Adaptive Morph-Patch Transformer for Aortic Vessel Segmentation

Zhenxi Zhang[1,2*], Fuchen Zheng[3,1], Adnan Iltaf[1], Yifei Han[2], Zhenyu Cheng[1], Yue Du[1], Bin Li[1†], Tianyong Liu[4,1†], Shoujun Zhou[1†]

[1] Institute of Scientific Instrumentation, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2] Department of Health Technology and Informatics, The Hong Kong Polytechnic University
[3] Department of Computer and Information Science, University of Macau
[4] School of Computer Science and Technology, Tongji University
{zx.zhang3, adnan, zy.cheng, yue.du2, b.li2, sj.zhou}@siat.ac.cn, yc37950@um.edu.mo, 20100916d@connect.polyu.hk, tianyong@tongji.edu.cn

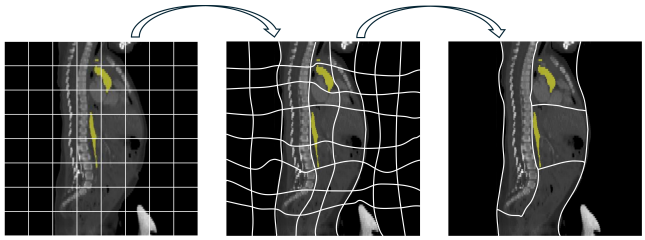
Abstract

Accurate segmentation of aortic vascular structures is critical for diagnosing and treating cardiovascular diseases. Traditional Transformer-based models have shown promise in this domain by capturing long-range dependencies between vascular features. However, their reliance on fixed-size rectangular patches often influences the integrity of complex vascular structures, leading to suboptimal segmentation accuracy. To address this challenge, we propose the adaptive Morph-Patch Transformer (MPT), a novel architecture specifically designed for aortic vascular segmentation. Specifically, MPT introduces an adaptive patch partitioning strategy that dynamically generates morphology-aware patches aligned with complex vascular structures. This starategy can preserve semantic integrity of complex vascular structures within individual patches. Moreover, a Semantic Clustering Attention (SCA) method is proposed to dynamically aggregate features from various patches with similar semantic characteristics. This method enhances the model's capability to segment vessels of varying sizes, preserving the integrity of vascular structures. Extensive experiments on three open-source datasets (AVT, AortaSeg24 and TBAD) demonstrate that MPT achieves state-of-the-art performance, with improvements in segmenting intricate vascular structures.

Code — https://github.com/iCherishxixixi/MPTransformer

Introduction

Cardiovascular diseases (CVDs) remain a primary cause of morbidity and mortality globally, emphasizing the necessity for precise and timely diagnosis to improve patient outcomes (Townsend et al. 2022; Tang et al. 2025). Aortic vascular segmentation, which involves delineating the structure of the aorta and its branches from medical images, plays a pivotal role in diagnosing and planning treatments for various cardiovascular conditions (Hahn, Baeumler, and Hsiao 2021), including aneurysms, dissections, and stenosis. The precision of vascular segmentation directly impacts the reliability of downstream tasks (Yagis et al. 2024; Xie et al. 2025), such as computational flow modeling, surgical planning, and disease progression monitoring. In recent years, deep learning models have revolutionized aorta segmentation, with Transformer-based architectures emerging as a dominant paradigm (Lin et al. 2023; Li et al. 2024). These models excel at capturing long-range dependencies and contextual information, making them well-suited for handling vascular structures, which extend across large spatial regions (Dosovitskiy et al. 2021; Zhang, Li, and Wang 2024).



(a) Traditional Fixed Patch Partition (b) Proposed Morph-Patch Partition (c) Proposed Semantic Clustering

Figure 1: Transformer improvements for tailoring vascular structure segmentation. (a) Traditional Transformers create fixed-size patches. (b) Our method creates morphed patches. (c) Morphed patches are grouped based on semantic similarities.

However, applying traditional Transformers (Hatamizadeh et al. 2022; Wang et al. 2021; Zhou et al. 2023) to vascular segmentation remains following challenges: (1) **Fixed patch partition hinders extraction of complex vascular shapes.** As shown in Fig. 1 (a), traditional Transformers divide images into fixed-size patches, which struggle to preserve the semantic integrity of intricate blood vessels. To accommodate complex morphological structures, DCN (Dai et al. 2017) introduces a learnable deformation field to capture morphology-aware receptive fields in CNNs. Building on this, DPT (Chen et al. 2021b) extends this method to Transformers, generating variable-sized rectangular patches in a data-driven manner. Although DPT is effective for natural images, it struggles to frame fragile and thin vessels within rectangular patches. Recent Transformer-based models (Zhu et al. 2024; Jian

*Work done during the internship at SIAT.
†Corresponding author.

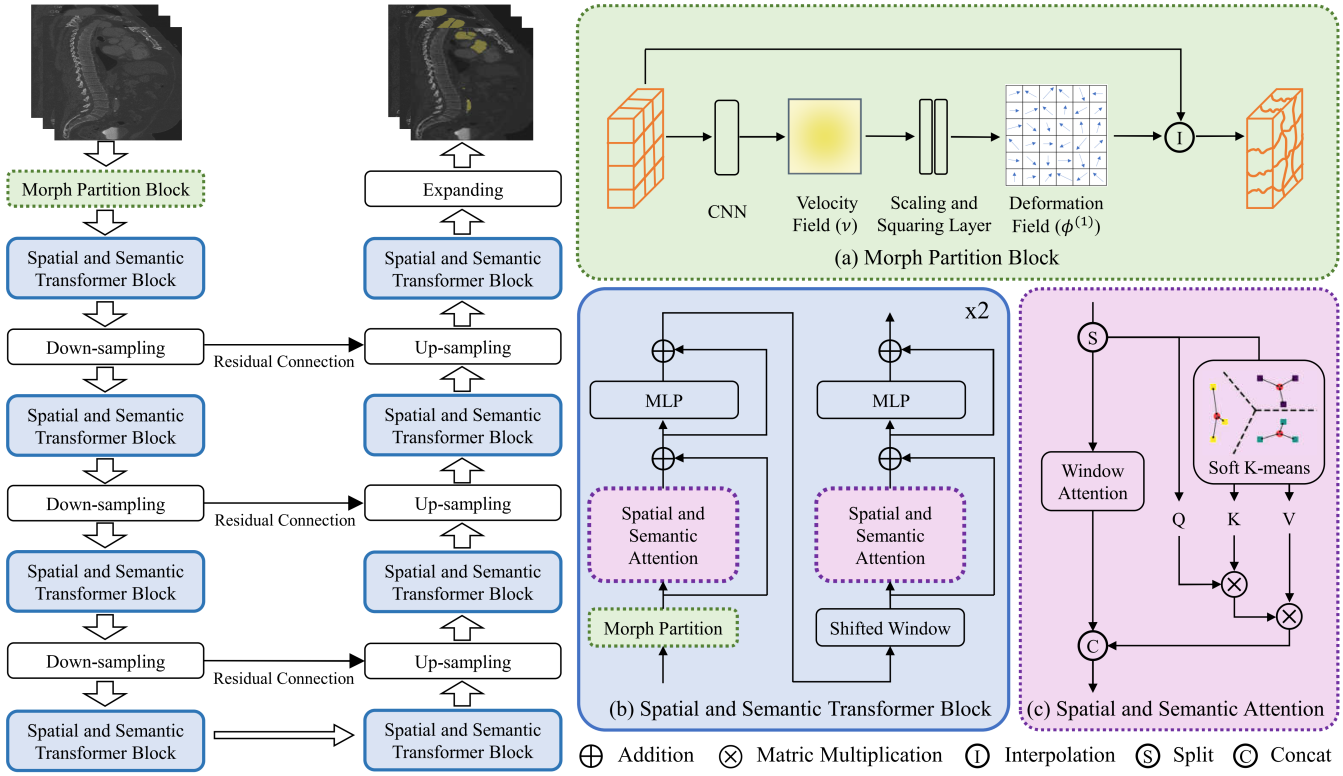


Figure 2: Overall architecture of the proposed Morph-Patch Transformer for vascular structure segmentation. (a) Morph partition block to create morphology-aware patches. (b) The Transformer block with spatial and semantic attention. (c) Spatial and semantic attention is combined by window attention and semantic clustering attention.

et al. 2025; Xian et al. 2025) adopt hybrid architectures to extract features of elongated vessels. Specifically, these Transformer-based models integrate Snake Convolution (Qi et al. 2023), which imposes prior constraints on the deformation field, to enhance feature extraction for elongated vessels. Although these methods adapt Transformers for vessel segmentation, they still face challenges in accurately capturing fine vessel details due to rectangular patch partition. (2) **Single-scale patches hinder the learning of semantic similarities at different scales.** Pyramid Transformer (Zhang et al. 2020) enables multi-scale feature extraction with its pyramid structure, but its fixed design limits adaptability to targets of varying scales. To address this, Swin-Transformer (Liu et al. 2021; Zhou et al. 2023) introduces hierarchical window attention, which not only captures multi-scale features more effectively but also reduces computational complexity. Despite these advancements, Swin-Transformer's reliance on fixed windows remains a limitation in modeling highly complex structures. Latest studies (Xia et al. 2022; Azad et al. 2024) have drawn inspiration from DCN to develop deformable attention mechanisms, enabling dynamic window adjustments. Nevertheless, these methods fail to incorporate semantic similarities, limiting their overall effectiveness.

To address these limitations, we propose the adaptive Morph-Patch Transformer (MPT), a novel model tailored for aortic vascular segmentation. MPT introduces an adaptive patch partitioning strategy guided by a velocity field. Unlike traditional deformable convolutions (Dai et al. 2017; Qi et al. 2023), which directly generate deformation fields, our method first constructs a velocity field and iteratively computes the final deformation shown in Fig. 1 (b). This method aligns patches more naturally with vascular structures, enabling the model to capture fine-grained vessel features without compromising topological integrity (Mukherjee et al. 2015). Furthermore, we propose a clustering attention mechanism that dynamically aggregates features from patches with similar semantic characteristics. This approach allows the model to effectively capture various vascular structures with semantic consistency as shown in Fig. 1 (c). By adaptively extracting vascular features at different scales, it enhances the segmentation of vessels, particularly for those with varying sizes. Extensive experiments on benchmark datasets demonstrate that MPT achieves state-of-the-art performance, outperforming other advanced Transformer-based models. The proposed method significantly improves segmentation accuracy for complex vascular structures, particularly in capturing fine-grained details and preserving topological integrity. Our contributions are summarized as follows:

- We introduce a novel **adaptive Morph-Patch Transformer (MPT)** that utilizes an adaptive patch partition-

ing strategy on a velocity field. This method effectively preserves vascular topology and enhances the alignment of patches with intricate blood vessel structures.

- We propose a **Semantic Clustering Attention (SCA)** method that dynamically aggregates features from patches with similar semantic characteristics, enhancing the segmentation of vessels across varying sizes.
- We conduct extensive experiments on three open-source dataset(AVT, AortaSeg24 and TBAD), demonstrating that MPT achieves **State-of-the-Art** performance, particularly in segmenting intricate vascular structures.

Proposed Method

Overall Morph-Patch Transformer Architecture

In this work, we propose the Morph-Patch Transformer (MPT), a novel architecture designed to address two key challenges in vascular structure segmentation. First, MPT tackles the complex and irregular morphological structures of vessels through morphology-aware patch partitioning, which adapts to the inherent geometry of vascular networks. Second, MPT addresses the need to capture multi-scale semantic features by introducing fusion attention, a mechanism that models both spatial and semantic contextual relationships across different scales. Built on a 3D UNet-like framework (Çiçek et al. 2016), MPT combines these innovations to achieve precise and robust segmentation. The key components are illustrated in Fig. 2 and detailed below.

- **Morph Partition Block:** This block first utilizes a CNN to predict the velocity field shown in Fig. 2 (a). The predicted velocity field is then transformed into a diffeomorphic deformation field through the scaling and squaring method (Arsigny et al. 2006), where each point in the field represents coordinate offsets. This diffeomorphic nature not only guarantees smooth and invertible transformations but also naturally preserves the continuity of vascular structures throughout the deformation process. Finally, the coordinate offsets are used to generate deformed features through bilinear interpolation of the original input. From these transformed features, the method extracts patches that effectively capture complex vessel structures while preserving their topological relationships.
- **Spatial and Semantic Transformer Block:** This block is designed to effectively integrate spatial and semantic contextual relations for vascular feature learning. As shown in Fig. 2 (b), the module builds upon SwinTransformer's (Liu et al. 2021) window attention and shift window strategies to capture spatial relationships. To address the complexity of vascular structures, we introduce the Morph Partition Block (Fig. 2 (a)), which dynamically adapts window shapes to fit varying morphological patterns. Furthermore, to enhance semantic understanding, Semantic Clustering Attention (SCA) (Fig. 2 (c)) is proposed, where a soft-Kmeans algorithm extracts key semantics and computes their relationships with patch features. These components are integrated into a 3D UNet-based framework (Çiçek et al. 2016), enabling multi-scale fusion of spatial and semantic relations.

Algorithm 1: Morph-Patch Feature Extraction with Diffeomorphic Deformation Field

Input: Feature map x, patch center p_0, neighborhood region $\mathcal{R}$, velocity field υ, number of steps n
Parameter: Spatial weighting function w
Output: Morph-patch feature $y(p_0)$

Let $\Delta t \leftarrow \frac{1}{2^n}$
Initialize deformation field $\phi^{(0)} \leftarrow Id$
Compute initial deformation: $\phi^{(\Delta t)} \leftarrow (Id + \upsilon(\Delta t)) \circ \phi^{(0)}$
for $i = n$ to 1 **do**
    $\phi^{(\frac{1}{2^{i-1}})} \leftarrow \phi^{(\frac{1}{2^i})} \circ \phi^{(\frac{1}{2^i})}$
end for
Set final deformation field $\phi \leftarrow \phi^{(1)}$
Initialize $y(p_0) \leftarrow 0$
for each $p_n \in \mathcal{R}$ **do**
    $y(p_0) \leftarrow y(p_0) + w(p_n) \cdot x(p_0 + p_n + \phi(p_0 + p_n))$
end for
return $y(p_0)$

Morph-Patch Feature

The morph-patch feature is derived from the morph partition block, as detailed in **Algorithm 1**. This feature is designed to extract complex vascular structures and is formulated as:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \phi(p_0 + p_n)), \quad (1)$$

where w is the feature weight, p_0 represents the center of a patch, p_n denotes its neighboring regions within the patch region $\mathcal{R}$, and $\phi(p_0 + p_n)$ represents the deformation field, which provides coordinate offsets to adaptively adjust the sampling locations. To ensure that the transformation preserves the topological properties of the features before and after deformation, we employ a stationary velocity field υ to generate a diffeomorphic mapping. The velocity field υ is integrated over time $t = [0, 1]$ to obtain the final deformation field $\varphi^{(1)}$, as described by the Ordinary Differential Equation (ODE):

$$\frac{\partial \phi^{(t)}}{\partial t} = \upsilon(\phi^{(t)}), \quad with\ \phi^{(0)} = Id, \quad (2)$$

where Id represents the identity transformation. This formulation ensures that $\phi^{(t)}$ is a diffeomorphism, meaning it is smooth, invertible, and preserves the topological structure of the input features.

The deformation field $\phi^{(1)}$ is computed numerically using the scaling and squaring method (Dalca et al. 2019). This approach leverages the properties of one-parameter subgroups of diffeomorphisms, where the velocity field υ lies in the Lie algebra and is exponentiated to generate the deformation field $\phi^{(1)}$ in the Lie group. To ensure numerical stability and accuracy, the computation begins by calculating the deformation field such that t is sufficiently small ($t = \frac{1}{2^n} \approx 0$). With the scaled velocity field, the initial deformation field is

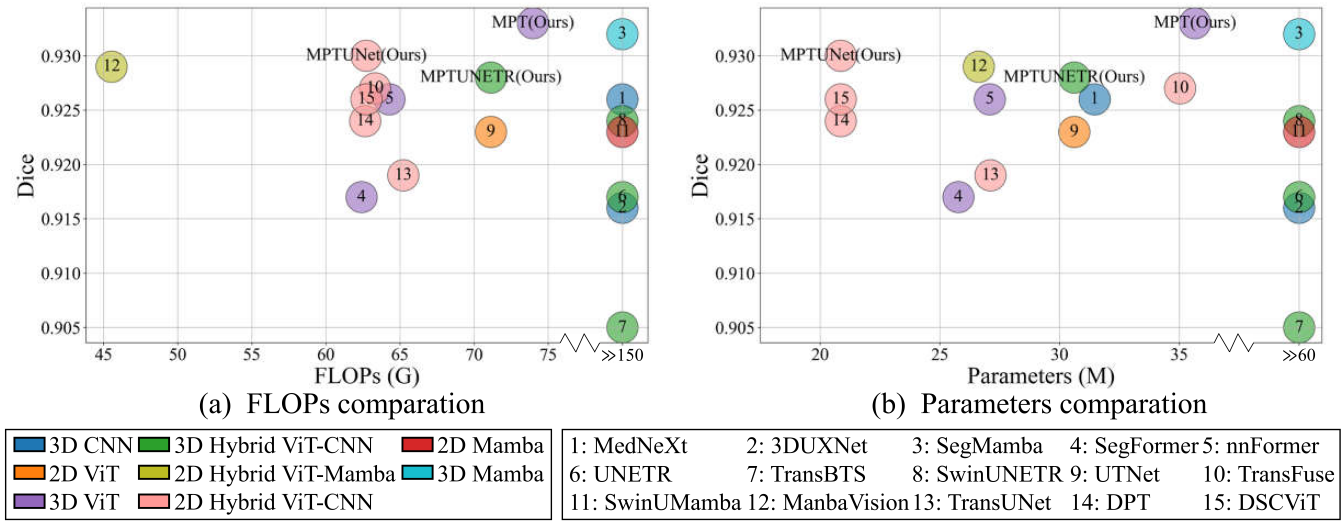


Figure 3: Accuracy–Efficiency Comparison on TBAD: (a) FLOPs, (b) Parameters; Colors Indicate Backbone Types.

represented as:

$$\phi^{(\frac{1}{2^n})} = (Id + v(\frac{1}{2^n})) \circ \phi^{(0)}, \tag{3}$$

where $\circ$ represents the composition operation. The deformation field is then iteratively refined using the recurrence relation:

$$\phi^{(\frac{1}{2^{n-1}})} = \phi^{(\frac{1}{2^n})} \circ \phi^{(\frac{1}{2^n})}. \tag{4}$$

By constructing $\phi^{(1)}$ in this manner, the transformation inherently preserves the topological integrity of the input features. This is particularly crucial in applications such as medical image analysis, where maintaining the structural continuity of complex vascular features is essential.

Semantic Clustering Attention with Soft K-means

In this section, we propose a Soft K-means module for semantic clustering, which extracts core semantic features to model semantic relationships in the data. The module is formulated as follows:

$$f^s_{newcore} = \sum_{i=1}^{m} g_s(f^i) d(f^i, f^s_{core}), \tag{5}$$

where $F = \{f^1, f^2, ..., f^m\}$ represents all patch features, $F_{core} = \{f^1_{core}, f^2_{core}, ..., f^n_{core}\}$ denotes the original core semantic features, $F_{newcore} = \{f^1_{newcore}, f^2_{newcore}, ..., f^n_{newcore}\}$ represents the new core semantic features, $g_s(\cdot)$ reflects the importance of all features F to the original core semantic feature f^s_{core}, and $d(\cdot, \cdot)$ measures the similarity between two features. In conventional K-means algorithm, $g_s(\cdot)$ would be a non-differentiable discrete function. Specifically, $g_s(\cdot)$ would be 1 if f^i belongs to the cluster corresponding to f^s_{core}, and 0 otherwise. To ensure differentiability, we design a smoothed version of $g_s(\cdot)$:

$$g_s(f^i) = \frac{e^{-\beta \left\| f^i - f^s_{core} \right\|^2}}{\sum_{k=1}^{n} e^{-\beta \| f^i - f^k_{core} \|^2}}, \tag{6}$$

By defining $\lambda^s = 2\beta f^s_{core}$, $\mu^s = -\beta \left\| f^s_{core} \right\|^2$, and $d(\cdot, \cdot)$ as vector subtraction, f^s_{core} can be simplified to:

$$f^s_{newcore} = \sum_{i=1}^{m} \frac{e^{\lambda^s f^i + \mu^s}}{\sum_{k=1}^{n} e^{\lambda^k f^i + \mu^k}} (f^i - f^s_{core}). \tag{7}$$

In practice, λ, μ and F_{core} are learned by the neural network. The updated semantic centers $f_{newcore}$ are then applied to compute Semantic Clustering Attention (SCA), formulated as:

$$\text{SCA}(F) = \text{softMax}\left(\frac{(FW_Q)(F_{\text{newcore}} W_K)^T}{\sqrt{d}}\right) \cdot (F_{\text{newcore}} W_V) \tag{8}$$

where W_Q, W_K and W_V are learnable weight matrices. This attention mechanism enables the model to effectively integrate semantic relationships, enhancing the model's ability to capture complex patterns in the data.

Experiments

Dataset and Implementation

In this section, we evaluate the efficacy of MPT on three widely used open-source aorta datasets: AVT (Radl et al. 2022), TBAD (Yao et al. 2021), and AortaSeg24 (Imran et al. 2025). The AVT dataset contains 56 cases and focuses on a single-class segmentation task, requiring the extraction of the aorta from CTA scans collected from three hospitals. TBAD includes 100 high-resolution 3D CTA images of Type-B Aortic Dissection (TBAD), with detailed annotations of the True Lumen (TL), False Lumen (FL), and False Lumen Thrombus (FLT). These annotations test the model's ability to identify diseased aortic structures. AortaSeg24 is one of the most detailed aorta segmentation datasets, encompassing 100 CTA scans annotated with 23 clinically meaningful aortic regions. Segmenting numerous and tiny vascular anatomical structures poses significant challenges. As

Model	AVT			TBAD		
	Dice	mIoU	clDice	Dice	mIoU	clDice
MedNeXt	0.809(0.198)	0.718(0.233)	0.724(0.185)	0.926(0.172)	0.871(0.190)	0.880(0.072)
3DUXNet	0.843(0.095)	0.740(0.136)	0.745(0.105)	0.916(0.187)	0.859(0.207)	0.892(0.042)
SegMamba	0.829(0.134)	0.730(0.179)	0.711(0.177)	0.932(0.164)	0.881(0.178)	0.918(0.039)
SegFormer	0.837(0.124)	0.737(0.159)	0.770(0.136)	0.912(0.168)	0.854(0.181)	0.893(0.062)
nnFormer	0.835(0.158)	0.743(0.201)	0.732(0.169)	0.926(0.158)	0.871(0.173)	0.895(0.039)
MPT(Ours)	0.856(0.106)	0.762(0.150)	0.757(0.152)	**0.933(0.150)**	**0.881(0.161)**	0.915(0.032)
UTNETR	0.791(0.145)	0.677(0.181)	0.674(0.150)	0.917(0.157)	0.864(0.170)	0.901(0.044)
TransBTS	0.822(0.149)	0.722(0.190)	0.770(0.163)	0.905(0.172)	0.841(0.191)	0.899(0.050)
SwinUNETR	0.811(0.224)	0.729(0.248)	0.740(0.222)	0.924(0.187)	0.870(0.204)	0.906(0.060)
MPTUNETR(Ours)	0.829(0.162)	0.736(0.203)	0.764(0.180)	0.928(0.159)	0.874(0.174)	0.911(0.041)
UTNet	0.861(0.072)	0.762(0.106)	0.780(0.087)	0.923(0.150)	0.865(0.161)	0.911(0.025)
TransFuse	0.880(0.079)	0.794(0.115)	0.796(0.107)	0.927(0.139)	0.872(0.145)	0.895(0.063)
SwinUMamba	0.861(0.067)	0.761(0.100)	0.771(0.093)	0.923(0.150)	0.867(0.158)	0.902(0.056)
MambaVision	0.882(0.064)	0.795(0.097)	0.795(0.099)	0.929(0.148)	0.874(0.151)	0.914(0.037)
TransUNet	0.874(0.064)	0.782(0.093)	0.801(0.091)	0.919(0.146)	0.860(0.159)	0.894(0.051)
DPT	0.886(0.055)	0.800(0.086)	0.825(0.089)	0.924(0.137)	0.868(0.144)	0.917(0.026)
DSCViT	0.877(0.060)	0.786(0.090)	0.782(0.085)	0.926(0.151)	0.870(0.159)	0.893(0.062)
MPTUNet(Ours)	**0.896(0.046)**	**0.815(0.073)**	**0.839(0.078)**	0.930(0.147)	0.877(0.156)	**0.920(0.031)**

Table 1: Experimental results of aortic segmentation on AVT and TBAD dataset. Values are reported as "mean (standard deviation)".

Model	Dice	mIoU	clDice
MedNeXt	0.758 (0.077)	0.631 (0.080)	0.963 (0.018)
3DUXNet	0.784 (0.077)	0.666 (0.083)	0.964 (0.011)
SegMamba	0.747 (0.076)	0.620 (0.080)	0.924 (0.022)
SegFormer	0.753 (0.104)	0.630 (0.105)	0.951 (0.023)
nnFormer	0.779 (0.056)	0.666 (0.056)	0.923 (0.025)
MPT	0.804 (0.021)	0.690 (0.024)	0.926 (0.032)
UTNETR	0.753 (0.096)	0.630 (0.099)	0.916 (0.024)
TransBTS	0.721 (0.078)	0.594 (0.075)	0.928 (0.040)
SwinUNETR	0.781 (0.091)	0.664 (0.097)	0.937 (0.027)
MPTUNETR	**0.809 (0.045)**	**0.695 (0.054)**	0.955 (0.015)
UTNet	0.715 (0.056)	0.593 (0.059)	0.905 (0.035)
TransFuse	0.739 (0.054)	0.630 (0.057)	0.890 (0.047)
SwinUMamba	0.766 (0.080)	0.648 (0.083)	0.961 (0.017)
MambaVision	0.795 (0.071)	0.682 (0.076)	0.960 (0.021)
TransUNet	0.751 (0.103)	0.633 (0.097)	0.922 (0.045)
DPT	0.778 (0.075)	0.662 (0.079)	0.959 (0.017)
DSCViT	0.788 (0.077)	0.673 (0.077)	0.965 (0.019)
MPTUNet	0.796 (0.075)	0.686 (0.077)	**0.966 (0.011)**

Table 2: Experimental results of aortic segmentation on AortaSeg24 dataset. Values are reported as "mean (standard deviation)".

such, AortaSeg24 serves as a robust benchmark for evaluating the fine-grained anatomical segmentation capabilities of aortic segmentation models.

In our experiments, all models are implemented in PyTorch and trained on NVIDIA GeForce RTX 3090 GPUs with Ubuntu 20.04. To ensure fair and accurate comparison, we adopt the officially released codes of the referenced methods and conduct all experiments within the nnU-Net framework (Isensee et al. 2021). Specifically, we follow the preprocessing steps of nnU-Net, including normalization, resampling, and cropping. Each dataset is split into training, validation, and test sets in an 8:1:1 ratio, and image sizes are standardized to 128×128×128 for 3D and 512×512 for 2D to ensure consistent input dimensions. To verify the effectiveness of our proposed models, we provide three versions: MPT, a pure 3D ViT-based architecture; MPT-UNETR, a hybrid 3D ViT-CNN framework; and MPT-UNet, a lightweight 2D model. Additionally, the number of clusters is set to 32 in these networks. All three models are optimized using the Adam optimizer with a learning rate of 5×10^{-5}. Training follows the nnU-Net strategy (Isensee et al. 2021) and is terminated after 1000 epochs. The Dice coefficient is employed as the loss function to guide segmentation performance.

Baselines

To demonstrate the superiority of our proposed models in segmenting complex aortic morphological structures, we compare MPTs with a diverse range of recent and competitive segmentation approaches. These methods are categorized into eight groups based on their backbone architectures, as illustrated in Fig. 3. Among them, 3D Vision Transformer (ViT)-based models—including SegFormer (Perera, Navard, and Yilmaz 2024), nnFormer (Zhou et al. 2023), and the proposed MPT—represent a prominent and widely studied category. These models tokenize volumetric inputs and employ self-attention mechanisms to capture long-range dependencies and global context effectively. However, pure ViT architectures typically demand significant computational resources and large-scale data to learn inductive biases effectively (Lavie, Gur-Ari, and Ringel 2024). To mitigate these limitations, several methods adopt hybrid ViT-CNN designs to improve training efficiency and model generalizability in the 3D setting. Notable exam-

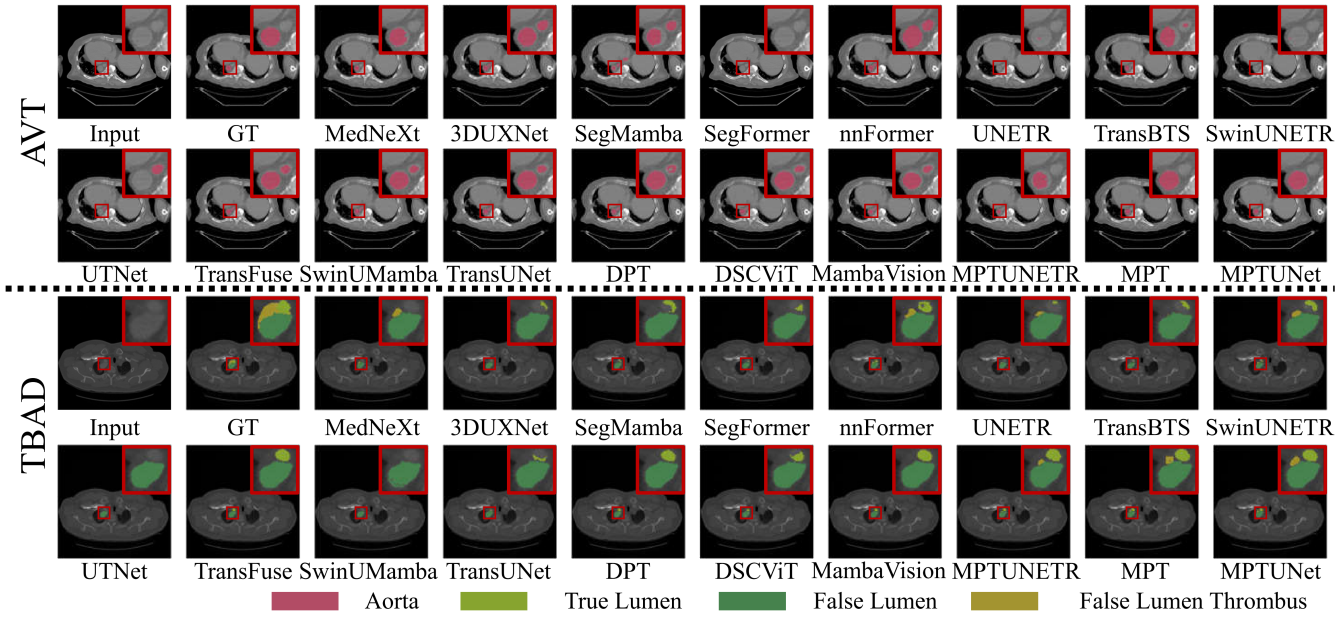


Figure 4: Visual results on the AVT and TBAD datasets. The first two rows illustrate the aortic segmentation performance of different models on the AVT dataset, while the bottom two rows present their ability to identify aortic dissection on the TBAD dataset. Red-boxed areas have been magnified to better visualize specific anatomical structures.

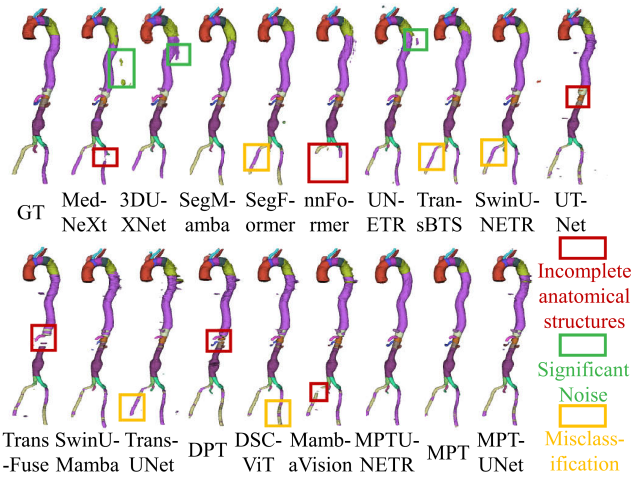


Figure 5: Comparison of different models in processing complex vascular structures on the AortaSeg24 dataset. Red boxes denote incomplete anatomical structures, green boxes highlight significant noise, and yellow boxes indicate misclassifications.

ples include SwinUNETR (He et al. 2023) and our MPTUNETR, which integrate convolutional priors to enhance spatial learning while maintaining ViT-based global modeling. On the 2D side, TransUNet (Chen et al. 2021a) serves as a representative hybrid ViT-CNN framework, wherein convolutional blocks extract local features and Transformer encoders capture global context. Recent developments such as DPT (Chen et al. 2021b) and DSCViT (Qi et al. 2023) further improve segmentation performance by incorporating deformable patch embedding and dynamic snake convolution, respectively. These architectural advances enhance the model's ability to delineate intricate, tubular anatomical structures. To verify the effectiveness of our 2D MPTUNet, we include these 2D hybrid models in our comparative evaluations. Besides, we also incorporate the Mamba family of models in our benchmarking. Approaches such as SegMamba (Xing et al. 2024), SwinUMamba (Liu et al. 2024), and latest MambaVision (Hatamizadeh and Kautz 2025) leverage state space models (SSMs) to efficiently model long-range dependencies while maintaining lower computational complexity compared to self-attention-based alternatives. This makes them particularly competitive under resource-constrained scenarios.

In Fig. 3, we further present a comparison of the parameter counts and computational costs across all baseline models. It is evident that our proposed MPT-based models achieve superior segmentation performance while maintaining significantly lower computational complexity and fewer parameters compared to many recent state-of-the-art methods. In addition to the Dice scores reported in Fig. 3, we utilize clDice(Shit et al. 2021), a topology-aware metric designed to evaluate vascular structural connectivity, on several aortic datasets.

Experiment

The experimental results in Table 1 and Table 2 demonstrate that the proposed MPT-based methods consistently achieve superior aortic segmentation performance compared to existing 2D and 3D segmentation models. Specifically,

among 3D approaches, the ViT-based MPT model achieves the highest scores on the AVT dataset (Dice 0.856, mIoU 0.762) shown in Table 1. MPT also performs competitively on AortaSeg24 and TBAD, surpassing established 3D CNNs, ViT, and Mamba-based models. Additionally, the hybrid MPTUNETR further enhances the performance of other 3D hybrid models,including latest SwinUNETR and TransBTS. Notably, on the AortaSeg24 dataset, which involves complex vascular structures, MPTUNETR achieves the best Dice (0.809) and mIoU (0.695) among all methods. For 2D architectures, MPTUNet achieves state-of-the-art results, with Dice 0.896, mIoU 0.815, and clDice 0.839 on the AVT dataset. Additionally, MPTUNet outperforms other 2D segmentation models, achieving Dice scores of 0.796 and 0.930 on AortaSeg24 and TBAD, respectively. These results highlight the excellent generalization ability of MPTUNet. When compared to previous hybrid models using deformable convolutions and TransUNet (including DPT and DSCViT), MPTUNet proves more effective for handling complex tubular structures on three datasets. Remarkably, MPTUNet achieves the highest clDice scores across all three datasets, with values of 0.839, 0.966, and 0.920 on AVT, AortaSeg24, and TBAD, respectively. This confirms that the proposed morph patch strategy effectively preserves the topology and continuity of the aortic structure. These results highlight the high performance and strong generalization capability of MPT-based models in aortic segmentation tasks.

Fig. 4 and Fig. 5 present the visual results of MPT-based methods across three datasets. To be specific, the first two rows of Fig. 4 show the visualization results of various models on the AVT and TBAD datasets. The MPT-based model successfully segments the aortic structure in the AVT dataset while effectively avoiding interference from surrounding small tissue. The last two rows of Fig. 4 display the performance of different models on the TBAD dataset. The results indicate that only nnFormer, UNETR, SwinUNETR, and the proposed MPT-based methods are capable of displaying three anatomical structures. Furthermore, the MPT-based methods offer more complete segmentation results, particularly for the True Lumen, compared to the other models. Fig. 5 shows the performance of different models on the AortaSeg24 dataset. It can be observed that the MPT-based methods handle complex vascular shapes more effectively, significantly reducing the occurrence of incomplete anatomical structures highlighted in the red boxes. Additionally, the MPT-based methods use SCA to aggregate similar semantic features. By obtaining representative features, these methods enhance the model's ability to accurately identify relevant structures, thereby reducing classification errors and minimizing segmented noise.

Ablation Study

Component Ablation Table 3 presents the ablation experiments of MPT on the AVT dataset, where MP (Morph-Patch strategy) and SCA (Semantic Clustering Attention) are key components. Without MP and SCA, the Transformer-based model achieves a Dice score of 0.834, an mIoU of 0.743, and a clDice of 0.732. Introducing MP significantly improves segmentation performance, with both Dice and mIoU scores showing noticeable gains. This indicates that the MP strategy adapts to the complex morphology of vessels and mitigates the structural damage caused by rigid patch partitioning. Moreover, the increase in clDice to 0.744 with MP further demonstrates MP's effectiveness in preserving the topological structure during vessel segmentation. When SCA is subsequently added, the model's ability to aggregate similar semantics is enhanced, further boosting performance. At this point, the model achieves a Dice score of 0.856, an mIoU of 0.762, and a clDice of 0.757.

ViT	Mamba	MP	SCA	Dice	mIoU	clDice
✓	–	–	–	0.834	0.743	0.732
✓	–	✓	–	0.839	0.748	0.744
✓	–	✓	✓	0.856	0.762	0.757
–	✓	–	–	0.829	0.730	0.711
–	✓	✓	–	0.831	0.730	0.725

Table 3: Ablation experiments of MPT on the AVT dataset. The MP and SCA are respectively brief expression for Morph-patch strategy and Semantic clustering attention.

To demonstrate the generalization ability of our proposed MP strategy across different frameworks, we integrate MP with the latest Mamba architecture (Liu et al. 2024), which also requires patch partitioning. Experimental results show a significant improvement in clDice, from 0.711 to 0.725, upon introducing MP. This indicates that the MP strategy effectively preserves the aortic topological structures in the Mamba framework.

DePatch	DefF	VecF	Dice	mIoU	clDice
–	–	–	0.874	0.782	0.801
✓	–	–	0.886	0.800	0.825
–	✓	–	0.877	0.786	0.782
–	–	✓	0.888	0.803	0.833

Table 4: Comparison of Patch Deformation Strategies on the AVT Dataset. DePatch generates adaptive rectangular patches with varying aspect ratios, while DefF and VecF create irregular-shaped patches using deformation fields. The key difference is that DefF directly generates the deformation field, whereas VecF creates it indirectly through a velocity field.

Patch Deformation Evaluation To ensure a fair evaluation of different patch deformation strategies, we modify the patch embedding layer of TransUNet, with results presented in Table 4. In general, all three patch deformation methods, implemented as the Morph-patch module, significantly improve the model's ability to recognize complex vascular structures compared to TransUNet. To be specific, DePatch (Chen et al. 2021b) improves the model's capacity to extract vessels at various scales by generating patches of different sizes, resulting in Dice and clDice scores of 0.866 and 0.801, respectively. DefF generates irregular-shaped patches with deformation fields to capture complex features, in-

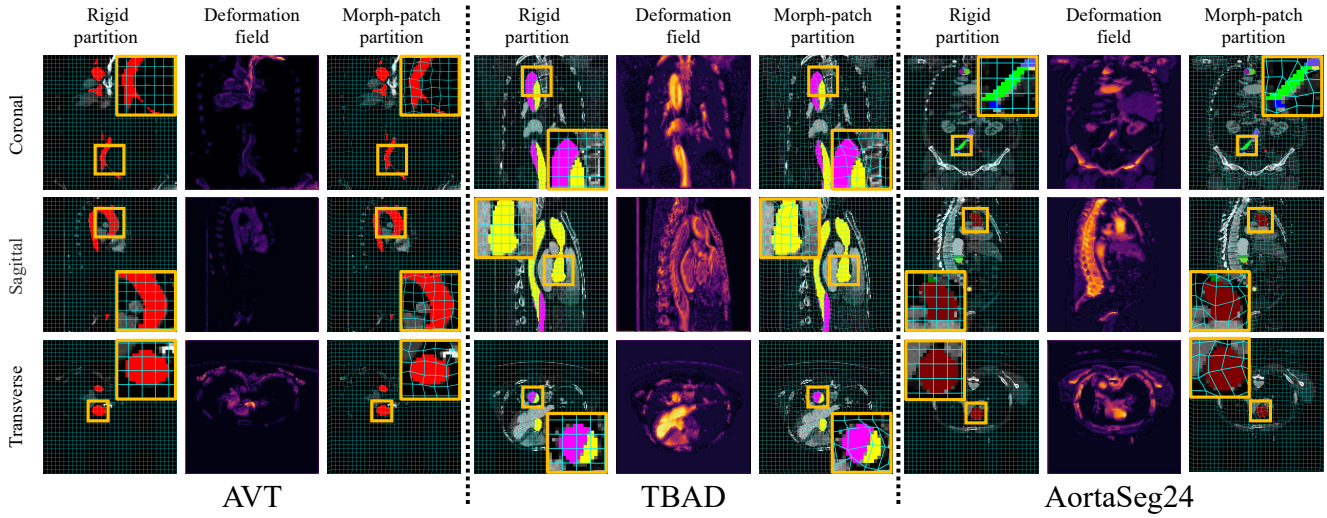


Figure 6: Morph-Patch Strategy preserves vascular morphology and enables cross-view anatomical alignment.

spired by the Dynamic Snake Convolution (Qi et al. 2023). As a result, DefF shows improvements in both Dice and mIoU compared to TransUNet. However, without topological constraints, such deformation may disrupt the topology of smaller vessels, leading to a decrease in clDice. To address this issue, we enforce the diffeomorphism constraint by iteratively generating the deformation field through a velocity field. Consequently, VecF outperforms both DePatch and DefF in extracting complex vascular features, achieving the highest Dice and mIoU scores. Moreover, due to the diffeomorphism constraint, the topology of the segmented vascular masks is better preserved. Experimental results show that, DefF significantly improved the clDice of TransUNet to 0.833, surpassing both DePatch and VecF.

Model Analysis

Our Morph-Patch strategy enables patches to effectively perceive complex vascular morphologies, as visualized in Fig. 6. In the coronal view of the AortaSeg24 case, the deformation field effectively guides the patch boundaries to closely follow the contours of slender vessels. This results in better preservation of fine morphological details, particularly at the vessel edges.

Additionally, the adaptively generated deformation field enables the model to perceive multiple anatomical structures, including both vertebrae and vessels. An inspection of nine representative deformation fields from three different datasets reveals that the deformations primarily concentrate around anatomically salient regions—especially the spine and major vascular structures. This behavior reflects a meaningful spatial bias: in the thoracoabdominal region, the aorta and its major branches typically run adjacent to the vertebral column. As a result, deformation fields that consistently concentrate in these areas suggest that the model is leveraging the stable spatial relationship between the spine and major vessels as an anatomical prior. This implicit guidance allows the model to better localize and delineate vascular structures by referencing nearby, morphologically salient landmarks. Such structure-aware deformation improves not only segmentation accuracy but also interpretability, as it demonstrates that the model's patch partitioning behavior aligns with well-established anatomical context.

Conclusion

In this study, we propose the adaptive Morph-Patch Transformer (MPT) to address the challenges of aortic vascular segmentation, specifically the limitations of fixed patch partitioning in traditional Transformer-based models. MPT incorporates two key innovations: a velocity field-guided Morph-Patch strategy that generates morphology-aware patches, and a Semantic Clustering Attention mechanism that aggregates features from semantically similar regions, enabling precise segmentation of fine-grained vessel structures while preserving topological continuity.

Extensive experiments on three public datasets across 2D and 3D settings demonstrate that MPT consistently outperforms existing approaches, achieving state-of-the-art performance in multiple segmentation metrics. The successful adaptation of the Morph-Patch strategy to the Mamba framework further validates its robustness and generalizability in tubular structure segmentation. These results highlight the potential of adaptive, structure-aware patch partitioning for accurate and reliable medical image analysis, with promising implications for clinical cardiovascular diagnostics. In future work, we aim to further evaluate MPT on a broader range of datasets and anatomical scenarios, paving the way for its practical adoption in clinical applications.

Acknowledgments

This work was supported by Shenzhen Medical Research Fund (No. D2404001), in part by the Key-Area Research and Development Program of Guangdong Province(No. 2025B1111020001), in part by the Natural Science Founda-

tion of Guangdong Province (No. 2023A1515010673), in part by the Shenzhen Science and Technology Innovation Bureau key project (No. JSGG20220831110400001, No. CJGJZD20230724093303007,KJZD20240903101259001), in part by Shenzhen Engineering Laboratory for Diagnosis & Treatment Key Technologies of Interventional Surgical Robots (XMHT20220104009), in part by Beijing Xisike Clinical Oncology Research Foundation (No.Y-2024AZ(NSCLC)MS-0156), and the Key Laboratory of Biomedical Imaging Science and System, CAS, for the Research platform support.

References

Arsigny, V.; Commowick, O.; Pennec, X.; and Ayache, N. 2006. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 924–931. Springer.

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

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

Chen, Z.; Zhu, Y.; Zhao, C.; Hu, G.; Zeng, W.; Wang, J.; and Tang, M. 2021b. Dpt: Deformable patch-based transformer for visual recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2899–2907.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 764–773.

Dalca, A. V.; Balakrishnan, G.; Guttag, J.; and Sabuncu, M. R. 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, 57: 226–236.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference On Learning Representations*.

Hahn, L. D.; Baeumler, K.; and Hsiao, A. 2021. Artificial intelligence and machine learning in aortic disease. *Current Opinion in Cardiology*, 36(6): 695–703.

Hatamizadeh, A.; and Kautz, J. 2025. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25261–25270.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584.

He, Y.; Nath, V.; Yang, D.; Tang, Y.; Myronenko, A.; and Xu, D. 2023. SwinUNETR-v2: Stronger swin transformers with stagewise convolutions for 3D medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 416–426. Springer.

Imran, M.; Krebs, J. R.; Sivaraman, V. B.; Zhang, T.; Kumar, A.; Ueland, W. R.; Fassler, M. J.; Huang, J.; Sun, X.; Wang, L.; et al. 2025. Multi-class segmentation of aortic branches and zones in computed tomography angiography: The aortaseg24 challenge. arXiv:2502.05330.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jian, M.; Xu, W.; Nie, C.; Li, S.; Yang, S.; and Li, X. 2025. DAU-Net: a novel U-Net with dual attention for retinal vessel segmentation. *Biomedical Physics & Engineering Express*, 11(2): 025009.

Lavie, I.; Gur-Ari, G.; and Ringel, Z. 2024. Towards Understanding Inductive Bias in Transformers: A View From Infinity. In *Forty-first International Conference on Machine Learning*.

Li, S.; Li, B.; Sun, B.; and Weng, Y. 2024. Towards visual-prompt temporal answer grounding in instructional video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, W.; Gao, Z.; Liu, H.; and Zhang, H. 2023. A deformable constraint transport network for optimal aortic segmentation from ct images. *IEEE Transactions on Medical Imaging*, 43(4): 1462–1475.

Liu, J.; Yang, H.; Zhou, H.-Y.; Xi, Y.; Yu, L.; Li, C.; Liang, Y.; Shi, G.; Yu, Y.; Zhang, S.; et al. 2024. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 615–625. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Mukherjee, A.; et al. 2015. *Differential topology*. Springer.

Perera, S.; Navard, P.; and Yilmaz, A. 2024. Segformer3d: an efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4981–4988.

Qi, Y.; He, Y.; Qi, X.; Zhang, Y.; and Yang, G. 2023. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6070–6079.

Radl, L.; Jin, Y.; Pepe, A.; Li, J.; Gsaxner, C.; Zhao, F.-h.; and Egger, J. 2022. AVT: Multicenter aortic vessel tree CTA dataset collection with ground truth segmentation masks. *Data in brief*, 40: 107801.

Shit, S.; Paetzold, J. C.; Sekuboyina, A.; Ezhov, I.; Unger, A.; Zhylka, A.; Pluim, J. P.; Bauer, U.; and Menze, B. H. 2021. clDice-a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16560–16569.

Tang, X.; Zhang, H.; Xie, B.; and Liu, X. 2025. Temporally consistent segmentation of main coronary artery in X-ray coronary angiography sequences. *Expert Systems with Applications*, 271: 126591.

Townsend, N.; Kazakiewicz, D.; Lucy Wright, F.; Timmis, A.; Huculeci, R.; Torbica, A.; Gale, C. P.; Achenbach, S.; Weidinger, F.; and Vardas, P. 2022. Epidemiology of cardiovascular disease in Europe. *Nature Reviews Cardiology*, 19(2): 133–143.

Wang, W.; Chen, C.; Ding, M.; Yu, H.; Zha, S.; and Li, J. 2021. TransBTS: multimodal brain tumor segmentation using transformer, Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 109–119.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4794–4803.

Xian, Y.; Zhao, G.; Chen, X.; and Wang, C. 2025. DCFU-Net: Rethinking an Effective Attention and Convolutional Architecture for Retinal Vessel Segmentation. *International Journal of Imaging Systems and Technology*, 35(1): e70003.

Xie, B.; Zhang, H.; Wang, A.; Liu, X.; and Gao, Z. 2025. Bivariational physics-informed operator network for fractional flow reserve curve assessment from coronary angiography. *Medical Image Analysis*, 103: 103564.

Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; and Zhu, L. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 578–588. Springer.

Yagis, E.; Aslani, S.; Jain, Y.; Zhou, Y.; Rahmani, S.; Brunet, J.; Bellier, A.; Werlein, C.; Ackermann, M.; Jonigk, D.; et al. 2024. Deep Learning for 3D Vascular Segmentation in Phase Contrast Tomography. *Research Square*, rs–3.

Yao, Z.; Xie, W.; Zhang, J.; Dong, Y.; Qiu, H.; Yuan, H.; Jia, Q.; Wang, T.; Shi, Y.; Zhuang, J.; et al. 2021. Imagetbad: A 3d computed tomography angiography image dataset for automatic segmentation of type-b aortic dissection. *Frontiers in Physiology*, 12: 732711.

Zhang, B.; Li, D.; and Wang, D. 2024. DCT based multi-head attention-BiGRU model for EEG source location. *Biomedical Signal Processing and Control*, 93: 106171.

Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020. Feature pyramid transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 323–339. Springer.

Zhou, H.-Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; and Yu, Y. 2023. nnFormer: volumetric medical image segmentation via a 3D transformer. *IEEE Transactions on Image Processing*, 32: 4036–4045.

Zhu, C.; Yang, Z.; Xiao, Y.; Wu, T.; Zou, B.; and Zhou, H. 2024. TTCNet: Transformer and Tubular Convolution Feature Attention Network for OCTA vessel segmentation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 479–484. IEEE.