: Tiny-scale encoder + decoder								
ConvNeXt v2 tiny	CNN	U-Net	CNN	add	41.53	7.48	65.45±0.20	74.01±0.36
PVT v2 b2	TF	EMCAD	CNN	add	26.58	5.24	**66.30**±0.19	**75.19**±0.18
Swin v1 tiny	TF	Swin-UNet	TF	add	39.74	7.41	66.08±0.26	73.22±0.35
VMamba tiny	VM	Mamba-UNet	VM	add	47.23	8.56	65.51±0.51	74.93±0.15
VMamba tiny	VM	MO-UNet	CNN	add	46.09	8.07	65.83±0.23	75.02±0.17
Group 2: Base-scale encoder + decoder								
ConvNeXt v2 base	CNN	U-Net	CNN	add	110.20	19.68	67.19±0.28	75.01±0.97
PVT v2 b5	TF	EMCAD	CNN	add	80.55	12.21	66.28±0.17	74.56±0.43
Swin v1 base	TF	Swin-UNet	TF	add	106.68	19.52	**67.36**±0.33	73.64±0.23
VMamba base	VM	Mamba-UNet	VM	add	116.93	20.62	66.97±0.16	75.87±0.27
VMamba base	VM	MO-UNet	CNN	add	115.07	19.87	66.83±0.13	**76.06**±0.13
Group 3: Swin v2 tiny + decoder								
Swin v2 tiny	TF	U-Net	CNN	add	41.25	10.56	68.52±0.20	76.73±0.21
Swin v2 tiny	TF	EMCAD	CNN	add	31.57	9.63	**68.85**±0.29	76.48±0.14
Swin v2 tiny	TF	Swin-UNet	TF	add	39.80	10.46	67.76±0.35	75.58±0.33
Swin v2 tiny	TF	Mamba-UNet	VM	add	45.42	11.52	68.02±0.34	76.30±0.31
Swin v2 tiny	TF	MO-UNet	CNN	add	44.28	10.87	68.25±0.19	**76.85**±0.15

IV. Experiments

We conduct experiments to answer the following research questions:

- **RQ1:** Do current Transformer- or Mamba-based methods really outperform CNNs in medical image segmentation?
- **RQ2:** Can the proposed BIDFormer effectively enhance CNN-based decoders?
- **RQ3:** Do the components in BIDFormer contribute to its performance gain?

A. Experiment Settings

1) Datasets: We conduct experiments using datasets of different modalities: ACDC [40] for cardiac MRI, AMOS [41] for abdomen MRI and BTCV [42] for abdomen CT. For AMOS, we keep only the axial images, and use the extended test set from U-Mamba [34].

2) Baselines: For both the U-Net empirical study and comparison with our method, we choose representative baselines of different types:

- CNN: U-Net [1] as a classical baseline, and EMCAD [37] as SOTA;
- Transformer: Swin-UNet [5] using shifted window attention [24], MISSFormer [43] using efficient attention [44], and MCADS [45] as SOTA combining CNN and standard attention;
- Mamba: Mamba-UNet [36] applying VSS, and Polyp-Mamba [7] combining CNN and VSS.

3) Implementation details: We implement and train all methods with PyTorch 2.6.0 + CUDA 12.6 in an NVIDIA GeForce RTX 4090 GPU with 24G memory, and replicate results using 3 different random seeds. For all datasets involved, we cut the 3D volumes into 2D slices along the axial direction, and resize them according to the input resolution of the pre-trained image encoder (256×256 for Swin v2 tiny, and 224×224 for others). For deep supervision, the decoder features from all stages are up-sampled to the original input size $H_0 \times W_0$ for pixel-wise classification.

We conduct all experiments with batch size 8, using the AdamW optimizer [46] as an enhanced version of Adam [47], with momentum coefficients $\beta_1 = 0.9, \beta_2 = 0.999$. Weight decay is set to 2×10^{-5} for AMOS and 1×10^{-5} for other datasets. As for learning rate, we first use 5 epochs for warmup, where the learning rate linearly increases from 2×10^{-5} to the initial learning rate 1×10^{-4}. For the 195 epochs left, the learning rate is scheduled with cosine annealing [48].

As a common practice, we use a combination of Dice loss [49] and cross entropy loss:

$$\mathcal{L} = \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{ce}}. \tag{10}$$

4) Evaluation metrics: We evaluate the methods using final Dice Similarity Coefficient (DSC) for each class, as well as the mean DSC (mDSC) across all classes except background.

B. U-Net Empirical Study (RQ1)

First, we conduct extensive experiments in AMOS and BTCV to show that current Transformer-based or Mamba-based methods do not necessarity surpass CNN-based methods under fair circumstances. Note that although nnUNet Revisited [50] made similar claims for auto-configuring 3D methods, we focus on comparing 2D methods using pre-trained image encoders. The mDSC results are listed in Table II, and the per-class DSC results are shown in Figure 3.

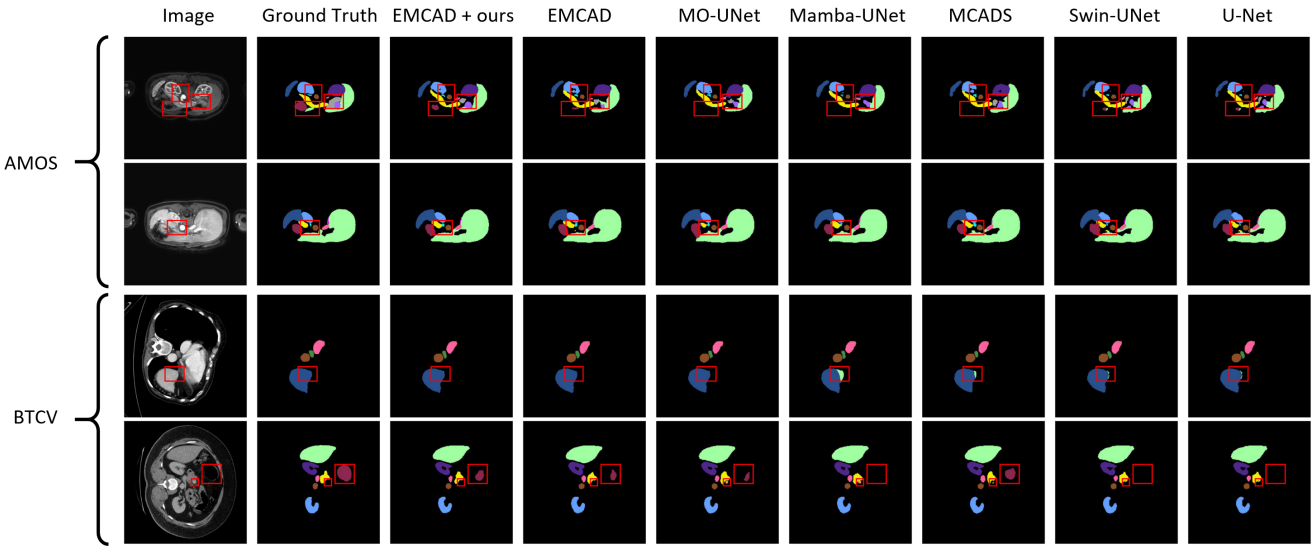
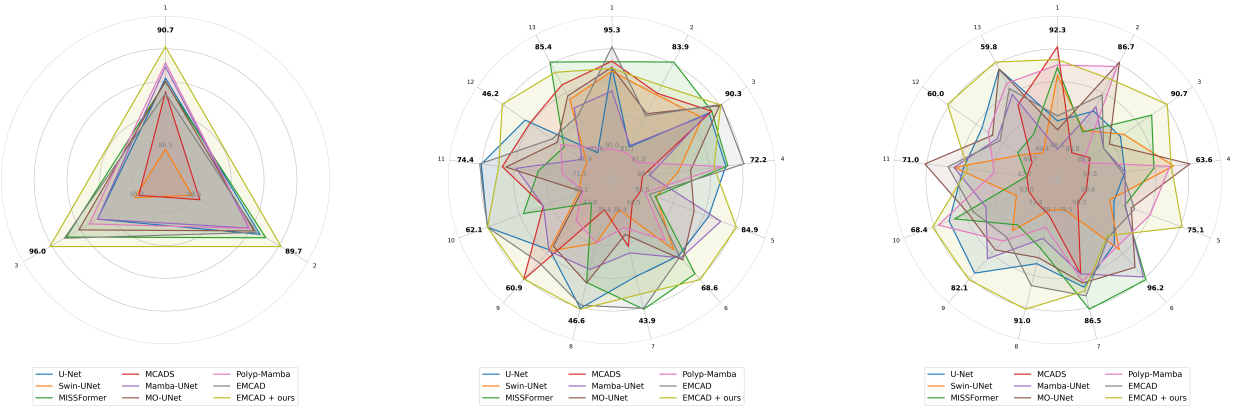


Fig. 4: Qualitative results for AMOS and BTCV. Differences among results from ACDC are barely recognizable, so they are not shown here. Best viewed in color, and zoom in for more details. From left to right, the columns correspond to the input images, ground truth masks, and results from different methods. The first two rows are from AMOS, and the last two rows are from BTCV. Note the differences marked with red boxes.



(a) ACDC (b) AMOS (c) BTCV

Fig. 5: Dice score per class for comparison with our proposed BIDFormer. *What the subfigures mean:* We compare segmentation decoders with our proposed BIDFormer, all using Swin v2 tiny as the encoder. From left to right are results for 3 datasets: ACDC, AMOS, and BTCV. *What the axes mean:* Each axis represents a class, and is scaled according to dice scores of all methods for the corresponding class in the dataset.

TABLE III: Results comparing our BIDFormer with baseline methods across 3 datasets: ACDC, AMOS, and BTCV. We implement all methods using Swin v2 tiny as the image encoder. Mean ± std over 3 seeds. **Bold**: best; underline: second best; green background: better than U-Net; yellow background: worse than U-Net.

Decoder	Decoder Type	Attn. Type	Source	Params (M)	FLOPs (G)	Mem. (GiB)	Time (ms)	AMOS	BTCV	ACDC
U-Net	CNN	-	MICCAI'15	41.25	10.56	6.22	25.8	68.52±0.20	76.73±0.21	91.53±0.15
Swin-UNet	TF	Swin attn.	ECCVW'22	39.80	10.46	6.77	31.0	67.76±0.35	75.58±0.33	90.90±0.08
MISSFormer	TF	efficient attn.	arxiv'21	52.12	21.07	8.14	42.2	68.26±0.13	76.18±0.47	91.59±0.08
MCADS	CNN + TF	standard attn.	CVPR'25	42.84	14.06	22.64	32.6	68.20±0.30	75.42±0.27	91.17±0.10
Mamba-UNet	VM	-	arxiv'24	45.42	11.52	8.40	36.0	68.02±0.34	76.30±0.31	91.54±0.05
MO-UNet	CNN	-	this paper	44.28	10.87	7.06	28.0	68.25±0.19	76.85±0.15	91.53±0.10
Polyp-Mamba	CNN + VM	-	MICCAI'24	50.43	12.12	8.68	48.6	66.94±0.16	76.73±0.48	91.57±0.11
EMCAD	CNN	-	CVPR'24	31.57	9.63	6.67	31.0	68.85±0.29	76.48±0.14	91.51±0.12
EMCAD + ours	CNN + TF	BLA	this paper	35.47	9.86	7.02	36.8	**69.22**±0.45	**77.58**±0.41	**91.84**±0.08

TABLE IV: Class names and visualization colors.

Class Name	Color	BTCV ID	AMOS ID
Background		0	0
Spleen		1	3
Right Kidney		2	2
Left Kidney		3	13
Gallbladder		4	9
Esophagus		5	10
Liver		6	1
Stomach		7	11
Aorta		8	5
Inferior Vena Cava		9	6
Portal Vein and Splenic Vein		10	-
Pancreas		11	4
Right Adrenal Gland		12	7
Left Adrenal Gland		13	8
Duodenum		-	12

1) Influence of image encoders: Empirically, we find that image encoders contribute substantially to segmentation performance, with consistent improvements for most decoders in both datasets. Hence, we find it necessary to equip all decoders with the same image encoder for fair comparison. For that purpose, we recommend Swin v2 as an outstanding option. Although scaling up from tiny-scale to base-scale encoders results in notable performance gains (except PVT v2 + EMCAD), they come at the cost of at least double computation. Instead, using Swin v2 tiny brings even higher Dice scores with limited computational overhead.

2) Ablation on SSM: MO-UNet removes the SSM from Mamba-UNet decoder and keeps everything else the same; we compare their performance together with other decoders in Table II and Figure 3. Interestingly, removing SSMs from the decoder (MO-UNet) does not necessarily degrade performance; in fact, it improves mDSC in most cases. The results suggest that for medical image segmentation, state-space models are not really superior to CNN-based methods, since the scanning routes in SSMs inevitably break the spatial relationships among image patches.

3) Decoder comparison: When comparing different decoders, we find that CNN-based methods, including vanilla U-Net, perform competitively or even better than Transformer-based and Mamba-based methods. Although getting lower scores under some circumstances, U-Net gains more advantages as the encoder gets stronger. Across all "encoder + decoder" settings in both datasets, Swin v2 tiny + U-Net consistently ranks within top 2 (68.52% in AMOS, and 76.73% in BTCV), only inferior to other CNN decoders (Swin v2 tiny + EMCAD in AMOS with 68.85%, and Swin v2 tiny + MO-UNet in BTCV with 76.85%).

C. Influence of BIDFormer (RQ2)

1) Qualitative comparison: We present qualitative comparisons of segmentation results in Figure 4. Our proposed BIDFormerintegrated into EMCAD demonstrates superior performance in accurately delineating organ boundaries compared to other methods. Corresponding class names and visualization colors are listed in Table IV.

2) Quantitative comparison: To demonstrate the effectiveness of our proposed BIDFormer, we integrate it into the EMCAD decoder, forming a hybrid architecture that combines CNN and Transformer components. We compare this modified EMCAD with the original EMCAD and other baselines using Swin v2 tiny as the encoder. As shown in Table III and Figure 5, adding BIDFormer to EMCAD leads to consistent performance improvements across all evaluated datasets. Specifically, the mDSC increases from 68.85% to 69.22% on AMOS, from 76.48% to 77.58% on BTCV, and from 91.51% to 91.84% on ACDC. These results highlight the capability of BIDFormer to enhance segmentation performance, with limited additional computational cost (approximately 3.9M more parameters, 0.23G more FLOPs, 0.35GiB more memory, and 5.8ms more inference time).

TABLE V: Ablation study on different components of BIDFormer using EMCAD as the baseline decoder. "Lin. Attn.": Linear Attention. "Bid.": Bidirectional. Mean ± std over 3 seeds in ACDC dataset.

NO.	Lin. Attn.	Bid.	DLC	Mem. (GiB)	mDSC
0	-	-	-	6.67	91.51±0.12
1	✗	✓	✗	6.97	91.55±0.22
2	✗	✓	✓	7.04	91.77±0.19
3	✓	✗	✓	7.00	91.76±0.08
4	✓	✓	✗	6.90	91.61±0.06
5	✓	✓	✓	7.02	**91.84**±0.08

D. Ablation Study (RQ3)

We conduct ablation studies in ACDC to validate the efficacy of our core components. From results shown in Table V, we make several observations:

- *Softmax attention (1,2) vs. linear attention (4,5):* As an efficient alternative to softmax attention, linear attention surprisingly gets higher mDSC in our setting. By eliminating the softmax operation, our method get better results using less computation.
- *Unidirenctional attention (3) vs. bidirectional attention (5):* Using only unidirectional CP attention and dynamic classifiers without PC attention gets 91.76 mDSC. It shows that PC attention is important, since it reallocates global context to patch tokens and enrich their information, so our method functions more similar to self-attention.
- *Static classifiers (1,4) vs. Dynamic classifiers (2,5):* Both softmax attention and linear attention can insert useful information into class embeddings, so they can serve as better classifiers than static ones.

V. Conclusion

In summary, we have conducted an empirical study on U-shaped networks for medical image segmentation, revealing that CNN-based methods can actually outperform

Transformer- or Mamba-based methods under fair comparison settings. To address the limitations of existing Transformer-based methods, we have proposed BIDFormer, a plug-and-play module that enhances CNNs with both long-range context modeling and in-depth semantic guidance. Extensive experiments on multiple public datasets have demonstrated the superior performance of BIDFormer over state-of-the-art methods.

ACKNOWLEDGMENT

This paper is supported by the Fundamental Research Funds for the Central Universities (project number YG2024ZD06).

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[2] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*. Cham: Springer International Publishing, 2014, pp. 818–833.

[3] W. Luo, Y. Li, R. Urtasun *et al.*, "Understanding the effective receptive field in deep convolutional neural networks," *NeurIPS*, vol. 29, 2016.

[4] J. Chen, Y. Lu, Q. Yu *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021.

[5] H. Cao, Y. Wang, J. Chen *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *ECCVW*, 2022.

[6] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," in *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[7] Z. Xu, F. Tang, Z. Chen *et al.*, " Polyp-Mamba: Polyp Segmentation with Visual Mamba ," in *MICCAI*, vol. LNCS 15008. Springer Nature Switzerland, 2024.

[8] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *NeurIPS*, 2017.

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *CoRR*, vol. abs/2312.00752, 2024.

[10] T. Dao and A. Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," in *ICML*, 2024.

[11] V. K. Vasa, W. Zhu *et al.*, "Sta-unet: Rethink the semantic redundant for medical imaging segmentation," *CoRR*, vol. abs/2410.11578, 2024.

[12] F. Dalvi, H. Sajjad *et al.*, "Analyzing redundancy in pretrained transformer models," in *EMNLP*, 2020.

[13] X. Liu, C. Zhang, and L. Zhang, "Vision mamba: A comprehensive survey and taxonomy," *CoRR*, vol. abs/2405.04404, 2024.

[14] R. Xu, S. Yang, Y. Wang *et al.*, "Visual mamba: A survey and new outlooks," *CoRR*, vol. abs/2404.18861, 2024.

[15] S. Bansal, S. A, M. P. J *et al.*, "A comprehensive survey of mamba architectures for medical image analysis: Classification, segmentation, restoration and beyond," *CoRR*, vol. abs/2410.02362, 2024.

[16] W. Yu and X. Wang, "Mambaout: Do we really need mamba for vision?" in *CVPR*, 2025, pp. 4484–4496.

[17] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016.

[18] Y. Liu, Y. Tian, Y. Zhao *et al.*, "Vmamba: Visual state space model," in *NeurIPS*, vol. 37. Curran Associates, Inc., 2024, pp. 103 031–103 063.

[19] C. Liang-Chieh, G. Papandreou, I. Kokkinos *et al.*, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[20] L.-C. Chen, Y. Zhu, G. Papandreou *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 801–818.

[21] H. Zhao, J. Shi, X. Qi *et al.*, "Pyramid scene parsing network," in *CVPR*, 2017, pp. 2881–2890.

[22] A. Kirillov, R. Girshick, K. He *et al.*, "Panoptic feature pyramid networks," in *CVPR*, 2019, pp. 6399–6408.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[24] Z. Liu, Y. Lin, Y. Cao *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[25] Z. Liu, H. Hu, Y. Lin *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *CVPR*, 2022, pp. 12 009–12 019.

[26] S. Zheng, J. Lu, H. Zhao *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021, pp. 6881–6890.

[27] E. Xie, W. Wang, Z. Yu *et al.*, "Segformer: Simple and efficient design for semantic segmentation with transformers," *NeurIPS*, vol. 34, pp. 12 077–12 090, 2021.

[28] A. Kirillov, E. Mintun, N. Ravi *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[29] R. Strudel, R. Garcia, I. Laptev *et al.*, "Segmenter: Transformer for semantic segmentation," in *ICCV*, 2021, pp. 7262–7272.

[30] B. Cheng, I. Misra, A. G. Schwing *et al.*, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022, pp. 1290–1299.

[31] B. Zhang, L. Liu, M. H. Phan *et al.*, "Segvitv2: Exploring efficient and continual semantic segmentation with plain vision transformers," *IJCV*, 2023.

[32] L. Zhu, B. Liao, Q. Zhang *et al.*, "Vision mamba: Efficient visual representation learning with bidirectional state space model," in *ICML*, ser. PMLR, vol. 235. PMLR, 2024, pp. 62 429–62 442.

[33] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," in *CVPR*, 2025, pp. 25 261–25 270.

[34] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *CoRR*, vol. abs/2401.04722, 2024.

[35] J. Liu, H. Yang, H.-Y. Zhou *et al.*, "Swin-umamba: Mamba-based unet with imagenet-based pretraining," in *MICCAI*. Springer, 2024, pp. 615–625.

[36] Z. Wang, J.-Q. Zheng, Y. Zhang *et al.*, "Mamba-unet: Unet-like pure visual mamba for medical image segmentation," *CoRR*, vol. abs/2402.05079, 2024.

[37] M. M. Rahman, M. Munir, and R. Marculescu, "Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation," in *CVPR*, 2024, pp. 11 769–11 779.

[38] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *CoRR*, vol. abs/1606.08415, 2023.

[39] A. Katharopoulos, A. Vyas, N. Pappas *et al.*, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *ICML*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 2020, pp. 5156–5165.

[40] O. Bernard, A. Lalande, C. Zotti *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *TMI*, 2018.

[41] Y. Ji, H. Bai, C. Ge *et al.*, "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," *NeurIPS*, 2022.

[42] B. Landman, Z. Xu, J. E. Igelsias *et al.*, "2015 miccai multi-atlas labeling beyond the cranial vault workshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[43] X. Huang, Z. Deng, D. Li *et al.*, "Missformer: An effective medical image segmentation transformer," *CoRR*, vol. abs/2109.07162, 2021.

[44] W. Wang, E. Xie, X. Li *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021, pp. 568–578.

[45] S. Wazir and D. Kim, "Rethinking decoder design: Improving biomarker segmentation using depth-to-space restoration and residual linear attention," in *CVPR*, 2025, pp. 30 861–30 871.

[46] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *CoRR*, vol. abs/1711.05101, 2019.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[48] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *CoRR*, vol. abs/1608.03983, 2017.

[49] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*. Ieee, 2016, pp. 565–571.

[50] F. Isensee, T. Wald, C. Ulrich *et al.*, "nnu-net revisited: A call for rigorous validation in 3d medical image segmentation," in *MICCAI*. Cham: Springer Nature Switzerland, 2024, pp. 488–498.