DEEP LEARNING FOR SKIN LESION SEGMENTATION: A REVIEW

Huy Hoang VU 202412982

ABSTRACT

Computer-Aided Diagnosis (CAD) systemes for melanoma require accurate skin lesion segmentation. Over the past decade, Deep Learning (DL) models have become the standard for this problem; howerver, in order to solve artifacts and fuzzy boundaries, there must be sophisticated optimizing strategies. This review shows a comprehensive look to computational techniques, which are categorized by their optimizing objectives. Firstly, we discuss the foundational Convolutional Neural Network (CNN) developments, from structural improvement models including Improved U-Net, Fully Convolutional Network (FCN)-based U-Net, and receptive field expansion models such as Deeplab, to automated hyperparameter management with Bayesian SegNet and attention mechanism integration models like Attention Gates, CBAMSNet. Secondly, we analyze the change to global context model FAT-Net utilising Hybrid Transformers solving the disadvantage of local inductive tendency of CNNs. Finally, we evaluate the new trend of 2024: Mamba. For instance, hybrid models such as AC-MambaSeg and VM-SwimUNet have shown their capability of balancing segmentation accuracy and linear computational efficiency ($O(N)$). The review also highlights the role of Generative Adversarial Networks (GANs) in data augmentation, which provides a perspective for developing next-generation clinical diagnostic systems.

Introduction

The skin plays a vital role as an interface between the human body and the environment, governing essential functions such as the body temperature regulation and fluid retention. In spite of the resilience, the skin is prone to a multitude of pathologies. It is estimated that there are over 3,000 distinct types of dermatological disorders, which makes skin diseases one of the most challenging health concerns worldwide. Global Cancer Statistics 2020 states that annually, fatal skin lesions claim thousands of lives[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 the accuracy; however, manual interpretation of dermoscopic images is labor-intensive, subjective, and dependending heavily on the clinician's expertise. Therefore, CAD systems turn out to be indispensable tools in clinical dermatology. Within the CAD pipeline, skin lesion segmentation, which is the process of accurately delineating the lesion boundary from the surrounding healthy skin, is the most critical prerequisite.

Accurate recognition of melanoma is regarded as a significant challenge due to several inherent complexities. To begin with, the low contrast between lesions and the surrounding healthy skin often creates ambiguous boundaries[3–5]. Secondly, high variability in patient-specific attributes, ranging from skin pigmentation and texture to lesion morphology, complicates the detection process[?, 4–6]. Furthermore, image usually contains various artifacts, including body hair, specular reflections, air bubbles, shadows, and inconsistent lighting conditions[3, 4, 6, 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, showing the complexity in precise boundary delineation.

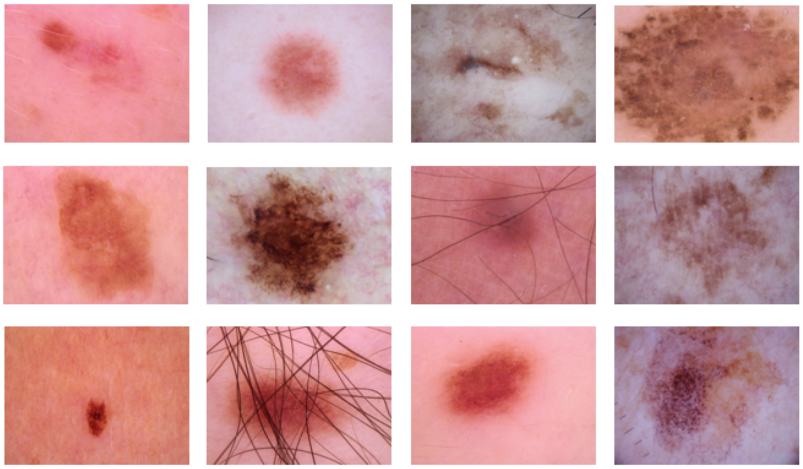

Figure 1. Dermoscopic images of some of the skin diseases from the ISIC 2016 dataset[8].

Early DL approaches such as FCN and U-Net, established a strong baseline but they would be struggled when there was a complex lesion heterogeneity. Consequently, recent research has moved beyond "vanilla" architectures, putting efforts on optimizing foundational models. For example, integrating structural designs (Improved U-Net), automating hyperparameter tuning (Bayesian SegNet), and evaluating attention mechanisms (Attention Gates, CBAMSNet) to enhance feature selection. Despite these solutions, CNN-based methods still remain limitation from their local receptive fields, failing to capture long-range semantic dependencies effectively.

So as to shorten the gap, Vision Transformers (ViTs) and its hybrid models like FAT-Net were introduced, leveraging self-attention to model global context. However, quadratic computational complexity ($O(N^2)$) of ViTs has resulted in a barrier to deploy on low resource clinical devices. This trade-off has catalyzed the emergence of Mamba in 2024, offering not only the global modeling capability of Transformers but also the linear efficiency ($O(N)$) of CNNs.

In this paper, we categorize each approach based on their optimization strategies: from data augmentation by Generative Adversarial Networks (GANs) and structural optimizations of CNNs, to the global context modeling of Transformers, and finally, the efficiency-driven era of Mamba models such as AC-MambaSeg, VM-SwinUnet. Our aim through this analysis is to identify the directions for future real-time diagnostic systems. To visualize this landscape, **Fig. 2** presents the taxonomy of the computational techniques reviewed in this study.

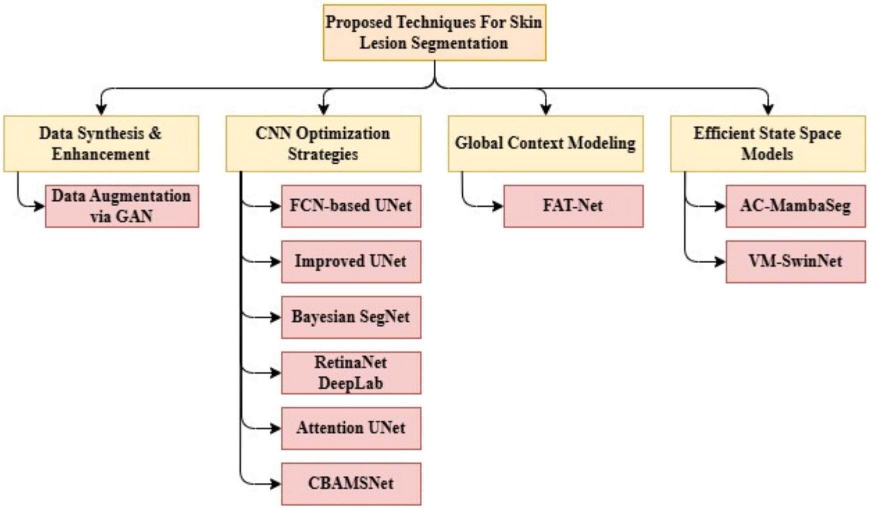
Proposed Techniques For Skin Lesion Segmentation
Data Synthesis & Enhancement
CNN Optimization Strategies
Global Context Modeling
Efficient State Space Models
Data Augmentation via GAN
FCN-based UNet
Improved UNet
Bayesian SegNet
RetinaNet DeepLab
Attention UNet
CBAMSNet
FAT-Net
AC-MambaSeg
VM-SwinNet

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 training data plays a decisive role in the performance of DL models that obey to the fundamental principle of "Garbage In, Garbage Out." particularly in dermatological analysis, the presence of artifacts and severe class imbalance necessitates sophisticated computational strategies before segmentation.

Artifact Removal and Normalization

Artifacts such as body hair, gel bubbles, ruler markers, and uneven illumination in dermoscopic images may prevent lesion boundaries and mislead feature extraction. To deal with these issues, multiple standard pre-processing pipelines are introduced to normalize the training data.

Artifact Removal: Hair occlusion is the most common impediment. The DullRazor algorithm is among the standard computational approaches for the task, 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: Dermoscopic images are taken under different conditions of lights, color inconsistency by a wide variety of devices, preventing model convergence. Techniques such as Shades of Gray or Gray World algorithms are used to standardize the illumination and normalize the color distribution. Such methods adjust the color channels based on the average intensity, ensuring that the skin tones are consistent across the dataset.

Advanced Data Augmentation Strategy via GANs

Class imbalance and lack of data diversity are noticing challenges in training robust skin lesion segmentation models, especially for malignant melanoma, which is always minor in public datasets. Traditional augmentation techniques, such as geometric transformations (rotation, flipping) and color jittering, are widely applied to 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 the limitation, GAN has been employed as advanced data augmentation strategy. To be more precise, a GAN model consists of two competing neural networks: a Generator (G) and a Discriminator (D). G tries to combine realistic lesion images from random noise distributions, while D works as a binary classifier, distinguishing between real images and the synthetic ones produced by G. Through a min-max adversarial training process, G will learn to produce excellent lesion samples capturing the complex texture and color characteristics of real dermoscopic images, which can consideralbly widen the training data.

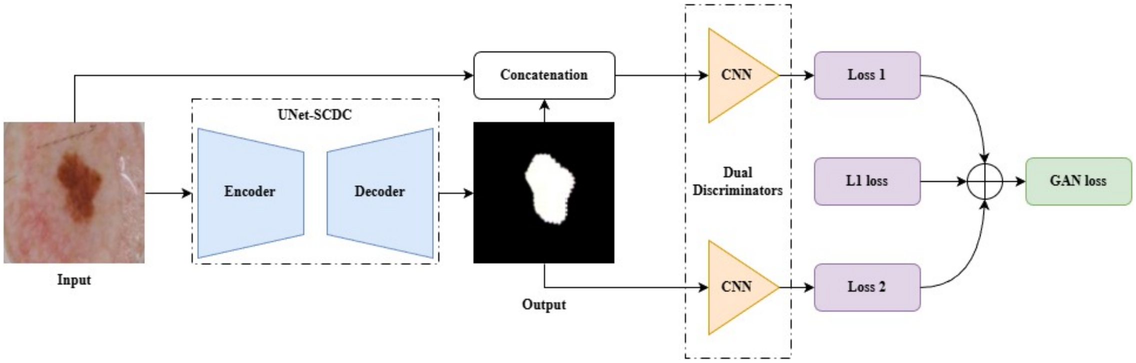

Figure 3. The flowchart of Dual Discriminator GAN architecture. After the generator module, there are two discrimination branches. One uses the concatenation of the generated mask and the original image as input, the integration is performed along the channel. The other just employs the generated mask as input.

Beyond pure data synthesis, adversarial learning concepts have succeeded in adapting to leverage 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 as **Fig. 3**:

- 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.[7] 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[11] 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.[6]. 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.

The Efficiency Revolution with State Space Models (Mamba)

Despite the success of Hybrid Transformers, they come with a significant computational cost: the self-attention mechanism scales quadratically with image size ($O(N^2)$). This complexity renders them challenging to deploy on resource-constrained devices, such as mobile dermoscopy systems. In 2024, a paradigm shift occurred with the emergence of State Space Models (SSMs), specifically the Mamba architecture, which offers the global modeling capability of Transformers with linear computational complexity ($O(N)$).

The Adaptive Hybrid (AC-MambaSeg)

Representing the fusion of CNN flexibility and Mamba efficiency, AC-MambaSeg[12] serves as a state-of-the-art benchmark. **Nguyen et al.** replaced the heavy convolutional backbone with Visual State Space (VSS) blocks. These blocks utilize a 2D Selective Scan (SS2D) mechanism to flatten the image into sequences and scan them in four directions, capturing global context without the quadratic cost of attention matrices.Crucially, AC-MambaSeg does not abandon the optimizations of the past. It integrates Attention Gates (from Era 1) in the skip connections and utilizes Selective Kernel (SK) modules to adaptively adjust the receptive field size. By synergizing the linear speed of Mamba with the feature-focusing capability of legacy attention mechanisms, AC-MambaSeg achieves a balance between high inference speed and segmentation accuracy.

The Ultimate Fusion (VM-SwinUnet)

While AC-MambaSeg targets efficiency, VM-SwinUnet represents the pursuit of maximum accuracy through a "Super-Hybrid" design. This architecture combines Swin Transformer blocks with VSS (Mamba) blocks.The computational rationale here is complementary: Swin Transformers excel at extracting hierarchical features via shifted windows (local-to-global), while Mamba blocks model long-range dependencies across the entire image (global). This fusion mitigates the "local forgetting" problem of pure Mamba models and the heavy computational load of pure Transformers. Although more complex, VM-SwinUnet demonstrates that for critical diagnostic tasks where accuracy is paramount, integrating the strengths of CNNs, Transformers, and SSMs into a unified framework yields the best performance metrics.

Discussion & Comparative Analysis

The evolution from foundational CNNs to modern Mamba architectures illustrates a clear trajectory in computational optimization.

Table 1. Comparative Analysis of Computational Paradigms in Skin Lesion Segmentation based on Optimization Strategies.

Paradigm	Model	Core Mechanism	Key Strength	Complexity
Optimized CNN	Improved U-Net	Structural Refinement	Precise Boundary Delineation	Low $\mathcal{O}(N)$
	DeepLab	Atrous Convolution	Expanded Receptive Field	Low $\mathcal{O}(N)$
Attention CNN	Attention U-Net	Soft Attention Gates	Artifact Suppression (Hair/Gel)	Low $\mathcal{O}(N)$
	CBAMSNet	ODConv + CBAM	Dynamic Feature Adaptation	Low $\mathcal{O}(N)$
Hybrid Transformer	FAT-Net	Self-Attention + FAM	Global Context Modeling	High $\mathcal{O}(N^2)$
Hybrid Mamba	AC-MambaSeg	VSS Block (Selective Scan)	Global Context with Linear Speed	**Linear** $\mathcal{O}(N)$
	VM-SwinUnet	Mamba + Swin Trans.	High-Performance Fusion	Moderate

Conclusion

This review has traversed the landscape of computational techniques for skin lesion segmentation, categorizing them by their optimization objectives rather than mere chronology. We analyzed how GANs address data scarcity, how Optimized CNNs (e.g., FCN, DeepLab) refined structural and contextual learning, and how Transformers (FAT-Net) bridged the global context gap. Finally, we highlighted the 2024 breakthrough of Mamba architectures (AC-MambaSeg, VM-SwinUnet), which achieve linear complexity without compromising accuracy. We conclude that the future of automated melanoma diagnosis lies in Hybrid Systems—architectures that leverage the speed of Mamba, the precision of Transformers, and the localized focus of attention mechanisms to deliver real-time, clinically reliable diagnoses.

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. 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).

7. 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).

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).

11. Wang, C. *et al.* Cbamsnet: A lightweight skin lesion segmentation network with omni-dimensional dynamic convolution and cbam-based multiscale attention. *Biomed. Signal Process. Control.* **113**, 109239, DOI: https://doi.org/10.1016/j.bspc.2025.109239 (2026).

12. Nguyen, V.-T., Pham, V.-T. & Tran, T.-T. Ac-mambaseg: An adaptive convolution and mamba-based architecture for enhanced skin lesion segmentation. *Comput. Intell. Methods for Green Technol. Sustain. Dev.* 13–26, DOI: https://doi.org/10.1007/978-3-031-76197-3_2 (2024).