Data Augmentation-Driven Segmentation of Ovarian Tumor Ultrasound Images using Vision Mamba

Thanh-Phuc Dao†, Huyen-Trang To†, Hoang-Son Bui†, Thi-Lan Le§‡.
‡ SigM lab, Hanoi University of Science and Technology, Hanoi, Vietnam
§ School of Electrical and Electronic Engineering (SEEE),
Hanoi University of Science and Technology, Hanoi, VietNam
Corresponding author e-mail: lan.lethi1@hust.edu.vn

***Abstract*—In this study, we investigate the impact of ultrasound-specific augmentation techniques on the segmentation of ovarian tumors in ultrasound images. These augmentations are designed to simulate common imaging artifacts such as shadowing, speckle noise, and contrast degradation, thereby enhancing model robustness and generalization. We integrate these augmentations into a training pipeline based on Vision Mamba, a state-space model known for its ability to model long-range dependencies with high computational efficiency. Experiments on the OTU2D dataset show notable improvements in segmentation performance. Compared to CNN-based baselines such as U-Net, our approach achieves higher Dice and Jaccard scores while maintaining a lower computational cost, making it suitable for real-time clinical deployment.**

***Index Terms*—Ovarian Tumor, Segmentation, Ultrasound Images, Data Augmentations, Vision Mamba.**

I. Introduction

Ovarian disorders - Including polycystic ovary syndrome (PCOS) and ovarian cancer remain significant health concerns for woman around the world, contributing to both reduced quality of life and increased mortality. Early detection not only aids in identifying the type of tumor but also enables healthcare providers to develop targeted treatment plans, reducing complications and improving reproductive patient survival rates. Accurate ovarian segmentation is a critical first step for diagnosis tasks like size measurement and pathology classification [1]. Automated segmentation reduces clinical workload and diagnostic variability.

In recent years, deep learning has transformed medical image analysis, becoming the gold standard for tasks like image segmentation. The evolution of medical image segmentation models began with the landmark U-Net architecture [2], which featured a symmetric structure with skip connections that effectively preserved both high-level semantic context and low-level spatial details.

More recently, the architectural paradigm in medical image segmentation has shifted towards models that capture global context more effectively. Inspired by their success in natural language processing, Transformer-based models like the Vision Transformer [3] have been adapted for vision tasks. With self-attention mechanism allows them to model long-range dependencies across an entire image, a capability particularly beneficial for segmenting large, complex structures such as tumors. Additionally, the Segment Anything Model [4], [5] has introduced zero-shot segmentation, allowing models to segment various objects without task-specific fine-tuning, further advancing the flexibility and adaptability of segmentation models. However, the quadratic computational complexity of self-attention presents a significant challenge for high-resolution medical images. To overcome this, newer approaches like State Space Models (SSMs) have emerged. Architectures like VM-UNet [6] utilize a Visual State Space (VSS) block to maintain the global receptive field of Transformers but with linear complexity. By combining this efficient long-range modeling with the proven U-Net structure, VM-UNet strikes a balance between performance and computational efficiency for medical image segmentation.

Despite advances in deep learning, model performance heavily depends on the quality and diversity of training data. These datasets are often small, imbalanced, and affected by artifacts. Standard augmentations help by increasing data variability and reducing overfitting, but they may fall short in capturing the complex patterns of tasks like ovarian tumor segmentation, where tumors vary widely in size, shape, and texture.

To improve robustness and relevance, we explore the integration of Vision Mamba [6] with ultrasound-specific augmentation methods [7], which include transformations that simulate shadowing, speckle noise, and haze effects. These augmentations are designed to emulate the physical and signal characteristics of ultrasound devices, making the training data more representative of real-world clinical settings. By applying these strategies on the MMOTU dataset [8], our goal is to create a more realistic training set, enhancing both segmentation performance and model generalizability for clinical application.

II. Related works

CNN-based architectures have been foundational in medical image segmentation, with U-Net [2] remaining one of the most widely used models due to its encoder-decoder structure and skip connections. Variants such as Attention U-Net [9] incorporate attention gates to highlight relevant regions, improving segmentation of small or ambiguous structures. In the context of ovarian tumor segmentation, SovaSegNet [10] introduces a hybrid loss combine of Focal, SSIM and IoU Loss and also using SPPF module to enhance multi-scale feature learning. Similarly, Bui et al. [11] combine ResNet50

979-8-3315-7206-8/25/$31.00

with U-Net3+ and apply marker-removal pre-processing to improve boundary clarity help the model focus on global information instead of the markers. Despite these advances, CNNs are inherently limited by their local receptive fields. CNNs often fail to capture long-range contextual cues and fine-grained boundaries, prompting interest in alternative architectures. Transformers, with their self-attention mechanism, have

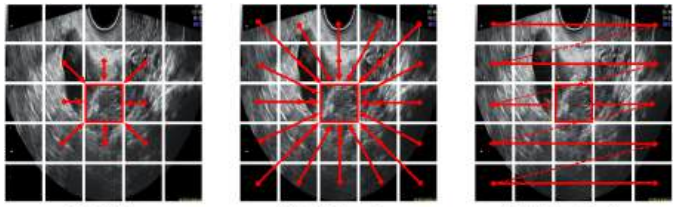

Figure 1: Comparison of long-range dependency modeling and computational complexity across CNNs, Self-Attention, and Vision Mamba.

recently shown promising results in medical image segmentation by modeling global dependencies. The Segment Anything Model (SAM) [4] enables prompt-based segmentation across domains. The authors in [5] adapt SAM for ovarian tumor segmentation by incorporating both automatic and manual box prompts, demonstrating improved accuracy in ultrasound settings. Their automatic prompting pipeline uses a two-stage detection approach based on YOLOv5, where candidate tumor regions are first localized and then refined to serve as effective prompts for SAM. However, the quadratic complexity of self-attention makes standard Transformers computationally intensive, particularly for high-resolution images. This has led to interest in more efficient alternatives capable of preserving global context.

State Space Models (SSMs) have emerged as an efficient alternative to self-attention for sequence modeling. Mamba [12], a recent SSM-based architecture, leverages linear recurrence to model long-range dependencies with reduced memory and computational cost. Vision Mamba [13] extends this idea to vision tasks, introducing a selective state-space mechanism that avoids the quadratic overhead of attention while preserving contextual awareness. It achieves competitive performance on natural image benchmarks with linear scalability. VM-UNet [14] presents a promising direction by balancing accuracy and efficiency in medical image segmentation.

However, while SAM and Mamba perform well in general vision tasks, their direct use in ultrasound segmentation is limited by domain-specific artifacts and small datasets. In this work, we investigate its integration into the segmentation of ovarian tumors in ultrasound images.

III. Proposed Methods

A. Preliminaries

In modern SSM-based models, i.e., Structured State Space Sequence Models (S4) [12] and Mamba, both rely on a continuous system that maps a one-dimensional input function or sequence, denoted as $x \in R^{1\times D}$, through latent space $h \in R^{1\times N}$, $(N > D)$ to compute and then, mapping h back to output $y \in R^{1\times D}$. SSM can be represented as a a linear Ordinary Differential Equation (ODE):

$$h'(t) = Ah(t) + Bx(t), \tag{1a}$$
$$y(t) = Ch(t) + Dx(t), \tag{1b}$$

where $A \in R^{N\times N}$, $B \in R^{N\times 1}$, $C \in R^{1\times N}$, $D \in R^{N\times 1}$ are parameter matrices learned by the model. Dx(t) serves a similar role to skip-connections in deep learning models and can be temporarily omitted to simplify the formulation of subsequent equations. However, this equation still cannot be directly applied to deep learning models because it is in a function-to-function (i.e., continuous) form. We need to discretize the equation using a step size (also known as timescale) and allow the model to learn this value. Typically, the zero-order hold (ZOH) is employed as the discretization rule and can be defined as follows:

$$\bar{A} = \exp(A\Delta), \quad \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - \mathbf{I}) \cdot \Delta B$$

where $\bar{A}$ and $\bar{B}$ are discrete parameters of A and B respectively. After discretization, SSM-basedmodels can be computed in two ways: linear recurrence or global convolution, defined as equations 3a, 3b, 4a and 4b, respectively.

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \tag{3a}$$
$$y_t = Ch_t + Dx_t, \tag{3b}$$
$$\mathbf{K} = (C\bar{B}, C\bar{A}\bar{B}, \ldots, C\bar{A}^{L-1}\bar{B}) \tag{4a}$$
$$y = \mathbf{K} * x \tag{4b}$$

where $\mathbf{K}$ represents a structured convolutional kernel, and L denotes the length of the input sequence x.

Our proposed method relies on VM-UNet [14]. However, we integrates a new ultrasound-specific augmentation step.

B. Segmentation Framework based-on Vision Mamba

1) Vision Mamba Unet (VM-UNet): Inheriting from the architecture of Unet, while keeping the encoder-decoder structure, VM-UNet represents for the replacement of the traditional convolution layer with VSS (Visual State Space) block. With the root of Vision Transformer, VM-UNet comprises a Patch Embedding layer, Final Linear Projection. The process begins with an input image of size $H \times W \times 3$, where H and W are the height and width, and 3 represents the RGB color channels.

After the initial layer - Patch Embedding, the input image is converted into a sequence of non-overlapping patches of size 4x4. This layer transforms the image resolution to $\frac{H}{4} \times \frac{W}{4}$ and increases the channel dimension to C. Learning from progressive papers, the default value of C is 96 which is greater than popular vision models using convolution block as a main contribution. The core of the encoder consists of four stages, particularly, the first three stages contain a VSS block followed by a Patch Merging layer. While VSS block plays an important

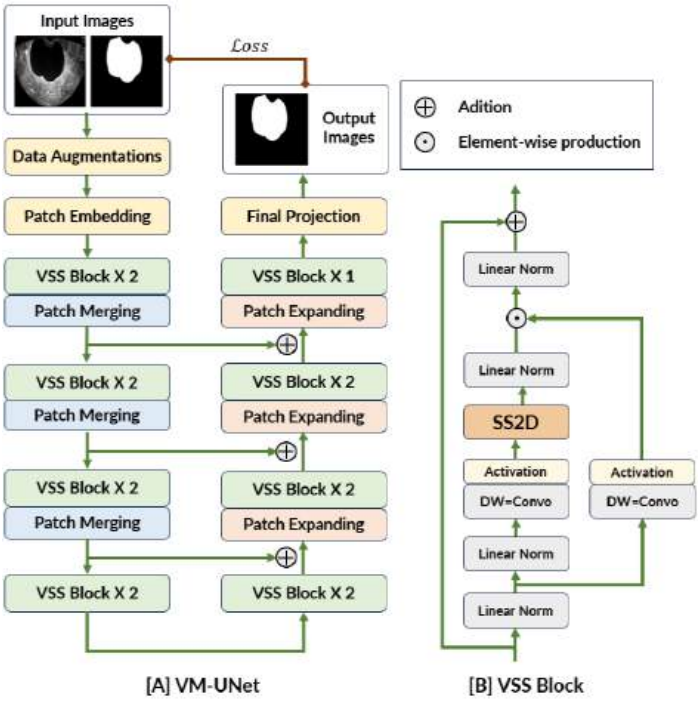

Figure 2: Overall architecture of VM-UNet

role in learning complex representations, Patch Merging layer downsamples the input feature map and increasing the number of channels. Specificially, there are [2, 2, 2, 2] VSS blocks across four stages with the channel counts for [C, 2C, 4C, 8C] respectively. We employed the VM-UNet architecture for the ovarian tumor segmentation task, following the paradigm of using only a single VSS block at the final stage.

The decoder mirrors the encoder in reverse, using Patch Expanding layers followed by VSS blocks at each stage to upsample features—doubling spatial resolution and halving channels. The final stage applies a single VSS block to restore the original channel dimension C, matching the encoder's first stage. To assess the VSS block's effectiveness, skip connections directly concatenate encoder features with their corresponding decoder outputs, without additional processing.

2) VSS Block: The VSS block (Fig. 2) is the core unit of VM-UNet [14], designed as a residual block with a dual-path structure. The first path applies a linear layer and SiLU activation. The second includes a linear layer, depthwise separable convolution, SiLU activation, and the SS2D module for capturing long-range spatial dependencies. Outputs from both paths are normalized and fused via a gating mechanism, then passed through a final linear layer. This output is added back to the original input via a residual connection, helping prevent vanishing gradients and enabling the network to learn more robust features. SiLU is used as the default activation throughout.

The SS2D (Fig. 3) comprises three components: a scan expanding operation, stacked S6 blocks from Mamba [12], and a scan merging operation. It starts by reordering patches across multiple directions (horizontal, vertical, diagonal) to enhance contextual diversity. These sequences are then processed by S6 blocks, which use state space models with learnable parameters to generate rich representations. The output is then reassembled into the original spatial layout via the scan merging step. This design enables SS2D to capture both local and global features. However, unlike convolution, the scanning operates with a stride equal to the kernel size, resulting in non-overlapping receptive fields and reduced local detail. To mitigate this, the channel size is set to 96—higher than typical convolutional models—to compensate for the lack of overlap and preserve expressiveness.

C. Ultrasound-Specific Data Augmentation

Ultrasound images are subject to unique artifacts such as speckle noise, haze bands, shadows, and depth-based intensity loss caused by reflection, interference, and attenuation of acoustic waves. These are not effectively replicated using traditional augmentations like rotation or brightness adjustments, which are tailored for natural images. To improve generalization and realism in ultrasound models, domain-specific augmentations are essential [7].

Depth attenuation simulates the exponential decay in ultrasound intensity with increasing depth. Assuming the probe is at the top-center of the image, an attenuation map $A(x, y)$ is created and applied as:

$$I'(x, y) = A(x, y) \odot S(x, y) \odot I(x, y),$$

$$A(x, y) = (1 - \lambda)e^{-\mu d} + \lambda, \quad d = \sqrt{(x - 0.5)^2 + y^2}.$$

Here, λ is the maximum attenuation (default 0), and μ is sampled from $[0, 3)$.

Acoustic haze is simulated by adding a Gaussian band of noise centered at a distance r from the probe:

$$H(x, y) = \frac{1}{2} u e^{\left(-\frac{(d-r)^2}{2\sigma^2}\right)}, \quad u \sim \mathcal{U}(0, 1),$$

with $r \sim \mathcal{U}(0.05, 0.95)$, and $\sigma \sim \mathcal{U}(0, 0.1)$.

Gaussian shadows, inspired by Smistad et al. (2018), mimic acoustic shadows using 2D Gaussian attenuation:

$$G(x, y) = 1 - s e^{\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right)}$$

with center (μ_x, μ_y) randomly placed, $\sigma_{x,y} \sim \mathcal{U}(0.1, 0.4)$, and strength $s \sim \mathcal{U}(0.25, 0.8)$.

Speckle noise is mitigated using a bilateral filter, with spatial and color parameters sampled from $[0.1, 2.0)$ and $[0, 1)$, respectively.

These augmentations help create realistic training data that better reflect clinical variability, improving model robustness in tasks like segmentation and classification. Fig. 4 illustrates the application of various domain-specific augmentations on ultrasound images.

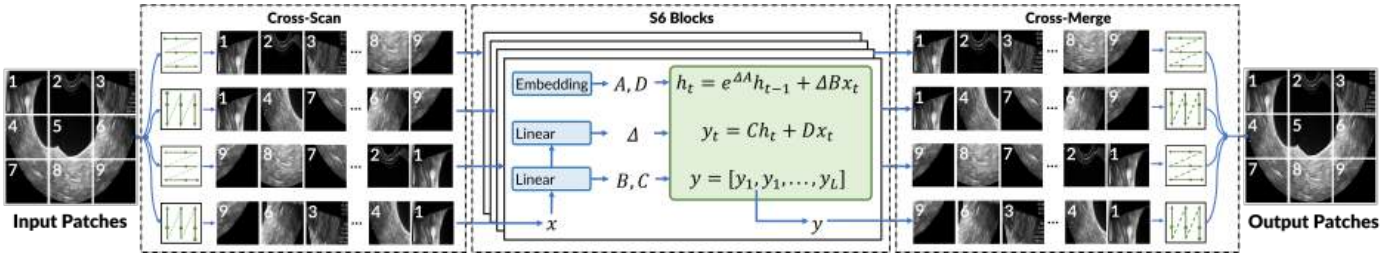

Figure 3: The scanning method in SS2D Block

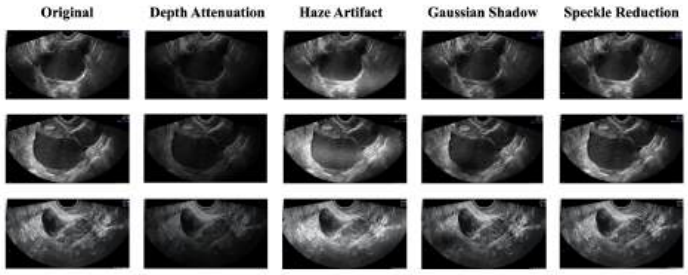

Figure 4: Each row shows one original ultrasound image and its augmented versions.

D. Loss Function

Loss functions critically influence how well the segmentation model learns to distinguish object contours and spatial areas. In this work, we adopt several losses function to address pixel-wise classification and overlap quality as follow:

a) Focal loss:

$$\mathcal{L}_{\text{Focal}} = -\alpha \cdot (1-p)^{\gamma} \cdot \log(p) \tag{1}$$

This loss is specialized function designed to address the issue of class imbalance in classification tasks. Specifically, Focal loss reduces the impact of easily classified samples and focuses more on hard-to-classify ones.

b) Jaccard:

$$\mathcal{L}_{\text{Jaccard}}(y, \hat{y}) = 1 - \frac{y * \hat{y}}{(y + \hat{y} - y * \hat{y})} \tag{2}$$

Jaccard loss, also known as IoU loss, is commonly used for image segmentation tasks. This loss measures the similarity between the predicted output and the ground truth.

c) Binary Cross-Entropy (BCE) Loss:

$$L_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N}\left(\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]\right) \tag{3}$$

This loss measures the pixel-wise error between predicted probability $\hat{y}_i$ and ground truth y_i, making it effective for binary classification tasks.

d) Dice Loss::

$$L_{\text{Dice}} = (y, \hat{y}) = 1 - \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|} \tag{4}$$

Dice loss evaluates the overlap between predicted and ground-truth masks, which reduces class imbalance and improves region-level accuracy.

$$\mathcal{L}_{\text{BCE-Dice}}(y, \hat{y}) = \alpha \cdot L_{\text{BCE}} + \beta \cdot L_{\text{Dice}} \tag{5}$$

To enhance robustness and generalization, we combine BCE and Dice loss with a weighted factor $\alpha = \beta = 1$, leveraging both pixel-wise precision and region consistency.

IV. Experiment and Results

A. Datasets

The evaluation was conducted on OTU2D set of the Multi-Modality Ovarian Tumor [8] (MMOTU) image dataset. The MMOTU dataset serves as the benchmark for validating performance of ovarian tumor segmentation task, containing 1469 ultrasound images. Given that pixel-wise and global-wise annotations are available. Following the previous work [15], images from the OTU 2D dataset were split into training, validation, and testing sets at an 80:10:10 ratio. This means 1177 images were allocated for training, with 146 images each for validation and testing. We also provide in details 4 evaluation metrics, including Mean Intersection over Union (mIoU), Dice Similarity Coefficient (DSC), Precision (P) and Recall (R).

B. Implementation Details

Following the prior works [14], we resize the images in the OTU2D dataset to 256x256. Several specialized data augmentation techniques are utilized to prevent overfitting. The BceDice loss function is used for training step. We set the batch size to 8 and employ AdamW optimizer with an initial learning rate of 1e-4. CosineAnnealingLR is also utilized as the scheduler with a maximum of 13 iterations and a minimum learning rate of 1e-5. Training epochs are set to 25.

C. Ablation Study

Through table I, we see that the BCE (Binary Cross Entropy) loss achieved the best and the most balanced result in terms of three metrics including Dice, mIoU and Precision. Only Dice loss has Recall higher than that of BCE loss which leads to the idea to combine BCE and Dice loss for better outcome. Although Bce loss achieves 3 highest value of DSC, mIoU and Precision among experiments, after experimenting, BceDice provides the best result using specific augmentations as shown

Table I: Performance of Our Proposed Method Comparison with Different Loss Functions (%). Best results are in bold

Loss	mIoU	DSC	Precision	Recall
Focal	71.85	83.62	84.33	82.92
BCE	**75.87**	**86.28**	**86.08**	86.47
Dice	73.49	84.72	81.21	**88.55**
Jaccard	72.62	84.13	85.07	83.22
BceDice	75.65	86.13	84.07	88.30

in the table III. Figure 5 further illustrates the advantages of combining BCE and Dice losses.

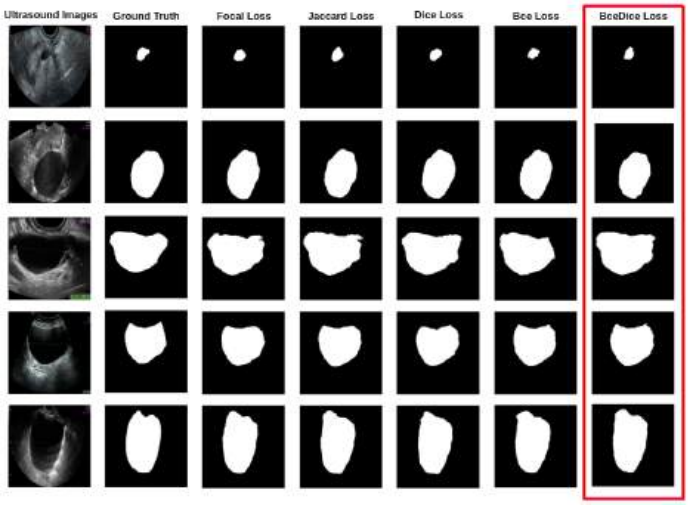

Figure 5: Example of segmentation results using different loss metrics.Among the metrics evaluated, the BceDice loss not only effectively preserves the overall tumor shape, a characteristic of the Dice loss but also captures the fine surface details, a strength of the BCE loss.

D. Comparision to the State-of-the-art models

As shown in table II, VM-UNet [14] with domain-specific augmentation techniques shows massive performance in ovarian tumor segmentation tasks in all metrics.

Compared to UNet, our method improves mIoU by 15.79%, DSC by 11.14%, Precision by 2.80%, and Recall by 9.99%, demonstrating both architectural benefits and the impact of domain-specific augmentations.

Against MU-Net, we observe gains of 14.10% in mIoU, 9.33% in DSC, 5.53% in Precision and 9.15% in Recall. Even when compared to strong baselines like DANet, our model achieves higher mIoU (+1.27%), DSC (+0.63%) and Recall (+1.25%), confirming its robustness, especially in boundary delineation.

While SAM achieves high Precision (87.34%), its low Recall (71.85%) suggests missed tumor regions. In contrast, our model balances both Precision (87.53%) and Recall (89.15%), enabling more complete and reliable segmentation.

These results confirm that our approach not only achieves state-of-the-art accuracy but also delivers stable and reliable performance across diverse evaluation criteria. Figure 6 further illustrates the robustness of VM-UNet.

Table II: Comparison with State-of-the-Art Models on the OTU2D Dataset (%). Best results are in bold.

Models	mIoU	DSC	Precision	Recall
Unet [2]	63.31	77.19	84.73	79.15
MU-Net [15]	65.00	79.00	82.00	80.00
DANet [16]	77.83	87.70	**88.20**	87.90
SAM [4]	64.07	75.60	87.34	71.85
Proposed method	**79.10**	**88.33**	87.53	**89.15**

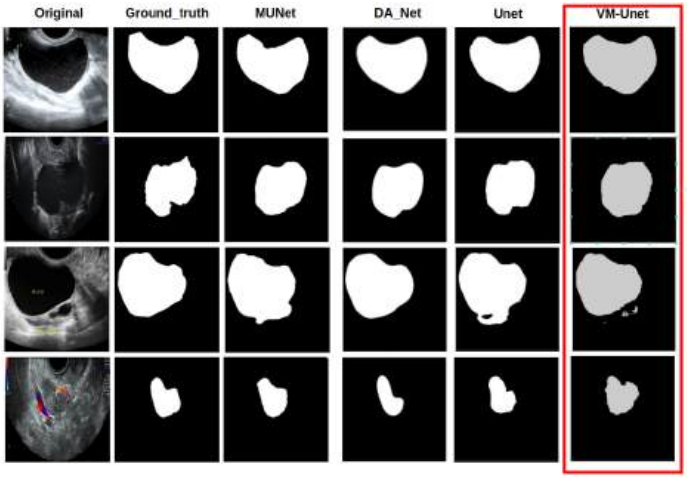

Figure 6: Example of segmentation results using different models. While preserving the global shape, VM-UNet also catches the tumor's rough surface better in popular cases.

E. Comparison VM-UNet with and without specific-domain augmentations

Table III: Performance of VM-UNet with Different Augmentation Strategies (%). Best results are in bold.

Strategies	mIoU	DSC	Precision	Recall
None [14]	75.36	85.95	81.97	90.33
Generic	75.65	86.13	84.07	88.30
Specific-domain	77.10	87.07	84.92	89.33
Combined	**79.10**	**88.33**	**87.53**	**89.15**

To improve model robustness and simulate real-world ultrasound variability, we apply four augmentations before training: Depth Attenuation mimics signal decay with depth, Gaussian Shadow simulates acoustic shadows, Speckle Reduction suppresses speckle noise to enhance structure clarity, and Haze Artifact introduces blur-like effects representing soft tissue

scattering. These transformations help enhance the semantic quality of ultrasound images by reducing machine-induced artifacts. We will conduct four experiments to compare VM-UNet performance:

- VM-UNet with Original Dataset: Baseline performance without augmentation.
- VM-UNet with Generic Augmentations: Using 2 standard techniques including rotation and horizontal flip. Precisely, horizontal flipping is applied to enhance data diversity while preserving anatomical plausibility, as ovarian tumors can appear on either side of the pelvic region without altering the semantic structure of ultrasound images.
- VM-UNet with Specific-Domain Augmentations: Employing 4 techniques tailored to our dataset's characteristics: Depth Attenuation, Gaussian Shadow, Speckle Reduction and Haze Artifact.
- VM-UNet with Combined Augmentations: Integrating both generic and specific-domain augmentations.

V. Conclusions

Based on the research, an effective method for ovarian tumor segmentation has been proposed, utilizing the VM-UNet architecture and tailored data augmentation strategies. This method has demonstrated superior performance compared to other state-of-the-art models on the OTU2D dataset. A key finding was that combining general augmentation techniques with those specific to ultrasound images yielded the best results, significantly improving the model's generalization and robustness.

Acknowledgment

This research is funded by Hanoi University of Science and Technology (HUST) under project number T2024-PC-044.

References

[1] H. M. Whitney, R. Yoeli-Bik, J. S. Abramowicz, *et al.*, "Ai-based automated segmentation for ovarian/adnexal masses and their internal components on ultrasound imaging", *Journal of Medical Imaging*, vol. 11, no. 4, pp. 044 505–044 505, 2024.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010 . 11929 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2010.11929.

[4] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment anything", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[5] N.-K. Nguyen, H.-S. Bui, T.-L. Pham, *et al.*, "A method for ovarian tumor segmentation based on segment anything model", in *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2024, pp. 1–6. DOI: 10 . 1109 / MAPR63514 . 2024 . 10660783.

[6] Y. Liu, Y. Tian, Y. Zhao, *et al.*, "Vmamba: Visual state space model", *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.

[7] A. Tupper and C. Gagné, "Revisiting data augmentation for ultrasound images", *arXiv preprint arXiv:2501.13193*, 2025.

[8] Q. Zhao, S. Lyu, W. Bai, *et al.*, "Mmotu: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation", *arXiv preprint arXiv:2207.06799*, 2022.

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

[10] H.-P. Luong, H.-S. Bui, N.-K. Nguyen, *et al.*, "Sovasegnet: Scale invariant ovarian tumors segmentation from ultrasound images", in *2024 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2024, pp. 2081–2087.

[11] H.-S. Bui, S.-H. Tran, T.-B. Nguyen, T.-H. Tran, H. Vu, and T.-L. Le, "Marker-aware ovarian tumor segmentation from ultrasound images", in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024, pp. 1–6. DOI: 10.1109/APSIPAASC63619.2025.10848960.

[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces", *arXiv preprint arXiv:2312.00752*, 2023.

[13] Z. Lianghui, L. Bencheng, Z. Qian, W. Xinlong, L. Wenyu, and W. Xinggang, "Vision mamba: Efficient visual representation learning with bidirectional state space model", *arXiv. Org*, arXiv–org, 2024.

[14] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation", *arXiv preprint arXiv:2402.02491*, 2024.

[15] T.-H. Nguyen, T.-L. Pham, Q.-V. Tran, *et al.*, "Systematic evaluation of loss functions for ovarian tumors segmentation from ultrasound images", in *2023 1st International Conference on Health Science and Technology (ICHST)*, IEEE, 2023, pp. 1–6.

[16] J. Fu, J. Liu, H. Tian, *et al.*, "Dual attention network for scene segmentation", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.