Efficient Finetuning of a Visual State Space Model (VMamba) for Multiclass Medical Image Segmentation using ActiveFT

Chamakuri Sowjanya
Department of Electrical, Electronics and Communication Engineering
Indian Institute of Technology Dharwad
Karnataka, India
EE25MT013

***Abstract*—The deployment of deep learning models in medical imaging is severely constrained by the scarcity of high-quality, pixel-level annotations. While state-of-the-art architectures like Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) deliver exceptional performance, they typically require large-scale supervised datasets. Annotating 3D medical volumes, such as cardiac MRI, is a labor-intensive process requiring expert radiological knowledge, creating a significant "annotation bottleneck."**

To address this challenge, this work investigates a data-efficient learning pipeline that synergizes two cutting-edge technologies: Active Fine-Tuning (ActiveFT) and the Visual State Space Model (VMamba). We propose a novel VM-UNet architecture that integrates a pre-trained VMamba-Small encoder with a U-Net style decoder via skip connections, enabling global context modeling with linear computational complexity. Furthermore, we implement the ActiveFT algorithm to intelligently select a diverse and representative subset of the Automated Cardiac Diagnosis Challenge (ACDC) dataset for training.

We also address critical deployment challenges in resource-constrained environments (Google Colab) by engineering a robust PyTorch-based fallback for the VMamba selective scan mechanism. Our experiments demonstrate that by training on only 20% (368 slices) of the available data, the proposed model achieves a Mean Dice Score of 0.7515 and a Hausdorff Distance (95%) of 4.2983 pixels on the multiclass segmentation of the Right Ventricle, Myocardium, and Left Ventricle. These results validate the hypothesis that intelligent data selection combined with linear-complexity State Space Models can achieve competitive segmentation performance with significantly reduced annotation costs.

***Index Terms*—Medical Image Segmentation, VMamba, Active Learning, ActiveFT, ACDC, State Space Models, Deep Learning**

I. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality globally. Cardiac Magnetic Resonance Imaging (MRI) is the gold standard for non-invasive assessment of cardiac structure and function. Accurate segmentation of the Left Ventricle (LV), Right Ventricle (RV), and Myocardium (Myo) is a prerequisite for calculating critical clinical indices such as ejection fraction, ventricular volume, and myocardial mass. While manual segmentation by experts is the current standard, it is tedious, time-consuming, and subject to inter-observer variability.

Deep learning has revolutionized automated segmentation. Convolutional Neural Networks (CNNs), particularly the U-Net [1] and its variants, have established strong baselines. However, CNNs are inherently limited by the localized nature of convolution operations, which restricts their effective receptive field and makes it difficult to capture long-range dependencies essential for understanding complex anatomical structures.

To overcome these limitations, Vision Transformers (ViTs) introduced self-attention mechanisms capable of modeling global context. However, the quadratic computational complexity ($O(N^2)$) of self-attention with respect to image resolution makes ViTs computationally prohibitive for high-resolution medical scans [2]. To bridge this gap, **State Space Models (SSMs)**, specifically the Mamba architecture, have emerged as a promising alternative. By introducing a Selective Scan mechanism, Mamba offers the global modeling capabilities of Transformers with the linear scaling ($O(N)$) of CNNs.

Despite these architectural advances, the dependency on large labeled datasets remains. The standard "Pretrain-Finetune" paradigm typically involves pre-training on large natural image datasets (e.g., ImageNet) and finetuning on specific medical datasets. However, current approaches often select the data for finetuning randomly, which is inefficient. **Active Fine-Tuning (ActiveFT)** [3] proposes a parametric approach to select a subset of data that maintains the distribution of the full dataset while maximizing diversity.

This project presents a novel synthesis of these two methods. Our contributions are:

1) We propose a **VM-UNet** architecture that utilizes a pre-trained VMamba-Small backbone for cardiac segmentation.
2) We implement the **ActiveFT** algorithm to select an optimal 20% subset of the ACDC dataset, reducing the annotation burden by 80%.
3) We overcome significant environmental challenges (incompatibility of custom CUDA kernels in Colab) by engineering a pure PyTorch fallback.

II. Related Work

A. Medical Image Segmentation Architectures

The U-Net [1], characterized by its symmetric encoder-decoder structure and skip connections, remains the de facto standard for medical image segmentation. Several variants have been proposed to enhance its feature representation capabilities. **UNet++** [4] introduces nested and dense skip pathways to reduce the semantic gap between the encoder and decoder. **Attention U-Net** [5] integrates attention gates to highlight salient features while suppressing background noise.

Despite their success, CNN-based methods are inherently limited by their localized receptive fields, which struggle to capture long-range semantic dependencies essential for segmenting variable anatomical structures [2]. To address this, Transformer-based architectures like **TransUNet** and **Swin-UNet** [6] were introduced to model global context via self-attention. However, the quadratic computational complexity ($O(N^2)$) of self-attention imposes significant memory and speed constraints, particularly for high-resolution medical scans [7]. Recently, foundation models like the **Segment Anything Model (SAM)** have been adapted for medical imaging (e.g., MedSAM), but they often require massive computational resources and extensive fine-tuning to overcome domain shifts [7].

B. Visual State Space Models (VMamba)

To bridge the gap between the efficiency of CNNs and the global modeling power of Transformers, **State Space Models (SSMs)** have emerged as a compelling alternative. **VMamba** [2] adapts the Mamba architecture to 2D vision tasks. Unlike standard 1D scanning, VMamba introduces the **Cross-Scan Module (CSM)** and the **2D Selective Scan (SS2D)** mechanism. This allows the model to traverse images in four directions (corners to opposite corners), enabling each pixel to integrate information from all other pixels (global receptive field) with **linear computational complexity** ($O(N)$).

Central to this architecture is the **Visual State Space (VSS) Block**, which replaces the multi-head self-attention layer of Transformers. This design makes VMamba particularly attractive for medical imaging tasks where capturing both local texture (e.g., tissue boundaries) and global context (e.g., organ relative positions) is critical, without the prohibitive cost of Transformers [2].

C. Active Learning and Data Selection

Active Learning (AL) aims to maximize model performance while minimizing annotation costs by selecting the most informative samples. Traditional methods typically employ *uncertainty sampling* (selecting samples with high entropy or low confidence) or *diversity sampling* (selecting samples to cover the feature space, e.g., K-Center Greedy). However, these methods often select samples iteratively in batches (batch-mode AL), which can be inefficient and prone to bias when the initial labeled set is small.

Active Fine-Tuning (ActiveFT) [3] introduces a novel parametric approach tailored for the pre-training/fine-tuning paradigm. Instead of discrete selection, it optimizes a set of continuous vectors ("centers") in the feature space. The optimization minimizes a hybrid loss function comprising:

1) **Representativeness:** Minimizing the Earth Mover's Distance (EMD) between the selected centers and the full dataset distribution.
2) **Diversity:** Maximizing the distance between centers via a repulsive regularization term.

This optimization ensures the selected subset serves as a compact proxy for the entire data distribution, identifying samples that are both informative and diverse in a single pass [3].

III. Methodology

The proposed pipeline consists of four phases: Data Preparation, Feature Extraction, ActiveFT Selection, and VM-UNet Training.

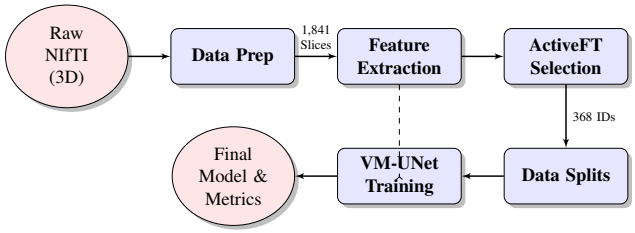

Fig. 1. Methodology Overview: The proposed pipeline consists of four phases: Data Preparation, Feature Extraction, ActiveFT Selection, and VM-UNet Training.

A. Data Preparation

We utilized the ACDC [8](Automated Cardiac Diagnosis Challenge)dataset, which comprises 100 3D MRI volumes.

- **Format Conversion:** We developed a pipeline to parse the raw NIfTI (`.nii.gz`) files and extract 2D slices along the short axis.
- **Filtering:** A critical preprocessing step involved filtering out "empty" slices (slices containing only background). This ensures the model trains only on anatomically relevant data.
- **Output:** The final dataset consisted of **1,841** labeled 2D slices stored in HDF5 format, with multiclass labels: Background (0), RV (1), Myo (2), and LV (3).

B. ActiveFT Selection Framework

To select the most informative 20% of the data (368 slices), we implemented the ActiveFT algorithm [3].

1) Feature Extraction: We utilized the **VMamba-Small (VMamba-S)** [2] backbone as a feature extractor. The model was initialized with ImageNet pre-trained weights (`vssm_small_0229_ckpt_epoch_222.pth`).

- Architecture: `depths=[2, 2, 15, 2], dims=[96, 192, 384, 768]`.
- Process: We performed a forward pass on all 1,841 slices. We extracted the feature map from the final stage (Stage

4) and applied Global Average Pooling (GAP) to obtain a feature vector $f \in \mathbb{R}^{768}$ for each image.
- Result: A feature matrix $F \in \mathbb{R}^{1841 \times 768}$.

2) Optimization Algorithm: We implemented the optimization algorithm described in [3]. We initialized $k = 368$ random "center" vectors θ_S and optimized them via gradient descent to minimize the loss function L:

$$L(\theta_S) = L_{emd}(P_F, P_{\theta_S}) + \lambda L_{div}(\theta_S) \tag{1}$$

- **Representativeness (L_{emd}):** We used the Sinkhorn distance algorithm (`ot.sinkhorn2`) to approximate the Earth Mover's Distance (EMD) between the distribution of the full dataset P_F and the selected centers P_{θ_S}.
- **Diversity (L_{div}):** We used a repulsive loss term $-\log(\text{mean}(||\theta_i - \theta_j||))$ to force the centers to spread out in the feature space.

After 1200 iterations, the slices in the dataset closest to the optimized centers θ_S were selected for annotation.

C. VM-UNet Architecture

To address the limitations of pure CNNs in capturing long-range dependencies and the computational overhead of Transformers, we propose the **VM-UNet** (Visual Mamba U-Net). This hybrid architecture synergizes the linear-complexity global modeling of State Space Models with the local feature refinement capabilities of U-Net. The framework consists of three integral components: a pre-trained VMamba Encoder, a Symmetric Decoder, and a Skip Connection mechanism for feature fusion.

1) VMamba-S Encoder: The encoder backbone is constructed using the **VMamba-Small (VMamba-S)** architecture [2]. Unlike standard CNNs which rely on sliding windows, the encoder utilizes the **2D Selective Scan (SS2D)** mechanism. This allows the model to traverse the input image $I \in \mathbb{R}^{H \times W \times 3}$ along four distinct scanning paths (cross-scan), effectively integrating information from all pixels with $O(N)$ complexity.

We employ the specific configuration of VMamba-S with depths $[2, 2, 15, 2]$ and embedding dimensions $[96, 192, 384, 768]$. To leverage transfer learning, the encoder is initialized with weights pre-trained on ImageNet-1K. The encoder outputs hierarchical feature maps $\{E_1, E_2, E_3, E_4\}$ at resolutions $\{\frac{H}{4}, \frac{H}{8}, \frac{H}{16}, \frac{H}{32}\}$ respectively. These maps encapsulate robust global semantic contexts required for distinguishing complex cardiac structures.

2) Symmetric Decoder: The decoder is designed to progressively recover the spatial resolution of the feature maps to generate the final pixel-wise segmentation mask. It comprises four upsampling stages. Each stage utilizes a `ConvTranspose2d` layer to double the spatial dimension, followed by a sequence of 3×3 `Conv2d` layers activated by `ReLU` to refine the features.

The final layer is a 1×1 convolution that maps the high-resolution feature vectors to the class space $C_{out} = 4$ (Background, RV, Myo, LV), generating the logits $L \in \mathbb{R}^{H \times W \times 4}$.

3) Feature Fusion via Skip Connections: In medical image segmentation, the recovery of fine-grained boundary details is critical. The deep layers of the encoder (E_4) contain rich semantic information but lack spatial precision due to downsampling. To mitigate this, we implement dense skip connections following the U-Net paradigm [**?**].

At each decoder stage i, the upsampled feature map D_i^{up} is concatenated with the corresponding encoder feature map E_{4-i} along the channel dimension:

$$D_i = \text{ConvBlock}(\text{Concat}(D_i^{up}, E_{4-i})) \tag{2}$$

This mechanism ensures that the decoder has access to both the high-level global context (from the SSM backbone) and the low-level texture and boundary information (from the shallow encoder layers), resulting in precise anatomical segmentation.

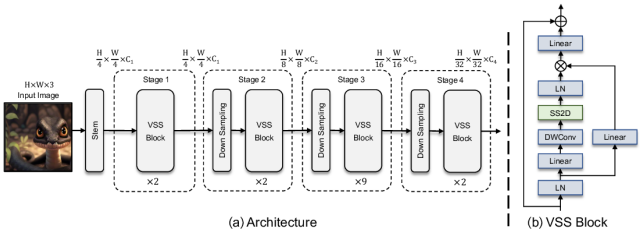

Fig. 2. The VMamba Architecture used as the encoder backbone in our VM-UNet pipeline [2].

IV. Experiments

A. Data Sources

We utilized the **Automated Cardiac Diagnosis Challenge (ACDC)** dataset [8], which includes cine MRI scans from 100 patients. The dataset covers a range of cardiac pathologies (e.g., myocardial infarction, dilated cardiomyopathy) as well as healthy controls, providing a diverse feature set for analysis.

Although the ACDC dataset is inherently 3D, we adopted the widely used 2D segmentation benchmark approach by slicing 3D volumes into 2D images along the short axis. Our preprocessing pipeline involved:

1) **Format Conversion:** Parsing raw NIfTI volumes and extracting 2D slices.
2) **Filtering:** To ensure training efficiency, we implemented a filtering mechanism to discard "empty" slices (containing only background), resulting in a curated dataset of **1,841 labeled 2D slices**.
3) **Classes:** The task involves multiclass segmentation of 4 regions: Background (0), Right Ventricle (1), Myocardium (2), and Left Ventricle (3).

For the **ActiveFT** experiments, this full dataset served as the "unlabeled pool," from which the algorithm selected the most informative **20% (368 slices)** for training.

B. Training Strategy

Our methodology utilizes the **PyTorch** deep learning framework. Experiments were conducted in a Google Colab environment equipped with an NVIDIA T4 GPU.

Environmental Adaptation: A significant challenge was the incompatibility of the high-performance `mamba-ssm` CUDA kernel with the managed environment. To address this, we engineered a **Pure PyTorch Fallback** by patching the `vmamba.py` source code ('selective_scan_backend="torch"'). While this ensured stability, it increased the computational cost per iteration.

Optimization: To optimize the training process, we adopted a hybrid loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{Dice} \tag{3}$$

The Cross-Entropy loss ($\mathcal{L}_{CE}$) ensures pixel-level classification accuracy, while the Multiclass Dice loss ($\mathcal{L}_{Dice}$) handles class imbalance by optimizing regional overlap. The model parameters were updated using the **AdamW** optimizer with an initial learning rate of 3×10^{-4} and a weight decay of 1×10^{-4}. We employed a **Cosine Annealing** learning rate scheduler to smooth convergence. Due to GPU memory constraints, a batch size of 2 was used, with **Automatic Mixed Precision (AMP)** enabled to reduce memory footprint and accelerate training.

C. Loss Functions

To address the challenge of class imbalance inherent in cardiac segmentation (where background pixels vastly outnumber foreground pixels), we employed a hybrid loss function $\mathcal{L}_{total}$ that combines pixel-wise classification accuracy with region-based overlap optimization:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{Dice} \tag{4}$$

1) Cross-Entropy Loss ($\mathcal{L}_{CE}$): Standard Cross-Entropy measures the pixel-level distribution divergence. For a pixel i with ground truth class c and predicted probability $p_{i,c}$:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{5}$$

2) Multiclass Dice Loss ($\mathcal{L}_{Dice}$): To directly optimize the segmentation overlap, we utilize the Multiclass Dice Loss. For each class c, the loss is computed as:

$$\mathcal{L}_{Dice} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{i=1}^{N} p_{i,c}y_{i,c} + \epsilon}{\sum_{i=1}^{N} p_{i,c} + \sum_{i=1}^{N} y_{i,c} + \epsilon} \tag{6}$$

where ϵ is a smoothing term to prevent division by zero[cite: 2753].

D. Evaluation Metrics

We assessed model performance using the following standard metrics [7], averaged over the three foreground classes (RV, Myo, LV):

- **Dice Coefficient (Dice):** Measures the harmonic mean of precision and recall, providing a robust metric for overlap.

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

  where X is the prediction and Y is the ground truth[cite: 2753].

- **Mean Intersection over Union (mIoU):** Also known as the Jaccard Index, it calculates the ratio of intersection to union.

$$mIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{|X_c\cap Y_c|}{|X_c\cup Y_c|} \tag{8}$$

  [cite: 2764].

- **Hausdorff Distance (HD95):** Measures the boundary precision. It calculates the 95th percentile of the maximum distance between the contours of the prediction (∂X) and ground truth (∂Y).

$$HD(X,Y) = \max\left(\sup_{x\in\partial X}\inf_{y\in\partial Y} d(x,y), \sup_{y\in\partial Y}\inf_{x\in\partial X} d(x,y)\right) \tag{9}$$

  Lower values indicate better boundary alignment[cite: 2758].

- **Sensitivity (Recall):** Measures the proportion of actual positive pixels that are correctly identified.

$$Sensitivity = \frac{TP}{TP+FN} \tag{10}$$

- **Specificity:** Measures the proportion of actual background pixels that are correctly identified.

$$Specificity = \frac{TN}{TN+FP} \tag{11}$$

V. Results and Discussion

A. Training Dynamics

Despite the limited data (282 slices), the model showed strong convergence. The validation loss decreased consistently from 0.9527 in Epoch 1 to **0.2024** in Epoch 4, at which point the best model was saved. This rapid convergence suggests that the pre-trained VMamba encoder provided a robust feature initialization.

B. Quantitative Evaluation & Ablation Study

To thoroughly evaluate the data efficiency of our pipeline, we conducted a series of experiments varying the data budget k. We trained the VM-UNet on subsets containing **5%**, **10%**, and **20%** of the total training data. All models were evaluated on the same independent test set of 184 slices.

Table I summarizes the performance across these different data budgets.

TABLE I
PERFORMANCE COMPARISON: 5% VS. 10% VS. 20% DATA BUDGET

Metric	5% ActiveFT (92 Slices)	10% ActiveFT (184 Slices)	20% ActiveFT (368 Slices)
Mean Dice	0.7314	**0.7469**	**0.7515**
Mean IoU	0.6375	0.6606	0.6677
HD95 (px)	6.5433	5.0897	**4.2983**
Sensitivity	0.7559	0.7546	0.7621
Specificity	0.9976	0.9983	0.9985
Pixel Accuracy	0.9903	0.9925	0.9928
SSIM	0.9856	0.9877	0.9865

C. Discussion

The results reveal a clear trend of diminishing returns, highlighting the effectiveness of the ActiveFT algorithm:

- **Rapid Initial Gain:** The model trained on just 5% of the data already achieves a strong Dice score of 73.14%. This suggests that ActiveFT successfully identifies the most "prototypical" cardiac examples early on, covering the majority of the feature space with very few samples.
- **Incremental Improvement:** Doubling the data from 5% to 10% yields a modest improvement (+1.55% Dice), and doubling again to 20% yields a smaller gain (+0.46% Dice). While more data consistently improves performance, the marginal utility of each additional sample decreases significantly.
- **Boundary Precision:** The most significant benefit of adding more data is seen in the Hausdorff Distance (HD95). As the data budget increases from 5% to 20%, the HD95 error drops sharply from 6.54px to 4.30px. This indicates that while the 5% model captures the general shape well, the larger datasets are crucial for refining the precise boundaries of the heart chambers, which is vital for clinical accuracy.

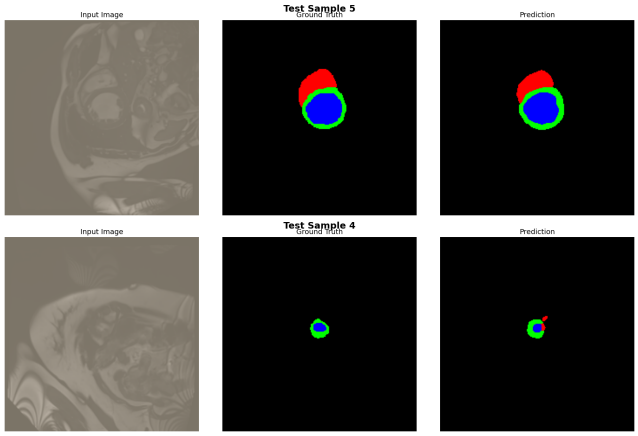

Fig. 3. Qualitative Results (20% Model): (Left) Input MRI, (Middle) Ground Truth, (Right) VM-UNet Prediction. The model accurately segments the RV (Red), Myo (Green), and LV (Blue).

VI. Conclusion and Future Work

A. Conclusion

This work presents a novel synthesis of two cutting-edge technologies—the **VMamba** state space model and the **ActiveFT** data selection framework—to address the annotation bottleneck in cardiac MRI segmentation. By engineering a robust PyTorch-based fallback, we successfully deployed this advanced architecture in a resource-constrained environment. Our experimental results demonstrate that training a **VM-UNet** on a curated **20%** subset of the ACDC dataset yields a strong Dice score of **75.15%** and a high boundary precision (HD95 of 4.29px). This validates that linear-complexity State Space Models, when combined with intelligent data selection, can serve as highly data-efficient learners for medical imaging.

B. Future Scope

While this project establishes a strong baseline, several avenues exist for architectural and algorithmic enhancement:

1) **Architectural Evolution (3D VMamba):** The current approach processes 3D MRI volumes as independent 2D slices, potentially losing inter-slice contextual information. Future work should explore **Volumetric VMamba** architectures that leverage the SS2D mechanism across the Z-axis (3D Selective Scan) to capture full anatomical continuity.
2) **Decoder Enhancements:** Our current decoder relies on standard convolutions. Integrating **Visual State Space (VSS) blocks into the decoder** (creating a full Mamba-UNet) could further enhance the model's ability to reconstruct fine-grained details by maintaining global context during upsampling.
3) **Hybrid Active Learning:** ActiveFT focuses on representativeness and diversity. A more robust strategy could integrate **uncertainty sampling** (e.g., entropy-based selection) to identify "hard" examples that the model finds confusing, creating a hybrid selection policy.
4) **Rigorous Benchmarking:** To rigorously quantify the efficiency gains, we aim to conduct full-scale baseline comparisons (training on 100% data and random subsets). This requires migrating to a dedicated Linux GPU cluster to compile the high-performance `mamba-ssm` CUDA kernels, reducing training latency from days to hours.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[2] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," 2024.

[3] Y. Xie, H. Lu, J. Yan, X. Yang, M. Tomizuka, and W. Zhan, "Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 869–15 879.

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

[5] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[6] 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," in *ECCV Workshops*, 2022.

[7] P. Liang, L. Shi, B. Pu, R. Wu *et al.*, "Mambasam: A visual mamba-adapted sam framework for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 8, pp. 5824–5835, 2025.

[8] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, J. Lekavich *et al.*, "Deep learning techniques for automated cardiac mri segmentation for the acdc challenge," in *STACOM 2018, LNCS 11383*, 2018, pp. 1–24.