Skin Lesion Segmentation Using a Selective State Space Network with Spatial Attention Refinement (VSS-SAR)

Saad Adnan Abed

Department of Computer Science, College of Computer Science and Information Technology, University of Anbar, Ramadi, Anbar 31001, Iraq

Abstract

Melanoma segmentation is the task of delineating lesion boundaries in dermoscopic images. This is a challenging task due to the inter-lesion heterogeneity. In this paper, we introduce a segmentation network named VSS-SAR built using the visual state space (VSS) block. The VSS is mainly composed of a selective state space model (SSM), which enables the proposed network to model long-range spatial relationships across the full image extent. Additionally, the proposed network incorporates spatial attention refinement (SAR) modules which are inserted after each VSS block.This is mainly used to preserve fine-grained structural detail that would otherwise be suppressed during long-range dependency modeling. To examine the effectiveness of the proposed model, we conducted set of ablation experiments on a benchmark dataset called ISIC 2017. The model achieved a melanoma segmentation with 92.3 %, and 91.5\%, for sensitivity and Dice score, respectively. In comparison to the literature, the proposed VSS-SAR outperforms competing segmentation methods in sensitivity and yields competitive results across other metrics. Thus, our model can be used as an effective solution to skin lesion segmentation.

Keywords: Medical image segmentation; Melanoma skin cancer; Selective state space model; Spatial attention refinement; Selective scan.

1. Introduction

Cancer is one of the leading causes of death worldwide, and its burden continues to grow. The most recent data from the WHO's International Agency for Research on Cancer (IARC) estimates approximately 20 million new cancer cases and 9.7 million cancer-related deaths globally in 2022 [1]. Looking further ahead, the IARC projects that by 2050, annual cases could reach 35 million — driven largely by population ageing and rising exposure to modifiable risk factors [1].

The skin is the body's largest organ, and its primary job is protecting internal systems from external harm. That same exposure, ironically, makes it one of the most disease- prone organs we have. UV radiation is a well-established driver of skin malignancies, and its effects are worsening as stratospheric ozone continues to deplete. Melanoma sits at the dangerous end of that spectrum. In the United States, the number of cases diagnosed with melanoma approached 97,610 in 2023. Based on gender categorization, 58,120 cases were recorded in men and 39,490 in women, with estimated fatalities reaching 5,420 and 2,570, respectively [2]. Hence, the necessity to effective diagnostic tools becomes urgent to mitigate cases at early stages.

In the context of deep learning developments, the clinical assessment of skin cancer has been substantially improved. Specifically, Lesion segmentation feeds directly into downstream

tasks including classification, treatment planning, and follow-up monitoring in automated diagnosis pipelines [3,4,5]. The real interest in automated approaches has been increased to avoid manual delineation which is slow, costly, and inconsistent across observers [3].
These challenges were addressed using the encoder-decoder architectures, where they work reasonably well for various tasks, especially ones that rely on sequences. However, the repeated downsampling operations in such architectures lead to loss in spatial detail, which cannot always fully recovered by the decoder. To overcome this issue, convolutional backbones with multi-scale feature fusion have been widely proposed in the literature of lesion segmentation. Nevertheless, kernel weights are fixed at inference time, regardless of input content. As a result, the model's ability to adapt to the shape and texture variability that characterizes melanoma lesions is fundamentally constrained by this static behavior.

In this work, we proposed VSS-SAR network which is mainly based on SSMs that was originally introduced by Gu and Dao [7]. In this network image patches are treated as sequences and processed in linear time. This enables the model to learn to selectively retain or discard spatial information depending on the input content. Long-range dependencies get captured this way, and crucially, the quadratic cost of full self-attention is avoided entirely. In this case the spatial resolution is preserved too, since no pooling is involved. On top of that, a spatial attention refinement (SAR) module is inserted after each selective scan block, which focuses the network's attention on boundary regions. This due to the parts that matter most for accurate lesion delineation are lied in these regions.

The rest of this paper is structured as follows. Section 2 covers related work on melanoma and skin lesion segmentation. Section 3 is dedicated for detailing the VSS- SAR architecture. Then, the obtained results, and comparisons to the related models are presented in Section 4. Finally, we conclude the work and suggests future directions in Section 5.

2. Related Works

In the literature of lesion segmentation, early efforts were heavily relied on classical image processing such as the narrowband region-fusion method [8]. In the work of Schaefer et al. [9], a combination of iterative non-lesion pixel estimation with cooperative neural networks was proposed to isolate lesion regions. Zhou et al. [10] introduced a lesion segmentation method based on gradient vector flow algorithm derived from mean-shift clustering. While Wong et al. [11] proposed an iterative random region-merging strategy for the segmentation task. These methods performed reasonably well under controlled conditions, but accuracy degraded noticeably when lesion heterogeneity or imaging artifacts were present. This make classical methods limited when the difficulties exist.

In the deep learning context, many of the aforementioned challenges were addresses, as several works have since been devoted to lesion segmentation. The study presented in the work [12] presented in dense pixel-level prediction without fully connected layers was enabled by the Fully Convolutional Network (FCN). Broad contextual information is aggregated in PSPNet [13] through large-kernel pooling at multiple scales. In [14], the instance segmentation capabilities extended to semantic tasks by the proposed Mask R-CNN which was built on the Fast R-CNN. Large convolutional kernels paired with graph convolutional decoders were explored by Peng et al. [15] in an encoder-decoder design. While SegNet was proposed by Badrinarayanan et al., in which the decoder upsamples via stored max-pooling indices rather than learned transposed convolutions [16]. Multi-scale contextual information is fused in ASPP [17] without sacrificing spatial resolution, and this approach has since been widely adopted in skin lesion segmentation architectures [18]. Bharathi et al. [19] proposed a parallel CNN architecture coupling color map histogram equalization with fuzzy edge detection. In [19], the GLCM and Law's texture features were subsequently refined by a Genetic Algorithm prior to classification.

Attention mechanisms started changing how researchers approached melanoma segmentation, particularly when boundary sensitivity became a priority. Ensemble models were developed by Kavitha et al. [20], built on Auto Correlogram and Binary Pyramid Pattern Filter descriptors. The squeeze-and-excitation module [22] was adapted for segmentation by Roy et al. [21], producing the spatial and channel squeeze-and-excitation (scSE) block. Feature maps get recalibrated along both spatial and channel dimensions simultaneously, which turns out to matter more than recalibrating either dimension alone. CA-Net was proposed by Gu et al. [18], applying spatial, channel, and scale attention jointly in a single network. On ISIC 2018, CA-Net raised the average Dice score from 87.77% to 92.08% relative to its baselines, and thus establishing a strong benchmark for networks based on attention mechanisms.

More recent work has converged on CNN-Transformer hybrid architectures. This was driven by the complementary strengths of local convolutional feature extraction and global self-attention. In [6], the authors proposed O-Net which integrates an attention class feature module (ACFM) with recurrent convolutional units (RCUs). In their proposed network, background noise is suppressed by the ACFM, which directs the network toward lesion-relevant regions. On the other hand, feature maps are iteratively refined by the RCUs through repeated application of the same convolutional layer, and fine spatial context is progressively captured through this process. A Dice score of 87.04% was achieved by O-Net on the ISIC 2017 dataset. The work in [23] adopted a dual-encoder design to compose the CTH-Net model. In this model, Res2Net50 is combined with a transformer branch equipped with dual attention. The two encoders are connected through a multi-domain feature fusion module, and local spatial features are merged with global contextual representations in a way that neither encoder could manage alone. Competitive results were produced by CTH-Net across four benchmark datasets.

Despite this progress, both convolutional and transformer-based architectures carry limitations. Convolutional models simply do not have the global receptive field needed to capture distant spatial correlations. On the other hand, full self-attention in transformer models brings substantial computational overhead that grows fast with image resolution. State space models (SSMs) offer a linear complexity to model global sequence [7]. This is a practical advantage over both aforementioned families. Nevertheless, their application to medical image segmentation is still in early stages, and what they can do for dermoscopic lesion analysis has barely been explored. That gap is what the present work sets out to address.

3. VSS-SAR Network

3.1. Background: State Space Models

SSMs have their roots in continuous-time dynamical systems. An input signal $x(t)$is mapped to an output $y(t)$through a hidden state $h(t)$, as given in Equation (1).

$$h'(t) = Ah(t) + Bx(t), y(t) = Ch(t) \qquad (1)$$

Where A, B, and $C$are learnable parameter matrices that govern how the system evolves and produces output. For discrete-time sequences, the continuous system is transformed via zero-order hold (ZOH) discretization as follows:

$$\bar{h}_t = \bar{A}h_{t-1} + \bar{B}x(t), y(t) = Ch(t) \qquad (2)$$

with $\bar{A} = \exp(\Delta A)$and $\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B$, where Δ is a learnable step size that controls how the continuous dynamics are sampled in discrete time.

The Mamba model [7] extends this framework through a selective scan mechanism (S6). Unlike earlier linear SSMs such as S4 [24], B, C, and Δ are computed as functions of the input rather than fixed globally. This input-dependence is what distinguishes Mamba from earlier linear SSMs. In Mamba, which portions of the sequence to retain and which to suppress are dynamically determined based on current content. For dermoscopic images, that selectivity was particularly valuable. Background skin, hair artifacts, and lesion pixels exhibited markedly different statistical signatures, and a mechanism that could not distinguish between them would have struggled to produce clean segmentation boundaries.

To apply this mechanism to 2D feature maps, spatial features are serialized into patch sequences by the VSS block [25]. Four complementary scan directions are then applied, which are the left-to-right, right-to-left, top-to-bottom, and bottom-to-top. Each direction produces an independent state representation. Before passing the result to the next stage, the network merges all four outputs, ensuring that spatial dependencies are captured regardless of lesion orientation or shape.

3.2. Spatial Attention Refinement

Selective scanning provides global context but does not inherently emphasize the boundary regions that are clinically most critical for accurate lesion delineation. The proposed VSS-SAR addresses this issue through a spatial attention refinement (SAR) module. We inserted SAR module after each VSS block as shown in Figure 1. The SAR module derives an attention map by applying average pooling and max pooling across the channel dimension, concatenating the two resulting descriptors, and passing them through a 7×7 convolutional layer followed by a sigmoid activation, as expressed in Equation 3.

Selective scanning provides global context but does not inherently emphasize the boundary regions that are clinically most critical for lesion delineation. VSS-SAR addresses this through a spatial attention refinement (SAR) module inserted after each VSS block. The SAR module computes an attention map by applying average pooling and max pooling across the channel dimension, concatenating the two descriptors, and passing them through a 7×7 convolutional layer followed by a sigmoid activation as shown in Equation 3.

$$M_{SAR} = \sigma(Conv_{7\times7}\,[AvgPool(F);\, MaxPool(F)]) \qquad (3)$$

where F denotes the input feature map and σ is the sigmoid function. The refined feature map is obtained as $\hat{F} = M_{\mathrm{SAR}} \odot F$, with $\odot$ indicating element-wise multiplication. Two pooling operations are used rather than one. Average pooling captures diffuse structural context, while max pooling responds to sharp, salient activations. These together give the attention map a richer basis than either descriptor alone. Background activations get suppressed, and boundary pixels get amplified. The SAR module has proven effective at recovering the fine boundary detail that selective scanning tends to smooth over.

3.3. Loss Function

VSS-SAR is trained using a combined loss function that weights binary cross-entropy (BCE) and Dice loss, as formulated in Equation 4. The weighting coefficients λ_1 and λ_2 were determined through ablation experiments, which will be detailed in Section 4.

$$\mathcal{L}_{\mathrm{total}} = \boldsymbol{\lambda_1} \cdot \mathcal{L}_{\mathrm{BCE}} + \boldsymbol{\lambda_2} \cdot \mathcal{L}_{\mathrm{Dice}} \qquad (4)$$

The per-pixel prediction errors are penalized as follows:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log\, \hat{y}_i + (1 - y_i)\log\,(1 - \hat{y}_i)] \qquad (5)$$

and the Dice loss directly optimizes the overlap-based evaluation metric:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \hat{y}_i + \varepsilon} \tag{6}$$

where $\hat{y}_i$is the predicted probability for pixel i, $y_i$is the ground-truth binary label, and $\varepsilon$is a small smoothing constant, which is used for numerical stability. Dice loss directly optimizes the overlap metric used at evaluation time. BCE, meanwhile, stabilizes gradient flow during early training, which matters more than it might seem. Class imbalance is typical of dermoscopic data, where lesion pixels represent a small fraction of the total image area, and without BCE the training at early stages tends to be unstable.

3.4. Network Architecture

VSS-SAR follows an encoder-bottleneck-decoder design. The conventional convolutional encoder is replaced with hierarchical VSS blocks as illustrated in Figure 1. The encoder goes through four stages of progressively reduced spatial resolution, and each stage comprised a stack of VSS blocks whose selective scan mechanism aggregates long-range spatial context. Between stages, patch merging layers halve the spatial dimensions while doubling the channel depth — a hierarchical strategy analogous to that used in Swin-style backbones, applied here to state space rather than self-attention computations.

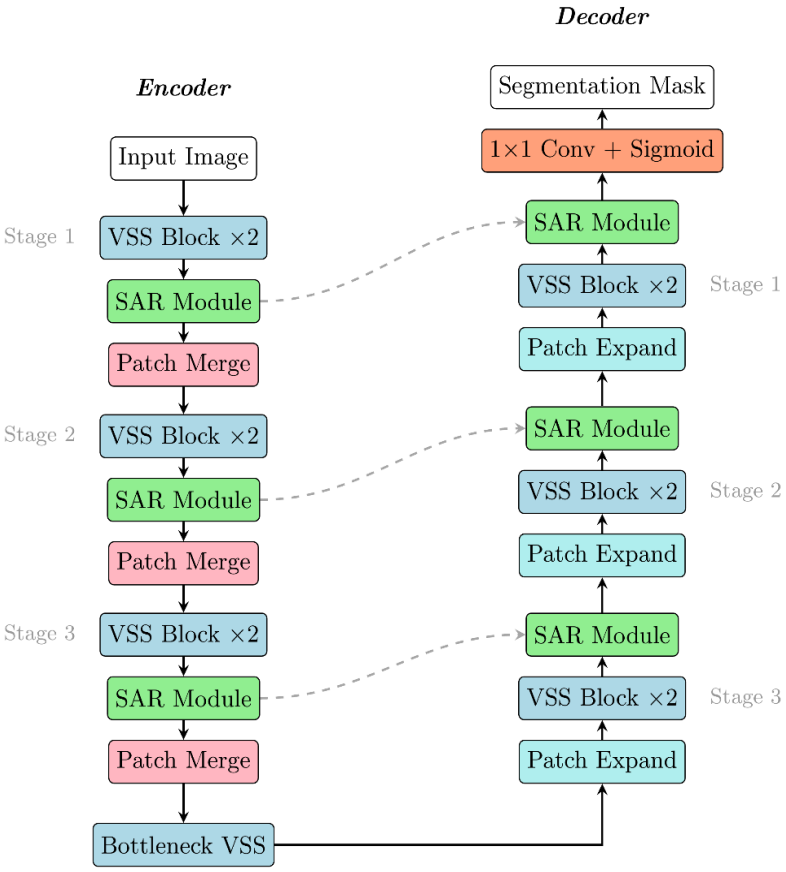

Figure 1. The proposed VSS-SAR network architecture.

The selective scan in each VSS block processes patch tokens in four directions, generating four separate state sequences that are summed and projected. Multi- directional aggregation means the network is not biased toward any particular spatial orientation — important given that melanoma lesions appear in arbitrary positions and orientations.

3.5. Data Preprocessing

Each network is trained on a consistently preprocessed version of the ISIC training data. Hair artifacts — a persistent nuisance in dermoscopic imaging — are suppressed through morphological closing, which combines erosion and dilation. Erosion is defined as:

$$M \ominus N = \{x, y \mid (N)_{xy} \subseteq M\} \qquad (7)$$

where N is a structuring element with origin at coordinate (x, y), and M is the binary region to be eroded. During traversal of N, a pixel in M is set to 1 only when all pixels of N are contained within M; otherwise it is set to 0. Dilation follows vector addition semantics as shown in Equation 8.

$$M \oplus N = \{x, y \mid (N)_{xy} \cap M \neq \emptyset\} \qquad (8)$$

where pixels of M are set to 1 wherever N and M intersect during traversal. Closing — dilation followed by erosion — fills the thin dark strands characteristic of hair without significantly altering the lesion boundary geometry.

Data Augmentation. Dermoscopic datasets are small by the standards of deep learning, and the ISIC 2017 training partition is no different. Without deliberate augmentation, a network of this capacity would overfit within a few epochs. In this work, we adopt a comprehensive strategy that targets the specific sources of variability in dermoscopic images: lesion orientation, scale, illumination, and colour tone. Each training image and its corresponding segmentation mask are subjected to the following transformations:

1. In this works, we applied random horizontal and vertical flipping independently along each axis with probability 0.5. This step is semantically neutral as Melanoma lesions carry no inherent left-right or top-bottom orientation.
2. Random rotation by a uniformly sampled angle from $[-45°, 45°]$was applied to each training sample. Continuous angular sampling was preferred over fixed discrete rotations because it produces genuinely distinct training examples and better covers the orientation space.
3. A scaling by a random factor was also applied to the data. The value of the factor is drawn from a uniform distribution in the period [0.8, 1.2]. This was followed by center cropping to restore the original resolution.
4. To address inconsistent lighting conditions across clinical acquisition sites, we used Brightness and contrast jitter, each adjusted by $\pm 20\%$. Shape and texture are what the network should be sensitive to, not absolute intensity values, and this perturbation pushes it in that direction.
5. Finally, Gaussian noise is injected into the image, with standard deviation $\sigma \in [0,0.05]$

The aforementioned geometric transformations are applied identically to both the image and its segmentation mask except for the Gaussian noise which was applied to the images only. Transformations that alter pixel statistics without changing the underlying lesion geometry, are applied to images only.

4. Experimental Results

This section details how the experiments were conducted in this work. This covers the dataset, evaluation metrics, hardware setup, and the obtained results. First, we report the ablation studies, thereafter, a comparison to recent methods was presented. The computation specifications used in this work, are a Core i9 Intel processor at 3.5 GHz and an NVIDIA GeForce RTX 2080 GPU under Ubuntu 18.04. The learning rate was fixed at 0.0001 across all runs.

4.1. Dataset

The conducted experiments in this work were based on the ISIC 2017 dataset. This and a well-established benchmark for skin lesion segmentation research, which was collected by the the international skin imaging collaboration (ISIC). The resolution of images is varied from 540 x 722 to 4499 x 6748 pixels. Additionally, part of these images contains substantial noise in the form of hair, ruler markings, and specular reflections. The training partition contains 2,000 dermoscopic images paired with expert-annotated binary masks. Figure 2 presents a representative sample of this dataset, which shows each image with its mask.

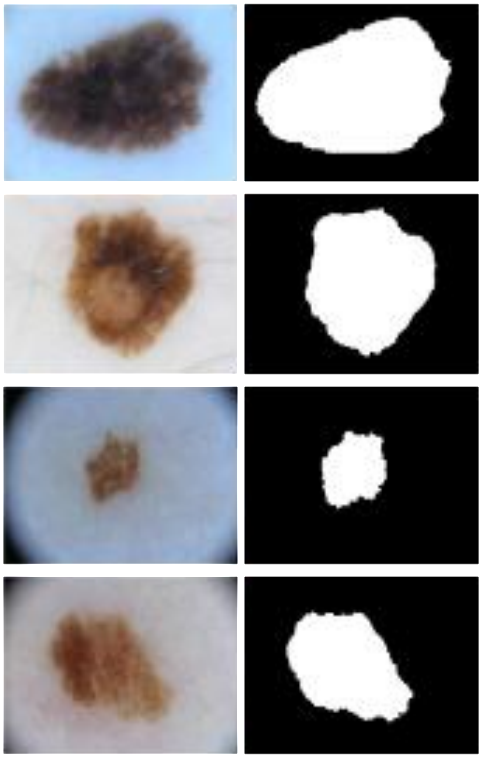

Figure 2. A snippet from the ISIC 2017 dataset.

4.2. Evaluation Metrics

The segmentation performance of the proposed VSS-SAR was evaluated based on five standard metrics. These include sensitivity, specificity, accuracy, JSI, and Dice which are conventionally used to evaluate the effectiveness of the melanoma segmentation. These metrics are mainly calculated based values derived from the confusion matrix. These include, the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The calculation of each of these metrics is expressed as follows:

1. **Sensitivity:** Sensitivity = TP / (TP + FN).
2. **Specificity:** Specificity = TN / (FP + TN).
3. **Accuracy:** Accuracy= (TP+TN)/ (TP+TN+FP+FN).
4. **Jaccard Similarity Index (JSI):** JSI = TP / (FP + TP + FN).
5. **Dice Similarity Coefficient (Dice):** Dice = 2×TP / (2×TP + FP + FN).

4.3. Results

In this work, we conducted two parts of the experiments, which are the ablation study, and comparisons to the state-of-the-art related methods. The ablation study demonstrates the impact of each individual architectural component. This is vital to show whether the used component have contribution to the total performance of the proposed model.

4.3.1. Ablation Study

Seven ablation sets were determined in this study. These are as follows:

1. VSS-SAR without multi-directional scanning (single-direction only).
2. VSS-SAR without the SAR module.
3. VSS-SAR using BCE loss only (λ_1=1, λ_2=0).
4. VSS-SAR using Dice loss only (λ_1=0, λ_2=1).
5. VSS-SAR with λ_1=0.25, λ_2=0.25 (under-weighted combined loss).
6. VSS-SAR with λ_1=0.75, λ_2=0.75 (over-weighted combined loss).
7. VSS-SAR with λ_1=0.5, λ_2=0.5 (balanced combined loss — full model).

For each ablation configuration, we report the sensitivity, specificity, accuracy, JSI, and Dice scores as reported in Table 1. A 4.2% reduction in Dice score was observed when multi-directional scanning was dropped. This confirms that spatial dependencies captured from non-horizontal scan directions carry real segmentation information. When the model is experimented without the SAR module, the dice value is decreased by 3.6%. Though the numerical difference appears small, it held consistently across samples. This indicates that the boundary-aware attention contributes meaningfully to segmentation performance.

Table 1. VSS-SAR ablation experiments on the ISIC 2017 dataset.

Strategy	**Evaluation Metrics**				
	Sensitivity	**Specificity**	**Accuracy**	**JSI**	**Dice**
Single-direction scan only	0.883	0.942	0.926	0.758	0.873
No SAR module	0.887	0.944	0.928	0.771	0.879
BCE loss only (λ_1=1, λ_2=0)	0.901	0.951	0.937	0.796	0.886
Dice loss only (λ_1=0, λ_2=1)	0.895	0.948	0.934	0.789	0.881
λ_1=0.25, λ_2=0.25	0.897	0.950	0.935	0.791	0.884
λ_1=0.75, λ_2=0.75	0.893	0.948	0.933	0.783	0.878
VSS-SAR (λ_1=0.5, λ_2=0.5)	**0.923**	**0.957**	**0.946**	**0.838**	**0.915**

From Table 1, loss function ablations revealed that neither BCE nor Dice alone gets the task

done. The balanced setting $\lambda_1 = \lambda_2 = 0.5$outperformed every other configuration tested. Relying heavily on one loss function on the count of the other one consistently decreases the performance of the model. The two losses complement each other, and the ablation made that clearly observable. Dice score distributions across all configurations are visualized in Figure 3. The full VSS-SAR model not only achieves the highest mean but also exhibited the lowest inter-sample variance. In a clinical setting, stable predictions across the dataset mattered more than strong average performance alone.

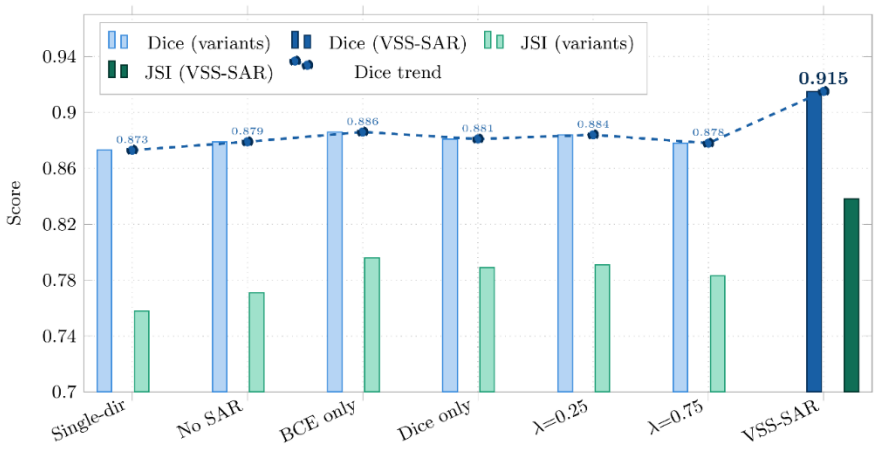

Figure 3. VSS-SAR ablation experiments.

Figure 3 presents Dice and JSI scores across all ablation configurations. Grouped bars are used to represent each configuration, and a dashed trend line is overlaid on the Dice series. Seven sets are arranged along the horizontal axis, with the full VSS-SAR model placed as the last group. Dice scores across the six incomplete configurations are relatively tightly clustered, ranging from 0.873 to 0.886. No single component removal alone produced a dramatic drop, but that does not mean the differences can be dismissed. The lowest Dice of 0.873 was recorded for the Single-dir variant, which removed multi-directional scanning, and the lowest JSI among all configurations at 0.758 was also observed for this same variant. When the SAR module is removed, a marginally better Dice of 0.879 is obtained, yet this remains below the full model.

4.3.2. Comparison with State-of-the-Art Methods

To show the effectiveness of the proposed model, we have compared the VSS-SAR against five recently published methods. These include SegNet [16], Mask R- CNN [14], CA-Net [18], O-Net [6], and CTH-Net [23]. These studies were selected to cover encoder-decoder architectures, attention-based CNNs, and recent CNN-transformer hybrids, and thus providing a broad and current benchmark.

Table 2. Comparison of VSS-SAR with state-of-the-art methods on the ISIC 2017 test set.

Network	**Evaluation Metrics**				
	Sensitivity	**Specificity**	**Accuracy**	**JSI**	**Dice**
SegNet [16]	0.801	0.954	0.918	0.696	0.821
Mask R-CNN [14]	0.848	0.960	0.935	0.743	0.853

CA-Net [18]	0.917	0.955	0.944	0.830	0.907
O-Net [6]	0.897	0.963	0.947	0.803	0.870
CTH-Net [23]	—	—	0.966	0.819	0.943
VSS-SAR	**0.923**	**0.957**	**0.946**	**0.838**	**0.915**

Table 2 and Figure 4 together show that VSS-SAR achieves the best Sensitivity and competitive performance across all other metrics. VSS-SAR achieves the best JSI and the second-best Dice among all compared methods. Against the classical baselines, VSS- SAR outperforms SegNet and Mask R-CNN in Dice by 9.4% and 6.2%, respectively. Among the more recent hybrid architectures, VSS-SAR surpasses O-Net in Dice (0.915 vs. 0.870) while remaining competitive with CTH-Net — a result that is notable given CTH-Net's substantially more complex dual-encoder design and its use of multiple bespoke modules. A small Specificity gap persists in favor of some baselines, but this is consistent with VSS-SAR's higher Sensitivity — a clinically preferable trade-off, since missed melanoma lesions carry far greater risk than false detections. Moreover, the VSS-SAR exhibits lower inter-sample variance than SegNet and Mask R-CNN based on the achieved Dice score as shown in Figure 4. This confirms robustness across the range of lesion types and imaging conditions in the ISIC 2017 test set.

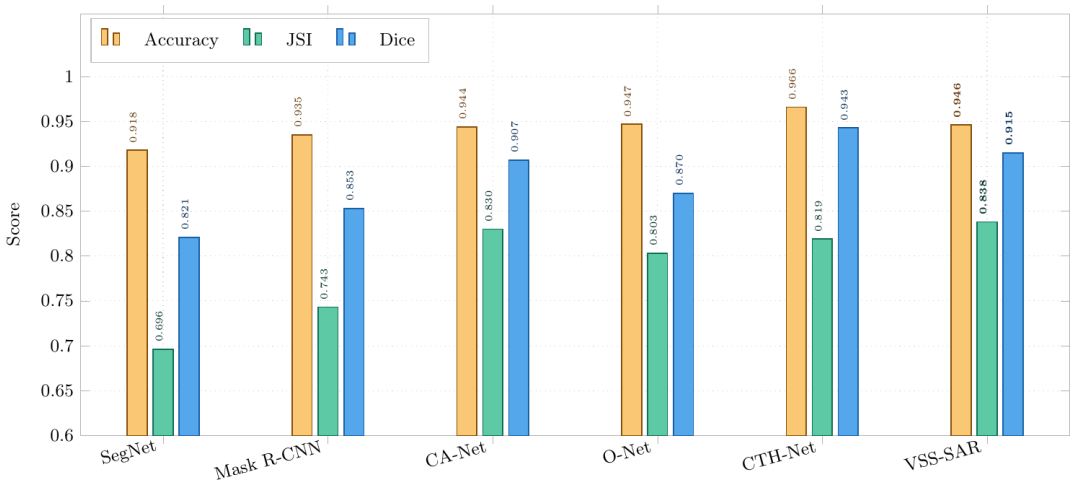

Figure 4. VSS-SAR comparison against other methods.

5. Conclusion

The proposed VSS-SAR was introduced based on SSM architecture for automatic melanoma segmentation. This model shows different computational routes from both convolutional and Transformer-based designs. Global spatial dependencies are modeled at linear time complexity through the selective scan mechanism, while lesion boundary predictions are explicitly sharpened by the spatial attention refinement module. Strong performance was demonstrated by VSS-SAR on the ISIC 2017 benchmark. However, the model behavior on datasets with substantially different imaging protocols or lesion distributions has not yet been characterized and warrants further study. SSM processing which is the base of VSS-SAR scales linearly with sequence length, and that scaling property becomes increasingly valuable as image resolution grows. This is especially important when image resolution increases. Expert labeling still caps how far training data can be scaled, and that bottleneck is not easily resolved under current annotation practices. Semi-automated annotation tools have been identified as a promising path forward. Larger and more diverse training sets could be assembled at a fraction of the manual effort currently required. Architectures like VSS-SAR

would benefit directly from the richer supervision such datasets provide.

References

[1] Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2024). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today, accessed Jan 12, 2024

[2] Melanoma Research Victoria. Key Statistics for Melanoma Skin Cancer, 2023. Available at: https://melanomaresearchvic.com.au/key-statistics-for-melanoma-skin-cancer. Accessed Nov 29, 2023.

[3] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari. Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer. *Neuroscience Informatics*, 2(4):100034, 2022. Elsevier.

[4] N. Melarkode, K. Srinivasan, S. M. Qaisar, and P. Plawiak. AI-powered diagnosis of skin cancer: A contemporary review, open challenges and future research directions. *Cancers*, 15(4):1183, 2023. MDPI.

[5] Y. Zhou, C. Koyuncu, C. Lu, R. Grobholz, I. Katz, A. Madabhushi, and A. Janowczyk. Multi-site cross-organ calibrated deep learning (MuSClD): Automated diagnosis of non-melanoma skin cancer. *Medical Image Analysis*, 84:102702, 2023. Elsevier.

[6] P. Chen, S. Huang, and Q. Yue. Skin lesion segmentation using recurrent attentional convolutional networks. *IEEE Access*, 2022. vol. 10, pp. 94007–94018. doi: ACCESS.2022.3204280.

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

[8] X. Yuan, N. Situ, and G. Zouridakis. A narrow band graph partitioning method for skin lesion segmentation. *Pattern Recognition*, 42(6):1017–1028, 2009. doi: 10.1016/j.patcog.2008.09.006.

[9] G. Schaefer, M. I. Rajab, M. E. Celebi, and H. Iyatomi. Colour and contrast enhancement for improved skin lesion segmentation. *Computerized Medical Imaging and Graphics*, 35(2):99–104, 2011. doi: 10.1016/j.compmedimag.2010.08.004.

[10] H. Zhou, G. Schaefer, M. E. Celebi, F. Lin, and T. Liu. Gradient vector flow with mean shift for skin lesion segmentation. *Computerized Medical Imaging and Graphics*, 35(2):121–127, 2011. doi: 10.1016/j.compmedimag.2010.08.002.

[11] A. Wong, J. Scharcanski, and P. Fieguth. Automatic skin lesion segmentation via iterative stochastic region merging. *IEEE Transactions on Information Technology in Biomedicine*, 15(6):929–936, 2011. doi: 10.1109/TITB.2011.2157829.

[12] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[15] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters -- improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4361, 2017.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. doi: 10.1109/TPAMI.2017.2699184.

[18] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang. CA-Net: Comprehensive attention convolutional neural networks for

explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):699–711, 2021. IEEE.

[19] G. Bharathi, M. Malleswaran, and V. Muthupriya. Detection and diagnosis of melanoma skin cancers in dermoscopic images using pipelined internal module architecture (PIMA) method. *Microscopy Research and Technique*, 86(6):701–713, 2023. doi: 10.1002/jemt.24307.

[20] P. Kavitha, G. Ayyappan, P. Jayagopal, S. K. Mathivanan, S. Mallik, A. Al-Rasheed, M. S. Alqahtani, and B. O. Soufiene. Detection for melanoma skin cancer through ACCF, BPPF, and CLF techniques with machine learning approach. *BMC Bioinformatics*, 24(1):458, 2023. doi: 10.1186/s12859-023-05584-7.

[21] A. G. Roy, N. Navab, and C. Wachinger. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2018*, pp. 421–429. Springer, 2018. doi: 10.1007/978-3-030-00928-1_48.

[22] X. Jin, Y. Xie, X.-S. Wei, B.-R. Zhao, Z.-M. Chen, and X. Tan. Delving deep into spatial pooling for squeeze-and-excitation networks. *Pattern Recognition*, 121:108159, 2022. doi: 10.1016/j.patcog.2021.108159.

[23] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li. CTH-Net: A CNN and Transformer hybrid network for skin lesion segmentation. *iScience*, 27(3):109273, 2024. Elsevier.

[24] J. Wang, W. Zhu, P. Wang, X. Yu, L. Liu, M. Omar, and R. Hamid, "Selective Structured State-Spaces for Long-Form Video Understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6387–6397.

[25] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. arXiv:2401.09417, Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.09417