DEEP LEARNING FOR SKIN LESION SEGMENTATION: A REVIEW

Duy Hung BUI[1,*] **and Huy Hoang VU**[2,+]

[1]202412987, Hanoi University of Science and Technology, Hanoi, Vietnam
[2]202412982, Hanoi University of Science and Technology, Hanoi, Vietnam

ABSTRACT

Accurate skin lesion segmentation is a critical prerequisite for Computer-Aided Diagnosis (CAD) systems for melanoma. Over the past decade, Deep Learning architectures have become the gold standard; however, their effective deployment requires sophisticated optimization strategies to address artifacts and fuzzy boundaries. This paper presents a comprehensive review of computational techniques, categorized by their optimization objectives. We first discuss foundational Convolutional Neural Network (CNN) optimization strategies, ranging from structural refinements (Improved U-Net, FCN-based U-Net) and receptive field expansion (DeepLab), to automated hyperparameter tuning via Bayesian SegNet and attention mechanism integration (Attention Gates, CBAMSNet). Subsequently, we analyze the shift towards global context modeling with Hybrid Transformers (FAT-Net), which address the local inductive bias of CNNs. Finally, we evaluate the emerging trend of 2024: State Space Models (Mamba). Specifically, hybrid architectures such as AC-MambaSeg and VM-SwinUnet are critically analyzed to demonstrate their ability to balance segmentation accuracy with linear computational efficiency ($O(N)$). The review also highlights the role of Generative Adversarial Networks (GANs) in data augmentation and boundary refinement, providing a holistic perspective for developing next-generation clinical diagnostic systems.

Introduction

The skin serves as a vital interface between the human body and the external environment, governing essential functions such as temperature regulation and fluid retention. Despite its resilience, the skin is prone to a multitude of pathologies. It is estimated that there are over 3,000 distinct types of dermatological disorders, making skin diseases one of the most prevalent and diagnostically challenging health concerns worldwide. Global Cancer Statistics 2020 states that fatal skin lesions claim thousands of lives annually[1]. More precisely, skin cancer ranks as the third most common human malignancy, with melanoma being its most aggressive and lethal form. Epidemiological data indicates a rapid surge in melanoma incidence over the last three decades. Notably, statistical projections estimated approximately 96,480 new diagnoses in the United States in 2019[2].

Dermoscopy, a non-invasive imaging technique, has improved diagnostic accuracy; however, manual interpretation of dermoscopic images is labor-intensive, subjective, and heavily dependent on the clinician's expertise. Consequently, CAD systems have become indispensable tools in clinical dermatology. Within the CAD pipeline, skin lesion segmentation, the process of accurately delineating the lesion boundary from the surrounding healthy skin, is the most critical prerequisite. A ccurate recognition of melanoma presents significant challenges due to several inherent complexities. Firstly, the low contrast between lesions and the surrounding healthy skin often creates ambiguous boundaries[3,4,5]. Secondly, high variability in patient-specific attributes, ranging from skin pigmentation and texture to lesion morphology, complicates the detection process[6,4,5,7]. Furthermore, image quality is frequently compromised by various artifacts, including body hair, specular reflections, air bubbles, shadows, and inconsistent lighting conditions[3,6,4,7]. Thirdly, the scarcity of high-quality annotated training data poses a severe constraint on the model's generalization capability. Fourthly, the class imbalance problem, where the lesion area is disproportionately smaller than the background, significantly impedes segmentation performance. Notably, these aforementioned occlusions and artifacts are pervasive in standard public dermoscopic datasets. Figure 1 visually exemplifies these impediments, highlighting the complexity involved in precise boundary delineation.

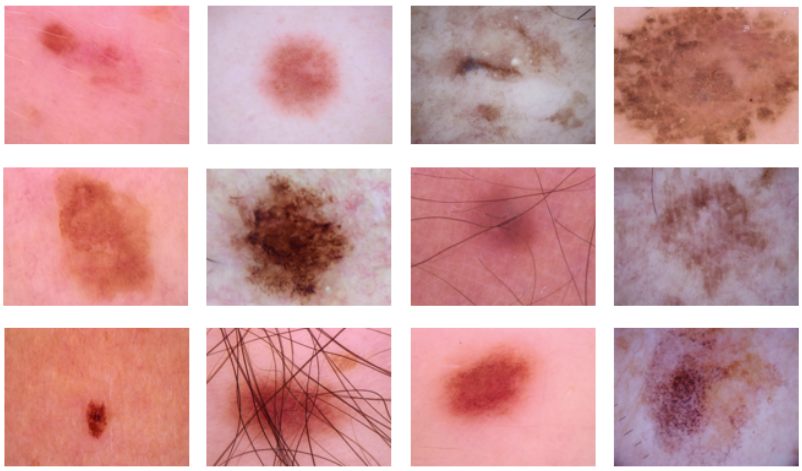

Figure 1. Dermoscopic images of some of the skin diseases from the ISIC 2016 dataset[8].

Early Deep Learning approaches, particularly Fully Convolutional Networks (FCN) and U-Net, established a strong baseline but often struggled with complex lesion heterogeneity. Consequently, recent research has moved beyond "vanilla" architectures, focusing instead on optimizing foundational models. Significant efforts have been made to refine structural designs (Improved U-Net), automate hyperparameter tuning (Bayesian SegNet), and integrate attention mechanisms (Attention Gates, CBAMSNet) to enhance feature selection. Despite these optimizations, CNN-based methods remain inherently limited by their local receptive fields, failing to capture long-range semantic dependencies effectively.

To bridge this gap, Vision Transformers (ViTs) and hybrid architectures like FAT-Net were introduced, leveraging self-attention to model global context. However, the quadratic computational complexity ($O(N^2)$) of Transformers poses a barrier to deployment on resource-constrained clinical devices. This trade-off has catalyzed the emergence of State Space Models (Mamba) in 2024. Offering the global modeling capability of Transformers with the linear efficiency ($O(N)$) of CNNs, Mamba represents a paradigm shift in medical image analysis.

This paper provides a systematic review of these computational techniques. Unlike chronological surveys, we categorize methods based on their optimization strategies: from data augmentation via Generative Adversarial Networks (GANs) and structural optimizations of CNNs, to the global context modeling of Transformers, and finally, the efficiency-driven era of Hybrid Mamba Architectures (e.g., AC-MambaSeg, VM-SwinUnet). Through this analysis, we aim to identify the most promising directions for future real-time diagnostic systems. To visualize this landscape, **Fig. 2** presents the taxonomy of the computational techniques reviewed in this study.

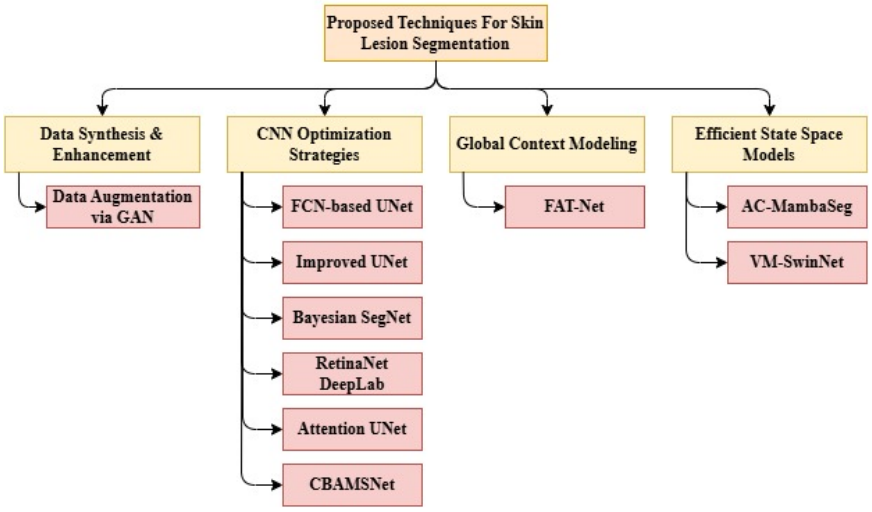

Figure 2. Taxonomy of computational techniques for skin lesion segmentation utilized in this survey, categorizing methods from foundational CNNs to state-of-the-art Hybrid Mamba architectures.

The Computational Pipeline & Data Preparation

The quality of input data plays a decisive role in the performance of deep learning models, adhering to the fundamental principle of "Garbage In, Garbage Out." particularly in dermatological analysis, the presence of extraneous artifacts and severe class imbalance necessitates sophisticated computational strategies before the primary segmentation stage.

Artifact Removal and Normalization

Dermoscopic images are frequently compromised by artifacts such as body hair, gel bubbles, ruler markers, and uneven illumination, which can obstruct lesion boundaries and mislead feature extraction. To mitigate these issues, standard pre-processing pipelines are employed to standardize the input data.

Artifact Removal: Hair occlusion is the most common impediment. The DullRazor algorithm is the standard computational approach for this task. It operates by identifying dark hair structures using generalized grayscale morphological closing operations and subsequently replacing the occluded pixels with values interpolated from the surrounding non-occluded tissue.

Color Normalization: Since dermoscopic images are acquired using various devices under different lighting conditions, color inconsistency can hinder model convergence. Techniques such as Shades of Gray or Gray World algorithms are utilized to correct the illumination and normalize the color distribution, ensuring that the model learns features based on the lesion's pathology rather than lighting variations.

Advanced Data Augmentation Strategy via GANs

A critical bottleneck in training medical image segmentation models is data scarcity and class imbalance. Malignant melanoma samples are often significantly underrepresented compared to benign nevi in public datasets. Traditional augmentation techniques, such as geometric transformations (rotation, flipping) and color jittering, are widely employed to artificially increase the dataset size. However, these methods only produce variations of existing samples and fail to introduce sufficient diversity into the data distribution.To address this limitation GANs have emerged as a breakthrough approach. Fundamentally, a GAN architecture consists of two competing neural networks: a Generator (G) and a Discriminator (D). The Generator aims to synthesize realistic lesion images from random noise distributions, while the Discriminator acts as a binary classifier, attempting to distinguish between real clinical images and the synthetic ones produced by G. Through a min-max adversarial training process, the Generator progressively learns to produce high-fidelity lesion samples that capture the complex texture and color characteristics of real dermoscopic images, thereby enriching the training set.

Beyond pure data synthesis, adversarial learning concepts have been successfully adapted to enhance segmentation performance directly. Dual Discriminators by Lei et al.[4] deploy the segmentation network functions as the Generator, while two distinct Discriminators operate in parallel to supervise the output:

- A Global Discriminator ensures the overall structural consistency of the predicted mask.
- A Local Discriminator focuses specifically on the boundary details, enforcing sharpness at the lesion edges.

By subjecting the model to this "double scrutiny," the system learns to generate segmentation maps that are not only pixel-wise accurate but also visually consistent with expert annotations. This study demonstrates the versatile potential of GANs: acting as both a tool for data augmentation and a segmentation mechanism through adversarial loss.

Optimizing Foundational CNN Architectures

While early CNNs marked a paradigm shift from manual feature extraction to automated learning, vanilla architectures often struggle with the specific challenges of dermoscopy, such as fuzzy boundaries and low contrast. Consequently, the first era of advancement was defined not merely by the application of CNNs, but by the rigorous optimization of foundational architectures. This section categorizes these computational improvements into three strategic domains: structural refinement, receptive field expansion, and attention-based feature selection.

Structural and Hyperparameter Optimization

The transition from patch-based classification to pixel-wise segmentation was pioneered by FCNs. However, standard implementations often faced resolution loss during upsampling. To address this, FCN-based model optimized with a U-Net-like architecture[3] was proposed. By adapting the FCN to learn end-to-end mappings while incorporating symmetric decoding paths, their approach effectively mitigates the loss of spatial information, demonstrating that FCNs can be structurally tuned to handle fine-grained lesion details.

Parallel to structural design, the optimization of training hyperparameters remains a critical yet often overlooked aspect. The SegNet architecture, known for its memory-efficient use of pooling indices, is highly sensitive to parameter initialization. Sahin et al.[9] introduced a Bayesian Optimized SegNet, replacing manual trial-and-error with a probabilistic approach. By employing Bayesian optimization to automatically search for the optimal learning rate and momentum within a complex search space, they achieved superior segmentation accuracy compared to standard SegNet implementations. This highlights that computational optimization in the training phase is as vital as the architectural design itself.

Furthermore, the standard U-Net, despite being the gold standard, can propagate noise through its skip connections. Liu et al.[6] proposed an Improved U-Net, which refines the encoder-decoder path to better preserve boundary information. By optimizing the internal feature concatenation process, this improved variant significantly reduces false positives at the lesion periphery compared to the classic U-Net.

Expanding Receptive Fields via Dilated Convolutions

A fundamental limitation of the aforementioned encoder-decoder models is the reduction of spatial resolution caused by pooling layers, which leads to coarse segmentation boundaries. To expand the receptive field without sacrificing resolution, Bagheri et al.[5] proposed a hybrid framework integrating RetinaNet with DeepLab.

The core innovation lies in the use of Atrous (Dilated) Convolution, which inserts holes (zeros) into the filters to enlarge the field of view. This allows the network to capture multi-scale contextual information—crucial for distinguishing large lesions from healthy skin—without increasing the number of parameters. Moreover, their approach goes beyond pure deep learning by incorporating a Graph-based refinement method in the post-processing stage. This hybrid pipeline, combining the object detection strength of RetinaNet, the semantic context of DeepLab, and the boundary precision of graph theory, exemplifies a sophisticated strategy to overcome the "local view" limitation of standard CNNs.

Feature Selection with Attention and Dynamic Mechanisms

As network depth increases, the risk of learning redundant features (e.g., background artifacts like hair or gel) rises. To enable the network to "focus" on relevant regions, Attention Mechanisms were introduced.

Arora et al.[10] integrated Attention Gates (AGs) into the U-Net architecture. Unlike hard attention which crops images, AGs employ soft attention to automatically learn weight coefficients that suppress irrelevant background regions while highlighting feature-rich lesion areas. This mechanism allows the model to inherently filter out artifacts without explicit pre-processing.

Pushing the envelope of optimization further, recent works like CBAMSNet[?] have combined attention with dynamic computation. This architecture integrates the Convolutional Block Attention Module (CBAM) to refine features along both channel and spatial dimensions. More importantly, it employs Omni-dimensional Dynamic Convolution (ODConv), where convolutional kernels are not static but dynamically adapt their weights based on the input image. This "dynamic" capability allows the network to be lightweight and computationally efficient while maintaining the flexibility to handle the high variability of skin lesion shapes.

The Shift to Global Context with Hybrid Transformers

While the optimized CNN architectures discussed in the previous section significantly improved boundary delineation, they remain bound by the inherent limitation of the convolution operation: the local receptive field. Even with dilated convolutions (as in DeepLab), CNNs struggle to capture long-range dependencies explicitly. In complex dermatological cases, where the diagnosis depends on the relationship between widely separated visual patterns, this local inductive bias becomes a bottleneck.

To address this, the second era of computational techniques is defined by the integration of Transformers, leveraging the Self-Attention mechanism to model global context.

Hybrid Architecture Strategy: The FAT-Net Case

Pure Transformers (like the original ViT) often suffer from localized feature loss due to patch partitioning and require massive datasets to converge. Therefore, the dominant strategy in medical imaging is the Hybrid Approach.

A prime example is FAT-Net (Feature Adaptive Transformers) proposed by Wu et al.[7]. Instead of discarding the CNN entirely, FAT-Net employs a Dual-Encoder Strategy:

- A ResNet-50 backbone is used to extract high-resolution spatial features.
- Transformer layers are embedded at the bottleneck to capture global semantic relationships.

The core computational innovation of FAT-Net is the Feature Adaptive Module (FAM). Recognizing that skin lesions often possess irregular shapes and varying sizes, the FAM utilizes channel attention to adaptively activate relevant feature maps while suppressing noise. Furthermore, to mitigate the high memory consumption typically associated with Transformers, the authors

designed a Memory-Efficient Decoder. This structural optimization allows FAT-Net to achieve a superior Dice score compared to U-Net and Att-Unet, proving that combining the "local detail" of CNNs with the "global vision" of Transformers is more effective than using either in isolation.

Methods

Topical subheadings are allowed. Authors must ensure that their Methods section includes adequate experimental and characterization data[3]necessary for others in the field to reproduce their work.

References

1. Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J. for Clin.* **71**, 209–249, DOI: https://doi.org/10.3322/caac.21660 (2021).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA: A Cancer J. for Clin.* **69**, 7–34, DOI: https://doi.org/10.3322/caac.21551 (2019).
3. Adegun, A. & Viriri, S. Deep learning model for skin lesion segmentation: Fully convolutional network. *Image Analysis Recognit.* 232–242, DOI: https://doi.org/10.1007/978-3-030-27272-2_20 (2019).
4. Lei, B. *et al.* Skin lesion segmentation via generative adversarial networks with dual discriminators. *Med. Image Analysis* **64**, 101716, DOI: https://doi.org/10.1016/j.media.2020.101716 (2020).
5. Bagheri, F., Tarokh, M. J. & Ziaratban, M. Skin lesion segmentation from dermoscopic images by using mask r-cnn, retina-deeplab, and graph-based methods. *Biomed. Signal Process. Control.* **67**, 102533, DOI: https://doi.org/10.1016/j.bspc.2021.102533 (2021).
6. Liu, L., Mou, L., Zhu, X. X. & Mandal, M. Skin lesion segmentation based on improved u-net. *2019 IEEE Can. Conf. Electr. Comput. Eng. (CCECE)* 1–4, DOI: https://doi.org/10.1109/CCECE.2019.8861848 (2019).
7. Wu, H. *et al.* Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Med. Image Analysis* **76**, 102327, DOI: https://doi.org/10.1016/j.media.2021.102327 (2022).
8. Codella, N. *et al.* Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC) (2019). 1902.03368.
9. Şahin, N., Alpaslan, N. & Hanbay, D. Robust optimization of segnet hyperparameters for skin lesion segmentation. *Multimed. Tools Appl.* **81**, 36031–36051, DOI: https://doi.org/10.1007/s11042-021-11032-6 (2022).
10. Arora, R., Raman, B., Nayyar, K. & Awasthi, R. Automated skin lesion segmentation using attention-based deep convolutional neural network. *Biomed. Signal Process. Control.* **65**, 102358, DOI: https://doi.org/10.1016/j.bspc.2020.102358 (2021).

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu[10]. Use the cite command for an inline citation,.

For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file.

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Figures and tables can be referenced in LaTeX using the ref command, e.g and Table 1.

Condition	n	p
A	5	0.1
B	10	0.01

Table 1. Legend (350 words max). Example legend text.