HVSS-Net: A Parallel Heterogeneous Feature Extraction Network for Dermoscopic Image Segmentation

Sitong Wu
Shenyang University of Technology, Shenyang 110870, China

Abstract: Accurate segmentation of lesion regions in dermoscopic images plays a critical role in computer-aided diagnosis of skin cancer. However, existing segmentation methods often struggle with insufficient global semantic modeling and limited fine-grained texture representation, which leads to blurred lesion boundaries and unstable segmentation performance. To address these challenges, this paper proposes HVSS-Net, a novel dermoscopic image segmentation network built upon an improved Mamba-UNet framework. In HVSS-Net, a dedicated HVSS feature extraction unit is designed to perform parallel heterogeneous feature extraction. This unit consists of three complementary sub-modules that focus on global contextual information, local texture characterization, and boundary-aware feature learning, respectively. The multi-path features are further integrated through a global fusion module to enhance representation robustness. Extensive experiments are conducted on three public dermoscopic datasets, namely ISIC2016, ISIC2018, and PH2. The proposed HVSS-Net achieves Dice scores of 89.95%, 87.95%, and 93.52%, respectively, demonstrating consistent performance across different datasets. Comparative results show that HVSS-Net outperforms multiple state-of-the-art methods in terms of segmentation accuracy and boundary detail preservation. These results indicate that HVSS-Net provides a stable and effective solution for accurate lesion segmentation in dermoscopic images and has strong potential for clinical auxiliary diagnosis applications.

Keywords: Dermoscopic image segmentation; Parallel heterogeneous feature extraction; Wavelet-based convolution; Boundary enhancement; Deep learning

1. Introduction

Dermoscopic image segmentation is a critical task in computer-aided diagnosis (CAD) systems for skin diseases, aiming to automatically and accurately delineate lesion regions in dermoscopic images to support the early detection and intervention of skin cancers such as melanoma [1, 2]. However, in practical clinical scenarios, dermoscopic image segmentation remains highly challenging due to low contrast between lesions and surrounding skin, blurred and irregular boundaries, diverse lesion shapes, and various interferences including hair occlusion, illumination variation, and image noise [2, 3]. These factors significantly degrade the robustness and generalization ability of segmentation models.

Traditional image segmentation approaches, such as thresholding, region growing, edge detection, and active contour models, rely heavily on hand-crafted low-level features and exhibit

limited capability in handling complex textures and irregular lesion boundaries, making them increasingly inadequate for modern clinical applications [4]. With the rapid development of deep learning, convolutional neural network (CNN)-based models have become the mainstream for medical image segmentation. Representative architectures such as U-Net [5] and its variants, including Attention U-Net [6], have demonstrated strong performance in lesion localization and boundary modeling. Nevertheless, CNNs primarily focus on local receptive fields and lack effective global contextual modeling, which often leads to under-segmentation of small lesions and inaccurate predictions in boundary-fuzzy regions.

To overcome these limitations, Transformer-based architectures, including Vision Transformer (ViT) [7] and Swin Transformer [8], have been introduced into medical image segmentation to enhance long-range dependency modeling through self-attention mechanisms. Hybrid CNN–Transformer models such as TransUNet [9], Swin-Unet [10], UCTransNet [11], and UNETR [12] further improve segmentation performance by combining local feature extraction with global contextual reasoning. Despite their effectiveness, Transformer-based models typically involve high computational costs, strong dependence on large-scale pretraining, and insufficient fine-grained structural modeling under limited medical data, which restrict their deployment in resource-constrained environments [7, 8].

Recently, state space models (SSMs) have emerged as an efficient alternative for long-range sequence modeling. The Mamba architecture introduces linear-time complexity and hardware-friendly design, achieving competitive performance and inference efficiency compared with Transformer models [13]. These characteristics indicate promising potential for medical image segmentation. However, relying solely on global modeling or local convolution remains insufficient to simultaneously preserve boundary accuracy, texture details, and small-lesion sensitivity. Consequently, designing a segmentation network that effectively integrates global semantics, local texture representations, and boundary-aware features remains a critical research challenge.

To address this issue, we propose HVSS-Net. The encoder employs a Hyper Visual State Space (HVSS) block to perform parallel multi-level feature extraction, where the Vision and Spatial-Spectral (VSS) submodule captures global contextual information, the Dense Wavelet Transform Block (Dense-WT Block) enhances local texture representations in the frequency domain, and the Enhanced Dynamic Snake Convolution Block (EDSC Block) strengthens boundary awareness. This collaborative design effectively preserves fine-grained texture details while enhancing global dependency modeling, enabling accurate small-lesion segmentation and improved boundary precision.

2. Related Work

2.1 Deep Learning-Based Dermoscopic Image Segmentation Methods

Early dermoscopic image segmentation mainly relied on traditional methods, including threshold-based approaches, region growing, and active contour models such as Snake and Level Set methods [4]. While these techniques are simple and effective for images with regular boundaries, their performance degrades significantly under complex backgrounds, low contrast, and noise interference, resulting in limited robustness and generalization. With the advent of deep learning, Long et al. proposed the fully convolutional network (FCN), enabling end-to-end pixel-wise prediction and significantly improving feature representation for semantic segmentation tasks [14]. Subsequently, Ranneberger et al. introduced U-Net [5], which employs an encoder—decoder structure with skip connections to preserve fine-grained spatial details and has become a

foundational architecture in medical image segmentation. Zhou et al. further extended this architecture by proposing U-Net++ [15], which enhances multi-scale feature fusion through nested skip connections; however, its performance remains limited when dealing with small lesions and blurred boundaries. To overcome the locality limitations of CNNs, Transformer-based architectures have been introduced into medical image segmentation. Chen et al. proposed TransUNet [9] by integrating a Transformer encoder with a U-Net decoder to capture long-range dependencies. Cao et al. further introduced Swin-Unet [10], which utilizes hierarchical window-based self-attention to improve global feature modeling. Despite these advances, Transformer-based methods often suffer from high computational costs and insufficient sensitivity to boundary and fine-grained structural details. To balance global context modeling and local feature representation, hybrid architectures have been extensively explored. Wang et al. proposed UCTransNet [11], which enhances multi-scale feature fusion by introducing channel-wise attention into skip connections. Huang et al. introduced MISS Former [16], incorporating boundary-aware mechanisms to improve segmentation performance in regions with ambiguous boundaries. However, such hybrid models typically lead to increased structural complexity, posing greater challenges for training stability and parameter optimization. In summary, existing segmentation methods still exhibit limitations in handling blurred boundaries, small lesions, and complex backgrounds. Recent studies have attempted to enhance boundary representation by incorporating geometry-aware or topological constraints. In this context, Mamba-based state space models have emerged as a promising alternative, offering efficient long-range dependency modeling while maintaining linear computational complexity. Nevertheless, current Mamba-based segmentation frameworks still face challenges in multi-scale feature integration and precise boundary geometry modeling, motivating further architectural improvements.

2.2 Multi-Feature Fusion and Attention Mechanisms

In recent years, multi-feature fusion and attention mechanisms have become key strategies for enhancing model representation capacity and robustness in medical image analysis. Multi-feature fusion methods aim to jointly model data from different scales, semantic levels, or even modalities, preserving both local texture details and global semantic structures. MedFuseNet [17] employs an encoder-decoder architecture combining convolutional neural networks (CNNs) with a Swin-Transformer backbone, which significantly improves global dependency modeling while maintaining fine-grained texture representations. However, MedFuseNet incurs high computational costs when processing high-resolution features and still struggles with boundary preservation. HiFuse [18] constructs a three-branch multi-scale fusion framework and integrates cross-scale semantics via a hierarchical feature fusion (HFF) module, achieving superior performance in classification and detection, but the precise localization of small targets and weak-boundary structures remains challenging.

To further enhance fine-grained structural representation, WTConv [19] introduces a wavelet-transform-based module to strengthen high-frequency texture responses, enabling multi-scale frequency-domain modeling with fewer parameters. Parallel CNN–Transformer fusion networks integrate local spatial details and global context by merging CNN and Transformer branch features via an attention-based fusion module, thereby enhancing region discrimination. A recent survey highlights common issues in multi-feature fusion methods, including semantic redundancy, insufficient scale alignment, and high computational cost under high-resolution conditions, particularly affecting weak-contrast lesions and discontinuous boundaries in dermoscopic images.

Moreover, the layer-wise feature reuse mechanism in DenseNet [20] provides structural inspiration for cross-layer semantic transmission in deep networks, enabling fine-grained feature representation without additional parameters. The dynamic snake convolution model [21] adapts convolution sampling positions based on geometric-topology constraints, effectively improving sensitivity to lesion boundary morphology. Inspired by these approaches, our proposed network introduces a local frequency-domain feature extraction module Dense-WT Block and a boundary-aware enhancement module EDSC Block, cooperating with the global modeling branch to form a parallel heterogeneous feature modeling structure. This design thus simultaneously enhances global semantic modeling, preserves texture fidelity, and improves boundary characterization, providing more stable and discriminative features for subsequent decoding and feature fusion (see Section 3).

3. Methodology

3.1 Network Architecture

Figure 1 illustrates the overall architecture of the proposed HVSS-Net. The input image first passes through the encoder-side Patch Embedding (PE), which divides the image into non-overlapping patches and maps them into high-dimensional feature representations. The encoder then extracts multi-scale features through multiple layers of the Hyper Visual State Space (HVSS) Block, while simultaneously integrating local frequency-domain features via the Dense-WT Block and boundary-aware features via the EDSC Block, providing rich semantic and structural information for downstream segmentation.

The deep features output by the encoder are first fed into the Bridge Module (BM), which fuses multi-scale and multi-type features to enhance global contextual information and discriminative capability, yielding higher-quality semantic representations for the decoder. The decoder consists of multiple upsampling stages. At each stage, Patch Expanding (PX) restores the spatial resolution of feature maps, and the HVSS Block further enhances boundary structures and global contextual information. In each decoding layer, the current features are fused with the corresponding encoder outputs via skip connections, integrating shallow details with deep semantics, ultimately generating high-resolution segmentation predictions of the same size as the input image.

Thus, the overall network collaboratively optimizes global semantic modeling, local texture enhancement, and boundary structure characterization, while maintaining computational efficiency.

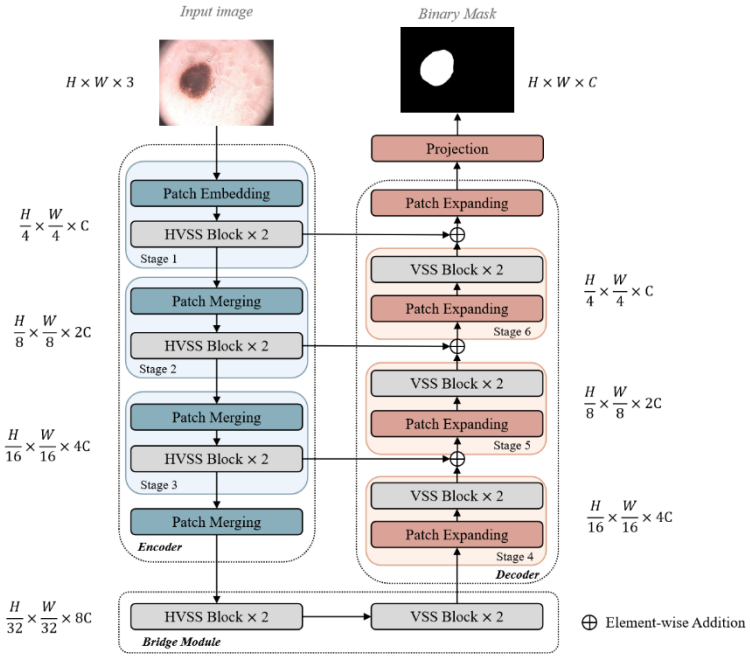

Figure 1: The Structure of the HVSS-Net Network.

3.2 Parallel Heterogeneous Feature Fusion Module

To enhance the model's ability to capture lesion details and texture information in dermoscopic images, while simultaneously improving the modeling of complex boundary structures, we propose the Parallel Heterogeneous Feature Fusion Module (HVSS Block). This module extracts heterogeneous features through three parallel paths. The VSS Block captures global contextual information, the Dense-WT Block enhances local frequency-domain texture representations, and the EDSC Block strengthens boundary awareness. This design enhances the richness and discriminative power of the extracted features, as illustrated in Figure 2.

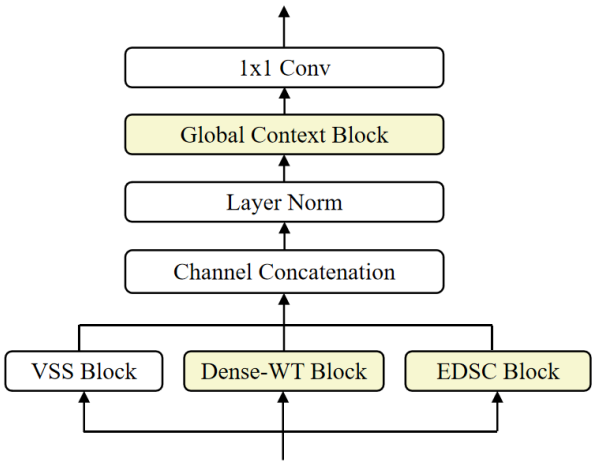

Figure 2: The Structure of the HVSS Block.

During the feature fusion stage, the heterogeneous features output by the three branches of the HVSS Block are first concatenated along the channel dimension to form a unified fused feature map. The fused feature is then normalized via a LayerNorm layer to enhance feature distribution

consistency and improve training stability. To further model spatial contextual dependencies within the feature map, a lightweight Global Context Block is introduced to perform global weighted fusion, thereby enhancing the overall consistency and representational capacity of the features. Finally, the fused feature is passed through a 1×1 convolution for channel compression and semantic integration, producing a unified feature representation that serves as input to the downstream decoder module.

This design effectively integrates multi-scale and multi-attribute information while preserving the distinctiveness of each branch's features, significantly enhancing the model's capacity for fine-grained texture representation and accurate boundary delineation. The three sub-modules of the HVSS Block are designed with specific functional contributions: the VSS Block is responsible for global semantic and context modeling, the Dense-WT Block enhances fine-grained texture representation, and the EDSC Block focuses on boundary awareness and multi-directional boundary information extraction. Together, they achieve a collaborative optimization of local texture, boundary structures, and global semantic features.

3.2.1 Local Frequency-Domain Feature Extraction Module

To enhance the model's ability to perceive fine-grained texture structures and blurred boundaries in dermoscopic images, we propose the Local Frequency-Domain Feature Extraction Module, Dense-WT Block (Figure 3). This module leverages the frequency-domain modeling capability of wavelet convolution and the multi-scale feature fusion advantages of dense connections, enhancing feature responsiveness to boundary regions and high-frequency textures.

The Dense-WT Block is optimized based on the traditional DenseNet architecture. Standard convolutions are replaced with wavelet transform convolutions (WT Conv), batch normalization (BN) is substituted with group normalization (GroupNorm), and channel concatenation is introduced, resulting in a multi-level fused frequency-domain feature representation.

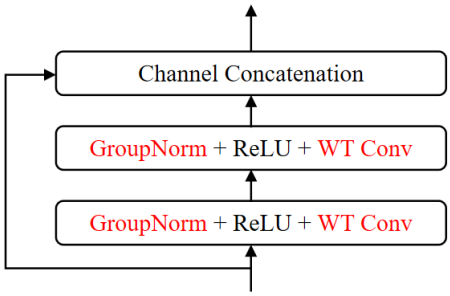

Figure 3: The structure of the Dense-WT Block.

In practice, the input feature map of the Dense-WT Block undergoes two successive wavelet modeling operations. Each operation consists of Group Normalization (GroupNorm), ReLU activation, and wavelet transform convolution (WT Conv). GroupNorm mitigates internal covariate shift and improves training stability. ReLU introduces nonlinearity to enhance feature representation. WT Conv extracts texture information from different frequency components based on multi-scale sub-band decomposition, enabling the module to effectively capture fine-grained textures and blurred boundaries.

After the two WT Conv operations, the resulting features are concatenated along the channel dimension (Channel Concatenation) to integrate multi-frequency texture information, resulting in the

final output of the Dense-WT Block. This module enhances local frequency-domain texture representation with high computational efficiency, providing rich boundary and detail information for subsequent segmentation.

3.2.2 Boundary-Aware Enhancement Module

The Boundary-Aware Enhancement Module, EDSC Block (Figure 4), is introduced to improve the model's multi-directional representation and fine-grained perception of lesion boundaries in dermoscopic images. Built upon a dynamic snake convolution mechanism, the EDSC Block comprises three main components: a Bottom Encoding Block, a Three-Branch Block for boundary feature extraction, and a Top Encoding Block. These components collaboratively enhance the extraction and representation of boundary features, enabling the network to better capture detailed and complex lesion edges.

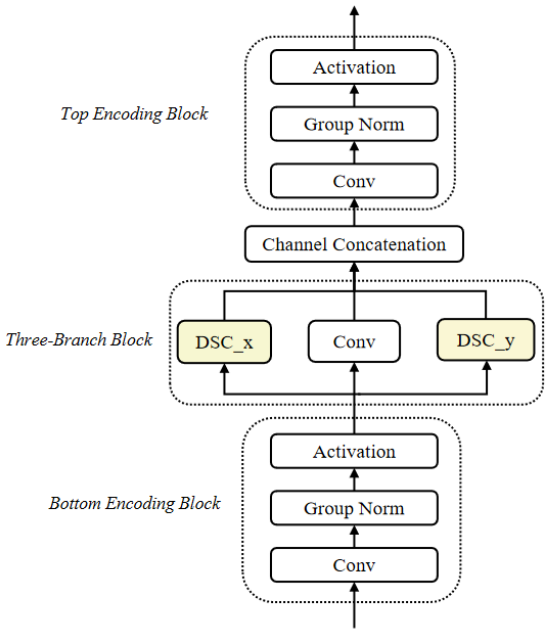

Figure 5: Structure of the EDSC Block.

The input features are first processed by the Bottom Encoding Block, where convolution, Group Normalization (GroupNorm), and ReLU activation are applied to extract fundamental texture information and reduce redundant dimensions, providing stable initial features for direction-sensitive boundary modeling. The features then pass through three parallel convolutional branches for boundary feature extraction: the horizontal dynamic snake convolution branch (DSC_x) adaptively captures horizontal boundary variations through deformable convolutions; the vertical dynamic snake convolution branch (DSC_y) enhances vertical boundary modeling; and the standard convolution branch (Conv) employs 3×3 convolutions to extract local contextual information while preserving spatial continuity. The outputs from the three branches are concatenated along the channel dimension and forwarded to the Top Encoding Block, where convolution, normalization, and a 1×1 convolution are applied for feature integration and channel compression, producing a final feature representation that consolidates multi-directional boundary information.

This module effectively enhances the model's representation of complex boundary structures by

leveraging multi-path directional awareness and feature fusion. It provides rich edge information for downstream segmentation and improves the recognition of small lesions and blurred boundaries. Both the Bottom and Top Encoding Blocks utilize 3×3 convolutions, GroupNorm, and ReLU activation to ensure the stability of initial features and the discriminability of the final output features.

4. Data and Experimental Results

4.1 Experimental Platform and Parameter Settings

All experiments were conducted on an Ubuntu 22.04 operating system, equipped with an Intel Core i9-12900K CPU and an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory, as summarized in Table 1. The experiments were implemented using the PyTorch deep learning framework with CUDA 11.8. During training, a batch size of 16 was used, and the AdamW optimizer was employed with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The maximum number of training epochs was set to 100.

Table 1: Experimental Environment Configuration.

Parameters	Configuration
CPU	Intel Core i9-12900K
GPU	NVIDIA GeForce RTX 3090
OS	Ubuntu 22.04
CUDA	11.8
Deep Learning Framework	PyTorch

4.2 Experimental Datasets and Preprocessing

Experiments in this study were conducted on three publicly available dermoscopic image segmentation datasets: ISIC2016, ISIC2018, and PH2. All datasets follow a unified directory structure. Specifically, the root directory is organized into train, val (optional), and test subdirectories, each containing two folders: image and label. The label folder stores binary lesion masks corresponding one-to-one with the original dermoscopic images.

To ensure reproducibility, training, validation, and testing splits are managed using a unified script. For datasets providing an official test set, it is directly used for evaluation, while the validation set is proportionally split from the training set. For datasets without official splits, the training data are randomly divided into training, validation, and testing sets according to predefined ratios (default: train_ratio = 0.8, val_ratio = 0.1, test_ratio = 0.1) under a fixed random seed (seed = 42). The resulting split configuration is saved in a splits.json file and reused in all subsequent experiments to guarantee consistent data partitioning across different models.

Data preprocessing is implemented through a customized PairedData class. All images and corresponding masks are first resized to a fixed resolution (default: 256 × 256 pixels), with bilinear interpolation applied to images and nearest-neighbor interpolation applied to masks to preserve label integrity. Image pixel values are normalized to floating-point tensors in the range [0, 1] and further standardized using a mean and standard deviation of 0.5, resulting in an approximate range of [−1, 1]. Segmentation masks are binarized into 0/1 labels. No online random data augmentation is applied during training, ensuring that performance differences can be attributed solely to variations in network architectures rather than stochastic effects.

4.3 Evaluation Metrics

To comprehensively evaluate the performance of segmentation models in dermoscopic image segmentation tasks, four widely used metrics are employed: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision, and Recall. These metrics provide complementary perspectives on segmentation performance, including region overlap, boundary prediction quality, and overall classification accuracy.

The Dice coefficient quantifies the overlap between the predicted mask and the ground truth mask, ranging from 0 to 1, with higher values indicating better agreement. Its computation is defined in Eq. (1). The Intersection over Union (IoU) evaluates the ratio of the intersection to the union of the predicted and ground truth masks, as given in Eq. (2). Precision measures the proportion of correctly predicted positive pixels among all predicted positive pixels, while Recall quantifies the proportion of ground truth positive pixels that are correctly identified. Their corresponding formulations are presented in Eq. (3) and Eq. (4), respectively.

By utilizing these evaluation metrics, the segmentation performance can be assessed from multiple complementary aspects, ensuring a thorough and reliable comparison among different models.

$$Dice = \frac{2|P \cap G|}{|P| + |G|} \tag{1}$$

$$IoU = \frac{|P \cap G|}{|P \cup G|} \tag{2}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

In these definitions, TP denotes the number of pixels correctly classified as belonging to the target region, FP denotes the number of pixels incorrectly classified as target, TN denotes the number of pixels correctly classified as non-target, and FN denotes the number of target-region pixels mistakenly classified as non-target. These quantities are used in the calculation of Dice, IoU, Precision, and Recall to quantify segmentation performance.

4.4 Comparative Experiments

To evaluate the effectiveness of the proposed HVSS-Net for dermoscopic image segmentation, comprehensive comparative experiments were conducted on three publicly available datasets: ISIC2016, ISIC2018, and PH2. The proposed method was benchmarked against eight representative segmentation models, including Attention U-Net, DCSAUNet, MALUNet, SwinUNet, UKAN, U-Net, U-Net++, and UTNet. All experiments were performed under consistent training configurations and identical hardware settings to ensure a fair and controlled comparison. Table 2 presents the quantitative results on the ISIC2016 dataset, including Dice coefficient, Intersection over Union (IoU), Precision, Recall, and training time for each method.

Table 2: Segmentation Results for Each Model on the ISIC2016 Dataset.

Model	Dice/%	IoU/%	Precision/%	Recall/%	train_time_hours
AttentionUNet	88.73	81.51	91.33	88.86	0.32

DCSAUNet	87.84	79.87	90.31	88.94	0.13
MALUNet	81.23	70.73	82.85	86.71	0.12
SwinUNet	88.55	80.93	89.14	90.29	0.14
UKAN	89.14	81.87	90.94	90.39	0.24
UNet	88.95	81.35	91.20	89.53	0.21
U-Net++	88.22	80.72	89.18	91.01	0.33
UTNet	87.62	80.15	88.44	91.08	0.30
HVSS-Net (Ours)	89.95	83.08	91.26	90.58	0.43

As shown in Table 2, the proposed HVSS-Net demonstrates superior segmentation performance on the ISIC2016 dataset, achieving the highest Dice and IoU scores of 89.95% and 83.08%, respectively. While CNN-based models such as U-Net and U-Net++ effectively capture local contextual features, they are limited in modeling global dependencies. Conversely, Transformer-based models like SwinUNet excel at long-range dependency modeling but often underperform in extracting fine-grained local details. By leveraging parallel HVSS blocks that emphasize global context modeling, local texture enhancement, and boundary awareness, HVSS-Net synergistically combines the global modeling capacity of state-space models (SSMs) with the local-detail sensitivity of convolutional networks, enabling deep feature fusion rather than simple feature aggregation. Compared with the U-Net baseline, HVSS-Net improves Dice, IoU, Precision, and Recall by 1.00, 1.73, 0.06, and 1.05 percentage points, respectively, and outperforms all other advanced models in the core metrics of Dice and IoU.

The training efficiency of the models was also analyzed. Although HVSS-Net requires relatively longer training due to the use of parallel heterogeneous modules and a more sophisticated feature fusion strategy, its substantial gains in segmentation accuracy justify the additional computational cost. In contrast, models such as MALUNet and DCSAUNet exhibit higher training efficiency but at the expense of reduced segmentation performance. These observations indicate that an appropriate balance between accuracy and computational efficiency is critical in medical image segmentation, and HVSS-Net achieves a favorable trade-off within acceptable training time limits.

Table 3 reports the segmentation results on the ISIC2018 dataset. HVSS-Net consistently achieves the best performance among all evaluated methods, with a Dice score of 87.95% and IoU of 80.97%, demonstrating its strong capability in pixel-level segmentation. Notably, HVSS-Net attains a Precision of 91.41%, surpassing the second-best model by 2.29 percentage points, reflecting more reliable predictions with lower false positive rates. Although its Recall (88.43%) is slightly below that of U-Net++ (90.00%), this difference represents a balanced trade-off between Precision and Recall rather than a deficiency. In terms of computational efficiency, HVSS-Net’s training time (1.01 hours) is comparable to UTNet and substantially shorter than U-Net, indicating competitive efficiency. Overall, HVSS-Net achieves an effective balance between segmentation accuracy and computational cost on the ISIC2018 dataset, yielding superior overall performance.

Table 3: Segmentation Results for Each Model on the ISIC2018 Dataset.

Model	Dice/%	IoU/%	Precision/%	Recall/%	train_time_hours
AttentionUNet	87.55	80.29	89.86	89.28	0.96
DCSAUNet	87.56	80.62	88.87	90.10	1.21

MALUNet	80.46	70.77	81.58	87.25	1.06
SwinUNet	85.80	78.28	88.93	87.58	1.34
UKAN	86.07	78.27	87.93	89.73	0.59
UNet	86.93	79.56	89.12	89.73	1.42
UNet++	86.56	79.03	87.51	90.00	1.20
UTNet	81.87	73.95	87.86	83.33	1.01
HVSS-Net (Ours)	87.95	80.97	91.41	88.43	1.01

The segmentation performance on the PH2 dataset is summarized in Table 4. The proposed HVSS-Net exhibits a clear superiority across all evaluation metrics. Specifically, it achieves the highest Dice coefficient of 93.52% and an Intersection over Union (IoU) of 88.01%, demonstrating the closest alignment between the predicted segmentation masks and the ground-truth annotations. Furthermore, HVSS-Net attains the highest Recall of 92.60%, substantially outperforming other compared methods, which indicates its strong capability in comprehensively identifying lesion regions and effectively reducing missed detections.

Although the Precision of HVSS-Net (94.99%) is slightly lower than that of certain models, such as UKAN (97.00%), the combination of leading Dice and Recall values reflects a more favorable overall balance between maximizing lesion detection (high Recall) and minimizing false positives (high Precision). This confirms that HVSS-Net achieves superior overall segmentation performance on the PH2 dataset.

Regarding computational efficiency, the training time of HVSS-Net is 0.11 hours, which is longer than that of most comparative models. Nevertheless, the significant gains in segmentation accuracy demonstrate that the network maintains a reasonable and well-considered trade-off between accuracy and computational cost, underscoring its practical effectiveness in dermoscopic image segmentation tasks.

Table 4: Segmentation Results for Each Model on the PH2 Dataset.

Model	Dice/%	IoU/%	Precision/%	Recall/%	train_time_hours
AttentionUNet	92.33	86.04	96.44	89.24	0.04
DCSAUNet	93.19	87.46	96.32	90.64	0.03
MALUNet	87.53	78.43	87.85	89.07	0.02
SwinUNet	86.47	77.80	96.08	80.94	0.01
UKAN	91.60	84.81	97.00	87.37	0.08
UNet	93.13	87.40	96.03	90.96	0.06
UNet++	93.26	87.62	95.16	91.88	0.06
UTNet	83.92	74.35	96.63	77.32	0.02
HVSS-Net (Ours)	93.52	88.01	94.99	92.60	0.11

A qualitative comparison of segmentation results produced by nine representative networks on the ISIC2018 dataset is presented in Figure 6, illustrating the advantages of HVSS-Net. Each row corresponds to a distinct lesion case, and each column shows the output of a specific model. The column labeled “Image” displays the original dermoscopic images, while “GT” denotes the

expert-annotated ground-truth masks.

In the first case, the lesion exhibits a regular shape with well-defined boundaries relative to the surrounding skin. While most models roughly capture the lesion region, MALUNet suffers from over-segmentation, producing spurious regions that incorrectly classify normal skin as lesion. Attention U-Net generates relatively coarse boundaries, whereas UKAN and DCSAUNet produce smoother contours but display slight blurring in local boundary details. UTNet slightly underestimates the lesion area. By contrast, HVSS-Net closely aligns with the ground-truth mask, with only minor deviations of 1–2 pixels along the upper-right boundary and no evident false positives or missed regions.

In the second case, the lesion is approximately circular with a pronounced color contrast against the surrounding skin. All models achieve basic lesion coverage; however, Attention U-Net's boundaries remain slightly jagged, and UKAN exhibits local boundary blurring at the top and right regions. Other models show moderate performance in capturing fine-grained boundary structures. HVSS-Net effectively preserves the overall lesion shape and produces smoother, more accurate boundaries that closely match the ground truth, with only minimal pixel-level deviations observed at the lower edge.

In the third case, the lesion covers a relatively large area with sharp boundaries. DCSAUNet shows minor roughness along small curved boundary segments, and UTNet slightly underestimates the lesion region. HVSS-Net accurately restores the lesion contour, generating smooth boundaries and maintaining a high degree of consistency with the ground-truth mask, with minimal deviations of 1–2 pixels at curvature inflection points.

Overall, HVSS-Net consistently achieves precise segmentation across all examined cases, characterized by smooth boundaries and rich structural details. Both false positives and missed regions are substantially reduced, and the predicted masks exhibit the closest agreement with expert annotations among all compared methods, while minor boundary imperfections remain.

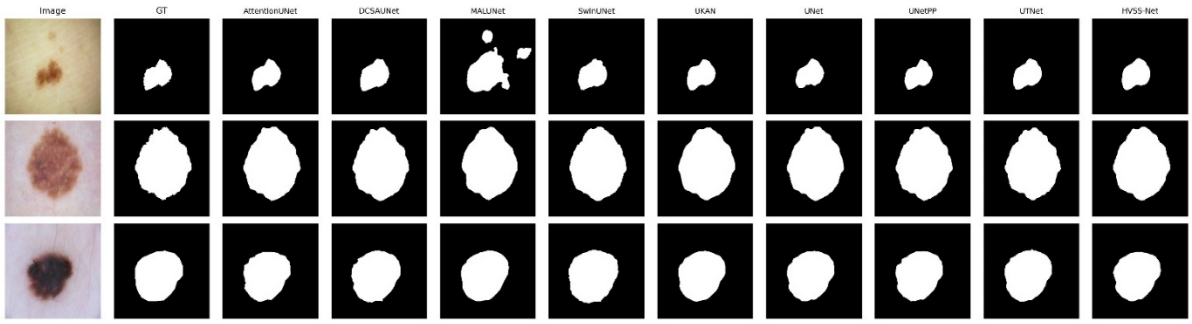

Figure 6: Visual Comparison of Segmentation Results by Different Models on the ISIC2018 Dataset.

4.5 Ablation Study

To validate the effectiveness of the proposed HVSS-Net architecture, systematic ablation experiments were conducted on the ISIC2018 dataset using the full model as the baseline. The study investigated the contributions of the three core modules—Dense-WT, VSS, and EDSC—and assessed the impact of the SDF loss function on segmentation performance. Table 5 summarizes the quantitative results for different module configurations. The complete model demonstrates superior overall segmentation capability, achieving a Dice score of 0.8813, IoU of 0.8124, and Precision and Recall of 0.9035 and 0.8987, respectively, indicating a well-balanced performance in distinguishing

positive and negative regions.

Analysis of individual module contributions reveals that removing the Dense-WT module (noDense-WT) reduces Dice to 0.8501 and Recall to 0.8285, underscoring its essential role in capturing multi-scale texture and fine-grained features. When using Dense-WT alone (Dense-WT_only), the model maintains strong performance with a Dice score of 0.8747. Employing the VSS module alone (VSS_only) achieves the highest Precision of 0.9208, although Recall slightly decreases to 0.8704. Conversely, removing VSS (noVSS) increases Recall to 0.9098 while reducing Precision, demonstrating the module's key function in balancing Precision and Recall. The EDSC module alone (EDSC_only) attains the highest Recall of 0.9109, highlighting its boundary-aware capability, whereas removing EDSC (noEDSC) results in a notable drop in Precision. Furthermore, omitting either the global context module (noGC) or the SDF loss (noSDF) leads to overall performance degradation, indicating that global feature aggregation and boundary supervision provide complementary benefits.

Regarding training efficiency, the VSS_only model converges fastest, requiring 26 epochs (0.48 h), whereas the noVSS configuration demands the longest training time, 44 epochs (0.83 h). The complete HVSS-Net achieves a reasonable compromise between segmentation performance and training time (0.78 h), demonstrating the effectiveness and efficiency of the proposed design.

Table 5: Ablation Study Results on the ISIC2018 Dataset.

Model	Dice (%)	IoU (%)	Precision (%)	Recall (%)	Epochs	Train_Time (s)
Mamba_EDSC_only	87.95	80.81	88.51	91.09	42	2735.90
Mamba_VSS_only	87.32	79.99	92.08	87.04	26	1728.50
Mamba_Dense-WT_only	87.47	80.45	89.68	89.69	41	2713.20
Mamba_noGC	87.64	80.49	89.84	89.72	27	2242.50
Mamba_noSDF	84.30	75.57	82.33	92.22	11	905.60
Mamba_noEDSC	87.03	79.50	87.96	90.44	34	2498.30
Mamba_noVSS	87.72	80.58	88.59	90.98	44	2985.80
Mamba_noDense-WT	85.01	78.02	91.37	82.85	18	1318.60
Ours	88.13	81.24	90.35	89.87	33	2795.30

The ablation study on the PH2 dataset, summarized in Table 6, demonstrates the contribution of each module to HVSS-Net's performance. The full model achieves a Dice of 0.9337, IoU of 0.8778, and Precision and Recall of 0.9564 and 0.9171, respectively, confirming its leading segmentation capability. Removing the Dense-WT module (noDense-WT) reduces Dice to 0.9270 and Recall to 0.9054, indicating its key role in capturing multi-scale texture. Excluding the VSS module (noVSS) slightly increases Recall to 0.9211 but decreases Precision to 0.9573, highlighting VSS's importance in reliable prediction. The EDSC module alone (EDSC_only) achieves a Recall of 0.9190, emphasizing its boundary-aware capability. The global context module (noGC) and SDF loss (noSDF) provide further incremental improvements.

In terms of training efficiency, the complete model requires 0.78 h, slightly longer than single-module variants, but the performance gains justify this computational cost. Overall, the Dense-WT, VSS, and EDSC modules complement each other, and their integration is essential for HVSS-Net's superior performance on PH2.

Table 6: Ablation Study Results on the PH2 Dataset.

Model	Dice (%)	IoU (%)	Precision (%)	Recall (%)	Epochs	Train Time (s)
Mamba_EDSC_only	93.54	88.06	95.77	91.90	69	266.80
Mamba_VSS_only	92.70	86.73	94.78	91.34	74	347.60
Mamba_Dense-WT_only	93.20	87.45	96.18	90.91	83	348.60
Mamba_noGC	93.25	87.59	95.31	91.90	64	524.40
Mamba_noSDF	93.45	87.90	95.25	92.19	74	573.70
Mamba_noEDSC	93.30	87.66	95.23	92.03	74	517.50
Mamba_noVSS	93.67	88.27	95.73	92.11	85	548.50
Mamba_noDense-WT	92.70	86.60	95.51	90.54	90	585.70
Ours	93.37	87.78	95.64	91.71	86	780.60

5. Conclusions

To enhance the segmentation accuracy of dermoscopic images and improve the delineation of lesion boundaries, this study proposes an end-to-end deep learning network that explicitly preserves boundary and texture information within lesion regions. Experimental results on the ISIC2016, ISIC2018, and PH2 datasets demonstrate that the proposed HVSS-Net achieves superior segmentation performance compared with baseline methods. The generated segmentation masks exhibit well-defined boundaries and rich texture details, leading to significant improvements in key evaluation metrics, including Dice and IoU.

Future work will focus on multi-task boundary optimization based on the signed distance function (SDF), aiming to address the limitations of conventional binary segmentation by enforcing boundary continuity constraints. This strategy is expected to further enhance boundary prediction accuracy and improve the model’s generalization capability across diverse dermoscopic datasets.

Acknowledgements

The authors would like to express their sincere gratitude to their advisors for valuable guidance and suggestions throughout this research, and to their classmates for assistance and support during experiments and discussions.

References

[1] Celebi M E, Iyatomi H, Schaefer G, et al. Lesion border detection in dermoscopy images. Computerized medical imaging and graphics, 2009,33(2):148-153.

[2] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis. Medical image analysis, 2017,42:60-88.

[3] Abbas Q, Celebi M E, García I F. Hair removal methods: A comparative study for dermoscopy images. Biomedical Signal Processing and Control, 2011,6(4):395-404.

[4] Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. International journal of computer vision, 1988,1(4):321-331.

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation: International Conference on Medical image computing and computer-assisted intervention, 2015. Springer.

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

[7] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[8] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows: Proceedings of the IEEE/CVF international conference on computer vision, 2021.

[9] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.

[10] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation: European conference on computer vision, 2022. Springer.

[11] Wang H, Cao P, Wang J, et al. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer: Proceedings of the AAAI conference on artificial intelligence, 2022.

[12] Hatamizadeh A, Tang Y, Nath V, et al. Unetr: Transformers for 3d medical image segmentation: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022.

[13] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces: First conference on language modeling, 2024.

[14] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[15] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation: International workshop on deep learning in medical image analysis, 2018. Springer.

[16] Huang X, Deng Z, Li D, et al. Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162, 2021.

[17] Chen R, He S, Xie J, et al. MedFuseNet: fusing local and global deep feature representations with hybrid attention mechanisms for medical image segmentation. Scientific Reports, 2025,15(1):5093.

[18] Huo X, Sun G, Tian S, et al. HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomedical Signal Processing and Control, 2024,87:105534.

[19] Finder S E, Amoyal R, Treister E, et al. Wavelet convolutions for large receptive fields: European Conference on Computer Vision, 2024. Springer.

[20] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[21] Qi Y, He Y, Qi X, et al. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation: Proceedings of the IEEE/CVF international conference on computer vision, 2023.