TvaraNet: A Lightweight Mamba Neural Network for Real-Time Medical Image Segmentation

Sridhatta Jayaram Aithal[1] , **Vandana Bharti**[1]

[1] School of Artificial Intelligence and Data Engineering, Indian Institute of Technology Ropar, India
sridhatta.2025aiz0030@iitrpr.ac.in, vandana@iitrpr.ac.in

Abstract

Medical image segmentation models based on Vision Mamba architectures have demonstrated strong performance, along with improved efficiency compared to convolutional and transformer-based approaches. However, existing lightweight and ultralight variants often suffer from boundary degradation and inconsistent shape prediction, thereby limiting their clinical reliability. We propose TvaraNet, an extremely lightweight segmentation network designed to preserve boundary fidelity under strict efficiency constraints. TvaraNet contains only 0.037M parameters and requires only 0.060 GFLOPs, enabling efficient deployment in resource-constrained environments. The architecture introduces two core lightweight modules. Parallel Skip Mamba Averaging (PASMA) enhances long-range dependency modeling by injecting globally aggregated channel context into mamba blocks. Dilated Asymmetric Spatial Mixer Attention (DASMA) improves boundary-aware feature refinement in skip connections through efficient multi-scale spatial modulation. In addition, training-time regularization strategies, including an Adaptive Gated Gradient Reversal Layer (AGGRL), an adaptive Singular Value Decomposition (SVD) loss, and a Boundary-weighted Cross-Entropy loss (BWCE), are employed to suppress redundant features and emphasize object boundaries without adding inference overhead. Experiments on ISIC-17, ISIC-18, and a spinal segmentation dataset demonstrate that TvaraNet consistently outperforms existing lightweight models and achieves competitive or superior boundary-aware performance compared to heavier architectures. Notably, it improves Mean Intersection over Union (mIoU) by +1.50, +1.42, and +3.35 over Ultralight VMUNet on the respective datasets, establishing TvaraNet as a practical solution for boundary-consistent medical image segmentation on edge and low-resource platforms. Code is available at: `https://github.com/MindsLab-GitHub/TvaraNet`

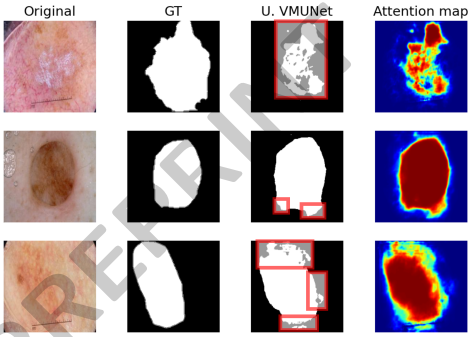

Figure 1: Segmentation mask produced by Ultralight VMUNet on the ISIC-17 dataset, highlighting boundary inconsistencies (red box). The corresponding attention map reveals spurious feature activations near the lesion boundary.

1 Introduction

Medical image segmentation is a fundamental yet challenging task due to anatomical variability, low contrast in tissue images, and the presence of noise commonly found in medical imaging modalities. Encoder-decoder architectures such as UNet [Ronneberger *et al.*, 2015] have established themselves as strong segmentation baselines by learning hierarchical feature representations. These architectures enable accurate delineation of anatomical structures and pathological regions.

Recent advances in backbone networks, including ResNet-based [Nisa and Ismail, 2022] and Transformer-based approaches [Chen *et al.*, 2021], further enhance feature extraction and contextual reasoning capabilities. However, these gains often come at the expense of high computational complexity and large parameter counts, which hinder deployment in real-time resource-constrained clinical environments. The emergence of state-space models has shifted research focus toward lightweight architectures, leading to designs such as VMUNet and its variants [Ruan *et al.*, 2024; Zhang *et al.*, 2024b; Wu *et al.*, 2025; Zhang *et al.*, 2024a]. Although these models reduce computational overhead, they remain insufficiently compact for edge deployment. Ultralight architectures [Wu *et al.*, 2024; Liao *et al.*, 2024] improve efficiency but still exhibit critical limitations, namely imprecise boundary extraction, false positives, and spurious island predictions in the output masks.

1.1 Motivation and Contribution

Existing lightweight medical image segmentation models, such as Ultralight VMUNet [Wu *et al.*, 2024], show poor shape fidelity and irregular feature activations around segmentation boundaries in predicted masks as shown in Fig. 1, which results in incomplete or distorted anatomical structures. In contrast, heavier segmentation networks mitigate this issue by leveraging deeper representations, but this comes with prohibitive computational and memory costs. These trade-offs motivate the following research question: *Can an accurate, shape-preserving segmentation mask be achieved under strict parameter and computational complexity constraints?*

To address this, we propose **TvaraNet**, an extremely lightweight architecture designed for accurate and boundary-consistent medical image segmentation. Our contributions are summarized as follows:

- PASMA enriches feature interactions across channel groups in Parallel Vision Mamba (PVM) [Wu *et al.*, 2024] by injecting globally aggregated channel-aware features prior to the Mamba update.
- Lightweight U-shaped segmentation networks propagate spurious features through skip connections, which lead to distorted shapes and irregular boundary activations. We address this using the DASMA block, which applies lightweight multi-scale spatial modulation to suppress irrelevant features and enhance spatially informative representations in skip pathways.
- Training-time feature suppression remains a challenging problem as spurious features are often reinforced during backpropagation. To address this, we introduce the AGGRL, which performs input-dependent, channel-wise gradient inversion to selectively suppress redundant features while promoting more discriminative representation learning.
- Adaptive SVD loss enforces low-rank regularization on encoder features to suppress noise and redundancy, resulting in more compact and discriminative feature representations.
- Standard region-based losses often underemphasize boundaries, resulting in imprecise contour localization. To address this, BWCE loss is introduced to impose higher penalties for incorrectly predicted boundary pixels during training.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 details the proposed methodology, Section 4 discusses the experimental results, and Section 5 concludes the study.

2 Related Work

Medical image segmentation is a critical field that integrates artificial intelligence into the medical domain [Yao *et al.*, 2024; Lee *et al.*, 2024; Shen *et al.*, 2023; Song *et al.*, 2024; Kim and Yoo, 2025]. This aids clinical studies in various applications, such as dermatology and neuroanatomy. Advancements in this area began with the development of the encoder-decoder based architecture called UNet [Ronneberger *et al.*, 2015], which inspired numerous subsequent variants [Huang *et al.*, 2020; Yan *et al.*, 2022; Zhou *et al.*, 2018; Peng *et al.*, 2025]. This evolution was followed by the insertion of recently developed backbones, which provided stronger feature extraction capabilities, such as Resnet [Nisa and Ismail, 2022], and ResNext [Xie *et al.*, 2020], which primarily rely on Convolutional Neural Network (CNN). Subsequently, research shifted toward transformers, which had the power of self-attention and cross-attention while maintaining the architecture style of UNet. This led to the development of transformer-based backbones such as TransUNet [Chen *et al.*, 2021], Segformer [Xie *et al.*, 2021], Segmenter [Strudel *et al.*, 2021], and Swin-UNet [Cao *et al.*, 2022]. Later, the emergence of state-space models shifted research toward lightweight architectures, leading to the development of models such as VM-UNet [Ruan *et al.*, 2024], VMUNet-v2 [Zhang *et al.*, 2024b], HMTUNet [Zhang *et al.*, 2024a], Swin-UMamba [Liu *et al.*, 2024], HVMUNet [Wu *et al.*, 2025], UNet-mamba [Wang *et al.*, 2024], which performed better than the earlier models while reducing the parameter count. Subsequently, research efforts focused on developing extremely lightweight architectures with minimal parameter counts and computational requirements, such as Ultralight VMUNet [Wu *et al.*, 2024] and LightM-UNet [Liao *et al.*, 2024]. Due to the reduction in parameters, these models often suffer from inconsistent boundary predictions in the segmentation masks of medical images. To address this issue, we propose TvaraNet in the following section.

3 Methodology

In this section, we present the architecture of TvaraNet, an ultra-lightweight segmentation network designed to preserve boundary fidelity under extreme efficiency constraints. TvaraNet follows an encoder–decoder architecture while introducing novel enhancements in the encoder, decoder, skip connections, and loss design to overcome common limitations of lightweight models, including inaccurate boundary localization and redundant feature propagation. We briefly describe each component of TvaraNet in the following subsections. A visualization of the TvaraNet model is shown in Fig. 2.

3.1 Parallel Skip Mamba Averaging

PVM processes channels independently, limiting its capability to capture semantics spread across channels. PASMA introduces globally aggregated features into the channel chunks of PVM. This is achieved by computing the channel-wise average and injecting it into the mamba layer. Let $X_{in} \in \mathbb{R}^{B\times C\times H\times W}$ be the input to PASMA, which is passed through the following operations,

$$Y_i^{C/4} = \mathrm{Sp}(\mathrm{LN}(Conv_{1\times1}(X_{\mathrm{in}}))) \tag{1}$$

In Eq. (1), the input is processed by a 1×1 convolution followed by layer normalization (LN). The normalized feature map is then split along the channel dimension into four equal groups of size $C/4$ using the split operation Sp, producing $Y_i^{C/4}$, where $i \in \{1, 2, 3, 4\}$.

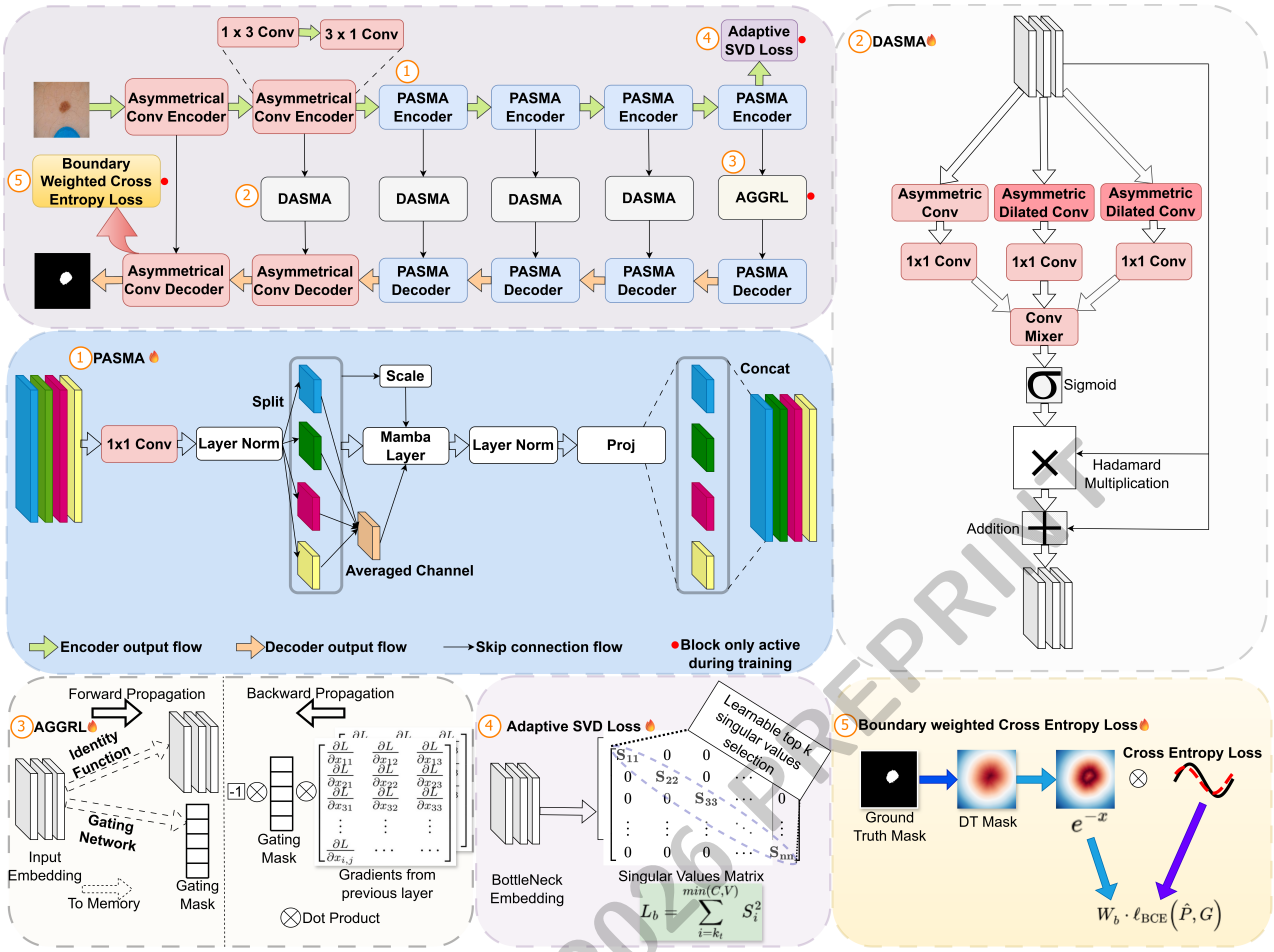

Figure 2: Architecture of TvaraNet. The network integrates asymmetric encoder–decoder pathways with ① PASMA, ② DASMA, and ③ AGGRL modules in the skip connections. During training, ④ the adaptive SVD loss leverages encoder features to enforce low-rank representations, while ⑤ the BWCE loss improves boundary delineation by assigning higher penalties to boundary pixels.

$$Y_{\text{avg}}^{C/4} = \frac{Y_1^{C/4} + Y_2^{C/4} + Y_3^{C/4} + Y_4^{C/4}}{4} \tag{2}$$

In Eq. (2), the average of the channel-split outputs is computed as $Y_{\text{avg}}^{C/4}$.

$$YM_i^{C/4} = Mamba(Y_i^{C/4} + Y_{\text{avg}}^{C/4}) + \theta \cdot Y_i^{C/4} \tag{3}$$

The averaged channel representation is then added to each channel chunk and passed through the Mamba layer, enabling inter-channel features across the split channel groups. To preserve the original feature representation, a skip connection with a scaling parameter θ is incorporated, as shown in Eq. (3). The resulting output is denoted as $YM_i^{C/4}$.

$$Y_{out} = Cat(YM_1^{C/4}, YM_2^{C/4}, YM_3^{C/4}, YM_4^{C/4}) \tag{4}$$

The outputs obtained from Eq. (3) are concatenated along the channel dimension, as shown in Eq. (4), to obtain Y_{out}.

$$X_{out} = Proj(LN(Y_{out})) \tag{5}$$

The concatenated output is then passed through the layer normalization LN to obtain the normalized output, and then it is projected to the required output dimension using the multilayer perceptron projection $(Proj)$ layer in Eq. (5).

3.2 Dilated Asymmetric Spatial Mixer Attention

DASMA enhances the skip connection by introducing scale-enhanced features. This is done by introducing three convolution branches with different dilation rates. Let Z_{in} be the input which belongs to $\mathbb{R}^{B\times C\times H\times W}$.

$$Z_1^1 = Conv_{1\times 3, pad=(0,1)}^{DW}(Z_{in}) \tag{6}$$

$$Z_2^1 = Conv_{3\times 1, pad=(1,0)}^{DW}(Z_1^1) \tag{7}$$

$$Z_{out}^1 = Conv_{1\times 1}^{PW}(Z_2^1) \tag{8}$$

The input Z_{in} is processed by the first convolution branch, which consists of three layers. Specifically, Z_{in} is processed by asymmetric depthwise convolutions with kernel sizes 1×3 and 3×1, each employing a unidimensional padding of 1, as defined in Eqs. (6) and (7). This is followed by a pointwise

convolution (PW) with a 1×1 kernel, which projects the incoming features to a single-channel representation, as shown in Eq. (8).

$$Z_1^2 = Conv_{1\times3,pad=(0,3),dil=(1,3)}^{DW}(Z_{in}) \tag{9}$$

$$Z_2^2 = Conv_{3\times1,pad=(3,0),dil=(3,1)}^{DW}(Z_1^2) \tag{10}$$

$$Z_{out}^2 = Conv_{1\times1}^{PW}(Z_2^2) \tag{11}$$

Similarly, depthwise asymmetric convolutions with dilation rates of $(1, 3)$ and $(3, 1)$ are applied to the input Z_{in}, as defined in Eqs. (9) and (10). The resulting dilated feature maps are then passed through a 1×1 pointwise convolution to produce a single-channel output, as shown in Eq. (11).

$$Z_1^3 = Conv_{1\times3,pad=(0,5),dil=(1,5)}^{DW}(Z_{in}) \tag{12}$$

$$Z_2^3 = Conv_{3\times1,pad=(5,0),dil=(5,1)}^{DW}(Z_1^3) \tag{13}$$

$$Z_{\text{out}}^3 = \text{Conv}_{1\times1}^{\text{PW}}\left(Z_2^3\right) \tag{14}$$

In the third branch of dilated asymmetric convolution, we pass the input Z_{in} through the dilated asymmetric depthwise convolution of kernels 1×3 and 3×1 with a unidimensional padding of 5 and unidimensional dilation of 5 in each case in Eq. (12) and (13). The dilated output is subsequently passed through a 1×1 convolution kernel to obtain the channel-modulated output, which is further modulated to a single-channel output, Z_{out}^3, as expressed in Eq. (14).

$$Z_{cat} = Cat(Z_{\text{out}}^1, Z_{\text{out}}^2, Z_{\text{out}}^3) \tag{15}$$

The outputs of the three convolution branches are concatenated along the channel dimension to form the three-channel feature map Z_{cat}, as defined in Eq. (15).

$$H_{BN}^{mix} = BN(Z_{cat}) \tag{16}$$

The concatenated output Z_{cat} is then passed through the batch normalization layer to obtain the batch-normalized output H_{BN}^{mix} in Eq. (16).

$$H_1^{mix} = Conv_{1\times1}(H_{BN}^{mix}) \tag{17}$$

The batch-normalized output H_{BN}^{mix} is then passed through a 1×1 convolution to increase the number of output channels to 4 times that of H_{BN}^{mix}, resulting in H_1^{mix}, as shown in Eq. (17).

$$H_2^{mix} = Conv_{3\times3,pad=1}^{DW}(H_1^{mix}) \tag{18}$$

The output of the 1×1 convolution layer H_1^{mix} is then passed through a 3×3 depthwise convolution layer to obtain H_2^{mix}, as shown in Eq. (18).

$$H_{Gelu}^{mix} = Gelu(H_2^{mix}) \tag{19}$$

The output H_2^{mix} is then passed through the Gelu activation layer to obtain H_{Gelu}^{mix} in Eq. (19).

$$H_3^{mix} = Conv_{1\times1}(H_{Gelu}^{mix}) \tag{20}$$

The output H_{Gelu}^{mix} is then passed through a 1×1 convolution for channel modulation, resulting in H_3^{mix}, as shown in Eq. (20). This is done to ensure that the extracted attention mask has dimensions similar to those of Z_{in}.

$$M = \sigma(H_3^{mix}) \tag{21}$$

The output H_3^{mix} is then passed through a sigmoid layer to obtain the mask M in Eq. (21) which ranges from 0 to 1.

$$Z_{out} = Z_{in} + Z_{in} \odot M \tag{22}$$

The attention mask M obtained is then multiplied element-wise with input Z_{in} (Hadamard multiplication) in Eq. (22) and then added with input Z_{in}. The output Z_{out} contains scale-enhanced features while suppressing irrelevant activations.

3.3 Adaptive Gated Gradient Reversal Layer

AGGRL contains a channel-wise modulated gradient reversal layer, which is driven by a gating layer in the forward propagation.

$$y = x \tag{23}$$

In the forward propagation, this layer acts as an identity function with no effect on the input as in Eq. (23).

$$x_{\text{flat}} = \frac{1}{H \cdot W} \sum_{h=1}^{H} \sum_{w=1}^{W} x[:, :, h, w] \tag{24}$$

The gating parameter provides channel-wise modulation in backpropagation, which is calculated in the forward propagation. This is done by applying global average pooling on the input x along the spatial dimensions H and W in Eq. (24).

$$g_{raw} = w_2 \cdot \sigma_{relu}(w_1 \cdot x_{flat}) \tag{25}$$

Further, the global averaged output is then passed through a gating layer, which consists of a multi-layer perceptron (MLP) layer followed by a ReLU activation (σ_{relu}) layer, followed by another MLP layer. This gives us the raw gating output g_{raw} in Eq. (25).

$$g = \sigma_{sigmoid}(g_{raw}) \tag{26}$$

The raw gating values g_{raw} are then passed through the sigmoid layer ($\sigma_{sigmoid}$) to obtain the gating mask (g) in Eq. (26). The g contains values between 0 to 1 and acts as a mask for the gradients in the backpropagation during training.

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{\partial \mathcal{L}}{\partial x} \tag{27}$$

$$\frac{\partial \mathcal{L}}{\partial x} = -g \cdot \frac{\partial \mathcal{L}}{\partial y} \tag{28}$$

During forward propagation, this layer behaves as an identity mapping, yielding the gradient relationship between the input and output shown in Eq. (27). During backpropagation, the incoming gradients are multiplied by the negative gating mask, as defined in Eq. (28), enabling channel-wise gradient reversal and selective suppression of redundant feature responses.

3.4 Adaptive Singular Value Decomposition Loss

The adaptive SVD loss leverages the singular values obtained from the bottleneck feature representation. Let F be the feature output of the encoder that belongs to $\mathbb{R}^{B\times C\times H\times W}$.

$$F_b = \text{reshape}(F[b]) \tag{29}$$

Each sample feature $F[b]$, which belongs to $\mathbb{R}^{C\times H\times W}$ in a batch of size B, is reshaped into $\mathbb{R}^{C\times V}$, where $V = H \times W$, as shown in Eq. (29).

$$F_b = U_1 \cdot S \cdot U_2^{\top} \tag{30}$$

In Eq. (30), S contains the singular values arranged in descending order.

$$R_b(k) = \frac{\sum_{i=1}^{k} S_i^2}{\sum_{j=1}^{min(C,V)} S_j^2} \tag{31}$$

The energy ratio $R_b(k)$ is computed using the cumulative squared singular values $\sum_{i=1}^{k} S_i^2$ and the sum of all squared singular values $\sum_{j=1}^{min(C,V)} S_j^2$ as per the Eq. (31).

$$R_b(k_t) \geq \tau \tag{32}$$

The smallest index k_t that satisfies Eq. (32) is then determined from R_b. The threshold parameter τ in Eq. (32) is learnable and constrained to the range $[0.5, 0.99]$ to ensure training stability.

$$L_b = \sum_{i=k_t}^{min(C,V)} S_i^2 \tag{33}$$

$$L_{svd} = \frac{1}{B}\sum_{b=1}^{B} L_b \tag{34}$$

The loss for each sample in Eq. (33) is computed as the cumulative sum of the squared singular values from index k_t, determined using Eq. (32), to $\min(C, V)$ according to the dimensions of S. This formulation penalizes smaller singular values, thereby suppressing redundant feature components and encouraging more compact feature representations. The final adaptive SVD loss is obtained by averaging the sample-wise losses L_b over the batch size B, as shown in Eq. (34).

3.5 Boundary-Weighted Cross-Entropy Loss

The binary ground-truth mask G is defined by $G \in \{0,1\}^{H\times W}$. To emphasize learning near object boundaries, we utilize the signed distance transform [Kervadec *et al.*, 2019]. For each pixel (i, j), the distance transform is defined as

$$DT(i,j) = \begin{cases} \min\limits_{G(m,n)=0} d_{ij,mn}, & \text{if } G(i,j) = 1, \\ -\min\limits_{G(m,n)=1} d_{ij,mn}, & \text{if } G(i,j) = 0 \end{cases} \tag{35}$$

where $d_{ij,mn} = \sqrt{(i-m)^2 + (j-n)^2}$ denotes the Euclidean distance between pixels (i, j) and (m, n). The distance map $\mathrm{DT}(i, j)$ obtained from Eq. (35) is normalized,

$$\widetilde{\mathrm{DT}}(i,j) = \frac{\mathrm{DT}(i,j)}{\max_{m,n} |\mathrm{DT}(m,n)| + \epsilon} \tag{36}$$

In Eq. (36), ϵ is added in the denominator for stability purposes, and its value is set to 10^{-6}. A boundary-aware weight map is defined as

$$W_b(i,j) = \exp\Big(-\Big|\widetilde{\mathrm{DT}}(i,j)\Big|\Big) \tag{37}$$

which assigns higher weights to pixels closer to the object boundary. The boundary-weighted binary cross-entropy loss L_{BWCE} is then formulated as

$$L_{\mathrm{BCE}}\Big(\hat{P}, G\Big) = -\,G \log \hat{P} - (1-G)\log(1-\hat{P}) \tag{38}$$

$$L_{\mathrm{BWCE}} = \frac{1}{B}\sum_{k=1}^{B} \frac{1}{M\cdot N} \sum_{i=1}^{M}\sum_{j=1}^{N} W_b^{(k)}\, L_{\mathrm{BCE}}\Big(\hat{P}^{(k)}, G^{(k)}\Big) \tag{39}$$

Eq. (38) gives pixel-wise binary cross-entropy loss L_{BCE}. In Eq. (39) B denotes the batch size, M and N denote the height and width of the image. Here, $W_b^{(k)}$ denotes the boundary-aware weight map for the k-th sample, while $\hat{P}^{(k)}$ and $G^{(k)}$ represent the predicted probability map and ground-truth mask, respectively. The weighting and loss computations are applied pixel-wise over spatial locations (i, j).

3.6 Final Objective Function

The final objective function is given by L. In this, we combine the weighted addition of losses from binary cross-entropy, dice loss [Milletari *et al.*, 2016], proposed adaptive SVD loss, and BWCE loss.

$$L_{\text{Dice}} = 1 - \frac{2|G \cap \hat{P}|}{|G| + |\hat{P}|} \tag{40}$$

$$L = L_{\text{BCE}} + L_{\text{Dice}} + \lambda_\theta \cdot L_{\text{svd}} + L_{\text{BWCE}} \tag{41}$$

The dice loss is given in Eq. (40). While the trainable weight λ_θ for adaptive SVD loss is clamped from 0 to 0.01. The final objective function for our proposed TvaraNet architecture is given by Eq. (41). BCE loss regulates global region-level probabilistic calibration. BWCE loss introduces spatially adaptive weighting to emphasize boundary pixels. Their combination enables joint optimization of region consistency and boundary precision. The theoretical convergence of the final objective is given in the supplementary material.

4 Experimental Results

4.1 Experimental Setup

The AdamW optimizer with a learning rate of 0.0001, a batch size of 32, and a weight decay of 0.00001 is used in all experiments. We used the CosineAnnealingLR approach to schedule the learning rate. The model was trained on an NVIDIA A100 GPU for a total of 400 epochs. During both training and inference, all input images were resized to $256\times256\times3$. During training, we applied augmentations including horizontal flip, vertical flip, and random rotation. Experimental results are reported using the metrics mIoU, Dice score (DSC), boundary intersection over union (B-IoU), and Hausdorff distance (HD95) [Taha and Hanbury, 2015; Crum *et al.*, 2006; Ribera *et al.*, 2018].

4.2 Dataset Description

ISIC 2017 [Codella *et al.*, 2018]: The ISIC 2017 dataset is a skin lesion segmentation benchmark comprising 1,500 training images and 650 validation images, each accompanied by an expert-annotated binary segmentation mask.

Type	Model	Params↓	FLOPs↓	ISIC-17				ISIC-18				Spinal segmentation			
				mIoU↑	DSC↑	B-IoU↑	HD95↓	mIoU↑	DSC↑	B-IoU↑	HD95↓	mIoU↑	DSC↑	B-IoU↑	HD95↓
Heavy	UNet	31.037	54.737	73.20	84.53	54.94	19.36	75.53	86.06	47.99	20.07	87.63	93.41	61.08	2.74
	UNet-V2	25.145	5.399	74.39	85.31	55.87	15.51	77.29	87.19	49.48	17.63	86.63	92.66	60.78	2.14
	VM-UNet	27.427	4.111	76.24	86.52	48.73	13.10	79.08	88.32	49.96	13.93	81.39	89.74	57.88	3.67
	HVMUNet	8.972	0.741	75.09	85.77	55.21	15.33	77.29	87.19	49.55	16.69	81.19	89.62	59.17	4.00
Lightweight	LightM-UNet	0.380	0.656	75.45	86.01	56.42	15.27	76.83	86.90	49.46	19.00	81.99	90.11	56.46	4.00
	U. VMUNet	0.049	0.064	76.26	86.53	55.41	16.28	77.89	87.57	49.65	16.42	80.35	89.11	59.57	3.74
	TvaraNet (Ours)	**0.037**	**0.060**	**77.76**	**87.49**	**57.56**	**14.65**	**79.31**	**88.46**	**51.85**	**16.30**	**83.70**	**91.13**	**61.32**	**2.25**

Table 1: Comparative results of heavy and lightweight models on three datasets. Bold values indicate the best performance among lightweight models, while underlined values indicate the best performance among heavy models for each evaluation metric. Model size is reported in millions of parameters (M), and computational complexity is reported in GFLOPs.

Type	Model	mIoU↑	DSC↑	B-IoU↑	HD95↓
Heavy	UNet	80.76	89.36	49.08	15.92
	UNet-V2	82.62	90.48	47.20	15.67
	VM-UNet	84.87	91.81	47.85	10.82
	HVMUNet	82.27	90.28	49.59	14.41
Lightweight	LightM-UNet	82.13	90.19	48.60	17.62
	U. VMUNet	82.68	90.52	46.74	16.10
	TvaraNet (Ours)	**85.74**	**92.32**	**53.13**	**14.65**

Table 2: Experimental results of segmentation models trained on the ISIC 2018 dataset and evaluated on the PH^2 dataset.

ISIC 2018 [Codella *et al.*, 2019]: The ISIC 2018 dataset comprises 2,694 dermoscopic images with corresponding ground-truth lesion masks. The dataset is divided into 1,886 training images and 808 validation images.

Spinal segmentation [Chu *et al.*, 2015]: The spinal segmentation dataset comprises 897 spine images and corresponding annotated masks of the lower spine region. The dataset is divided into 819 training images and 78 validation images.

$\mathbf{PH^2}$ [Mendonça *et al.*, 2015]: The PH^2 dataset is a dermoscopic image dataset consisting of 200 high-resolution images, each accompanied by a pixel-level lesion segmentation mask. The dataset is used for cross-dataset evaluation of the proposed model.

Models	Latency (ms)↓	FPS ↑
LightM-UNet	23.02	43.43
U. VMUNet	20.56	48.61
TvaraNet (Ours)	**18.45**	**54.17**

Table 3: Comparison of segmentation models in terms of latency and inference speed on the ISIC-18 dataset using an Nvidia-A100 GPU. Latency is reported in milliseconds (ms), and inference speed in frames per second (FPS).

4.3 Results and Discussion

Quantitative Analysis

In Table 1, TvaraNet is compared with two categories of models. The first category comprises heavy models (parameters $\geq$ 1 million), including UNet [Ronneberger *et al.*, 2015], UNet-V2 [Peng *et al.*, 2025], VM-UNet [Ruan *et al.*, 2024], and HVMUNet [Wu *et al.*, 2025]. The second category comprises recent lightweight baseline models (parameters $\leq$ 1 million), including LightM-UNet [Liao *et al.*, 2024] and Ultralight VMUNet (U. VMUNet) [Wu *et al.*, 2024]. On ISIC-17, TvaraNet outperforms both lightweight and heavy models, as presented in Table 1. Specifically, it achieves superior performance in mIoU (77.76 vs. 76.24), Dice score (87.49 vs. 86.52), and B-IoU (57.56 vs. 55.87). A similar trend is observed on ISIC-18, where TvaraNet outperforms both lightweight and heavy models in terms of mIoU, Dice score, and B-IoU (51.85 vs. 49.96).

Despite being the lightest model at only 0.060 GFLOPs, TvaraNet consistently outperforms both lightweight and heavier segmentation architectures. This indicates that its gains stem from effective architectural design rather than model capacity. The consistent improvements across region and boundary-based metrics indicate robust representation learning and improved contour preservation. TvaraNet attains superior performance among lightweight models on the spinal segmentation dataset across all evaluation metrics. Nonetheless, it does not surpass heavier models in region-based measures such as mIoU and Dice score, presumably due to the presence of multiple anatomical structures within a single spinal image, which may require greater model capacity. Despite its compact design of only 0.037 million parameters, TvaraNet surpasses heavier models in boundary-sensitive evaluation. It attains a superior B-IoU (61.32 vs. 61.08), indicating better boundary delineation. The comparison of latency and inference speed between TvaraNet and other lightweight segmentation models is reported in Table 3. Additional comparison results with other lightweight models are provided in the supplementary material.

Cross-Dataset Evaluation

The model trained on the ISIC-18 dataset is evaluated on the PH^2 dataset as part of cross-dataset benchmarking, as presented in Table 2. We observe that TvaraNet consistently outperforms not only lightweight models but also heavier models by a considerable margin. It achieves superior performance in mIoU (85.74 vs. 84.87), Dice score (92.32 vs. 91.81), and B-IoU (53.13 vs. 49.59). These results validate the model's generalization capability beyond the dataset on which it was trained.

Qualitative Analysis

Fig. 3, Fig. 4, and Fig. 5 present qualitative comparisons of the segmentation masks predicted by TvaraNet and the corresponding baseline models on the ISIC-17, ISIC-18, and Spinal segmentation datasets. On both ISIC-17 and ISIC-18, TvaraNet performs better than the baseline models, LightM-UNet and Ultralight VMUNet. This is evidenced by more accurately delineated boundaries and fewer spurious activations in the attention maps near object boundaries, indicating the suppression of irregular feature activations. Similarly,

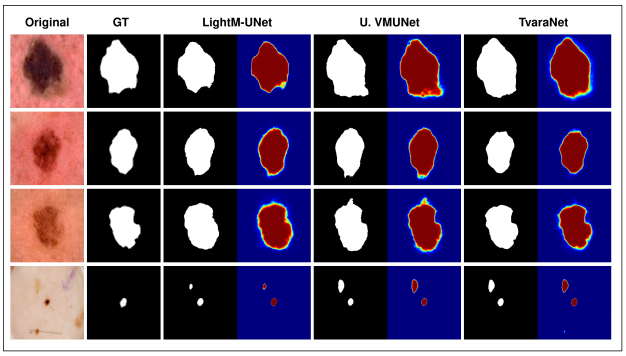

Figure 3: Comparison of prediction results of LightM-UNet, Ultralight VMUNet, and TvaraNet on the ISIC-17 dataset. TvaraNet achieves better lesion shape prediction than Ultralight VMUNet, although false-positive predictions remain in challenging cases, as highlighted in row 4.

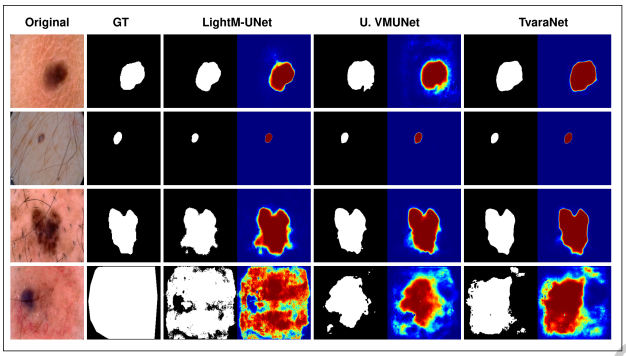

Figure 4: Visualization comparison of prediction results from LightM-UNet, Ultralight VMUNet, and TvaraNet on the ISIC-18 dataset. TvaraNet achieves better lesion shape preservation and suppresses false predictions more effectively than Ultralight VMUNet. However, TvaraNet struggles in cases where the lesion occupies nearly the entire image frame, as shown in row 4.

in Fig. 5, TvaraNet demonstrates improved shape prediction along with reduced false-positive predictions in the presence of multiple anatomical structures. This highlights the model's effectiveness in handling multiple objects of interest within a single image.

Challenges in Prediction

TvaraNet faces challenges in predicting object masks when the lesion occupies a large portion of the image frame. This limitation primarily arises from the limited number of such samples in the training dataset. Another challenge is the generation of false-positive spurious regions in the predictions. This issue occurs because certain non-target regions exhibit visual characteristics similar to those of the medical object of interest, leading to incorrect predictions.

Ablation Study

In Table 4, the ablation study of TvaraNet trained and evaluated on ISIC-18 is presented. The experiment begins with the

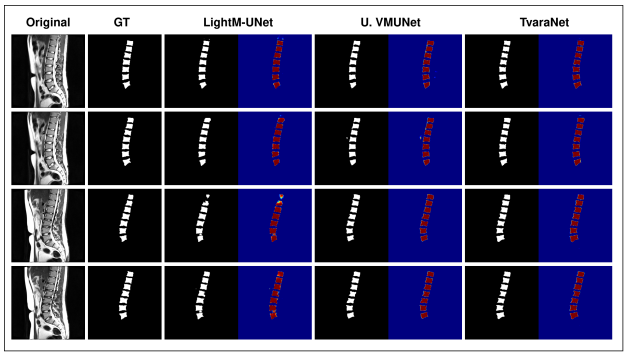

Figure 5: Predicted segmentation masks on the spinal segmentation dataset generated by Ultralight VMUNet and TvaraNet. TvaraNet achieves better shape prediction and boundary consistency than the baseline model.

Configuration	mIoU↑	DSC↑	B-IoU↑	HD95↓
PASMA	78.24	87.79	**53.42**	16.70
PASMA + DASMA	78.42	87.91	51.42	16.91
PASMA + DASMA + AGGRL	78.77	88.12	51.56	16.69
PASMA + DASMA + AGGRL + SVD	78.83	88.16	52.63	16.82
PASMA + DASMA + AGGRL + SVD + BWCE	**79.31**	**88.46**	51.85	**16.30**

Table 4: Ablation study of TvaraNet. Results obtained through the stage-wise addition of each component.

stage-wise incorporation of each component. Starting with only the PASMA module, the baseline performance improves mIoU (78.24 to 78.42) and DSC (87.79 to 87.91) with the addition of the DASMA module, which enhances the skip connections. However, performance gains in boundary-aware metrics are limited at this stage, suggesting that DASMA primarily enhances region-level representations. Furthermore, incorporating AGGRL improves performance by suppressing irrelevant channel-wise features during backpropagation. The addition of the adaptive SVD loss further improves mIoU (78.77 to 78.83), DSC (88.12 to 88.16), and B-IoU (51.56 to 52.63) by reducing the influence of non-dominant and irrelevant features in the bottleneck representation. Finally, the introduction of the BWCE loss achieves the best overall segmentation performance, yielding the highest mIoU (79.31) and Dice score (88.46) while reducing HD95 to 16.30. These results demonstrate that the proposed components contribute complementary benefits toward improving segmentation accuracy and boundary localization.

5 Conclusion

We present TvaraNet, an ultra-lightweight medical image segmentation network with 0.037M parameters and 0.060 GFLOPs that achieves boundary-consistent predictions. By integrating PASMA, DASMA, AGGRL, adaptive SVD loss, and BWCE loss, TvaraNet effectively mitigates boundary distortion, feature redundancy, and shape inconsistency. Extensive experiments and ablation studies demonstrate its effectiveness across multiple datasets while maintaining suitability for resource-constrained deployment. Future work will focus on reducing false-positive predictions and improving model interpretability.

References

[Cao *et al.*, 2022] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Chu *et al.*, 2015] Chengwen Chu, Daniel L. Belavy, Gabriele Armbrecht, Martin Bansmann, Dieter Felsenberg, and Guoyan Zheng. Annotated t2-weighted mr images of the lower spine, July 2015.

[Codella *et al.*, 2018] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[Codella *et al.*, 2019] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[Crum *et al.*, 2006] William R Crum, Oscar Camara, and Derek LG Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, 25(11):1451–1461, 2006.

[Huang *et al.*, 2020] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. Ieee, 2020.

[Kervadec *et al.*, 2019] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International conference on medical imaging with deep learning*, pages 285–296. PMLR, 2019.

[Kim and Yoo, 2025] Min Hyuk Kim and Seok Bong Yoo. Data poisoning attack defense and evolutionary domain adaptation for federated medical image segmentation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1305–1313, 2025.

[Lee *et al.*, 2024] Ho Hin Lee, Yu Gu, Theodore Zhao, Yanbo Xu, Jianwei Yang, Naoto Usuyama, Cliff Wong, Mu Wei, Bennett A Landman, Yuankai Huo, et al. Foundation models for biomedical image segmentation: A survey. *arXiv preprint arXiv:2401.07654*, 2024.

[Liao *et al.*, 2024] Weibin Liao, Yinghao Zhu, Xinyuan Wang, Chengwei Pan, Yasha Wang, and Liantao Ma. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint arXiv:2403.05246*, 2024.

[Liu *et al.*, 2024] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, pages 615–625. Springer, 2024.

[Mendonça *et al.*, 2015] Teresa Mendonça, M Celebi, T Mendonca, and J Marques. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy image analysis*, 2, 2015.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[Nisa and Ismail, 2022] Syed Qamrun Nisa and Amelia Ritahani Ismail. Dual u-net with resnet encoder for segmentation of medical images. *International Journal of Advanced Computer Science and Applications*, 13(12), 2022.

[Peng *et al.*, 2025] Yaopeng Peng, Danny Z Chen, and Milan Sonka. U-net v2: Rethinking the skip connections of u-net for medical image segmentation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025.

[Ribera *et al.*, 2018] Javier Ribera, David Güera, Yuhao Chen, and Edward Delp. Weighted hausdorff distance: A loss function for object localization. *arXiv preprint arXiv:1806.07564*, 2(1):496, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Ruan *et al.*, 2024] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[Shen *et al.*, 2023] Zhiqiang Shen, Peng Cao, Hua Yang, Xiaoli Liu, Jinzhu Yang, and Osmar R Zaiane. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4199–4207, 2023.

[Song *et al.*, 2024] Zhengxuan Song, Xun Liu, Wenhao Zhang, Yongyi Gong, Tianyong Hao, and Kun Zeng.

Spgnet: A shape-prior guided network for medical image segmentation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1263–1271, 2024.

[Strudel *et al.*, 2021] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[Taha and Hanbury, 2015] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):29, 2015.

[Wang *et al.*, 2024] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024.

[Wu *et al.*, 2024] Renkai Wu, Yinghao Liu, Guochen Ning, Pengchen Liang, and Qing Chang. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *Patterns*, 2024.

[Wu *et al.*, 2025] Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. H-vmunet: High-order vision mamba unet for medical image segmentation. *Neurocomputing*, 624:129447, 2025.

[Xie *et al.*, 2020] Xiuzhen Xie, Lei Li, Sheng Lian, Shaohao Chen, and Zhiming Luo. Seru: A cascaded se-resnext u-net for kidney and tumor segmentation. *Concurrency and Computation: Practice and Experience*, 32(14):e5738, 2020.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[Yan *et al.*, 2022] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3971–3981, 2022.

[Yao *et al.*, 2024] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From cnn to transformer: A review of medical image segmentation models. *Journal of Imaging Informatics in Medicine*, 37(4):1529–1547, 2024.

[Zhang *et al.*, 2024a] Mingya Zhang, Zhihao Chen, Yiyuan Ge, and Xianping Tao. Hmt-unet: a hybird mamba-transformer vision unet for medical image segmentation. *arXiv preprint arXiv:2408.11289*, 2024.

[Zhang *et al.*, 2024b] Mingya Zhang, Yue Yu, Sun Jin, Limei Gu, Tingsheng Ling, and Xianping Tao. Vm-unet-v2: rethinking vision mamba unet for medical image segmentation. In *International symposium on bioinformatics research and applications*, pages 335–346. Springer, 2024.

[Zhou *et al.*, 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018.