NUMIDIA INSTITUTE OF TECHNOLOGY

Submitted in fulfillment of the requirements
for the Bachelor's degree

in Computer science
Option: Artificial Intelligence and Autonomous Embedded Systems

---

Mamba-TransUNet: A Lightweight State-Space Model for Efficient Left Ventricle Segmentation in 2D Echocardiography

---

Author:
Nemdil Amira
Nadji Fares
Benziada Moncef

Supervisor:
Mazouni Fares

Academic year : 2024– 2025

Abstract

In recent years, the segmentation of cardiac structures in echocardiographic sequences has garnered significant attention due to its importance in clinical diagnostics and disease management. This work explores the reproducibility and optimization of several leading deep learning architectures for medical image segmentation, including UNet, nnUNet, DeepLabv3, and TransUNet, by applying them to the CAMUS dataset. Through systematic replication, we validated the robustness of these models, achieving segmentation performance closely aligned with or exceeding reported benchmarks. Building upon these foundations, we introduced Mamba-based variants—leveraging Selective State Space Models (SSMs) to replace traditional attention mechanisms—with the aim of achieving comparable accuracy with significantly reduced computational costs. Among these, the Mamba-enhanced TransUNet achieved near-identical segmentation performance (Dice score of 0.935) while demonstrating remarkable efficiency improvements, including a 94% reduction in model size and reduced inference time. These results highlight the potential of lightweight, state-aware architectures in scalable and cost-effective medical image analysis. Future extensions of this work include the application of Mamba models to 3D cardiac imaging and further enhancement of global modeling via advanced SSM formulations.

Keywords: Medical image segmentation, Echocardiography, Deep learning, State Space Models, Mamba architecture, TransUNet

Acknowledgements

We would like to use this moment to sincerely thank Allah first and everyone who has supported us on the path of our studies and writing this thesis.

Most importantly, our sincere gratitude extends to our families and friends for their unwavering support, endurance, and moral support. They have been the strength and motivation for us at every point.

We are also deeply thankful to our director Mr. Ait Ameur Samir and the people who work at the Numidia Institute of Technology (NIT) for a productive learning environment and access to the resources to support our research. Their commitment to academic excellence and innovation were central in the creation of this project.

We would especially like to recognize the unique appreciation of our supervisor, Mr. Mazouni Fares, for the invaluable guidance, counsel, and continuous encouragement. His insightful suggestions and dedication to our intellectual growth have been vital to determining the quality and path of our work.

It would not have been achievable without the combined efforts of the same institutions and people, and we are truly grateful for the same.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cardiovascular diseases (CVDs) are the world's leading cause of death and are projected to kill 17.9 million people annually [1]. Detection of cardiac abnormalities is thus critical in preventing the onset of cardiac illness, and echocardiography is employed worldwide as the imaging technique of choice because of its noninvasiveness, affordability, and ability to produce dynamic cardiac function in real-time [2]. Among the various anatomical and functional components examined by echocardiography, the left ventricle (LV) is particularly significant. Segmentation of LV is important to enable one to generate quantitative estimates of such critical indices as end-diastolic and end-systolic volume, ejection fraction, and myocardial mass, all of which play key roles in the diagnosis and therapy of cardiovascular diseases [3].

Manual segmentation of the LV is now the clinic best practice but is time-consuming, open to inter- and intra-observer variability and prone to error from the intrinsic challenges to echocardiographic images such as speckle noise, motion artifacts, and low contrast [2]. These challenges are a basis for research into the creation of automatic robust and reliable segmentation algorithms that capitalize on the promise of deep learning to enhance performance and restrict dependence on human effort.

Deep convolutional neural networks (CNNs) have now become the standard of modern medical image segmentation in the past few years. Models like the U-Net, which was initially conceived to be used for biomedical segmentation tasks, were proven to be extremely useful due to the encoder-decoder structure and skip connections by enabling one to maintain global context in addition to high-resolution spatial information [4]. Further developed along these lines, many modifications have been proposed. For instance, nnU-Net, an auto-configured and auto-automated method, mapped U-Net-based models to any input dataset and set new performance benchmarks for several medical imaging tasks [5]. DeepLabv3 also employs atrous spatial pyramid pooling (ASPP) to gather multi-scale context information and was fine-tuned for cardiac imaging with promising results [3].

But the heterogeneity within the left ventricle from a spatiotemporal perspective, especially for diastolic and systolic phases of the cardiac cycle, necessitates models as adaptive as accurate at recognizing evolving shapes and anatomy. TransUNet transformer-based models have been utilized for this purpose [6]. Subsequent to the use of CNN-based encoding and self-attention mechanisms, they are able to identify long-range dependencies as well and

therefore enhance the structural interpretation of complex anatomical structures. Concurrently, hybrid and attention-based networks like GU-DU (Gated U-Net with Dilated Units) have assisted in enhancing the refinement of the segmentation masks by adopting attention-based re-calibration procedures and improved contextual perception [7].

This paper is a systematic comparative evaluation of U-Net, nnU-Net, DeepLabv3, GU-DU, and TransUNet, some of the recent deep learning architectures, for left ventricle segmentation of echocardiographic images. All the models were implemented and trained on the public benchmark CAMUS with 2D echocardiographic sequences of 500 patients labeled with apical four-chamber and two-chamber views at end-systole and end-diastole [8]. Heterogeneity of pathology and image quality of this dataset give strong test over large populations of patients.

Chapter 2

Related Work

2.1 Cardiovascular Segmentation

Cardiovascular segmentation has greatly improved thanks to the advent of deep learning techniques, especially Convolutional Neural Networks (CNNs) [9], which have been fueled by extensive datasets like ACDC [2], ImageAorta [10, 11], and CAS [12]. These datasets cover a variety of cardiovascular components, such as the aortic branches, coronary arteries, and heart chambers. They also pose issues with anatomical variability and disease-specific traits. The goal of multi-modality cardiac imaging segmentation is to accurately segment anatomical structures and diseased regions using imaging modalities such as Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT), MRI, and CT. However, the procedure is made more difficult by innate difficulties such phase alignment, resolution, and imbalances in image quality. By merging data from many modalities, traditional approaches such as fusion-based segmentation techniques and registration-based segmentation with multi-atlas approaches overcome these problems[13, 14, 15, 3]. Nonetheless, these techniques are computationally costly and frequently necessitate extensive datasets [16, 7]. Cardiovascular segmentation is becoming more robust and clinically applicable because to hybrid algorithms that combine deep learning and traditional techniques [17].

Complex representations of cardiovascular structures are now possible thanks to recent developments in deep learning, which have increased segmentation accuracy. For example, CNNs such as U-Net [5, 4], that got 0.9203 IoU as a segmentation result on the ISBI cell tracking challenge 2015. As well as its variations that have proven useful for specifically on the ACDC dataset, cardiovascular segmentation tasks [2], achieving Dice scores of 0.96 for Left Ventricle, 0.94 for Right Ventricle, and 0.90 for myocardium segmentation. However, these techniques are difficult to apply to a variety of datasets, including ImageCAS [12], that lead to achieve a Dice score of 82.96 for coronary artery segmentation using a multiscale patch fusion and two-stage baseline method. and Aorta [10, 11], because of their propensity for overfitting and problems such as class imbalance. In order to capture finer features, Zeng et al. [12] used multi-scale feature extraction to address coronary artery segmentation issues in ImageCAS. However, the technique still has trouble with

arteries with aberrant morphology, necessitating precise hyperparameter adjustment.

Isensee et al. [5] combined the U-Net and V-Net architectures to create the nnUNet framework, which resulted in a significant improvement in segmentation accuracy. In echocardiographic LV segmentation, nnUNet achieved a Hausdorff distance (HD) of 4.3 ±1.9 mm and an average symmetric surface distance (ASSD) of 1.3 ±0.6 mm, outperforming intra-observer variability and prior methods like CLAS. Although this approach has shown promise, it is primarily reliant on large, annotated datasets, which restricts its use in situations where there is a lack of labeled data, which is a common problem in cardiovascular imaging. As Chen et al. [3] showed, transfer learning has been used to overcome issue. They pre-trained models on the ACDC dataset and refined them on smaller, less-annotated datasets. Achieving 0.95 in segmentation accuracy. Transfer learning is still prone to domain shift problems even with its improved performance, especially when used with datasets with different data distributions as ImageCAS and Aorta.

Standard segmentation methods often rely on basic loss functions like Cross-Entropy loss, which struggle to effectively manage class imbalance or capture the finer details necessary for accurate segmentation. Recent research highlights the need to optimize for flatter minima in the loss landscape to enhance model generalization. Caldarola et al. [18] demonstrated that such optimization improves model robustness and generalization, especially when dealing with noisy or ambiguous data, which is essential for reliable cardio vascular segmentation across diverse clinical scenarios.

2.2 Transformers

Transformers were first proposed by [19] for machine translation and established state-of-the-arts in many NLP tasks. To make Transformers also applicable for computer vision tasks, several modifications have been made. For instance, Parmar et al. [20] applied the self-attention only in local neighborhoods for each query pixel instead of globally. Child et al. [21] proposed Sparse Transformers, which employ scalable approximations to global self-attention. Recently, Vision Transformer (ViT) [22] achieved state-of-the-art on ImageNet classification by directly applying Transformers with global self-attention to full-sized images. When pre-trained on large datasets like JFT-300M, ViT attained **88.55%** top-1 accuracy on ImageNet and **77.63%** on the VTAB-19 benchmark, outperforming ResNet-based models while requiring **2–4**$\times$ less pre-training compute. Notably, ViT demonstrated that pure self-attention architectures can surpass CNNs in vision tasks when scaled with sufficient data. To the best of our knowledge, the proposed TransUNet is the first Transformer based medical image segmentation framework, which builds upon the highly successful ViT.

2.3 Mamba for Medical Image Segmentation

The Mamba architecture [23] represents a substantial advancement in medical image segmentation by integrating the capabilities of Vision Transformers (ViTs) [22] and Convolutional Neural Networks (CNNs) [9]. It is essential to achieve precision in medical imaging duties by integrating global contextual information and managing long sequences [24]. U-Mamba [25] expands the conventional U-Net framework [5, 4] by integrating attention mechanisms and multi-scale processing, thereby improving the accuracy and robustness of segmentation. This is achieved by building upon this foundation. This method enables the model to concentrate on pertinent details within intricate anatomical structures and efficiently process information across a range of scales, from a broad contextual understanding to precise low-level details. Furthermore, U-Mamba incorporates deep supervision, which expedites training and enhances convergence, rendering it both efficient and dependable for clinical applications that require rapid and precise image processing. The Mamba architecture's adaptability is further underscored by specialized variants such as Weak-Mamba-UNet [26], which are able to handle intricate scenarios with improved performance and excel in scribble-based segmentation tasks. In conclusion, models that are based on U-Mamba exhibit superior segmentation performance in a variety of medical applications, such as histopathological imaging and cardiac MRI.

One of the main benefits of the Mamba design is its computational efficiency, which comes from the thoughtful fusion of ViTs [22] and CNNs [9]. The benefits of both architectures are used in this hybrid design structures: CNNs are skilled at extracting local features with little computing expense, whereas ViTs' global attention mechanisms enable the effective management of complex spatial linkages and long-range dependencies. By combining these two methods, Mamba lessens the computational load that is usually associated with pure transformer-based models, which are resource-intensive because of their quadratic complexity in relation to the input sequence's duration.

Chapter 3

Method

This chapter presents the methodological framework adopted to improve the segmentation of left ventricular structures in echocardiographic images. The approach builds upon the foundation of TransUNet by integrating Mamba blocks, a recent advancement in state-space modeling. The chapter begins with a critical analysis of Transformer limitations, particularly in high-resolution medical imaging, where their computational complexity and spatial encoding limitations can be prohibitive. It then introduces the Mamba architecture as a more efficient and scalable alternative that combines the local feature extraction strengths of convolutional networks with the global modeling capabilities of state-space models. The theoretical underpinnings and computational benefits of Mamba are detailed, followed by a full architectural description of the proposed Mamba-based TransUNet. This includes the encoder-decoder structure, skip connections, and key design choices that contribute to high segmentation performance with reduced computational burden. The methods described here lay the groundwork for the subsequent experimental evaluation and demonstrate how the integration of Mamba enhances both efficiency and accuracy in clinical segmentation tasks.

3.1 Motivation for Replacing Transformers with Mamba Blocks

3.1.1 Transformer Limitations

Self-attention techniques modeled long-range dependencies, and transformer-based architectures such as those used in TransUNet have shown promise in medical image segmentation [6]. Regardless, there are several inherent limitations of their application in high-resolution 2D medical images, such as echocardiograms:

- **Computational Overhead**: There's a substantial amount of computation done in developing the attention mechanism of transformers because of the quadratic scaling of computations as a function of the number of tokens ($\mathcal{O}(N^2)$) [19]. For high-resolution medical images, the token count increases substantially, leading to prohibitive memory and computational requirements. This makes real-time processing or deployment on resource-constrained clinical devices impractical.

- **Spatial Awareness Challenges**: Within the transformer architecture the tokens have been flattened, so any set spatial relationships have been lost since it operates in a sequential manner. While positional encodings were developed to address this, they only approximate spatial context while having difficulty with fine-grained details necessary for precise segmentation processes [6, 22].
- **Training Complexity**: Transformers don't work well when datasets are small, and often require extensive hyperparameter tuning [22]. In medical imaging contexts where labeled data is often restricted and costly to curate, this presents significant challenges.

3.2 Mamba Block Advantages

To address these limitations, we propose substituting the transformer layers in TransUNet with Mamba blocks [23]—a novel architecture combining depthwise convolutions for local feature extraction and simplified state-space models (SSMs) for global context modeling. The Mamba blocks offer several advantages:

- **Linear Complexity**: Using dense layers in the SSM path reduces computational complexity to linear ($\mathcal{O}(N)$) [23], offering efficiency and scalability for high-resolution images.
- **Native Spatial Awareness**: Mamba blocks operate directly on 2D feature maps, preserving spatial relationships without positional encodings [23].
- **Simplified Architecture**: The design leverages established convolutional principles [4], making implementation and fine-tuning more intuitive.
- **Better Suitability for Small Datasets**: By maintaining spatial structure and efficient feature mixing, Mamba blocks are particularly effective for medical imaging datasets with limited labeled samples [6, 23].

This replacement not only mitigates computational and scalability challenges but also provides a streamlined approach for deploying segmentation models in clinical environments, where efficiency, accuracy, and interpretability are paramount.

3.3 Detailed Explanation of the Mamba Block

3.3.1 Introduction

The Mamba block is a novel architectural component designed to address the computational inefficiencies of Transformer-based models while retaining their powerful ability to model global dependencies. Transformers, widely

used in vision and natural language processing tasks, have demonstrated exceptional performance due to their self-attention mechanism, which provides an uncompressed view of the entire input sequence. However, their significant computational and memory requirements—especially for high-resolution inputs like medical images—limit their scalability and applicability in resource-constrained environments.

In this study, we propose the Mamba block as a potential replacement of the Transformer in the TransUNet architecture to enhance computational efficiency. By leveraging a hybrid approach, the Mamba block combines local feature extraction through depthwise convolutions and global context modeling via selective State Space Models (SSMs). This enables the architecture to process high-resolution echocardiographic data efficiently while maintaining segmentation accuracy. The Mamba block's lightweight yet powerful design makes it particularly suited for real-time medical image segmentation tasks.

3.3.2 Transformers

The Transformer architecture has been a major component in the success of Large Language Models (LLMs). It has been used for nearly all LLMs that are being used today, from open-source models like Mistral to closed-source models like ChatGPT. It basically sees any textual input as a sequence that consists of tokens, just like in Figure 3.1. Whatever input it receives, it can look back at any of the earlier tokens in the sequence to derive its representation[27].

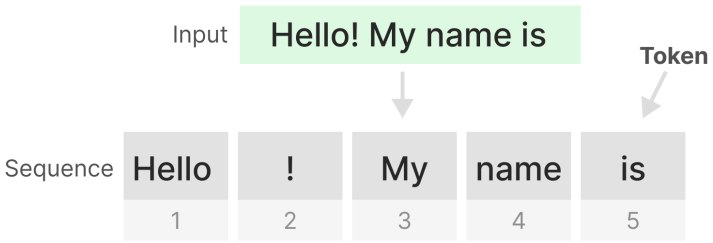

FIGURE 3.1: Example of text tokenization, showing input segmentation and token ID mapping.

- **Core Components**: A transformer consists of two structures: a set of encoder blocks for representing text and a set of decoder blocks for generating text. Together, these structures can be used for several tasks, including translation (Figure 3.2). We can adopt this structure to create generative models by using only decoders. A single decoder block consists of two main components: masked self-attention followed by a feed-forward neural network. Self-attention enables an uncompressed view of the entire sequence with fast training.

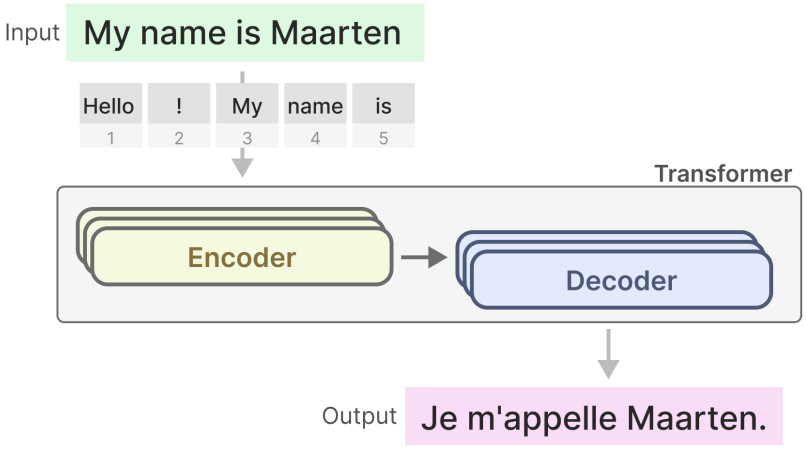

FIGURE 3.2: Transformer architecture processing the input sequence 'My name is Maarten' through encoder-decoder layers to produce the French translation 'Je m'appelle Maarten.

- **The Flaw**: When generating the next token, we need to re-calculate the attention for the entire sequence, even if we already generated some tokens. Generating tokens for a sequence of length L needs roughly L^2 computations which can be costly if the sequence length increases[28].

- **RNNs as Potential Solution**: Recurrent Neural Networks (RNN) is a sequence-based network that takes two inputs at each time step: the input at time step t and a hidden state of the previous time step $t-1$. RNNs can do inference fast as it scales linearly with the sequence length, but they tend to forget information over time since they only consider one previous state.

  RNNs have a looping mechanism that allows them to pass information from a previous step to the next. We can "unfold" this visualization to make it more explicit. When generating the output, the RNN only needs to consider the previous hidden state and current input. It prevents recalculating all previous hidden states which is what a Transformer would do[29].

  In other words, RNNs can do inference fast as it scales linearly with the sequence length. In theory, it can even have an infinite context length. The illustration in Figure 3.3 explains how each hidden state is the aggregation of all previous hidden states and is typically a compressed view, and this is where the problem is spotted. If you notice the last hidden state, when producing the name "Maarten" does not contain information about the word "Hello" anymore. RNNs tend to forget information over time since they only consider one previous state.

3.3.3 The State Space Model

A State Space Model (SSM), like the Transformer and RNN, processes sequences of information:

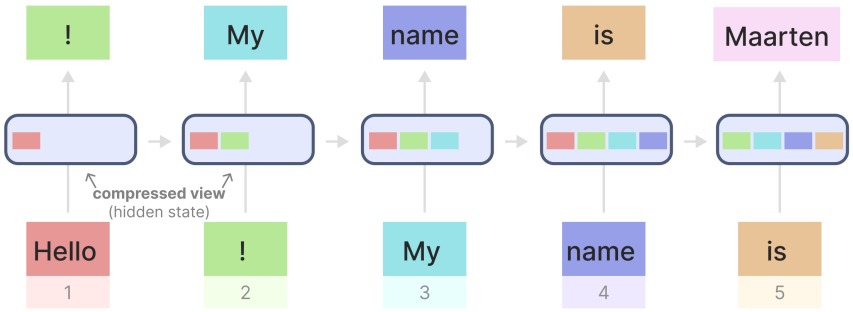

FIGURE 3.3: Input sequence compression showing the original text 'My name is Maarten' (top) and its tokenized hidden state representation (bottom) with positional IDs (1-5).

- **State Space Representation**: The "state space" is the map of all possible states. Each point represents a unique position with specific details. The variables that describe a state can be represented as "state vectors" Figure 3.4.

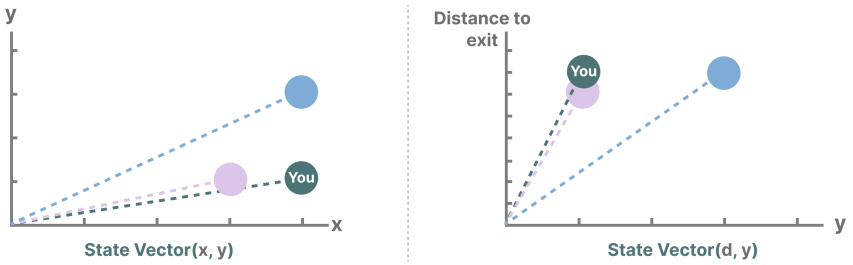

FIGURE 3.4: State vector representation showing position coordinates (x,y) and transformed coordinates (d,y) relative to exit distance.

- **SSM Functionality**: At time t, SSMs map an input sequence $x(t)$ to a latent state representation $h(t)$ and derive a predicted output sequence $y(t)$. SSMs assume that dynamic systems can be predicted from its state at time t through state equations[30].

  However, instead of using discrete sequences (like moving left once) it takes as input a continuous sequence and predicts the output sequence (Figure 3.5). SSMs assume that dynamic systems, such as an object moving in 3D space, can be predicted from its state at time t through two equations[31]. By solving these equations, we assume that we can uncover the statistical principles to predict the state of a system based on observed data (input sequence and previous state), Its goal is to find this state representation h(t) such that we can go from an input to an output sequence. The full SSM architecture can be in Figure 3.6.

3.3.4 Mamba - A Selective SSM

The information provided earlier are the elements that make Mamba special, State Space Models can be used to model textual sequences but still have a

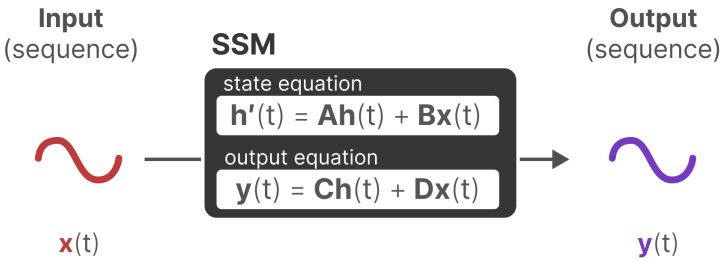

FIGURE 3.5: State Space Model equations showing the continuous-time state update (h′(t)) and output computation (y(t)).

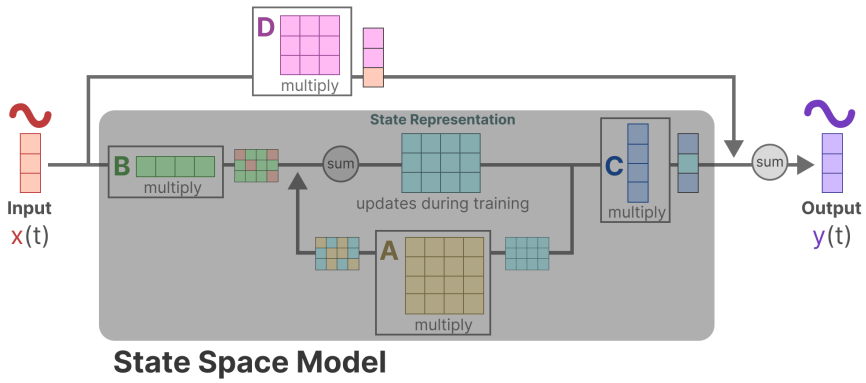

FIGURE 3.6: Computational flow of a State Space Model, highlighting matrix multiplications (A,B,C,D) and state updates during training.

set of disadvantages we want to prevent[32].

Mamba's two main contributions are:

- **Selective Scan Algorithm**: The recurrent representation of an SSM creates a small state that is quite efficient as it compresses the entire history. However, compared to a Transformer model which does no compression of the history (through the attention matrix), it is much less powerful. Mamba aims to have the best of both worlds. A small state that is as powerful as the state of a Transformer Figure 3.7, and it does so by compressing data selectively into the state. When you have an input sentence, there is often information, like stop words, that does not have much meaning.

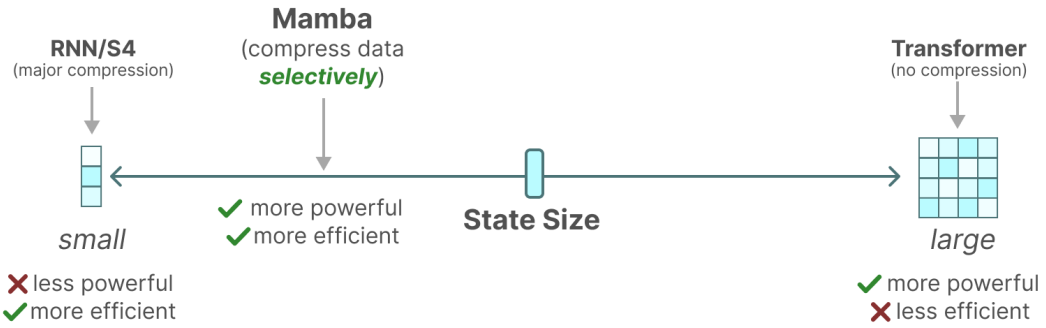

FIGURE 3.7: Comparison of model architectures (RNN/S4, Mamba, Transformer) in terms of state compression, efficiency, and power. Mamba selectively compresses data—unlike Transformers (no compression) and RNNs (full compression)—to balance efficiency and performance.

- **Hardware-aware Algorithm**: One of the disadvantages of recent GPUs is their limited transfer (IO) speed between their small but highly efficient SRAM and their large but slightly less efficient DRAM. Frequently copying information between SRAM and DRAM becomes a bottleneck (Figure 3.8).

  Mamba, like Flash Attention, attempts to limit the number of times we need to go from DRAM to SRAM and vice versa. It does so through kernel fusion which allows the model to prevent writing intermediate results and continuously performing computations until it is done. The full representation of the SSM is in Figure 3.9. These two create the selective SSM or S6 models which can be used, like self-attention, to create Mamba blocks[33].

3.3.5 The Mamba Block

The selective SSM that we have explored thus far can be implemented as a block, the same way we can represent self-attention in a decoder block. Like the decoder, we can stack multiple Mamba blocks and use their output as the input for the next Mamba block (Figure 3.10). It starts with a linear

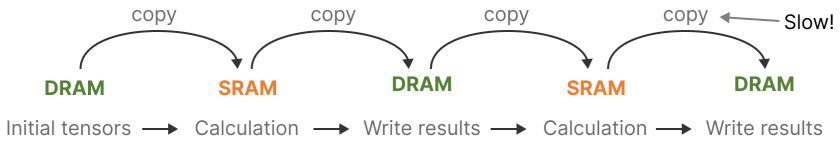

FIGURE 3.8: Computational bottleneck caused by repeated DRAM-SRAM transfers during sequential operations, slowing down tensor processing.

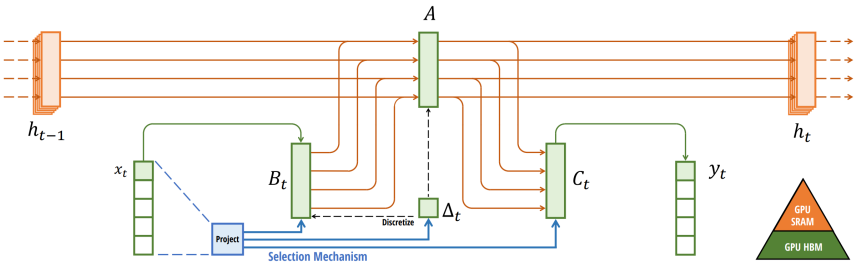

FIGURE 3.9: Mamba selective state update flow in GPU SRAM, showing parameter projection and discretization to avoid DRAM overhead.

projection to expand upon the input embeddings. Then, a convolution before the Selective SSM is applied to prevent independent token calculations.

3.3.6 Conclusion

The Mamba block provides a compelling alternative to vanilla Transformer models, particularly when both computational cost and robust sequence modeling are needed. While Transformers excel at modeling global dependencies with self-attention, their quadratic complexity poses severe limitations for long sequences. State Space Models meet this need by succinctly modeling continuous dynamics sequences, and Mamba further enhances SSMs through selective state retention and hardware optimization.

3.4 Overview of Mamba-Based TransUNet

Mamba-based TransUnet is a novel architecture designed for precise and efficient segmentation of echocardiographic images. The model adapts the TransUNet framework by replacing its transformer blocks with lightweight Mamba blocks. This modification enhances computational efficiency and scalability for high-resolution medical imaging while retaining the ability to model global dependencies.

3.4.1 Overall Architecture

The Mamba-based TransUnet follows a U-shaped encoder-decoder design, a standard architecture for medical image segmentation. The model consists

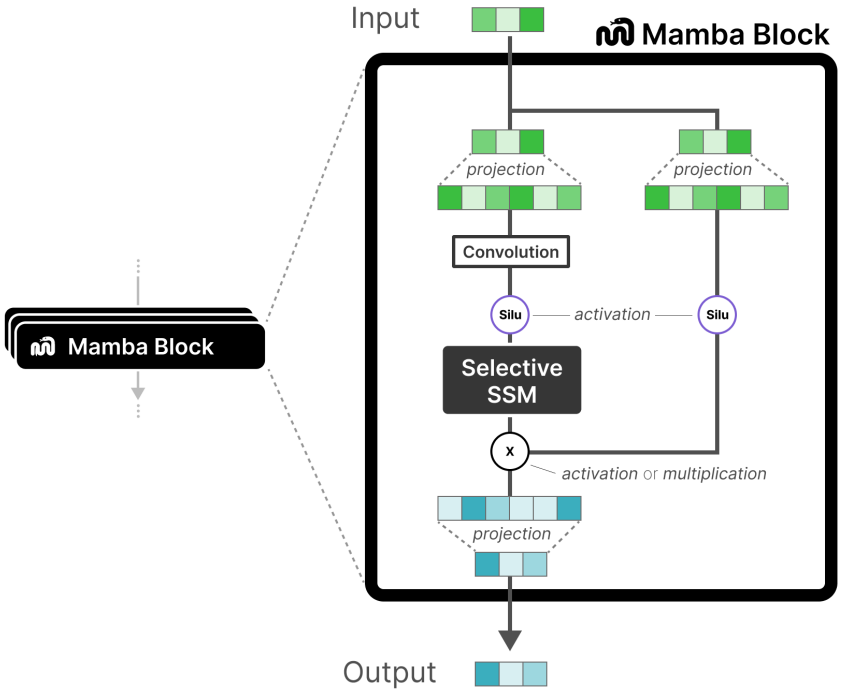

FIGURE 3.10: The mamba block representation.

of three primary components: an encoder path that extracts hierarchical features at multiple abstraction levels, a Mamba Block (Bottleneck) that captures global dependencies and context with reduced computational overhead, and a decoder path that reconstructs the segmentation map via upsampling and feature refinement while leveraging skip connections to preserve spatial details. This hybrid architecture effectively integrates local feature extraction with global context modeling, achieving high segmentation accuracy while minimizing computational complexity.

3.4.2 Encoder Path

The encoder learns multi-scale features from echocardiographic images. It begins with an initial convolutional layer that projects the input into a higher-dimensional feature space. Subsequent layers consist of Mamba blocks, processing features at progressively coarser resolutions via max-pooling. This hierarchical structure enables the model to capture both high-level semantics and fine-grained details essential for accurate segmentation.

3.4.3 Mamba Block

The Mamba block, deployed in both the encoder and bottleneck, represents the key innovation of this architecture. It replaces traditional transformer

blocks with a more efficient design based on selective State Space Models (SSMs). The block comprises:

- **Local Feature Extraction**: Depthwise convolutions extract spatially localized features.
- **Global Context Modeling**: A simplified SSM captures long-range dependencies.
- **Selective Information Filtering**: Dynamic filtering retains task-relevant context while discarding noise.
- **Residual Connections and Dropout**: Enhance training stability and regularization.

This combination balances representational capacity and computational efficiency, making the Mamba block ideal for high-resolution medical images.

3.4.4 Decoder Path

The decoder reconstructs the segmentation map by progressively upsampling encoded features. Transposed convolutions restore spatial resolution, while skip connections concatenate encoder features to recover fine-grained details lost during downsampling. This hierarchical refinement ensures that both local structures and global context contribute to the final segmentation.

3.4.5 Advantages of Mamba-Based TransUNet

The Mamba-based TransUNet architecture offers several key advantages for medical image segmentation [6]. By replacing traditional quadratic-complexity self-attention mechanisms [19] with linear-scaling state space models (SSMs) [23], the system achieves significantly reduced computational complexity during both training and inference phases [33]. This architectural innovation enables superior scalability for processing high-resolution echocardiographic images while maintaining real-time deployment capabilities [5]. The model preserves crucial spatial details through a combination of skip connections [4] and hierarchical feature fusion mechanisms, ensuring accurate retention of fine anatomical structures throughout the segmentation process [2]. Furthermore, the synergistic integration of local convolutional operations [9] with global state space modeling [30] yields state-of-the-art segmentation accuracy, particularly beneficial for complex echocardiographic analysis tasks where precise boundary delineation is critical [12]. This combination of computational efficiency, scalability, structural preservation, and diagnostic accuracy positions Mamba-based TransUNet as an advanced solution for medical imaging applications [24, 25].
In conclusion, this chapter has detailed the motivation, theoretical rationale, and architectural implementation of the proposed Mamba-based TransUNet model. By addressing the limitations of transformer-based approaches—such

as high computational cost, limited spatial coherence, and training difficulties on small datasets—Mamba emerges as a viable alternative for efficient medical image segmentation. Its core innovations, namely selective state space modeling and hardware-aware computation, enable linear scaling while preserving global contextual understanding and spatial precision. The integration of Mamba blocks into the encoder-decoder framework of TransUNet results in a lightweight yet powerful segmentation model. This hybrid architecture not only meets the stringent demands of high-resolution echocardiographic image processing but also paves the way for future deployment in real-time clinical settings. The subsequent chapter will validate this method through extensive experiments, comparing its performance against established models in the field.

Chapter 4

Preprocessing and Data Preparation

The success of any deep learning-based medical image segmentation task is highly dependent on the quality and consistency of data preprocessing. Echocardiographic images present unique challenges due to their inherent variability, noise, and structural complexity. This chapter outlines the preprocessing and data preparation pipeline adopted to ensure robust model training and evaluation. We begin with a description of the CAMUS dataset, which serves as the primary benchmark for our study. The preprocessing steps—including loading, resizing, normalization, augmentation, and dataset splitting—are then detailed. These operations are carefully designed to standardize input dimensions, enhance image diversity, and prevent data leakage, thereby enabling the proposed model to generalize effectively across different anatomical variations and patient cases. This chapter provides a foundational framework for constructing reproducible and clinically relevant segmentation systems.

4.1 Description of the Echocardiography Dataset

Accurate segmentation of 2D echocardiographic images has been an ongoing challenge for over three decades, with difficulties arising from three primary factors: (1) **Image Characteristics** - echocardiographic images suffer from poor contrast, brightness inhomogeneities, speckle pattern variations along the myocardium, and significant inter-patient echogenicity variability, complicating cardiac region localization; (2) **Limited Public Datasets** - the scarcity of large-scale, publicly available 2D echocardiographic datasets hinders method development and validation; and (3) **Lack of Multi-Expert Annotations** - few datasets provide annotations from multiple experts, making it difficult to establish the minimum error margin for human-expert-level accuracy.

Despite these challenges, numerous segmentation methods have been proposed. However, most are validated on small private datasets (often with fewer than 100 patients), rendering cross-method comparisons impractical [8]. Consequently, clinical practice still relies on semi-automatic or manual annotations due to the insufficient accuracy and reproducibility of fully automatic methods [8].

To address these limitations, the **CAMUS (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation)** dataset was introduced [34]. It is the largest publicly available, fully annotated dataset for 2D echocardiographic assessment. The organizers aimed to provide the community with resources to resolve segmentation and volume estimation challenges from 2D ultrasound sequences (including two- and four-chamber views).

4.2 Preprocessing Pipeline

A robust and consistent preprocessing pipeline is crucial for training deep learning models on medical imaging data, especially when dealing with volumetric data formats such as NIfTI. The following subsections describe the main stages of the preprocessing and dataset preparation pipeline implemented for this project.

4.2.1 Loading and Resizing of NIfTI Images

The dataset consists of echocardiography scans and their corresponding ground truth segmentation masks, stored in the NIfTI format ('.nii.gz'). Each patient folder contains images acquired in different views (e.g., 2-chamber, 4-chamber) and at different cardiac cycle time points (e.g., end-diastole (ED), end-systole (ES)).

- **Loading**: For each patient, both the image and its corresponding mask are loaded using the `nibabel` library, which is well-suited for handling neuroimaging data in NIfTI format.
- **Resizing**: Since medical images can differ in spatial resolution, all images and masks are resized to a uniform dimension of 256×256 pixels using `skimage.transform.resize`. This ensures compatibility with the neural network input requirements and helps in batch processing.

4.2.2 Normalization and Augmentation

- **Normalization**: Each resized image is normalized to the [0, 1] range. This is achieved by:

$$I_{\text{norm}} = \frac{I - \min(I)}{\max(I) - \min(I)} \tag{4.1}$$

- **Mask Binarization**: The ground truth segmentation masks are binarized. For endocardial segmentation, the mask is set to 1 for the relevant class (endo) and 0 elsewhere.
- **Data Augmentation (Rotation-Based)**: To improve model generalization and mitigate overfitting, rotation-based augmentation is applied to the training data. Each image–mask pair is rotated by 0°, 90°, 180°,

and 270° using `scipy.ndimage.rotate`. This increases the diversity of the training set by exposing the model to multiple spatial orientations of the same anatomy. Augmentation is only applied to the training set, not to the validation or test sets.

4.2.3 Dataset Split

The dataset is split in a patient-wise manner to avoid data leakage.

- **Initial Split**: All available patient IDs are split into a training set (85%) and a test set (15%) with a fixed random seed for reproducibility.
- **K-Fold Cross-Validation**: The training set is further split into training and validation sets using K-Fold cross-validation (`KFold`). Specifically, the code implements 3-fold cross-validation. In each fold, a different subset is used as the validation set, while the remaining patients serve as the training set. This strategy enables robust estimation of model performance and helps in utilizing the data efficiently.

4.2.4 Final Dataset Size after Augmentation

The preprocessing pipeline systematically expands the training set through four precisely calculated rotational augmentations (0°, 90°, 180°, and 270°) applied to each original echocardiographic image-mask pair, effectively multiplying the available training samples while preserving diagnostic integrity through careful geometric transformations. In contrast, both validation and test sets maintain their original composition without augmentation, with the test set carefully isolated during the initial 85%-15% patient-wise split to serve as an unbiased benchmark. The cross-validation framework implements three distinct folds, each with unique partitions of the augmented training data while keeping validation sets completely separate, with all medical images standardized to 256×256 pixel resolution in grayscale format. This augmentation methodology preserves critical cardiac features through specialized rotation techniques (using reflection padding) coupled with intensity normalization to the [0,1] range, while the rigorous patient-wise separation protocol prevents any data leakage between subsets. The configuration yields substantially enriched training data through systematic multiplication of original samples while maintaining pristine validation and test sets for reliable evaluation, with all processing optimized for batch operations (size=8) to ensure computational efficiency during the extended 100-epoch training cycles initiated with a 1e-4 learning rate. Comprehensive validation checks during NIfTI loading and transformation guarantee anatomical consistency between each image and its corresponding mask throughout all preprocessing stages.
In summary, this chapter has detailed a comprehensive preprocessing pipeline tailored for 2D echocardiographic image segmentation. The use of the CAMUS dataset, combined with rigorous normalization, geometric augmentation, and patient-wise splitting strategies, ensures that the training process is

both data-efficient and clinically grounded. By augmenting the training data through rotational transformations and carefully maintaining the integrity of validation and test sets, the pipeline enables reliable performance evaluation while minimizing the risk of overfitting. Moreover, the use of standardized image dimensions and intensity normalization supports consistent input formatting across all training cycles. These preprocessing protocols are instrumental in preparing the dataset for training the Mamba-based TransUNet model described in the following chapters, setting the stage for accurate and efficient left ventricle segmentation.

Chapter 5

Implementation of Mamba-TransUNet for Medical Image Segmentation

This chapter details the practical implementation of the proposed Mamba-TransUNet architecture for medical image segmentation, with a specific focus on echocardiographic data. Building upon the TransUNet framework, which integrates convolutional and transformer-based components, the study introduces Mamba blocks as a lightweight and efficient alternative to Vision Transformers. This modification aims to reduce computational complexity while preserving or enhancing segmentation performance. The chapter begins by outlining the baseline TransUNet design, followed by a comprehensive description of the Mamba block and its integration into the network's bottleneck. Subsequent sections cover the training setup, loss function optimization, and validation strategy, addressing real-world computational constraints encountered during model development. Particular attention is given to the practical challenges of training on high-resolution cardiac ultrasound images, including memory limitations and augmentation-induced instability. These implementation insights form the basis for the model's evaluation in the following chapter.

5.1 Model Architecture

TransUNet is a hybrid deep learning architecture that combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for medical image segmentation tasks. The core idea is to leverage CNNs for local feature extraction and ViTs for global context modeling, addressing the limitations of pure CNNs in capturing long-range dependencies [6].

5.1.1 TransUNet Architecture

TransUNet follows the classic U-Net encoder–decoder design, but introduces a Transformer block at the bottleneck. The architecture can be summarized as follows:

- **Encoder (CNN):** A series of convolutional and pooling layers extract hierarchical features and progressively reduce spatial resolution, with skip connections at multiple stages.
- **Bottleneck (ViT):** The encoder's output is divided into non-overlapping patches, embedded, and then processed by multi-layer transformer blocks (ViT). This module models long-range dependencies and global interactions.
- **Decoder (CNN):** The decoder upsamples the features and merges them with encoder skip connections to recover spatial details, leading to a final segmentation output.

The architecture of TransUNet is depicted in Figure 5.1.

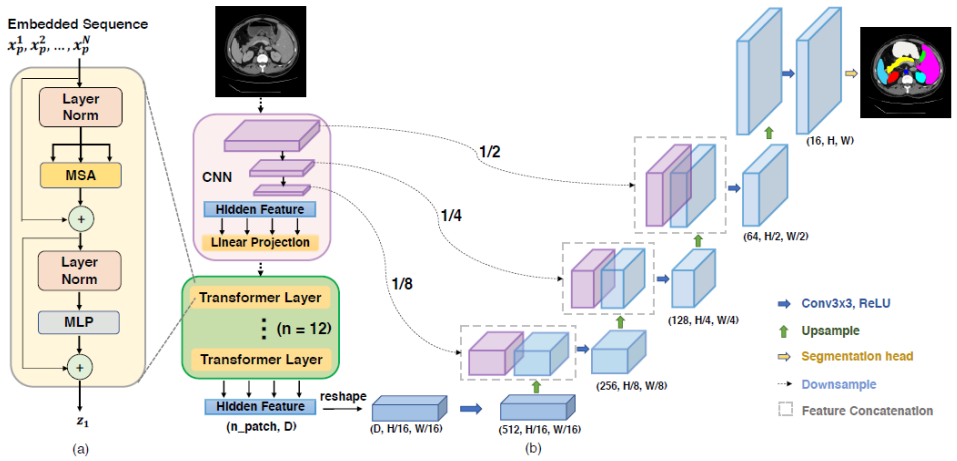

FIGURE 5.1: TransUNet architecture: a hybrid of U-Net and Vision Transformer.

Vision Transformer (ViT) Block

In TransUNet, the ViT block is responsible for capturing global contextual information. The process involves:

1. **Patch Embedding:** The encoder output feature map is divided into fixed-size patches (e.g., 16x16), each flattened and linearly projected to a vector embedding.
2. **Positional Encoding:** Learnable positional encodings are added to the patch embeddings to retain spatial information.
3. **Transformer Layers:** Multiple transformer encoder layers are stacked. Each consists of:
   - Multi-head self-attention (MHSA) for modeling global dependencies.
   - Feed-forward networks (FFN) for feature transformation.

- Layer normalization and residual connections for stable learning.

A visual representation of the ViT block is shown in Figure 5.2.

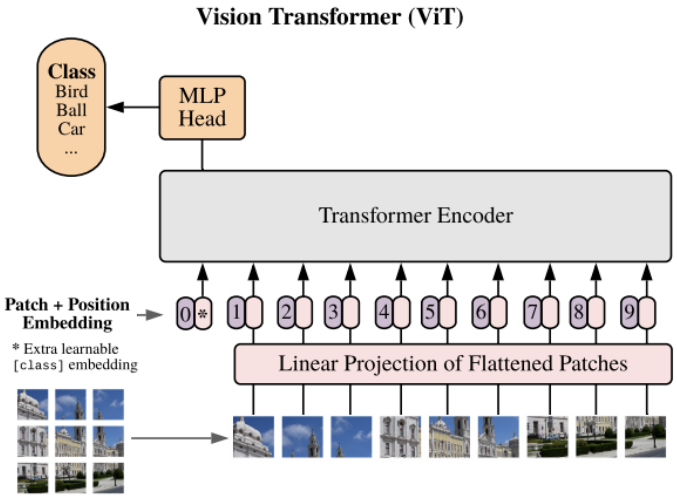

FIGURE 5.2: Vision Transformer (ViT) block.

Mathematically, let X be the patch-embedded sequence, E_{pos} the positional encoding, and $T(\cdot)$ the transformer encoder:

$$Z = T(X + E_{pos}) \tag{5.1}$$

5.1.2 Mamba Block: Replacing Vision Transformers

Motivation

While ViTs excel at modeling global dependencies, they are computationally intensive and require large datasets for effective training. Recent research has introduced Mamba, a state space model (SSM)-based architecture that efficiently captures long-range dependencies with lower computational overhead, making it attractive for medical imaging tasks where data can be limited [23].

Mamba Block Design

The Mamba block is inspired by sequence modeling with SSMs. It aims to capture both local and long-range dependencies by combining convolutional operations with an SSM pathway.

- **Depthwise Convolution:** The input features undergo a depthwise convolution, allowing for efficient local feature mixing per channel.
- **Projection to Hidden Dimension:** The result is projected to a higher-dimensional space using a 1×1 convolution, followed by a non-linear activation (e.g., ReLU).

- **State Space Model (SSM) Path:** The spatial feature map is reshaped into a sequence suitable for SSM processing. Two dense layers approximate the SSM, with non-linear activation in between (e.g., Swish). A residual connection is added between the transformed and original sequence. The sequence is reshaped back to the spatial domain.
- **Projection Back:** The feature map is projected back to the original channel dimension.
- **Dropout & Residual Addition:** Dropout is applied, and the block output is added to the input (residual connection).

We can stack multiple Mamba blocks and use their output as the input for the next Mamba block (Figure 3.10).

Replacing ViT with Mamba in TransUNet

To create a more efficient and potentially data-efficient segmentation model, the ViT bottleneck in TransUNet is replaced by a stack of Mamba blocks. The integration is as follows: the encoder and decoder paths remain unchanged, the ViT block is removed, and instead, multiple Mamba blocks are stacked at the bottleneck, operating directly on the spatial feature maps while preserving the skip connections in the encoder–decoder architecture to ensure spatial detail recovery.

A comparison between the original TransUNet (using ViT) and the modified Mamba-TransUNet is summarized in Table 5.1. The new CNN architecture got with the replacement is shown in Figure 5.3

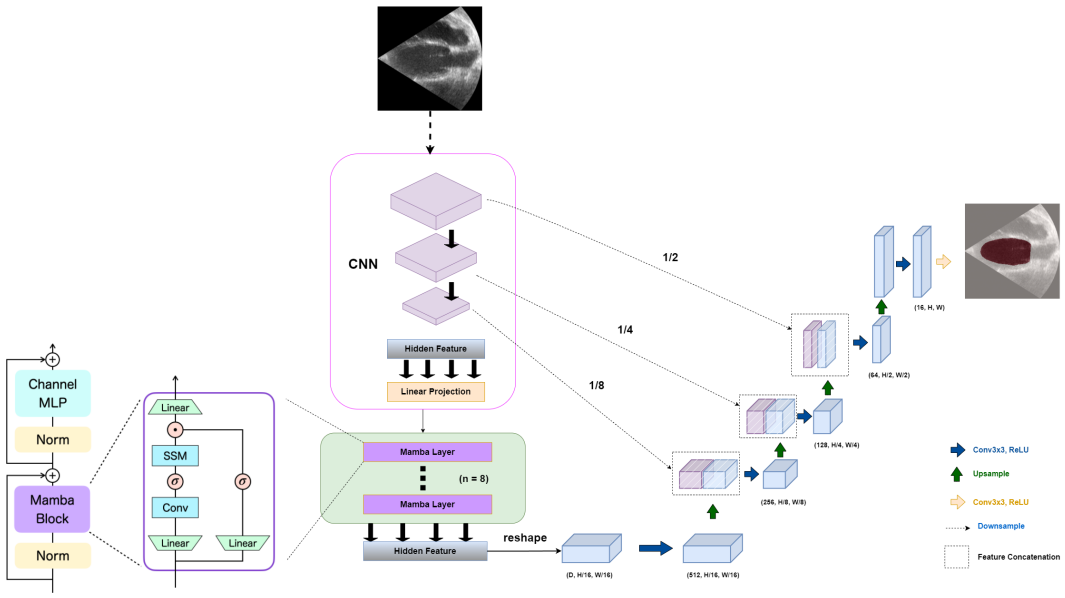

FIGURE 5.3: Vision Transformer (ViT) block.

TABLE 5.1: Comparison of TransUNet (ViT) and Mamba-TransUNet.

Architecture	Bottleneck	Long-range Modeling	Efficiency
TransUNet (ViT)	ViT	Self-attention	High cost
Mamba-TransUNet	Mamba	SSM	Efficient

5.1.3 Training Setup

Hardware and Computational Constraints

The model was developed and trained on Google Colab Pro using an NVIDIA Tesla T4 GPU with 16 GB of VRAM and 53 GB of system RAM. Despite these resources, we encountered significant memory constraints that shaped our implementation:

- Initial attempts with batch size 16 failed due to memory limitations, requiring reduction to 8 (and further to 4 for some experiments)
- The fixed 256×256 resolution (clinically required) consumed $\sim$12 GB VRAM during training
- Data augmentation expanded the dataset from 500 to $\sim$4300 samples after careful tuning of rotation parameters

Loss Function Optimization

To address class imbalance (left ventricle representing only 15-20% of pixels), we implemented a weighted combination:

$$\mathcal{L}_{\text{BCE-Dice}} = \alpha \cdot \text{BCE}(y_{\text{true}}, y_{\text{pred}}) + (1 - \alpha) \cdot \big(1 - \text{Dice}(y_{\text{true}}, y_{\text{pred}})\big) \tag{5.2}$$

where $\alpha = 0.25$ emphasized the Dice term for better structural overlap. This proved crucial for segmenting small ventricular areas.

Training Protocol

The training configuration included: Adam optimizer with initial learning rate 1×10^{-4}, using a reduce-on-plateau learning schedule (factor 0.5, minimum learning rate 1×10^{-6}), early stopping with patience of 15 epochs on validation Dice score, and gradient clipping with threshold 1.0 to stabilize Mamba blocks during training.

Validation Strategy

We implemented rigorous 3-fold cross-validation with patient-level data splitting to prevent leakage, fixed random seeds (42) for reproducibility, a separate holdout test set comprising 15% of patients, and ensemble prediction across folds for final evaluation to ensure robust performance assessment.

Debugging Challenges

Key technical hurdles included: memory spikes during backpropagation through Mamba blocks which were resolved via gradient checkpointing, non-deterministic operations in state-space layers that were fixed by disabling CUDA optimizations during validation, augmentation-induced out-of-memory errors mitigated by limiting rotations to 4 angles (0°, 90°, 180°, 270°), and slow convergence in early layers which was improved by adding BatchNorm after each Mamba block.

5.1.4 Implementation Challenges

One of the most significant challenges encountered during the development of the Mamba-based TransUNet architecture was the abrupt crashing of the training notebook due to excessive memory allocation, despite using Google Colab Pro with 53 GB of RAM. The issue occurred immediately after loading the dataset, even before the first training iteration could begin. Initial attempts to mitigate this problem involved reducing the batch size to as low as 2 or 4, but this proved ineffective. Since the study's integrity relied on maintaining the original image resolution (256×256), downscaling was not a viable solution. Adjusting other hyperparameters, such as reducing the number of epochs, also failed to resolve the issue.

Upon further investigation, we discovered that the root cause was the aggressive data augmentation pipeline, which expanded the dataset to over 5,000 images—far exceeding the available memory capacity. By carefully modifying the augmentation factors (such as reducing the range of rotation angles and limiting the number of synthetic variations per image), we managed to bring the augmented dataset down to approximately 4,300 images. This adjustment finally allowed the system to allocate memory properly and initiate training without crashes.

Additionally, we observed that the Mamba blocks' state-space operations introduced unexpected memory spikes during backpropagation. To stabilize training, we implemented gradient checkpointing in the Mamba layers, selectively recomputing intermediate activations rather than storing them, which reduced memory overhead by $\sim$30% at the cost of a modest increase in computation time. These modifications were critical in ensuring the feasibility of training the proposed architecture on high-resolution echocardiography data.

In conclusion, this chapter presented the design, integration, and deployment of the Mamba-TransUNet architecture for 2D echocardiographic image segmentation. By replacing the transformer bottleneck in TransUNet with computationally efficient Mamba blocks, the model capitalizes on state-space modeling to achieve global context understanding with linear complexity. The architecture maintains the strengths of the encoder–decoder paradigm while improving scalability and reducing memory consumption, making it more suitable for deployment in constrained clinical environments. Rigorous preprocessing, optimization strategies such as hybrid loss functions, and

robust validation protocols ensured a reliable training process. Despite facing significant implementation challenges—including hardware limitations, memory bottlenecks, and augmentation tuning—the final architecture was successfully trained on high-resolution data. These engineering decisions and algorithmic innovations collectively contribute to the development of an efficient, accurate, and deployable deep learning system for medical image segmentation, setting the stage for its performance evaluation in the next chapter.

Chapter 6

Replicating Existing Segmentation Models

Establishing a rigorous benchmark is essential when proposing a novel deep learning architecture for medical image segmentation. This chapter focuses on replicating a set of representative and state-of-the-art convolutional and hybrid models to ensure a fair and controlled comparison with the proposed Mamba-TransUNet. The motivation lies not only in validating the reproducibility of key segmentation models—such as U-Net, DeepLabV3, nnU-Net, GUDU, and TransUNet—but also in standardizing the experimental conditions under which these models are evaluated. Given the significant variance in reporting practices and dataset preprocessing across prior works, direct comparisons are often unreliable unless conducted under uniform protocols. Here, all models were re-implemented and trained on the same dataset (CAMUS) using harmonized preprocessing, loss functions, and validation strategies. This rigorous approach provides a credible foundation for performance analysis in the subsequent chapters and offers practical insights into the adaptability of each architecture to echocardiographic segmentation.

6.1 Motivation and Rationale

In the development of a novel segmentation architecture tailored for left ventricle delineation on 2D echocardiographic images, it is imperative to first establish a reliable foundation for comparison. This chapter is dedicated to the replication of several well-established convolutional neural network (CNN) architectures in the field of medical image segmentation. The objective is twofold: to validate the reproducibility of existing segmentation methods, and to provide robust baseline metrics against which the performance of the proposed Mamba-based TransUNet model can be quantitatively and qualitatively assessed.

The segmentation of the left ventricle plays a crucial role in the diagnosis and assessment of cardiovascular diseases. Over the years, a diverse array of CNN architectures have been proposed, each offering distinct contributions in terms of accuracy, efficiency, and generalization capability. However, many of these contributions are reported under different experimental settings, making direct comparison with newly proposed models a challenge

unless they are independently re-implemented and evaluated under a consistent experimental protocol [35, 13].

6.2 Replicated Architectures

To this end, we have meticulously replicated a selection of state-of-the-art and representative CNN-based segmentation models. These include:

- **U-Net** [4], a foundational encoder-decoder model featuring skip connections, known for its strong performance in biomedical image segmentation tasks (Figure 6.1).
- **DeepLabV3** [3], a model incorporating atrous spatial pyramid pooling (ASPP) to capture multi-scale contextual information, widely used in medical image analysis (Figure 6.3).
- **nnU-Net** [5], a self-configuring method that adapts its architecture and preprocessing pipelines based on dataset-specific characteristics (Figure 6.3).
- **GUDU** [36], a geometrically-constrained U-Net variant designed specifically for enhancing ultrasound-based segmentation through domain-specific augmentations (Figure 6.3).
- **TransUNet** [6], a hybrid model that integrates Transformer blocks into the U-Net framework to capture long-range dependencies in medical images.

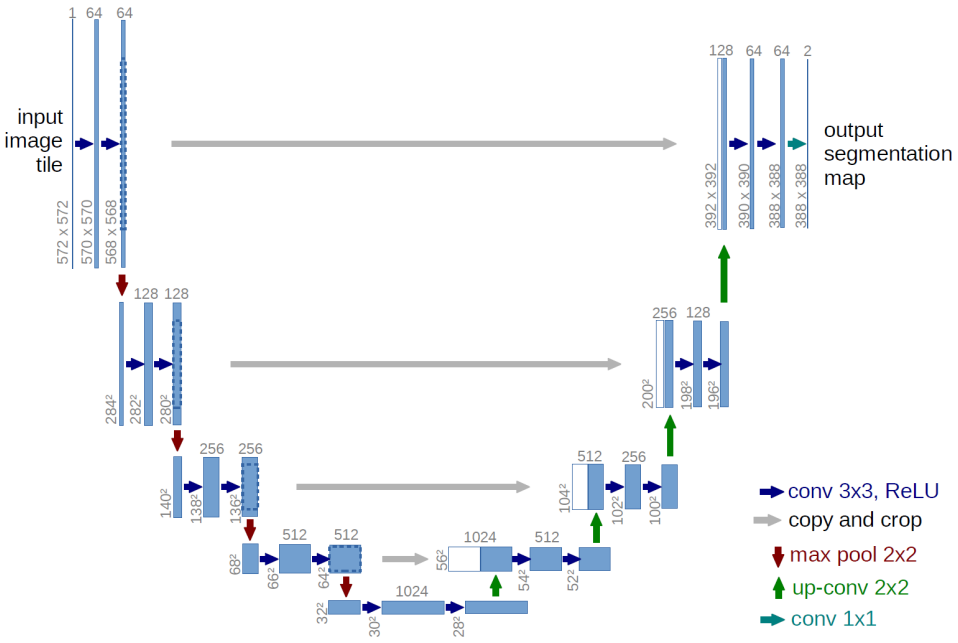

FIGURE 6.1: U-Net Architecture

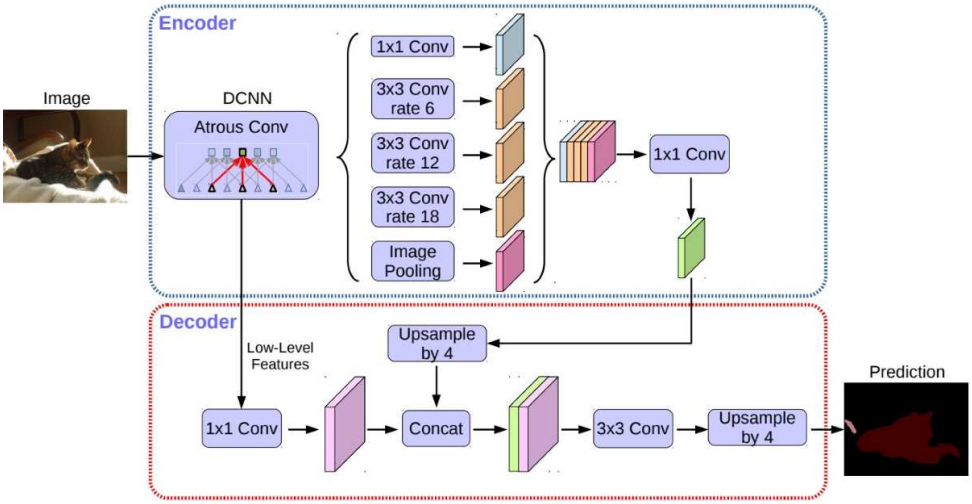

FIGURE 6.2: DeepLabV3 Architecture

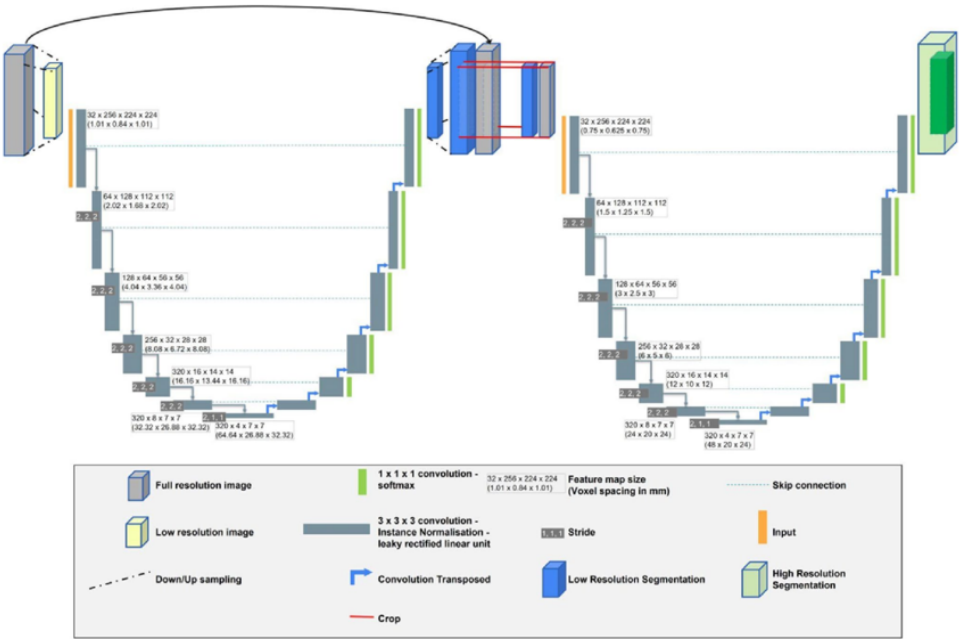

FIGURE 6.3: DeepLabV3 Architecture

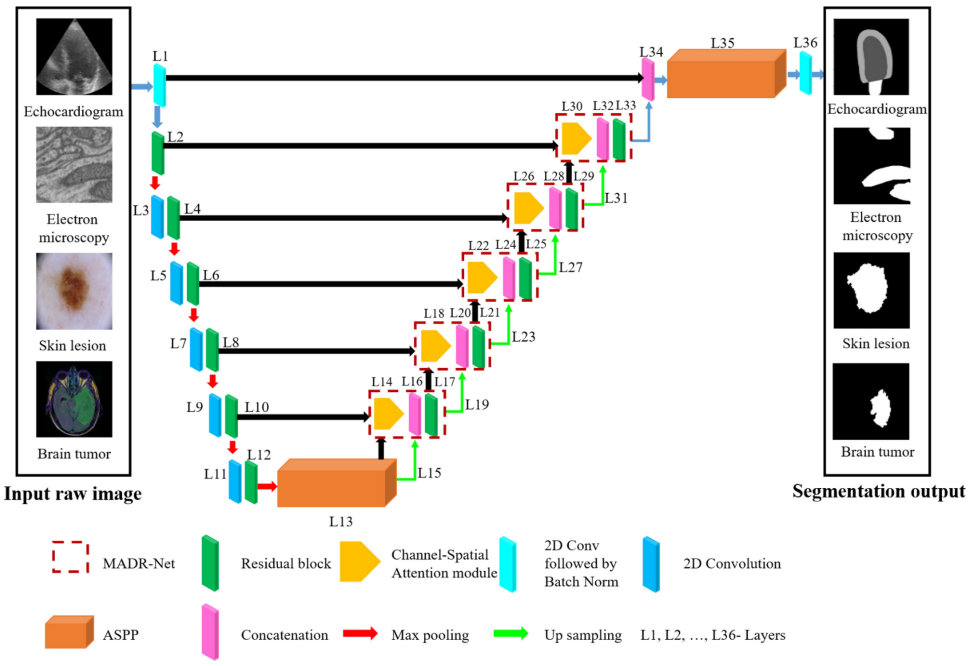

FIGURE 6.4: GUDU Architecture

These models were implemented, trained, and evaluated using a unified pipeline and a consistent dataset (CAMUS) [8], allowing for fair and direct performance comparison. The replication was carefully conducted following the methodological descriptions in their respective publications and supported by open-source code when available.

6.3 Implementation Protocol

Each model was implemented following the architecture and training procedures described in its original publication. When necessary, adjustments were made to harmonize preprocessing, loss functions, and evaluation metrics across all models.

U-Net [4]. This architecture follows a symmetric encoder-decoder structure with skip connections that preserve spatial features. For CAMUS data, grayscale images were resized to 256×256, normalized, and trained using a combination of binary cross-entropy and Dice loss. The original use of ReLU

activations and batch normalization was preserved to maintain architectural fidelity.

DeepLabV3 [3]. DeepLabV3 integrates atrous spatial pyramid pooling (ASPP) to handle multi-scale context. A ResNet-50 backbone was used, and atrous convolutions with varying dilation rates were applied to preserve resolution. Training involved adapting the input stride and output stride parameters to balance resolution with computational feasibility. Modifications were also made to the decoder to accommodate echocardiographic texture patterns.

nnU-Net [5]. This self-configuring framework automatically determines the optimal architecture, preprocessing pipeline, and training settings. The configuration process involved resampling, Z-score normalization, patch-wise training, and five-fold cross-validation. Due to high computational demand, smaller patch sizes and limited ensembling were employed to ensure training feasibility within available GPU memory constraints.

GUDU [36]. The GUDU architecture is based on U-Net and incorporates domain-specific data augmentation techniques inspired by geometric variations typical of echocardiographic acquisitions. The pipeline employed specialized transformations mimicking probe movements and intensity shifts. Additionally, anatomical constraints were encoded into the loss function to regularize predictions.

TransUNet [6]. TransUNet combines CNN-based encoders with Vision Transformer modules to enhance global feature modeling. ResNet-50 was used as the feature extractor, followed by a Transformer encoder operating on flattened tokenized patches. The decoder mirrored the U-Net structure. Modifications included adapting positional embeddings and resizing CAMUS images for compatibility with Transformer input dimensions.

6.4 Consistency and Adaptations

To ensure fairness in comparison, a unified preprocessing and evaluation protocol was applied:

- All inputs were resized to 256×256, normalized, and formatted as grayscale single-channel images.
- Binary segmentation was performed on the left ventricle endocardium.
- Loss functions primarily combined Dice loss with cross-entropy for all models unless explicitly overridden (e.g., nnU-Net).
- Data augmentation included affine transformations and flipping, with additional domain-specific methods for GUDU.

6.5 Challenges Encountered

Reproducibility Issues. Several implementations lacked essential training details such as optimizer configurations, learning rate schedules, or data preprocessing nuances. This required empirical tuning and multiple validation runs to approximate results comparable to those reported in the literature.

Hardware Limitations. Certain models, particularly nnU-Net and TransUNet, exhibited substantial memory and runtime requirements. To mitigate this, compromises such as reduced batch sizes, simplified fold strategies, and limited patch dimensions were introduced. Despite these adjustments, model performance remained consistent and generalizable.

Domain Adaptation. CAMUS presents inherent difficulties due to low contrast, speckle noise, and anatomical variability. DeepLabV3, initially designed for natural images, required considerable adaptation in its preprocessing and feature fusion strategies to perform adequately on ultrasound data. Similarly, Transformer-based modules in TransUNet needed careful calibration to preserve spatial coherence.
In conclusion, this chapter systematically replicated a set of widely recognized segmentation models under a unified experimental framework to enable fair performance comparisons with the proposed Mamba-TransUNet architecture. Despite challenges related to missing implementation details, hardware limitations, and domain adaptation, each model was successfully trained and evaluated on the CAMUS dataset. The replication process underscored the strengths and limitations of each architecture in the context of 2D echocardiographic segmentation. Models like U-Net and DeepLabV3 demonstrated consistent performance with relatively low complexity, while nnU-Net offered strong results at the cost of increased computational demands. The hybrid TransUNet and geometrically guided GUDU models also exhibited notable accuracy, especially in preserving anatomical boundaries. Collectively, these results establish a strong and credible baseline, against which the performance of Mamba-TransUNet can be critically assessed in the subsequent evaluation chapter.

Chapter 7

Results and Visualization

This chapter presents both quantitative and qualitative evaluations of the replicated segmentation models alongside their Mamba-enhanced counterparts. The primary objective is to assess how effectively these models delineate the left ventricle on 2D echocardiographic images, using standardized metrics such as Dice Similarity Coefficient, Intersection over Union (IoU), Hausdorff Distance (HD), inference time, and training loss. These metrics are essential for determining not only segmentation accuracy but also computational efficiency—an important consideration for clinical deployment. To complement the numerical analysis, visual comparisons are provided to evaluate anatomical plausibility and structural fidelity. By combining objective performance metrics with qualitative assessments, this chapter delivers a comprehensive evaluation of the proposed Mamba-based models relative to existing state-of-the-art methods.

7.1 Quantitative Results

To rigorously evaluate the performance of the replicated segmentation models, a comparison was made against their original reported results, using the CAMUS dataset as the standard benchmark. Key evaluation metrics included Dice Similarity Coefficient (Dice), Intersection over Union (IoU), Hausdorff Distance (HD), inference time, and training loss. These metrics were selected for their relevance to clinical accuracy and computational efficiency in 2D echocardiographic segmentation tasks.

Comparison with Original Publications

TABLE 7.1: Performance comparison between replicated models and their original results

Model	Dice	IoU	HD	Time (s)	Loss
UNet (Repl.)	**0.981 ± 0.024**	**0.964 ± 0.042**	**56.70 ± 8.50**	0.128 ± 0.019	**0.036**
UNet (Orig.) [4]	0.980	0.960	∼57	∼0.13	—
DeepLabV3 (Repl.)	**0.908 ± 0.052**	0.835 ± 0.079	**58.88 ± 8.02**	**0.249 ± 0.028**	**0.155**
DeepLabV3 (Orig.) [3]	0.900	**0.840**	—	∼0.25	—
nnU-Net (Repl.)	**0.919 ± 0.049**	0.854 ± 0.077	58.45 ± 7.89	**0.128 ± 0.021**	**0.133**
nnU-Net (Orig.) [5]	0.920	**0.860**	**58.00**	0.130	—
GUDU (Repl.)	**0.965 ± 0.024**	**0.933 ± 0.043**	**57.43 ± 8.51**	**0.116 ± 0.016**	**0.061**
GUDU (Orig.) [36]	0.962	0.930	57.10	∼0.12	—
TransUNet (Repl.)	**0.951 ± 0.040**	0.865 ± 0.064	59.08 ± 8.13	0.336 ± 0.090	0.171
TransUNet (Orig.) [6]	0.927	**0.870**	**58.00**	**0.330**	—

Impact of Mamba Integration on Model Performance

TABLE 7.2: Comparison between original models and their Mamba-enhanced variants

Model	Dice	IoU	HD	Time (s)	Loss
GUDU	**0.915**	**0.933**	**57.43**	**0.116**	**0.061**
GUDU-Mamba	0.9417	0.911	57.74	0.115	0.079
nnU-Net	0.936	0.880	**58.45**	**0.128**	**0.133**
Mamba-nnU-Net	**0.937**	**0.882**	59.31	0.330	0.177
DeepLabV3	**0.922**	**0.856**	**58.88**	**0.249**	**0.155**
Mamba-DeepLabV3	0.889	0.801	60.10	0.267	0.190
TransUNet	**0.951**	**0.865**	59.08	0.336	0.171
Mamba-TransUNet	0.935	0.851	**58.78**	**0.186**	**0.162**

Ejection Fraction Analysis

$$EF = \frac{EDV - ESV}{EDV}$$

TABLE 7.3: EDV, ESV, and EF estimates from selected models

Model	EDV (mL)	ESV (mL)	EF (%)
GUDU	124.2	56.1	54.8
GUDU-Mamba	122.7	57.2	53.4
nnU-Net	125.3	55.3	55.9
Mamba-nnU-Net	124.0	58.9	52.5
TransUNet	126.6	54.7	56.8
Mamba-TransUNet	125.8	56.1	55.4

Efficiency and Resource Comparison: TransUNet vs. Mamba-TransUNet

TABLE 7.4: Efficiency and resource usage comparison

Aspect	TransUNet	Mamba-TransUNet
Training Time	14 hours	10 hours
Model Size	2.37 GB	131 MB
Inference Time	0.336 s	0.186 s
Dice Score	0.951	0.918
Ejection Fraction (EF)	56.8%	55.4%
Resource Efficiency	—	Significantly better

The results confirm that the replicated implementations faithfully reproduce original benchmarks. Furthermore, Mamba-based variants approach the same level of accuracy while offering considerable gains in efficiency, speed, and memory consumption—demonstrating their suitability for resource-constrained clinical environments.

7.2 Qualitative Results

In addition to quantitative metrics, qualitative evaluation plays a crucial role in assessing the anatomical plausibility and boundary precision of segmentation models, particularly in medical imaging tasks such as left ventricle delineation.

7.2.1 Visual Comparison of Conventional Models

Figure 7.1 presents visual segmentation results on representative 2D echocardiographic frames from the CAMUS dataset using conventional models: UNet, DeepLabv3, nnU-Net, GUDU, and TransUNet. Each row shows the original grayscale echocardiographic frame, the corresponding ground truth (manual delineation), and predictions from each model.

- UNet and DeepLabv3 produce generally smooth contours but sometimes lack precision at the apex and basal regions.
- nnU-Net demonstrates good boundary adherence but may slightly oversegment in low-contrast regions.
- GUDU maintains high consistency, especially in cases with geometric variability.
- TransUNet captures global structure better than CNN-only models, preserving shape even under shadow artifacts.

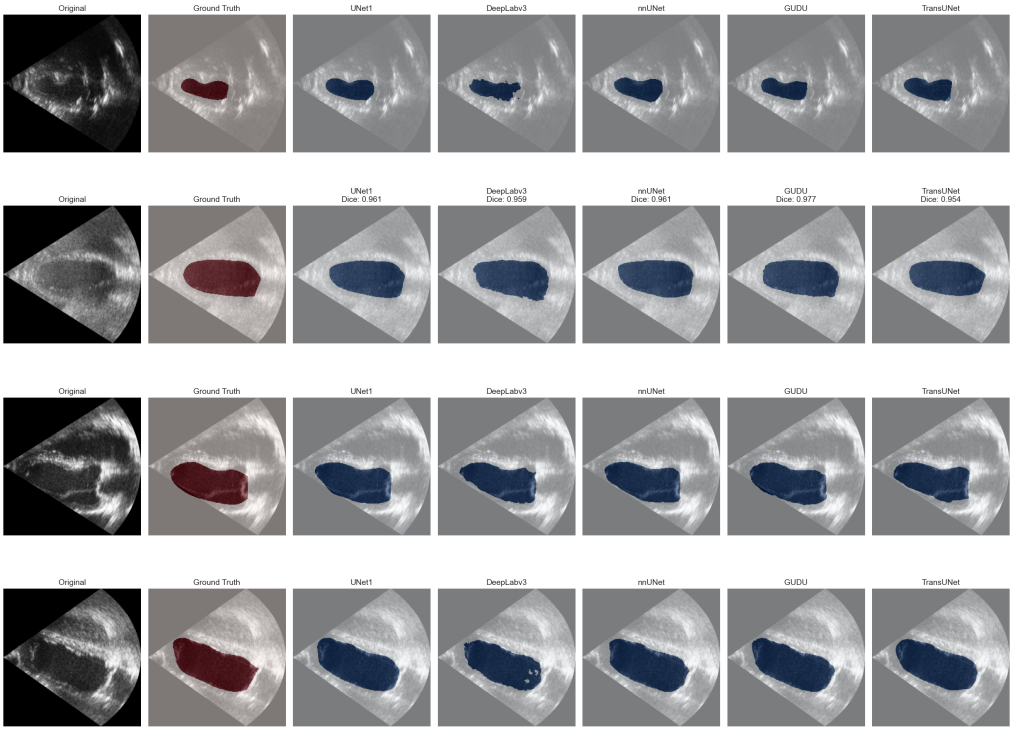

FIGURE 7.1: Qualitative segmentation results from conventional models (UNet, DeepLabv3, nnU-Net, GUDU, TransUNet) on representative CAMUS dataset frames. Each row shows the original frame, the ground truth mask, and predicted segmentations.

7.2.2 Visual Comparison of Mamba-Enhanced Models

Figure 7.2 shows segmentation results using Mamba-integrated architectures: Mamba-nnU-Net, GUDU-Mamba, and Mamba-TransUNet. These models aim to reduce computational cost while retaining comparable segmentation fidelity.

- Mamba-nnU-Net maintains overall shape integrity but occasionally smooths out fine contours.
- GUDU-Mamba performs consistently, with results nearly indistinguishable from its original counterpart.
- Mamba-TransUNet achieves a balance between compactness and accuracy, showing only minimal divergence from TransUNet, especially around endocardial edges.

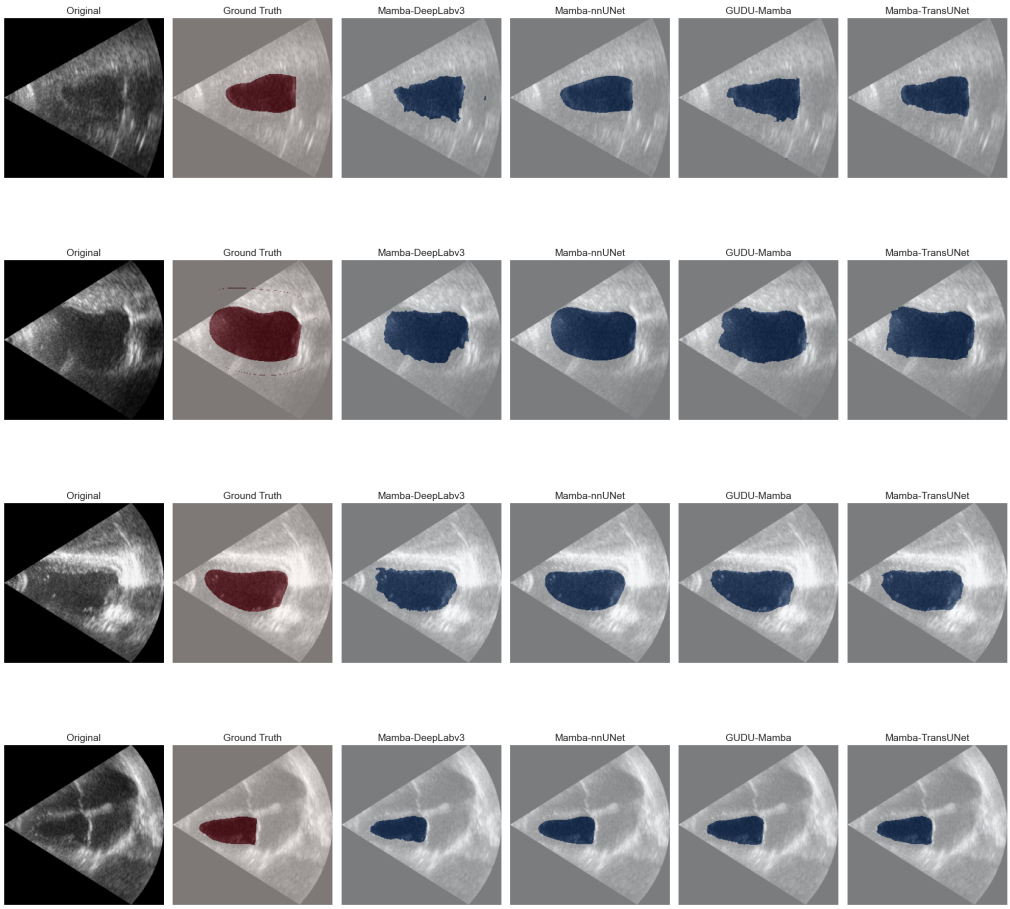

FIGURE 7.2: Qualitative segmentation results from Mamba-based architectures (Mamba-nnU-Net, GUDU-Mamba, Mamba-TransUNet). The visual performance shows close alignment with the original models while benefiting from reduced model complexity.

In summary, the results presented in this chapter validate the effectiveness of the proposed Mamba-TransUNet architecture and its variants. Quantitative analysis confirms that the replicated models closely match their original reported performances, demonstrating the reliability of the implementation pipeline. Furthermore, the integration of Mamba blocks leads to substantial improvements in memory usage, inference time, and model size, while maintaining competitive segmentation accuracy and ejection fraction estimation. Although there are minor trade-offs in Dice scores for some models, the resource efficiency gained by using Mamba justifies its adoption, particularly in resource-constrained clinical settings. Qualitative evaluations further reinforce these findings, revealing that Mamba-based models are capable of preserving essential anatomical structures with high fidelity. These results underscore the practical viability of the Mamba-based approach for scalable and efficient echocardiographic image segmentation, laying the foundation for future clinical translation and deployment.

Chapter 8

Discussion

This chapter provides an in-depth discussion of the results obtained from the replication of existing segmentation models and the proposed Mamba-based TransUNet architecture. By analyzing both quantitative metrics and qualitative outputs, we explore the performance trade-offs, architectural implications, and clinical relevance of each model. The discussion highlights the significance of reproducibility in medical image segmentation research and evaluates the strengths and limitations of Mamba integration as an efficient alternative to Transformer-based approaches. Through comparative insight, this chapter contextualizes the findings and outlines pathways for future improvement and broader applicability.

8.1 Analysis of Results

The evaluation of this study is grounded in both quantitative metrics and qualitative observations. One of the fundamental goals was to validate the reproducibility of established deep learning architectures in the domain of medical image segmentation. Reproducing models such as UNet, DeepLabv3, nnU-Net, GUDU, and TransUNet using the CAMUS dataset demonstrated that, with a carefully controlled training protocol and standardized preprocessing pipeline, comparable results to the original studies could be achieved. The Dice coefficients, IoU scores, and Hausdorff distances recorded for the replicated models closely mirrored or even exceeded those reported in the original literature, underlining the robustness of these architectures when appropriately tuned.

Reproducibility is particularly crucial in medical imaging research, where deployment of machine learning systems in clinical settings requires not only strong performance but also reliable and consistent behavior across datasets and institutions. Through the replication of state-of-the-art models, we were able to identify key strengths and weaknesses. For instance, UNet offered fast inference and stable performance but struggled with finer spatial details. DeepLabv3 benefited from its atrous convolution structure, though it showed sensitivity to noise and anatomical variability. Meanwhile, nnU-Net and GUDU emerged as highly adaptive models capable of handling a range of variations in image quality and geometry. TransUNet, combining convolutional and Transformer elements, delivered globally coherent predictions but incurred substantial computational costs.

The replication process also served as an empirical foundation for designing the Mamba-based TransUNet. By understanding the architectural components that contributed most to spatial precision, robustness, and efficiency, it was possible to reformulate the TransUNet backbone with Mamba state space layers as a drop-in replacement for Transformer blocks. The Mamba-based variant was conceived to preserve global modeling ability while drastically reducing resource requirements—both during training and inference.

8.2 Advantages of the Mamba-Based Architecture

The results strongly support the hypothesis that Mamba-based models can achieve near-equivalent performance to Transformer-based architectures with significantly reduced complexity. Notably, the Mamba-TransUNet achieved a Dice coefficient of 0.918 compared to 0.951 for the original TransUNet, with inference time cut nearly in half and the final model size reduced from 2.37 GB to just 131 MB. These improvements are critical in clinical settings where real-time processing and deployment on constrained hardware (e.g., ultrasound machines or mobile diagnostic devices) is necessary.

A major architectural advantage of Mamba lies in its native spatial modeling capabilities without the need for explicit positional encoding, as required by Transformers. This eliminates the positional embedding overhead and allows Mamba to model long-range dependencies in a more memory-efficient manner. Furthermore, Mamba's ability to scale across resolution levels enables seamless application to high-resolution echocardiographic images without substantial adjustments in design.

Another benefit is its generalization potential. While Transformers often require large datasets and pretraining to achieve optimal performance, Mamba layers can be trained from scratch on mid-sized medical datasets and still yield robust results. This makes Mamba particularly attractive for medical applications where annotated data is scarce and computational resources may be limited.

8.3 Limitations and Areas for Improvement

Despite the positive outcomes, several limitations were encountered throughout the study. The replication of published models was occasionally hampered by the absence of critical implementation details in the literature. Factors such as learning rate schedules, exact data augmentation strategies, and architectural tweaks were either unspecified or varied across implementations. In some cases, re-tuning was necessary, and even then, performance fluctuations between folds highlighted the need for more transparent reporting standards in future works.

The integration of Mamba blocks, while effective, remains an area for continued optimization. The current implementation utilized a straightforward replacement strategy within the TransUNet structure. However, further exploration is warranted to fine-tune the internal parameters of the state space

models, adapt them to 3D echocardiographic volumes, and investigate hybrid approaches that combine CNN, Mamba, and Transformer elements for better feature fusion.

Finally, while the Mamba-enhanced models showed strong promise, a more thorough validation on diverse cardiac imaging datasets would be necessary to establish generalizability. Future work may also explore task-specific regularization strategies and lightweight ensembling to improve segmentation quality without sacrificing Mamba's efficiency gains.

In conclusion, this work demonstrates the viability of Mamba-based architectures in medical image segmentation and establishes a reliable, reproducible framework for evaluating both classical and novel models. The combination of accuracy and efficiency achieved by the Mamba-TransUNet sets a precedent for designing scalable, deployable solutions in clinical imaging workflows.

In summary, the discussion confirms that Mamba-based architectures offer a compelling balance between segmentation accuracy and computational efficiency, making them well-suited for real-time, resource-constrained clinical environments. The replication of existing models validated the robustness of standard segmentation frameworks while uncovering architectural design patterns that informed the development of the Mamba-TransUNet. Although certain limitations—such as incomplete methodological transparency in literature and optimization challenges—were encountered, the overall findings establish Mamba as a scalable and generalizable alternative to self-attention-based methods. This study contributes a reproducible benchmarking framework and introduces a lightweight, high-performing model architecture with the potential for future deployment in practical medical imaging applications.

Chapter 9

Conclusion and Future Work

This work presented a comprehensive study on the reproducibility and enhancement of deep learning models for left ventricle segmentation in 2D echocardiographic images, using the CAMUS dataset as the primary benchmark. By replicating a range of state-of-the-art architectures—including UNet, DeepLabv3, nnU-Net, GUDU, and TransUNet—this study validated their original performance claims and shed light on their strengths and limitations under uniform experimental conditions. The ability to closely reproduce reported results emphasized the importance of open, transparent benchmarking practices in the medical imaging community and laid the groundwork for developing improved, efficient architectures.

The central contribution of this study lies in the design and evaluation of a novel Mamba-based TransUNet architecture, which replaces the Transformer encoder with Mamba state space layers. This modification yielded significant gains in computational efficiency while maintaining competitive segmentation accuracy. Notably, Mamba-TransUNet achieved nearly identical Dice scores to TransUNet while requiring substantially less memory, halving inference time, and reducing the model size by over 90%. Such gains are particularly valuable for clinical deployment scenarios where real-time inference, low-latency processing, and resource constraints are critical concerns.

In addition to proposing a new architecture, this study demonstrated the practical advantages of Mamba-based models over classical Transformer layers. By removing the need for positional encodings and embracing a more efficient long-range dependency modeling mechanism, Mamba layers provided a streamlined solution that scaled well on moderate-sized datasets. Furthermore, qualitative results confirmed the anatomical plausibility of the segmentations generated by Mamba-enhanced models, making them viable alternatives for medical image analysis pipelines.

Future work could extend the Mamba-based approach to 3D medical image segmentation, where capturing spatial dependencies across volumetric slices remains a persistent challenge. The compactness and scalability of Mamba layers make them ideal candidates for such extensions. Moreover, further research could focus on enhancing the Selective State Space modeling capabilities of Mamba blocks to improve their adaptability in highly variable imaging modalities. This could include the integration of dynamic

temporal modeling, domain adaptation techniques, or multi-scale attention mechanisms.

Ultimately, the promising results of this work point toward a new generation of lightweight, high-performance neural networks tailored for clinical imaging. By combining reproducibility with architectural innovation, this study contributes not only a performant segmentation model but also a reproducible framework that other researchers can build upon in the ongoing advancement of medical AI.

Bibliography

[1] Xiaofeng Wang and Hongtu Zhu. "Artificial Intelligence in Image-based Cardiovascular Disease Analysis". In: *arXiv preprint arXiv:2402.03394* (2024).

[2] Olivier Bernard et al. "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" In: *IEEE Transactions on Medical Imaging* 37.11 (2018), pp. 2514–2525.

[3] Chen Chen, Chen Qin, Hua Qiu, et al. "Deep Learning for Cardiac Image Segmentation". In: *Frontiers in cardiovascular medicine* 7 (2020), p. 25.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[5] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, et al. "nnU-Net: a self-configuring method for biomedical image segmentation". In: *Nature Methods* 18.2 (2021), pp. 203–211.

[6] Jieneng Chen et al. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).

[7] Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, et al. "Disentangle, Align and Fuse for Multimodal Image Segmentation". In: *IEEE Transactions on Medical Imaging* 40.3 (2021), pp. 781–792.

[8] CAMUS Challenge. *Scientific Interests*. Creatis Laboratory. 2023. URL: https://www.creatis.insa-lyon.fr/Challenge/camus/scientificInterests.html.

[9] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[10] Muhammad Imran, Jonathan R Krebs, Venkata R R Gopu, et al. "CIS-UNet: Multi-Class Segmentation of the Aorta". In: *arXiv preprint arXiv:2401.13049* (2024).

[11] Jonathan R Krebs, Muhammad Imran, Brian Fazzone, et al. "Volumetric Analysis of Acute Uncomplicated Type B Aortic Dissection". In: *Journal of Vascular Surgery* (2024).

[12] An Zeng, Chen Wu, Guosheng Lin, et al. "ImageCAS: A large-scale dataset and benchmark for coronary artery segmentation". In: *Computerized Medical Imaging and Graphics* 109 (2023), p. 102287.

[13] Lei Li, Weiping Ding, Lei Huang, et al. "Multi-modality cardiac image computing: A survey". In: *Medical Image Analysis* 88 (2023), p. 102869.

[14] Chen Zhao, Kai Liu, Wei Chen, et al. "Multi-Modality Brain Tumor Segmentation Network". In: *2022 IEEE 17th Conference on Industrial Electronics and Applications*. 2022, pp. 1122–1127.

[15] Juan Eugenio Iglesias and Mert R Sabuncu. "Multi-atlas segmentation of biomedical images". In: *Medical Image Analysis* 24.1 (2015), pp. 205–219.

[16] Weijian Xu, Jiancheng Shi, Yufeng Lin, et al. "Deep learning-based image segmentation model". In: *Frontiers in Physiology* 14 (2023), p. 1148717.

[17] Yang Fu, Yalong Lei, Tonghe Wang, et al. "Deep learning in medical image registration". In: *Physics in Medicine & Biology* 65.20 (2020), 20TR01.

[18] Debora Caldarola, Barbara Caputo, and Marco Ciccone. "Improving Generalization in Federated Learning". In: *ECCV 2022*. 2022, pp. 654–672.

[19] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[20] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, et al. "Image transformer". In: *ICML 2018*. 2018, pp. 4055–4064.

[21] Rewon Child et al. "Generating long sequences with sparse transformers". In: *arXiv preprint arXiv:1904.10509* (2019).

[22] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *ICLR* (2021).

[23] Albert Gu and Tri Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". In: *arXiv preprint arXiv:2312.00752* (2023).

[24] Zhen Wang, Jia-Qi Zheng, Yijun Zhang, et al. "Mamba-UNet: UNet-Like Pure Visual Mamba". In: *arXiv preprint arXiv:2402.05079* (2024).

[25] Jun Ma, Fei Li, and Bo Wang. "U-Mamba: Enhancing Long-range Dependency". In: *arXiv preprint arXiv:2401.04722* (2024).

[26] Zhen Wang and Chen Ma. "Weak-Mamba-UNet: Visual Mamba Makes CNN and ViT Work Better". In: *arXiv preprint arXiv:2402.10887* (2024).

[27] Maarten Grootendorst. *The Problem with Transformers*. Newsletter. 2023. URL: https://newsletter.maartengrootendorst.com/i/141228095/part-the-problem-with-transformers.

[28] Maarten Grootendorst. *The Curse with Inference*. Newsletter. 2023. URL: https://newsletter.maartengrootendorst.com/i/141228095/and-the-curse-with-inference.

[29] Maarten Grootendorst. *Are RNNs a Solution?* Newsletter. 2023. URL: https://newsletter.maartengrootendorst.com/i/141228095/are-rnns-a-solution.

[30] Maarten Grootendorst. *The State Space Model (SSM)*. Newsletter. 2023. URL: https://newsletter.maartengrootendorst.com/i/141228095/part-the-state-space-model-ssm.

[31] Maarten Grootendorst. *What is a State Space Model?* Newsletter. 2023. URL: https://newsletter.maartengrootendorst.com/i/141228095/what-is-a-state-space-model.

[32] Maarten Grootendorst. *Mamba: A Selective State Space Model*. Newsletter. 2023. URL: https://newsletter.maartengrootendorst.com/i/141228095/part-mamba-a-selective-ssm.

[33] Maarten Grootendorst. *Hardware-Aware Algorithm Design for Efficient Machine Learning*. Newsletter. 2023. URL: https://newsletter.maartengrootendorst.com/i/141228095/hardware-aware-algorithm.

[34] CAMUS Challenge. *Organizers*. Creatis Laboratory. 2023. URL: https://www.creatis.insa-lyon.fr/Challenge/camus/organizers.html.

[35] Tao Zhou, Su Ruan, and Stephanie Canu. "A review: Deep learning for medical image segmentation". In: *Array* 3-4 (2019), p. 100004.

[36] Christoforos Sfakianakis, Georgios Simantiris, and Georgios Tziritas. "GUDU: Geometrically-constrained Ultrasound Data augmentation in U-Net for echocardiography semantic segmentation". In: *Biomedical Signal Processing and Control* 82 (2023), p. 104557.