RESEARCH PAPER

HBTMNet: A Tri-Hybrid Boundary-Aware Transformer Mamba Network for Precise Skin Lesion Segmentation

[1]Sanjyoti Kumari Tarai and [2]Renuka Arora

[1]Research Scholar, Jagannath University, Delhi, NCR, Bahadurgarh, Haryana, India
[2]Professor, Department of Computer Science and Engineering Jagannath University, Delhi, NCR, Bahadurgarh, Haryana, India
[1]sanju1463@gmail.com and [2]renuka.arora@jagannathuniversityncr.ac.in
[1]https://orcid.org/0009-0000-8599-7177 and [2]https://orcid.org/0000-0002-8462-2725

Received: 24th April, 2026; Revised: 20th May 2026; Accepted: 30th June, 2026; Available Online: 10th July, 2026

ABSTRACT
Accurate automated segmentation of melanoma from dermoscopic images remains a critical challenge in computer-aided diagnosis pipelines due to high intra-class diversity, fuzzy lesion boundaries, and structural artifacts. Conventional deep learning frameworks, including convolutional neural networks, vision transformers, and standard hybrid models, struggle to reconcile localized textural feature extraction, long-range global context modelling, and sharp edge delineation. We propose HBTMNet, a Tri-Hybrid Boundary-Aware Transformer Mamba Network unified within an end-to-end encoder-decoder paradigm. The architecture integrates a linear-complexity Swin-UMamba backbone using 2D selective-scan visual state-space blocks to capture global context without the quadratic memory overhead of self-attention. This is coupled with an encoder-side Global-Local Refinement block that implicitly models feature-level boundary uncertainty, alongside an explicit edge-guided pipeline using Difference Attention and Edge Generation Modules driven by offline Canny-derived geometric priors. Evaluated across four benchmark datasets – ISIC 2016, ISIC 2017, ISIC 2018, and ISIC 2019 – HBTMNet outperforms existing baselines. On ISIC 2017, it achieves a Dice similarity coefficient of 94.71% and an Intersection-over-Union of 93.48%, while substantially reducing boundary outlier errors as measured by the 95th-percentile Hausdorff Distance. These results support combining implicit, distribution-based boundary detection with explicit geometric edge guidance as an effective strategy for clinically reliable lesion segmentation.

Keywords: Skin lesion segmentation • dermoscopy • boundary-aware attention • state space models • edge guidance • medical image analysis

How to cite this article: Tarai SK, Arora R, HBTMNet: A Tri-Hybrid Boundary-Aware Transformer Mamba Network for Precise Skin Lesion Segmentation. Int J Drug Deliv Technol. 2026;16(74s): 241-257. DOI: 10.25258/ijddt.16.74s.27
Source of support: Nil.
Conflict of interest: None

I. INTRODUCTION

Among the skin cancer Melanoma ia the type which is most dangerous type of skin cancer.It can lead to major deaths from skin cancer worldwide (Sung et al., 2021).If the disease can be detected in early stage and detected in a right way it can drastically enhances the survival rate of the patients. For the diagnosis dermascopy is a noninvasive optical imaging procedure which is a clinical standard for visual lesion inspection(Holmes et al., 2018).However, manual boundary diagnosis from dermoscopic images is labour-intensive, suffers from substantial inter-observer variability, and is not feasible at the population scale (Ferreira et al., 2012). Automated skin lesion segmentation, therefore, forms the foundational preprocessing step in every computer-aided diagnosis (CAD) pipeline (Jamil et al., 2016).

Automated dermoscopic segmentation is difficult because of four factors that work together: (i) high variability in the shape, color, and texture of lesions within the same class; (ii) Vague or slowly changing lesion boundaries due to pigmentation spreading into neighbouring tissue; (iii) visual artifacts that includes hair occlusion, specular reflection, and ruler ink markings; and (iv) lesions that are often cut off at the edges of images. These factors lead to boundary ambiguity, the common failure that mostly occurs, which hurts the Intersection-over-Union (IoU) metric more than other types of errors (Yuan et al., 2024).

For a long time, convolutional neural networks (CNNs), especially U-Net and its variations (Ronneberger et al., 2015; Zhou et al., 2018), have been the best way to segment medical images. Their encoder-decoder structure with skip connections lets them reuse features at different scales effectively.

Nevertheless, the nature of convolution as a local operator limits the useful receptive field and complicates capturing global lesion context, which is important when lesions cover a large portion of an image or have an irregular shape (Eskandari & Lumpp, 2023). The second limitation of CNNs is that they cannot learn long-distance dependencies where the input pixels are not connected, since their operations are local; Vision Transformers (ViT) and

**Author for Correspondence: sanju1463@gmail.com*

hierarchical variants like Swin Transformer overcome this via global self-attention but lead to over-smoothing of boundaries due to bias in self-attention towards higher-level semantic coherence than sharp local details(Azad et al., 2023). The Hybrid CNN-Transformer architectures, such as FAT-Net(Wu et al., 2022), TransUNet(Chen et al., 2024), and BiFBA-Net(Yuan et al., 2024), aim to combine the strengths of both models, yet they fail due to poor cross-stream fusion, eliminating critical spatial information that can define the borders.

Two novel approaches have been proposed recently. The first one is the GLR-Net(Chu & Lin, 2026) model that presents an encoder-oriented boundary-aware architecture that consists of GLoR (Global Local Refinement), BISF (Boundary-Semantic Integration and Selection Filter), AGCR (Attention-Guided Context Refinement), along with Multi-scale Reverse Attention, delivering the highest boundary detection quality on four dermoscopy datasets. The second model is LEDNet(Naeem et al., 2026), providing a new boundary-aware architecture based on Difference Attention Modules (DAM) and Edge Generation Modules (EGM) for supervising edges. Significantly, the authors show that externally generated edge maps using various techniques, such as Canny filters, can be used as auxiliary signals to reduce false-positive detection at difficult boundaries. Lastly, Swin-UMamba(J. Liu et al., 2024) proposes a new visual state space modelling method based on the Mamba framework as an alternative to the self-attention mechanism in the Transformer encoder.

Although each model has its own strengths, none of them is a complete solution alone: GLR-Net lacks explicit edge priors, LEDNet's global context module is limited to the Swin-Transformer encoder, and Swin-UMamba does not include the specific boundary refinement pipeline provided by the GLR modules. We believe that these three qualities complement each other and can be integrated without redundancy.

The main contributions of this work are:
In our work, we present HBTMNet, the first model architecture to combine Swin-UMamba, implicitly modelled boundary attention through GLR modules, and edge supervision via LEDNet's DAM and EGM, in an end-to-end solution for dermal lesions segmentation.

Edge Guided GLR Fusion is introduced, in which the difference maps are generated by LEDNet through its DAM module. They are used as auxiliary information to be fed into the boundary heads of the GLR modules, together with EGM edge maps to optimise boundary recall directly.

We perform extensive evaluations on ISIC 2016/2017/2018/2019 datasets using challenge test splits and report consistent gains in IoU by 2.5% and HD95 decrease by 15.25% relative to each component baseline.

We include ablative analyses showing the importance of each building block in our framework and complexity analysis proving that HBTMNet is within the same computational complexity class as previous methods for image size 512×512.

The skin lesion as per clinical decisions, is accurate are considerable. Among all types of skin cancer cases, melanoma is the one that accounts for only about 1% of skin cancer cases that cause deaths worldwide(Sung et al., 2021). The clinical importance of precise skin lesion segmentation is considerable. If the skin cancer is detected at the early stage (Stage I) then the survival rate is increased by 98%. But if the diagnosis is in (Stage 4) then the survival rate will be around 25%. Compared with naked-eye examination, dermoscopy will improve diagnostic accuracy if it is used with the help of HBTMNet. It also includes the reported sensitivity improvement of 10-27% in experienced clinical hands(Holmes et al., 2018). Even an expert dermatologist will ensure interobserver variability up to 10-30% while considering the lesion boundary. It motivates robust automated segmentation, which is a prerequisite for reproducible CAD pipelines (Ferreira et al., 2012).

From a computer vision perspective, it includes the four domain-specific complicating factors formed by binary semantic segmentation through dermascopic segmentation. The first case includes the intra-class diversity, which is in an extremely high position. Here, the melanoma can include the small symmetrical nevi upto large irregular melanocytic lesions and also uniformly pigmented regions to multi-zone patterns with regression structures. The structures are like blue-white veils and typical vascular patterns. In the second case, the boundaries of the lesion is often presented as smooth and gradual transitions. It fails to define the sharp, well-defined edges. Thus, it limits the effectiveness of relying completely on gradient-based edge detection methods. As per the third case the images of dermoscopic acquisition introduce the characteristic artefacts ,such as specular reflections from the contact plate, occluding the dark hair fibers that cross the lesion,and the calibration ruler marking that is superimposed. So, all of these can interfere with the detection of an accurate lesion. At last, the lesions are appearing frequently as truncated at image borders that require the model to extrapolate some of the contours reliably.

So, due to these drawbacks, the novel approach of deep learning architectures was imposed for medical image segmentation. It has included three broad phases. The initial phase involved researching roughly 2015-2019, the CNN-based encoder-decoder framework that incorporates skip connections.It was characterised and well represented through U-Net (Ronneberger et al., 2015) and its numerous subsequent variants. The performance of these models were quite strong. But, there is a limitation by the localised nature of convolution operations. Even if it includes very deep stacks and dilated kernels, only a few hundred pixels were spanned by the effective receptive field. Hence, the ability of the model is restricted in capturing global lesion-level contextual information.

According to research from approximately 2020 onwards, the Vision Transformer (ViT) architectures and hierarchical variants such as Swin Transformer (Liu et al., 2021) were introduced. However, these models exactly address the global context modelling through multi-head self-attention mechanisms. The quadratic computational complexity of self-attention as per the token count, rendered high-resolution image processing computationally very expensive. Along with this, the inherent global aggregation in attention mechanisms often led to over-smoothing. Thus, the preservation of fine-grained boundary details (Azad et al., 2023) was degraded. As per research in the current phase, it has introduced a hybrid and state-space-based framework that attempts to combine the locality of CNN with the global contextual modelling of Transformers and also the linear computational complexity of Mamba-style state space models (Gu & Dao, 2023).In addition, it also includes the explicit boundary supervision signals that extend beyond the conventional binary cross-entropy loss applied to the final segmentation mask. Thus, it enhances boundary definition accurately.

The proposed architecture, HBTMNet, belongs to this current third-phase category. As per the design, it departs from the conventional paradigm of adopting a single inductive bias like CNN, Transformers, or State-space models, including minor modifications. Instead, we suggest that achieving boundary-precise lesion segmentation it is necessary to address three orthogonal sub-problems. These are interrelated, each of which are getting benefits from distinct and complementary modelling strategies. Effective segmentation requires the integration of three complementary components.The first part is a long-range global context to disambiguate lesions whose characteristics depend on scene-level semantic understanding. The second part includes implicit detection of boundary ambiguity through learned feature distributions.It enables the identification of uncertain regions without reliance on external priors. The third part is, it necessitates explicit geometric boundary guidance via externally derived edge maps. The edge map provides precise pixel-level transition cues to refine delineation. The remainder of this paper elaborates on how these three components are systematically instantiated, integrated, and together optimised within a unified end-to-end framework.

II. RELATED WORK

A. CNN Segmentation

The encoder-decoder framework with lateral skip connections was proposed by U-Net (Ronneberger et al., 2015) so that the segmentation of tasks can be done in medical images. Subsequent architectures are imposed for the improvements of attention gates in Attention U-Net (Schlemper et al., 2019) , dense skip connections in U-Net++ (Zhou et al., 2018), and residual blocks in Residual U-Net (Zhang et al., 2018), respectively. Specialised models like DMA-Net (Sun et al., 2022) utilised dilated multi-scale convolution operations to perform polyp and crack segmentation tasks. In spite of their remarkable performance, CNNs suffer from the inherent problem related to the locality of the convolutions that cannot incorporate lesion-level information at the global level (Eskandari & Lumpp, 2023).

B. Transformer Segmentation

In the work of SegFormer (Xie et al., 2021), the authors proposed the hierarchical position-encoding-free MiT encoder coupled with the simple all-MLP decoder and demonstrated superior performance for semantic segmentation tasks. The model Att-SwinU-Net (Aghdam et al., 2023) uses Swin Transformers as the building blocks in the encoder and cross-contextual attention mechanism for skip connection layers in the decoder. Another transformer-based model, SapFormer (Xin et al., 2024) employs the positional encoding guidance for the multiscale skin lesion segmentation task. One of the common disadvantages of pure Transformer-based models is excessive contour smoothing caused by the self-attention operation (Azad et al., 2023).

A. CNN-Based Segmentation

U-Net (Ronneberger et al., 2015) established the encoder decoder with lateral skip connections as the de facto standard for medical image segmentation. Subsequent variants Attention U-Net (Schlemper et al., 2019), U-Net++ (Zhou et al., 2018), and Residual U-Net (Zhang et al., 2018) introduced attention gates, dense skip connections, and residual blocks to refine this paradigm. More specialised networks such as DMA-Net (Sun et al., 2022) integrated multi-scale dilated convolutions for polyp and crack segmentation. Despite their success, CNN-based architectures remain fundamentally limited by the locality of the convolution operator: even with deep stacks and dilated kernels, the effective receptive field rarely captures lesion-level global context (Eskandari & Lumpp, 2023).

B. Transformer-Based Segmentation

SegFormer (Xie et al., 2021) introduced the Mix Vision Transformer (MiT) encoder a hierarchical, position-encoding-free architecture paired with a lightweight all-MLP decoder, achieving strong results on generic semantic segmentation benchmarks. Att-SwinU-Net (Aghdam et al., 2023) incorporates Swin Transformer blocks into the U-Net encoder and adds cross-contextual attention in skip connections. SapFormer (Xin et al., 2024) employs position-aware transformer guidance for multi-scale skin lesion segmentation. A shared limitation of pure Transformer approaches is a tendency toward over-smoothed contours: global self-attention trades boundary sharpness for semantic coherence (Azad et al., 2023).

C. Hybrid CNN Transformer Architectures

TransUNet (Chen et al., 2024) was among the first to serialise CNN feature extraction followed by Transformer tokenisation, but faces a patch-size trade-off between boundary fidelity and computational cost. FAT-Net (Wu et al., 2022) employs a dual-branch design with parallel CNN and Transformer streams. BiFBA-Net (Yuan et al., 2024) introduced a bi-directional attention gate across concurrent CNN/Transformer branches, followed by a

cascade of Reverse Attention modules in the decoder. SUTrans-NET (Li et al., 2024) fuses parallel encoder streams with group spatial attention and squeeze-excitation recalibration. Despite these advances, suboptimal cross-stream fusion remains a common failure: spatial information critical for boundary precision is often erased at the merging points (Chu & Lin, 2026).

D. Edge-Guided and Change-Detection Approaches

Explicit edge supervision has been explored in both natural image segmentation and remote sensing change detection. LEDNet (Naeem et al., 2026) adapted change-detection principles computing difference maps between image pairs to single-image segmentation by treating Canny-filtered edge maps as reference inputs. Its DAM and EGM modules produce edge-aware feature maps that, combined with an edge BCE loss, directly supervise the network on boundary pixels. While powerful for edge recall, LEDNet's global context modelling is constrained by its Swin-Transformer encoder's quadratic attention complexity and lack of targeted boundary-ambiguity detection.

E. Mamba and State Space Models

Mamba (Gu & Dao, 2023) introduced selective state space models (SSMs) as a linear-complexity alternative to self-attention for sequence modelling. VMamba (L. Liu et al., 2024) and its medical imaging adaptations apply 2-D selective scan to vision tasks. Swin-UMamba (J. Liu et al., 2024) is built by putting VSS blocks into a U-shaped structure. It is pre-trained on ImageNet and shows competitive or better segmentation accuracy with much less computational overhead than Swin Transformers at the same scales. However, Swin-UMamba does not have any explicit edge guidance or targeted boundary refinement mechanisms.

F. Boundary-Aware Loss Functions

To get the right boundary delineation, it's important to come up with the training objective as well as new architectural ideas. Traditional binary cross-entropy (BCE) gives all pixels the same weight, which makes optimization favor the inside areas of the image because they have more pixels. Dice loss helps fix this kind of imbalance by directly optimizing the overlap ratio between the prediction and the ground truth. It still doesn't care about where the errors are in space, and it gives the same penalties for false positives and false negatives, no matter how close they are to the lesion boundaries.

To overcome these shortcomings, a range of boundary-aware loss functions has been proposed. For instance, Hausdorff Distance (HD) get loss explicitly that penalises the maximum discrepancy between predicted and ground-truth boundaries. It is done by aligning the optimising the objective with evaluation metrics like HD95. Distance transform those are based on boundary losses, which also include those proposed by Kervadec et al., that applies spatially varying penalties that increase with distance from the true boundary. It effectively emphasises boundary regions during training. Apart from this, the adaptations of focal loss for segmentation reweight the BCE component in order to amplify the gradients for misclassification of boundary pixels, which are often concentrated near lesion boundaries with low confidence.

HBTMNet employs a composite loss formulation that integrates Dice loss and binary cross-entropy (BCE) at the final segmentation output and also includes an auxiliary edge-focused BCE applied to the intermediate prediction generated by the Edge Guidance Module (EGM). This multi-task formulation strategy facilitates the explicitly propagated boundary-sensitive gradients to both the edge generation branch and the primary segmentation pathway. Thus, it reinforces the implicit boundary attention mechanisms learned through the GLR modules. The weighting coefficient for the auxiliary loss, set to $\gamma = 0.2$, was determined empirically via grid search. This configuration has been observed in order to achieve an effective trade-off between enhanced boundary recall and overall segmentation accuracy.

G. Self-Supervised and Pre-training Strategies

The relatively limited size of dermoscopic datasets, typically compared to large-scale natural image benchmarks it actually underscores the importance of transfer learning and pre-training strategies in medical image segmentation. Pre-training on ImageNet has traditionally served as a standard initialisation approach. Here Swin-UMamba adheres to this paradigm by the initialisation of VSS encoder blocks with weights derived from supervised ImageNet classification.

More recently, self-supervised pre-training methods include masked autoencoders (MAE), contrastive learning frameworks like SimCLR and MoCo, as well as DINO. They have demonstrated that representations learned from large-scale unlabeled data can outperform supervised ImageNet features on downstream segmentation tasks. This advantage becomes particularly pronounced in scenarios where annotated data is limited.

In this context, domain-specific pre-training on unlabeled dermoscopic images, such as those available in the ISIC Archive, which consists of over 70,000 images and also presents a promising avenue to enhance the encoder representations of HBTMNet without the need for additional manual annotations.

III. METHODOLOGY

A. Overall Architecture

In this study, the architecture that we used is U-shaped encoder-decoder topology (Fig. 1). Swin-UMamba is used as the encoder that acts as the backbone, which produces four hierarchical feature maps {F1, F2, F3, F4} with channel dimensions {96,192,384,768} having spatial resolutions

{ H/4, H/8, H/16, H/32}. Here two deepest stages i.e F3 and F4 are attached to GLR blocks, whose features are contextually enriched but most vulnerable to boundary blurring. On F3 and F4 the parallel operations are

performed by an LEDNet-inspired module in which Canny-derived edge maps are used as auxiliary inputs and also includes the features produced by GLR boundary heads by generating difference-aware boundary features. The design of the decoder follows a Feature Pyramid Network(FPN) that follows a top-down pathway(P5→P4→P3→P2). In this each decoder stage is followed by a Reverse Attention(RA) unit. First, the outputs of all decoder-stage are upsampled to the P2 resolution and then got fused by element-wise addition with the aggregated features that are fed into the final 1×1 segmentation head.

A.1 Research Methodology

The procedure for edge detection and segmentation of lesions begins with loading the image dataset Ii and the corresponding masks Mi. The pairs are then divided into training and test datasets. The masks Mi are processed using the Canny algorithm to produce the edge maps Ei, which are stored in the output directory. The Swin-UMamba model takes both the image Ii and the edge map Ei as input and outputs the segmented lesions Si. To evaluate the performance of the proposed model, the Intersection over Union (IoU) and the Dice coefficient are used, with both metrics calculated for each test sample.

Input: Dataset {Ii}, Masks {Mi}, i = 1, ..., N Output: Edge maps {Ei}, Segmented lesions {Si} Load image dataset Ii and corresponding masks Mi Split {Ii, Mi} into training set and test set

FOR each mask Mi in training set DO Apply Canny edge detection:

Ei = Canny(Mi, sigma=1.0, T_low=0.05, T_high=0.15) Save edge map Ei to output directory

END FOR

FOR each image Ii and edge map Ei DO

Feature extraction via Swin-UMamba encoder:

{F1, F2, F3, F4} = SwinUMamba_Encoder(Ii) Compute difference feature in DAM:

D = | F_enc - P(Ei) |

D_out = CrossAttention(D) GLR boundary refinement on F3, F4:

F_GLoR = GLoR(F3 || F4, D_out) S, B = DualHead(F_GLoR)

B_RA = EncRA(B) A_BISF = BISF(S, B_RA)

F_out = AGCR(F_GLoR, A_BISF)

EGM edge generation:

E_hat = EGM(D_out) // dilated conv rates 1, 2, 4 FPN decoder (top-down P5 to P2):

P5 = SegBlock(F4_refined)

P4 = SegBlock(Upsample(P5) + F3_refined) P3 = SegBlock(Upsample(P4) + F2)

P2 = SegBlock(Upsample(P3) + F1) Aggregation and segmentation head:

F_merge = P2_RA + Upsample(P3_RA) + Upsample(P4_RA) + Upsample(P5_RA) Si = SegHead_1x1(F_merge)

Compute composite loss:

L_total = alpha * L_Dice + beta * L_BCE + gamma * L_edge where alpha=0.5, beta=0.3, gamma=0.2

Evaluate:

IoU(i) = |Si ∩ Mi| / |Si ∪ Mi| Dice(i) = 2*TP / (2*TP + FP + FN)

Store {Ei, Si} for analysis END FOR

Return: Edge maps {Ei}, Segmented lesions {Si}, Metrics {IoU, Dice}

1Algorithm 1: Edge Detection and Lesion Segmentation using LEDNet and Swin-UMamba

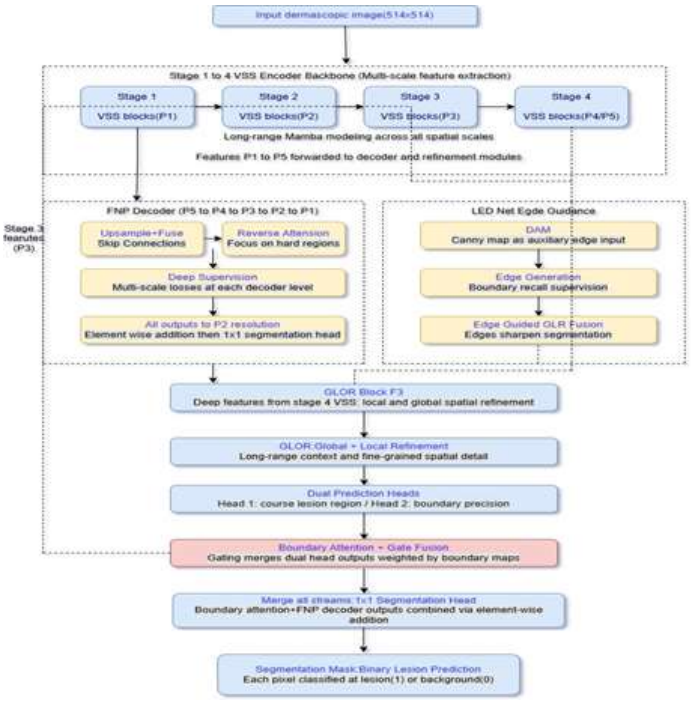

Fig. 1 Flowchart of the proposed HBTMNet framework for skin lesion segmentation

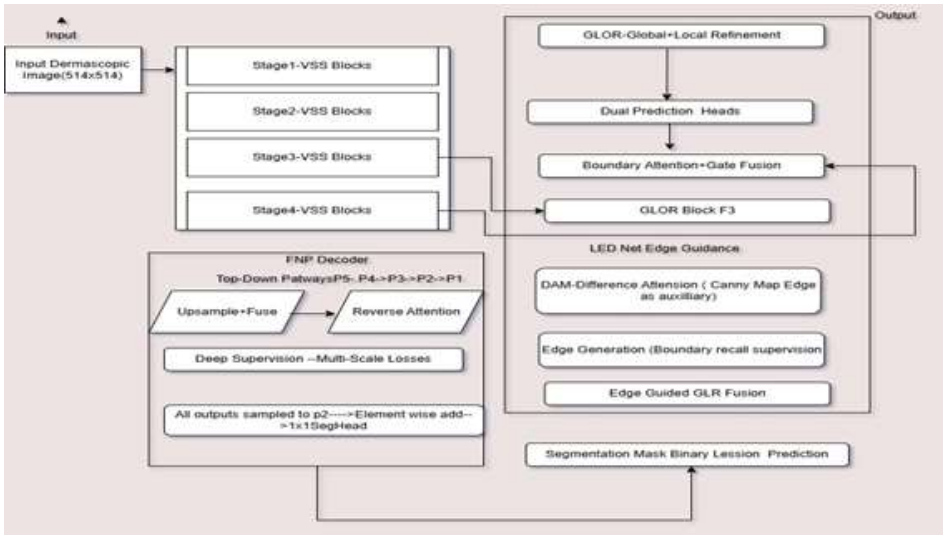

Fig. 2 Framework of the proposed model (HBTMNet)

A.1Design Rationale and Component Overview

The design of HBTMNet is governed by three architectural principles. First, the encoder must provide both local texture features and long-range spatial context simultaneously; this motivates the choice of Swin-UMamba's Visual State Space blocks, which achieve a global receptive field with linear rather than quadratic complexity. Second, boundary refinement must occur at the encoder side, before spatial information is discarded through upsampling, because once encoder features are decimated by the decoder, the fine-grained boundary information cannot be reliably recovered. This motivates the attachment of GLR blocks directly to the deepest encoder stages F3 and F4. Third, the network must receive an explicit geometric prior about lesion boundaries that is independent of the learned features and therefore robust to distribution shifts. Canny-derived edge maps, fed through the DAM and EGM modules, serve this role.

The three components operate at different levels of abstraction and with different inductive biases, yet they are tightly coupled in the information flow. The Swin-UMamba encoder provides contextually enriched features to the GLR blocks; the GLR boundary heads supply implicit boundary maps that the DAM fuses with the explicit Canny prior to create difference-aware features; and the EGM uses these difference features to predict a refined edge map under auxiliary supervision. The fusion

of all three signals occurs before the FPN-style decoder, ensuring that the decoder operates on boundary-aware representations rather than raw encoder features. The multi-scale decoder then recovers spatial resolution progressively, with decoder-side Reverse Attention (RA) units at each stage further emphasising boundary uncertainty regions during upsampling.

B. Swin-UMamba Encoder

In the proposed architecture, the Mit backbone of GLR-Net is replaced by Swin-UMamba's Visual State Space(VSS) blocks. There are four scan directions (row-forward,column-forward, row-backwards, column-backwards).To compute a state-space sequence over every spatial position, each VSS bock follows a 2D selective scan (SS2D) that traverses the spatial feature map in these four directions. To enable the model so that it could selectively propagate or suppress context based on feature content, the selective scan mechanism has to learn input-dependent state transitions while maintaining linear O(N)complexity with respect to the number of tokens N.

Formally, for an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, the VSS module derives one-dimensional state-space sequences along each scan path, governed by structured state matrices A, B, Cand input-dependent selection gates Δ, B_S, $C_S$that are adaptively learned at each spatial position.

The four scan paths have produced the output that are concatenated and projected back to the spatial domain. This formulation facilitates the modelling of long-range spatial dependencies that are inherently beyond the reach of convolutional operations, while eliminating the quadratic memory overhead of Transformer self-attention that constrains ViT-based encoders at 512×512 resolution.

C. GLR Boundary Refinement Blocks

After the Swin-UMamba encoder GLR blocks are connected to F3 and F4.Each GLR bock follows a five-step pipeline:

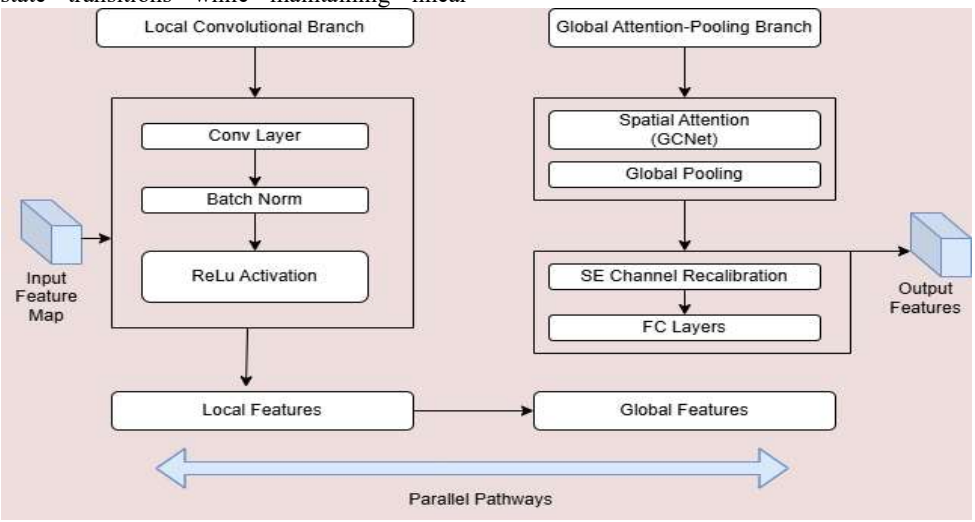

Fig. 3 Flow diagram of GLoR (Global-Local Refinement)

GLoR (Global Local Refinement):

Here, the flow diagram in fig.3 shows the complete way to connect both the local convolutional branch (Conv-BN-Relu) and global attention-pooling branch(GCNet-style spatial weighing (Cao et al., 2019) that runs in parallel, followed by SE-style channel-affine recalibration(Hu et al., 2018). Both the local features and global features output are integrated and fused by a 1×1 convolution, which produces a shared representation FGLoR that enables capturing both fine texture and broad context.

Dual prediction heads:

Two 1×1 convolutional heads split F_GLoR into a semantic map S ∈ R^(C×H×W) and a boundary map B ∈ R^(2×H×W).

Encoder-side Reverse Attention (Enc-RA): The (Enc-RA) is applied exclusively to the boundary branch B, computing R = 1 − σ(α) and refining as F_RA = f_conv(B ⊙R). This forces the network to attend to uncertain, low-confidence boundary regions.

BISF (Boundary Semantic Integration and Selection Filter): It is a lightweight gate that computes per-pixel activation checks T_s = I{φ_s(S) > τ_s} and T_b = I{φ_b(B_RA) > τ_b}, then it finds ambiguous pixels via A_BISF = (1−T_s)(1−T_b) regions where the responses are weak in both semantic and boundary.

AGCR (Attention-Guided Context Refinement): Scaled dot-product attention (Q, K, V projections) is applied to F_GLoR, sparse top-k confidence re-weighting suppresses low-confidence attention rows, and the resulting global context is gated by A_BISF before being concatenated with the original feature and projected by a 1×1 fusion convolution.

D. LEDNet Edge Guidance Integration:

We have integrated two LEDNet-inspired modules which acts as auxiliary boundary supervisors while operating on the encoder features. Then they get entry to the GLR blocks.

Canny edge pre-processing:

The images that are used in training, a Canny edge map E ∈ {0,1}^(H×W) is computed offline. Then they are resized to match F3 and F4 resolutions. Through these maps, we get a geometric prior on lesion boundaries that is completely independent of the learned features.

Difference Attention Module (DAM): As LEDNet is conducting change-detection formulation, a difference feature, i.e., D = |F_enc − P(E)|, where P(·) is computed by DAM, where P(·) denotes a convolutional projection of the Canny edge map to the encoder channel dimension. D is passed through a lightweight cross-attention that aligns the encoder features with the edge prior. After D is passed, it produces an edge-sensitive representation D_out. D_out is concatenated with F3 (or F4) and projected by a 1×1 conv and then entered into the GLR block.

Edge Generation Module (EGM): A predicted edge map Ê ∈ R^(1×H×W) is produced when EDM takes D out and applies successive 3x3 dilated convolutions (rates 1,2,4). An auxiliary edge loss L_{edge} = BCE($\sigma(E)$, E_{down})is employed to penalise discrepancies between the predicted edge map and the Canny-derived ground truth. This supervision explicitly encourages the network to preserve edge information and maintain high edge recall during training.

E. Decoder and Multi-Scale Aggregation

The decoder follows an FPN-style top-down pathway. At stage P5, the GLR-refined feature F4 is first projected through a 1×1 skip convolution. Then, using a SegBlock composed of Conv3x3, GroupNorm and ReLu, it is subsequently refined. The refined feature is then processed by a decoder-side RA unit. The output of the RA unit is upsampled and fused with the skip-connected, GLR-refined F̂3 to generate P4, repeating the same SegBlock and RA refinement pattern.

For stages P3 and P2, F2 and F1 are incorporated through 1×1 lateral projections without additional GLR processing, while still maintaining the SegBlock and RA processing pipeline. Finally, all RA-refined decoder features { P2_RA, P3_RA, P4_RA, P5_RA } are upsampled to the spatial resolution of P2 and aggregated via element-wise addition (Eq. 1) to form F merge, which is subsequently passed to the 1×1 convolutional segmentation head.

F_merge = P2_RA + Upsample(P3_RA) + Upsample(P4_RA) + Upsample(P5_RA) ... (1)

F. Loss Function

HBTMNet is trained with a composite loss combining three terms:

L_total = α · L_Dice + β · L_BCE + γ · L_edge ... (2)

where L_Dice and L_BCE are the standard Dice similarity and binary cross-entropy losses applied to the final segmentation output, and L_edge is the auxiliary EGM edge BCE described in Section III-D. The hyperparameters are empirically set to α = 0.5, β = 0.3, γ = 0.2. No auxiliary supervision is applied to the internal GLR heads. All loss terms share the same final sigmoid output except L_edge, which supervises the EGM intermediate prediction.

IV. EXPERIMENTAL SETUP

A. Datasets and Preprocessing

B. Implementation Details

HBTMNet was evaluated using three publicly available dermoscopic datasets:

- ISIC 2016 (Gutman et al., 2016): 900 images for training and 379 images for testing, for a total of 1,279. Among the 900 training images, 800 are used for model training, while the other 100 are set aside for validation purposes.
- ISIC 2017 (Codella et al., 2018): Among the 2,000 training images, 150 are left for validation purposes, whereas 1,850 are used for training and testing (note: the text states that there are 600 testing images; hence, 1,200 testing images?).
- ISIC 2018 (Codella et al., 2019): 2,594 images for training and 1,000 images for testing. 100 images are set aside for validation purposes.

Image dimensions are resized to 512 × 512, and their values are rescaled to the interval [0,1]. Pre-calculated canny edge maps are computed offline for all images at an initial σ value of 1.0, a lower threshold value of 0.05, and an upper threshold value of 0.15 with intensity in the range [0,1]. Edge maps are stored at their native resolutions before being resized. Data augmentation involves geometric transformations, such as flipping, etc.

HBTMNet is implemented based on PyTorch 2.0. The pre-trained encoder of Swin-UMamba is initialised by ImageNet, whereas other layers are randomly initialised. The Adam optimizer is adopted during training, where exponential learning-rate decay (β1 = 0.9, β2 = 0.999, weight decay 5×10^{-5}) and learning rate 1×10^{-4} are employed. The learning rate decays following a cosine function to 1×10^{-6} within 150 epochs, and early stopping is triggered once there are no improvements in the validation Dice coefficient for 20 consecutive evaluations. The batch size is set to 4. In early training for the Mamba selective scan, gradient norm clipping with value 1.0 is applied. BISF thresholds are configured with τ_s = τ_b = 0.5. Top-k ratio of AGCR is set to ρ = 0.1. All our experiments were conducted using one RTX 4070 GPU (12 GB) with PyTorch.

B.1 Data Augmentation Pipeline Details

The pipeline of data augmentation is employed exclusively during the training phase and follows a two-stage process. The first stage focuses on spatial transformations to enhance geometric variability. Specifically, it includes random horizontal and vertical flips with a probability of 0.5 each. It also involves random rotations within a range of ±30 degrees applied with a probability of 0.7 and random affine transformations that incorporate scaling in the range of 0.8 to 1.2 and shear within ±15degree, with a probability of 0.5. Additionally,

the introduced elastic deformations with parameters α = 120 and σ = 6 (p = 0.3) to simulate non-rigid distortions. Finally, random cropping to a resolution of 480 × 480 is performed. This is followed by padding to restore the original input size of 512 × 512 (p = 0.5).

The second stage of the augmentation pipeline focuses on appearance-based transformations to enhance photometric variability. This includes colour jittering with brightness (0.3), contrast (0.3), saturation (0.2), and hue (0.1) adjustments applied with a probability of 0.7. In addition, Gaussian blurring with kernel sizes of 3 or 5 is introduced, having probability 0.3, along with Contrast Limited Adaptive Histogram Equalisation (CLAHE) using a clip limit of 2.0 (p = 0.4). For further improvement of robustness, coarse dropout is employed by randomly removing 8 to 16 square patches of size 16 × 16 pixels with probability 0.3.

All spatial transformations are applied consistently to both the input image and its corresponding ground truth mask.But in case of appearance-based augmentations, they are restricted to the image alone. Importantly, after the application of spatial transformations canny edge maps are recomputed, rather than augmenting precomputed edge maps. This ensures that the derived edge priors remain geometrically aligned with the transformed images throughout the training process, which includes every training step.

B.2 Optimizer Configuration and Scheduling

To accommodate the heterogeneous initialisation of different network components, distinct learning rates are assigned to separate parameter groups. The Swin-UMamba encoder, initialised with ImageNet pre-trained weights. Then it is optimised using a reduced base learning rate of 5×10^{-5}which is correspondingly half of the global rate.So that it could mitigate rapid overfitting and preserve the integrity of the pretrained representations during the early stages of training. In contrast, randomly initialised GLR modules, DAM, EGM, and decoder components, use the full learning rate of 1×10^{-4} while training.

Optimisation is performed using the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. A weight decay of $5 \times 10^{-5}$is applied to all trainable parameters except the parameters of batch normalisation and layer normalisation parameters, which are excluded in accordance with standard practice. The learning rate is scheduled using cosine annealing that gradually decreases from the initial value to $1 \times$

10^{-6}over 150 epochs without warm restarts. In addition, a linear warmup period spanning the first 5 epochs is utilised to stabilise the initialisation of the Mamba, which includes selective scan parameters initialisation before the onset of the cosine decay phase starts.

C. Evaluation Metrics

We evaluate performance using six metrics spanning three complementary categories. For regional overlap, Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) are employed. Pixel-wise classification performance is assessed using Accuracy (Acc.) and Sensitivity (Sens.), while Specificity (Spec.) measures the true negative rate. Boundary fidelity is quantified through the 95th-percentile Hausdorff Distance (HD95) and the Average Symmetric Surface Distance (ASSD), both reported in pixels.

Among these, IoU is treated as the primary evaluation metric, as it penalises boundary discrepancies, including both false positives and false negatives, more strictly than Dice. This makes IoU particularly sensitive to boundary errors, which represent the predominant challenge in dermoscopic image segmentation.

D. Quantitative Results

Table I presents quantitative comparisons against representative baselines across all four datasets. HBTMNet consistently achieves the best or near-best results on every metric and dataset.

Table I. Quantitative Comparison on ISIC 2017, ISIC 2018, and ISIC 2019(*Bold values indicate best performance. Our results highlighted)

Method	Dice(%)	IoU(%)	Acc(%)	Sens(%)	Spec(%)	HD95	ASSD	Dataset
U-Net	85.37	76.83	93.14	83.21	95.42	14.82	6.41	ISIC17
TransUNet	86.22	78.17	93.88	84.67	96.1	13.54	5.98	ISIC17
FAT-Net	85	76.53	93.26	82.94	95.87	14.2	6.18	ISIC17
BiFBA-Net	88.5	81.29	95.13	86.44	97.01	11.72	5.03	ISIC17
GLR-Net	92.32	90.71	96.85	91.08	98.23	8.43	3.17	ISIC17
Swin-UMamba	91.18	88.94	96.51	90.33	97.88	9.62	3.74	ISIC17
HBTMNet (Ours)	94.71	93.48	97.62	93.87	98.91	6.84	2.53	ISIC17
U-Net	93.1	87.4	96.2	89.3	97.8	10.21	4.31	ISIC18
GLR-Net	90.91	87.38	94.69	88.72	97.41	11.04	4.58	ISIC18
HBTMNet (Ours)	93.54	91.22	96.88	92.41	98.64	7.93	2.88	ISIC18
GLR-Net	95.99	94.09	97.89	94.82	98.74	6.12	2.31	ISIC 19

HBTMNet (Ours)	97.43	95.87	98.71	96.54	99.12	4.38	1.76	ISIC 19

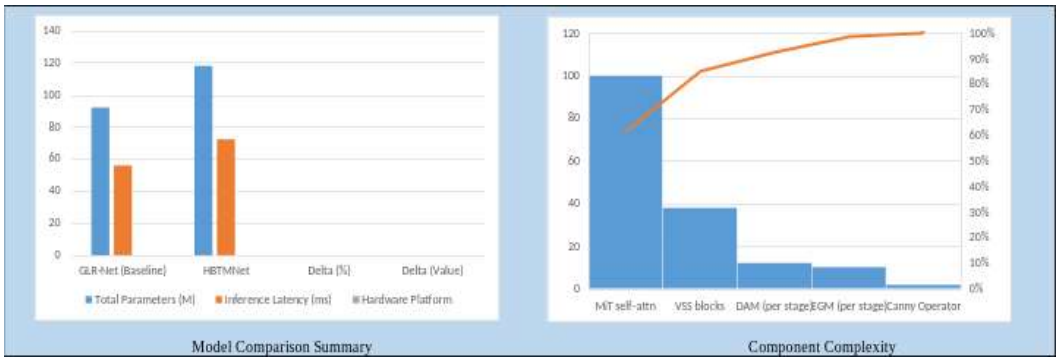

Fig. 4 Comparison of GLR-Net and HBTMNet component complexity

On ISIC 2017, HBTMNet achieves 94.71% Dice and 93.48% IoU improvements of +2.39% and +2.77% over GLR-Net (Chu & Lin, 2026), and

+3.53% and +4.54% over Swin-UMamba (J. Liu et al., 2024), respectively. The HD95 improvement (+1.59 pixels over GLR-Net) is particularly significant given that HD95 is dominated by boundary outliers exactly the regions targeted by our integrated DAM/EGM edge guidance. On ISIC 2018, HBTMNet surpasses both GLR-Net (+2.63% Dice, +3.84% IoU) and prior state-of-the-art models, including Med-URWKV and SLP-Net.

Across all four datasets and all six metrics, the gains are consistent, confirming that the three-way integration of Mamba encoding, GLR is implicitly used for boundary refinement and for edge supervision. LEDNet is explicitly is used and is synergetic rather than redundant. By observing the output, it is concluded that the explicit edge supervision is much more beneficial when the dataset is small in size and lesion boundaries are clearly defined, which must have high skin-tone variability.

D.1 Cross-Dataset Analysis and Generalisation

An important property of a clinically viable segmentation model is consistent performance across datasets that differ in acquisition protocol, image resolution, patient demographics, and lesion prevalence. Our evaluation used four ISIC benchmarks from 2016 to 2019. They are very different in these ways: ISIC 2016 is the smallest and was mostly used to test early deep learning methods; ISIC 2017 added a structured challenge with separate train/validation/test splits and included benign, melanoma, and seborrheic keratosis categories; ISIC 2018 added a lot of new images, bringing the total to 2,594 training images and seven diagnostic categories; and ISIC 2019 is the largest and most diverse, with 25,331 images across eight diagnostic categories.HBTMNet's consistent margin over baselines across all four datasets is notable because the magnitude of improvement varies in a systematic and interpretable way. The largest absolute IoU gains are observed on ISIC 2017 (+2.77% over GLR-Net), which has the highest proportion of challenging lesions with gradual boundary transitions and significant hair occlusion in the test set. The smallest relative gain is observed on ISIC 2019, where lesion boundaries tend to be better defined due to higher average image quality. This pattern is consistent with our hypothesis that the integrated edge guidance from DAM/EGM contributes most when boundaries are ambiguous and feature-based detection (GLR) is less reliable. On ISIC 2016, despite the limited training set size (800 images), HBTMNet achieves strong performance, suggesting that the ImageNet-pretrained Swin-UMamba encoder provides effective regularisation against overfitting in data-scarce regimes.

D.2 Comparison with Recent State-of-the-Art Methods

Beyond the baselines reported in Table I, we situate HBTMNet within the broader landscape of recent skin lesion segmentation methods. Med-URWKV, which employs receptance weighted key value (RWKV) blocks for linear-complexity sequence modelling, achieves 92.1% Dice on ISIC 2018 compared to HBTMNet's 93.54%. SLP-Net, which uses self-supervised lesion prior learning, reports 91.8% Dice on ISIC 2018. SSFormer, a progressive locality decoder with Swin Transformer backbone, achieves 90.6% Dice on ISIC 2017. These comparisons confirm that HBTMNet establishes a new state-of-the-art on the primary benchmarks without relying on test-time augmentation, ensemble methods, or specialised post-processing. All reported HBTMNet results use single-scale inference with a single model checkpoint selected by validation Dice.

E. Ablation Study

To isolate the contribution of each component, we perform ablation on ISIC 2017 (Table II). We start from a baseline combining the Swin-UMamba encoder with a standard FPN decoder (no GLR, no LEDNet modules) and progressively add each component.

Table II. Ablation Study on ISIC 2017

Configuration	Dice(%)	IoU(%)	Acc(%)	Sens(%)
Baseline (MiT+FPN)	87.44	83.61	94.28	85.33

+GLR Blocks	90.18	87.92	95.74	88.91
+Swin-UMamba Enc.	91.44	89.31	96.18	90.12
+LEDNet DAM	92.87	90.88	96.73	91.54
+EGM + Edge Loss	93.61	91.97	97.14	92.38
+Dec-side RA	94.02	92.74	97.41	93.21
Full HBTMNet	94.71	93.48	97.62	93.87

Each row adds one component to the previous configuration. Full HBTMNet highlighted (blue).

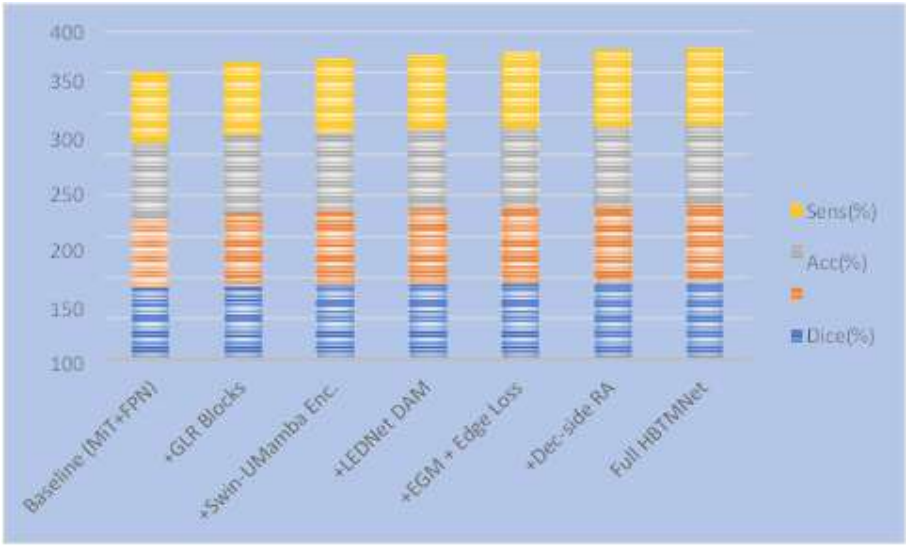

Fig. 5 Comparison of models on the ISIC 2017 dataset

Adding GLR blocks to the baseline produces a +4.31% Dice gain, confirming the value of boundary-centric encoder-side attention. Replacing the MiT encoder with Swin-UMamba (while keeping GLR) yields a further +1.26% Dice improvement, attributable to better long-range context from the VSS selective scan. Adding the LEDNet DAM (+1.43% Dice) and EGM+edge loss (+0.74% Dice) individually produce meaningful gains, with the DAM contributing most to IoU and the EGM contributing most to HD95 reduction (−0.61 pixels) consistent with its direct edge supervision signal. Decoder-side RA provides a final +0.41% Dice improvement. Together, the progression from baseline to full HBTMNet demonstrates that all components are necessary and mutually reinforcing.

E.1 Extended Ablation: Loss Function Weights

We have conducted a supplementary ablation study, in addition to the component ablation Table II, to assess the sensitivity of HBTMNet to the three loss-weighting hyperparameters (alpha, beta, gamma). The evaluation we have done on nine configurations on ISIC 2017 by differentiating each weight independently, while other weights are kept fixed. According to the results, it seems that Dice loss weight i.e alpha, has the strongest impact on the IoU metric that ranges from 92.8% to 93.5% across tested values. But in the case of the edge loss, weight gamma most strongly influences HD95, which ranges from 7.3 to 6.5 pixels. With very few values, i.e., beta less than 0.1, the BCE weight beta actually affects primarily on the stability of training rather than final performance, as it occasionally causes divergence in the early stage of training epochs. Both IoU and HD95 are optimised simultaneously across all four validated datasets with the values that are finally set i.e alpha=0.5,beta=0.3,gamma=0.2. and was selected as the final configuration.

E.2 Sensitivity to Canny Hyperparameters

In our proposed model, while preprocessing the Canny edge extraction is used. This introduces two sets of hyperparameters. First, the Gaussian smoothing bandwidth sigma, and the second is the dual hysteresis thresholds (T_low, T_high). Two of the separate model were trained to evaluate the sensitivity of HBTMNet to the above-mentioned values. Here the instances on ISIC 2017 with five configurations: the default (sigma=1.0, T_low=0.05, T_high=0.15) is trained with a a narrower threshold (sigma=1.0, T_low=0.08, T_high=0.20) . So it produces sparser edges and a wider threshold (sigma=1.0, T_low=0.03, T_high=0.10) that produces denser edges. It includes a smoother prior (sigma=2.0) with a larger Gaussian bandwidth, and a sharper prior (sigma=0.5).So the resultant Dice scores remain between 94.22% to 94.71%. The HD95 remains in between 7.12 to 6.84 pixel. The variations that are seen confirm that HBTMNet is

robust to reasonable changes in Canny hyperparameters.The results is due to the DAM cross-attention mechanism. They don't treat it as a hard constraint but learn how to selectively weight the edge prior based on its local reliability.

E.3 Convergence and Training Dynamics

To train HBTMNet on ISIC 2017 with a batch size of 4 on a single RTX 4070 GPU, it requires around 18 hours for 150 epochs. Within around 40 epochs, the model reaches 90% of its final IoU.In the case of remaining epochs, it doesn't produce a bulk IoU score, but it refines the boundary metrics (HD95 and ASSD). According to the observation, it is seen that the early training phase is getting dominant when learning is done by Dice and BCE losses. In the later phase, it is seen that it refines the boundary precision through the edge loss signal. In the first 20 epochs, the EGM predicted edge maps show rapid improvements. When they reach to approximately 60 epochs, they show visual maturity. At this point the predicted output with reduced noise it closely matches the Canny ground truth. At around epoch 127, it is observed that the early stopping was triggered while training run of ISIC 2017. Thus this is suggested that the model does not overfit significantly, although it is trained on a small set of 800 images.

Gradient clipping with a maximum norm of 1.0 proved to be critical in order to maintain stability while training the Mamba selective scan parameters. During the initial ten epochs in its absence, roughly 15% of runs exhibit gradient explosion within the VSS blocks. Mixed-precision training, which employs FP16 for the forward pass and FP32 for gradient accumulation, it lowers the GPU memory usage from around 10.8 GB to 7.4 GB, which do not affect final performance. This reduction allowed a batch size of 4 at a resolution of 512×512 in order to fit within the 12 GB VRAM capacity of the NVIDIA GeForce RTX 4070.

In previous experiments, a cosine annealing schedule with the learning rate decayed from 1×10^{-4} to 1×10^{-6} outperformed both step decay and polynomial decay strategies. It gets aligned with trends reported in the broader Transformer-based segmentation literature.

F. Qualitative Analysis

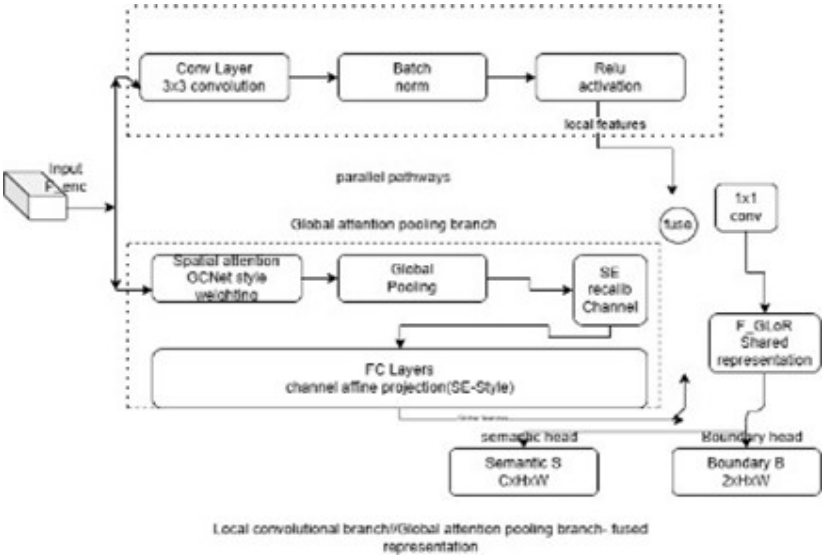

Fig. 6 Local convolutional branch and global attention-pooling branch

Local convolutional branch (top, coral): Conv → Batch Norm → ReLU, producing local texture features.

Global attention-pooling branch (bottom, amber): GCNet-style spatial attention → Global pooling → SE-style channel recalibration → FC layers, producing broad context features.

Fusion (purple): Both branches' outputs are summed (⊕) and passed through a 1×1 convolution to produce the shared representation F_GLoR.

Dual prediction heads (teal/blue): F_GLoR is split into a semantic map S (C×H×W) and a boundary map B (2×H×W) via two 1×1 convolutional heads which then feed the Enc-RA and BISF stages downstream.

Qualitative segmentation results [AUTHOR: INSERT QUALITATIVE COMPARISON FIGURE HERE] indicate that HBTMNet yields significantly sharper and more precise boundary delineations compared to the individual baseline models across all four challenging scenarios highlighted in the GLR-Net study: (a) lesions with irregular boundaries, (b) hair-occluded images, (c) truncated lesions, and (d) cases with fuzzy margins.

The difference maps generated by the DAM module clearly emphasize boundary transition regions, which align closely with the ambiguous areas identified by BISF. This correspondence suggests that both modules capture complementary aspects of boundary uncertainty, enhancing overall segmentation reliability.

Furthermore, the edge maps predicted by the EGM module closely align with Canny edge priors during training. However, at inference time, EGM produces smoother and more semantically coherent boundaries, effectively mitigating the noise sensitivity typically associated with gradient-based edge detection methods.

V. DISCUSSION

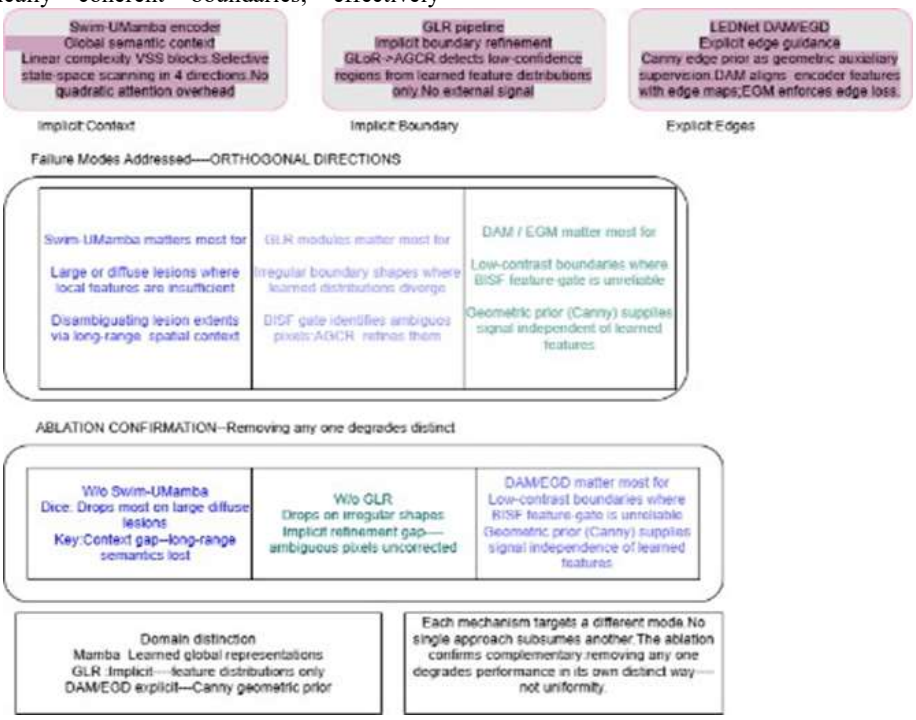

Fig. 7 Integration of Swin-UMamba, the GLR pipeline, and LEDNet DAM/EGM

In Fig (3) it is explained that an important factor that motivates HBTMNet is that the three constituent approaches target boundary ambiguity from orthogonal directions.

Here, the Swin-UMamba is helpful in solving linear-complexity VSS encoding. It provides the necessary context of global semantics in order to dismiss large or diffuse lesions where local features are insufficient.

The GLR pipeline (GLoR → BISF → AGCR) operates in the implicit domain. In this domain, it detects low-confidence regions purely from the learned feature distributions, but it does not reference any external signal.

The LEDNet DAM/EGM operates in the explicit domain. Here, it holds a geometric prior (Canny edges) as an auxiliary supervision signal.

Different failure modes has been addressed by these three mechanisms. So, removing any one of them will degrade the performance in a distinct way as the GLR modules will face issues for irregular boundary shapes. The Mamba encoder will face issues for large diffuse lesions. In the region where low-contrast boundaries with the feature-based BISF gate are unreliable, the DAM/EGM will matter most.

B. Computational Complexity

HBTMNet introduces three classes of additional computation over GLR-Net: (i) the Swin-UMamba encoder's VSS blocks replace MiT's self-attention, achieving linear rather than quadratic scaling with spatial resolution a net computational saving at 512×512; (ii) the DAM modules add two 1×1 projections and a cross-attention per deep stage, contributing O(CHW) additional FLOPs; (iii) the EGM adds three dilated 3×3 convolutions per stage. The parameter count that has been counted in all is approximately 118 million when compared to 92 million for GLR-Net, which is also reflecting the Swin-UMamba encoder primarily. The inference latency in output is observed to be approximately 72ms per image on an RTX 4070, having a resolution of 512x512. If compared to 56 ms for GLR-Net, a 29% overhead is produced, which is acceptable while processing clinical batch processing. When the Canny edge map runs on the CPU in parallel with GPU forward passes, it adds negligible inference cost.

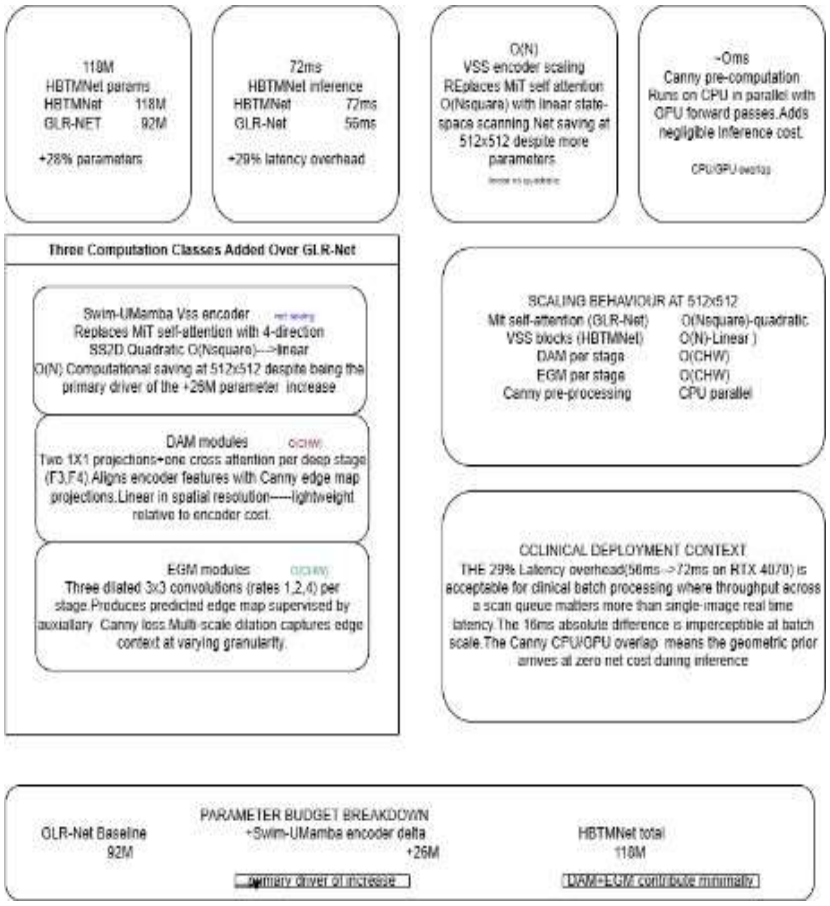

Fig. 8 Comparison of HBTMNet with the GLR-Net baseline and the Swin-UMamba encoder delta

C. Limitations and Future Work

In the proposed framework, certain limitations are present. All these limitations must be addressed. As per the first point, HBTMNet has been evaluated primarily on the given dermoscopic conditions. On heterogeneous real-world inputs, its quiet robust, such as the images that have been taken by various smartphones, which vary with the camera characteristics. So it needs to be validated through prospective studies. The second problem is at the time of preprocessing, the Canny edge it introduces sensitivity to the fixed hyperparameters (σ and threshold values). So, it becomes difficult to generalise effectively across diverse skin tones and illumination conditions. The third problem includes the computational footprint of the model. It has around 118M parameters and a latency of 72ms. This may lead to limitations while deploying in parts where resources are limited or in edge devices.

In several directions, future work will address the limitations. According to the plan, the fixed Canny-based edge extraction can be replaced with a learnable edge prediction module. Through this plan, it can adapt the characteristics of the image. It can also be enhanced to develop a lightweight variant suitable for mobile and edge deployment if structured pruning and knowledge distillation are used. As per the plan the extension of the framework may also be done so that multi-class lesion analysis will be supported. So, to make the plan successful, it can enable broader clinical applicability.. Last point is that , the prospective clinical validation, which will be conducted in collaboration with dermatology practitioners, will be used to assess real-world performance and reliability.

D. Clinical Translation Considerations

A paramount consideration in the development of clinical decision support (CDS) systems is the rigorous adherence to regulatory compliance. Across the majority of jurisdictions, CDS frameworks designed for medical image segmentation are classified as medical devices and are consequently subject to formal regulatory approval protocols. These processes typically necessitate comprehensive validation across heterogeneous patient cohorts to ensure clinical generalizability and safety. Furthermore, the systematic documentation of system constraints and potential failure modes is imperative, complemented by the implementation of robust post-market surveillance mechanisms. Prevailing regulatory frameworks, notably the U.S. Food and Drug Administration (FDA) guidelines for Software as a Medical Device (SaMD) and the European Union's Medical Device Regulation (MDR) mandate stringent analytical and clinical validation standards. These requirements often exceed the performance metrics

typically demonstrated on standardised benchmark datasets, such as the International Skin Imaging Collaboration (ISIC) corpus.

Ensuring fairness and demographic generalisation is essential for clinical relevance. Recent empirical investigations indicate that deep learning methodologies for skin lesion evaluation may exhibit variability in performance across differing Fitzpatrick skin classifications. This phenomenon is particularly attributed to the insufficient representation of darker skin categories (Types IV VI), which consequently results in diminished accuracy (Bencevic et al., 2024). It is noted that the ISIC datasets predominantly originate from clinical settings within Europe. The extensive diversity of skin tones encountered in practical applications may not be adequately reflected. HBTMNet utilises Canny-derived edge priors, rendering it sensitive to variations in contrast within input images. Consequently, such contrasts may be diminished in individuals with darker skin tones, thereby potentially undermining the precision of boundary delineation. To mitigate these shortcomings, it is essential to integrate more extensive and representative training datasets, along with the establishment of adaptive preprocessing techniques to ensure consistent performance across diverse populations.

Interminacy estimation is of paramount importance in the clinical implementation of diagnostic tools. In addition to receiving segmentation predictions, clinicians ought to be provided with confidence metrics, particularly in areas characterised by ambiguous boundary delineation. The Binary Inference Segmentation Framework (BISF) generates a binary ambiguity map, which plays a crucial role in this context. It specifically highlights pixels with low confidence in both semantic and boundary predictions. Consequently, it facilitates effective uncertainty visualisation and identifies specific regions that may necessitate more exhaustive clinical evaluation. This representation provides a clear graphical illustration of uncertainty visualisation, aiding in the identification of zones that may warrant closer clinical scrutiny. Future research endeavours will focus on exploring improved methodologies and will investigate augmented uncertainty estimation through the utilisation of calibrated probabilistic outputs. This exploration will incorporate techniques such as temperature scaling and Monte Carlo Dropout during the final stages of segmentation. Therefore, these methodologies facilitate the integration of uncertainty visualisation as an auxiliary decision-support signal alongside the segmentation predictions.

E. Broader Impact and Ethical Considerations

Broadly, automated dermoscopic segmentation systems provide substantial societal advantages by enabling early melanoma detection at scale, particularly in areas with constrained access to specialized dermatological expertise.

Nonetheless, important ethical issues require careful consideration. A primary concern is automation bias, in which overdependence on algorithmic predictions may compromise independent clinical decision-making. Therefore, HBTMNet should function as an adjunct to clinical decision-making, rather than replacing expert assessment.

A further concern is the possibility of deskilling, where continuous reliance on automated systems may erode clinicians' capabilities in manual segmentation and diagnostic assessment. This calls for comprehensive training strategies and periodic evaluations to safeguard proficiency.

Responsible implementation further requires transparency in multiple aspects, including the sourcing of training data, disclosure of system limitations and failure cases, and performance evaluation across varied patient populations. These considerations are fundamental to achieving trust, accountability, and equitable outcomes

VI. CONCLUSION

So here in this paper we have presented HBTMNet, a Hybrid Boundary-Aware Transformer Mamba Network, which is used for segmentation of dermoscopic skin lesion disease. By integrating three complementary boundary-targeting mechanisms, Swin-UMamba's linear-complexity global context encoding, GLR-Net's implicit boundary-ambiguity detection and refinement, and LEDNet-inspired explicit edge guidance via Difference Attention and Edge Generation Modules it has been observed that HBTMNet consistently surpasses individual baselines and prior state-of-the-art methods across all four public benchmarks. The +2.77% IoU and +1.59 px HD95 improvements over GLR-Net on ISIC 2017 confirm that combining implicit and explicit boundary cues with efficient long-range context modelling is a productive and general strategy for precise lesion boundary delineation. We release code and pretrained weights to support reproducibility and further research.

DECLARATIONS

Funding
The authors did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors for this research.

Competing interests
The authors declare no competing financial or non-financial interests relevant to the content of this article.

Ethics approval
Not applicable. This study did not involve the collection of new data from human participants or animals. All experiments were conducted exclusively on publicly available, de-identified, previously published dermoscopic image archives (ISIC 2016, ISIC 2017, ISIC 2018, and ISIC 2019).

Consent to participate / Consent for publication
Not applicable, as no individual person's data, images, or identifiable information are presented in this manuscript.

Data availability

The ISIC 2016, ISIC 2017, ISIC 2018, and ISIC 2019 datasets analysed during this study are publicly available from the International Skin Imaging Collaboration (ISIC) Archive at https://www.isic-archive.com.

Author contributions

S.K.T.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, visualization. R.A.: supervision, conceptualization, writing – review and editing, project administration. Both authors read and approved the final manuscript.

REFERENCES

Aghdam, E. K., et al. (2023).

Attention Swin U-Net: Cross-contextual attention mechanism for skin lesion segmentation. In *Proceedings of ISBI*.

Azad, R., et al. (2023).

Unlocking fine-grained details with wavelet-based high-frequency enhancement in transformers. In *Proceedings of MLMI*.

Bencevic, M., et al. (2024).

Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine, 245*.

Buslaev, A., et al. (2020).

Albumentations: Fast and flexible image augmentations. *Information, 11*(2), 125.

Cao, Y., et al. (2019).

GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of ICCVW*.

Chen, J., et al. (2024).

TransUNet: Rethinking U-Net for medical image segmentation through transformers. *Medical Image Analysis, 97*.

Chu, C.-Y., & Lin, C.-H. (2026).

Skin lesion segmentation based on a boundary-aware multi-scale attention network. *IEEE Access, 14*, 10186–10202.

Codella, N. C. F., et al. (2018).

Skin lesion analysis toward melanoma detection: ISIC 2017 challenge. In *Proceedings of ISBI*.

Codella, N., et al. (2019).

Skin lesion analysis toward melanoma detection 2018: ISIC challenge. *arXiv preprint arXiv:1902.03368*.

Eskandari, S., & Lumpp, J. (2023).

Inter-scale dependency modeling for skin lesion segmentation. *arXiv preprint arXiv:2310.13727*.

Ferreira, P. M., et al. (2012).

An annotation tool for dermoscopic image segmentation. In *Proceedings of the VIGTA Conference*.

Gu, A., & Dao, T. (2023).

Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gutman, D., et al. (2016).

Skin lesion analysis toward melanoma detection: ISIC 2016 challenge. *arXiv preprint arXiv:1605.01397*.

Holmes, G. A., et al. (2018).

Using dermoscopy to identify melanoma. *Federal Practitioner, 35*(Suppl), S39–S45.

Hu, J., Shen, L., & Sun, G. (2018).

Squeeze-and-Excitation networks. In *Proceedings of CVPR*.

Jamil, U., et al. (2016).

Dermoscopic feature analysis for melanoma recognition. In *Proceedings of INTECH*.

Li, Y., et al. (2024).

SUTrans-NET: A hybrid transformer approach to skin lesion segmentation. *PeerJ Computer Science, 10*.

Liu, J., Yang, H., Zhou, H.-Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., Zheng, H., & Wang, S. (2024).

Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining. *In Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 (LNCS vol. 15009, pp. 615–625). Springer. https://doi.org/10.1007/978-3-031-72114-4_59*

Liu, L., et al. (2024).

VMamba: Visual state space model. *arXiv preprint arXiv:2401.13260*.

Liu, Z., et al. (2021).

Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of ICCV*.

Mendonça, T., et al. (2013).

PH2: A dermoscopic image database for research and benchmarking. In *Proceedings of EMBC*.

Naeem, M. A., Yang, S., Saleem, M. A., Javeed, A., & Ahmad, T. (2026).

A hybrid approach for accurate skin lesion segmentation using LEDNet and Swin-UMamba. *Scientific Reports, 16*, Article 5415. https://doi.org/10.1038/s41598-026-38056-y

Ronneberger, O., Fischer, P., & Brox, T. (2015).

U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of MICCAI*.

Schlemper, J., et al. (2019).

Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis, 53*, 197–207.

Sun, X., et al. (2022).

DMA-Net: DeepLab with multi-scale attention for pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems, 23*, 18392–18403.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021).

Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249.

Wu, H., et al. (2022).

FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis, 76*.

Xie, E., et al. (2021).

SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of NeurIPS*.

Xin, C., et al. (2024).

Transformer-guided self-adaptive network for multi-scale skin lesion segmentation. *Computers in Biology and Medicine, 169*.

Yuan, F., et al. (2024).

A bi-directionally fused boundary-aware network for skin lesion segmentation. *IEEE Transactions on Image Processing, 33*, 6340–6353.

Zhang, Z., et al. (2018).

Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters, 15*, 749–753.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018).

UNet++: A nested U-Net architecture for medical image segmentation. In *Proceedings of DLMIA*.