# YOLO-SCSA: Enhanced YOLOv8 with Spatially Coordinated Shuffling Attention Mechanisms for Skin Cancer Detection

1st Jinyoon Kim
*Department of Computer Science*
*Pennsylvania State University Harrisburg*
Middletown, USA
juk481@psu.edu

2nd Tianjie Chen
*Department of Computer Science*
*New Mexico State University*
Las Cruces, USA
tvc5586@nmsu.edu

3rd Hien Nguyen
*Department of Computer Science*
*Pennsylvania State University Harrisburg*
Middletown, USA
nguyen.hien@psu.edu

4th Md Faisal Kabir
*Department of Computer Science*
*Pennsylvania State University Harrisburg*
Middletown, USA
mpk5904@psu.edu

***Abstract*—Skin cancer is one of the most prevalent and deadliest diseases worldwide. Traditional detection methods, relying on visual examination and biopsy, are time-consuming. Early detection is crucial, as delays can significantly risk patients' lives. Advances in machine learning, particularly in computer vision, have enabled faster and more accurate detection of skin cancer. YOLO (You Only Look Once) is a state-of-the-art model for object detection, known for its high accuracy and speed. In 2023, Ultralytics released the latest version, YOLOv8. This research proposes the YOLO-SCSA model, which enhances performance by integrating both general and domain-specific attention modules. Our SCSA attention module combines mechanisms from previous attention modules and introduces a new branch for richer feature understanding. Additionally, the Center Weighted Masking module improves focus on crucial parts of the feature map, enhancing performance on the skin cancer dataset within the YOLOv8 architecture.**

***Index Terms*—Computer Vision, Skin Lesion Detection, YOLOv8, Object Detection, Spatial Average Pooling, Global Average Pooling, Coordinate Attention, Shuffle Attention, HAM10000**

## I. Introduction

There has been extensive research on deep neural networks (DNNs) for skin cancer detection since the emergence of computer vision in artificial intelligence [**?**], [1]–[4]. DNN-based models have shown strong performance in detecting skin cancer, often matching or surpassing human experts in various studies [3], [5]. These models have been used for both classifying and localizing skin cancer, demonstrating their viability for industrial applications in dermatology.

YOLO (You Only Look Once) [6] is a state-of-the-art object detection model that has been effectively used for skin cancer detection. Prior studies have trained YOLO on skin cancer datasets with impressive results. We chose YOLOv8 [7], the latest version by Ultralytics, as the baseline for our architecture due to its advancements over previous versions.

The attention mechanism is a crucial innovation in computer vision, enhancing neural networks' ability to process images by focusing on relevant features. Originally developed for machine translation [**?**], attention mechanisms have been adapted for various vision tasks, including the Convolutional Block Attention Module (CBAM), which sequentially applies channel and spatial attention to improve model performance [8].

Our research integrates the baseline YOLOv8 model with novel attention modules, combining efficient mechanisms from prior attention modules with a new branch to enhance performance in skin cancer detection. We also explore a center weighted masking module to improve focus on key sections of the feature map, enhancing detection in dermoscopic images. The main contributions of this paper are:

1) Develop a novel attention module integrating mechanisms from prior attention modules with a new branch.
2) Implement a center weighted masking module to enhance YOLOv8's focus on key feature map sections, particularly for dermoscopic skin images.
3) Demonstrate that the integrated attention module and YOLOv8 baseline model outperform the original YOLOv8 and other state-of-the-art attention modules across various performance metrics.

The remainder of this article is organized as follows: Section II reviews recent works on skin cancer detection with DNNs and XAI methods. Section III details our methodology, presenting our proposed architecture

and existing attention modules used for comparison. Section IV describes the environmental settings and experimental results. Section V discusses the comparison, research implications, and limitations. Finally, the conclusion summarizes the paper.

## II. Related Works

Deep learning techniques, such as DNNs, have been widely explored for detecting skin cancer since the mid-1990s [**?**]. These models typically utilize artificial neural networks (ANN) or convolutional neural networks (CNN). For instance, despite a limited dataset, researchers in [9] trained a CNN model for binary skin cancer diagnosis and achieved over 95% accuracy. More recently, YOLO, a CNN model originally developed for object detection, has gained popularity for skin cancer detection. In [10], YOLOv7 demonstrated superior performance for early skin cancer diagnosis, while YOLO-based models in [2] achieved over 98% accuracy in classifying nine types of skin cancer. YOLO's efficiency and real-time processing capability make it suitable for advanced detection tools [11]. Notably, the latest version, YOLOv8, was not utilized in these studies, motivating us to use YOLOv8 as the baseline for our architecture due to its state-of-the-art performance.

Attention mechanisms, such as CBAM, GAM, Shuffle Attention, Coordinate Attention, and ECA, have enhanced feature representation in computer vision models. CBAM applies channel and spatial attention sequentially [8], GAM uses gradient information to focus on salient regions [12], Shuffle Attention combines attention with a shuffle operation to capture cross-dimensional interactions [13], Coordinate Attention encodes spatial information into channel attention [14], and ECA introduces a lightweight mechanism to capture local cross-channel interactions without dimensionality reduction, using a 1D convolution whose kernel size is adaptively determined by the channel dimension [15]. These modules inspired us to integrate powerful attention mechanisms into the YOLOv8 architecture for improved object detection in skin lesions.

Previous research has integrated YOLO with attention modules outside skin cancer detection. For example, in [16], YOLOv3 combined with an improved squeeze-and-excitation [17] module was used for blood cell detection. SC-YOLO [18] enhanced feature extraction for small objects with a cross-stage attention network module. YOLOv8-AM [19] integrated various attention modules for detecting pediatric wrist fractures in X-ray images, improving performance despite increased computational work. These studies suggest that integrating attention modules with YOLO can boost performance.

In summary, YOLO-SCSA integrates a novel attention module, inspired by previous mechanisms, into the YOLOv8 architecture. This combination leverages YOLOv8's strengths and advanced attention techniques to enhance efficiency and performance in skin cancer detection tasks.

## III. Methodology

### A. Baseline Model: YOLOv8

The YOLO series is renowned for its exceptional efficiency and accuracy in object detection due to its single forward pass architecture, which significantly reduces processing time while maintaining high accuracy. YOLOv8, the latest iteration, follows this efficient structure and is divided into three main sections: Backbone, Neck, and Head. The Backbone extracts essential features, the Neck generates multi-scale feature maps, and the Head predicts bounding boxes, objectness scores, and class probabilities using Distribution Focal Loss (DFL) [20] and Complete Intersection over Union (CIoU) for bounding box loss, and Binary Cross-Entropy (BCE) for class loss.

YOLOv8's backbone is based on CSPDarknet53 (CSPNet), originally from YOLOv4, which incorporates Cross Stage Partial (CSP) connections to improve learning capability and reduce computational complexity without compromising accuracy [**?**]. CSPNet's residual connections help mitigate the vanishing gradient problem, enhancing convergence speed and allowing deeper network training. Additionally, YOLOv8 replaces the C3 module of YOLOv5 with the faster C2f module, improving execution speed while maintaining performance [**?**].

The Neck, derived from Path Aggregation Network (PANet), concatenates feature maps from different backbone layers and processes them for the Head [**?**]. PANet enhances multi-scale feature propagation, improving object localization across varying scales. The Spatial Pyramid Pooling Fast (SPPF) module, an optimized version of SPP, is integrated into the backbone, providing multi-scale feature representation with fewer floating-point operations.

YOLOv8's detection head uses BCE loss for classification and a combination of CIoU and DFL for bounding box regression. BCE loss calculates the error between predicted and ground truth class probabilities:

$$\text{BCE}(p_s, t_s) = -\frac{1}{N}\sum_{i=1}^{N}\left[t_s^i \log(p_s^i) + (1 - t_s^i)\log(1 - p_s^i)\right]$$

CIoU loss measures the overlap between predicted and ground truth boxes, considering their distance and aspect ratio:

$$\text{CIoU}(p_b, t_b) = 1 - \text{IoU}(p_b, t_b) + \frac{\rho^2(\mathbf{c}_{p_b}, \mathbf{c}_{t_b})}{c^2} + \alpha v$$

DFL refines bounding box predictions by focusing on the distribution of regression targets:

$$\text{DFL}(p_d, t_b) = \sum_{\text{batches}}\sum_{\text{anchors}} \text{softmax}(p_d) \cdot \text{proj}$$

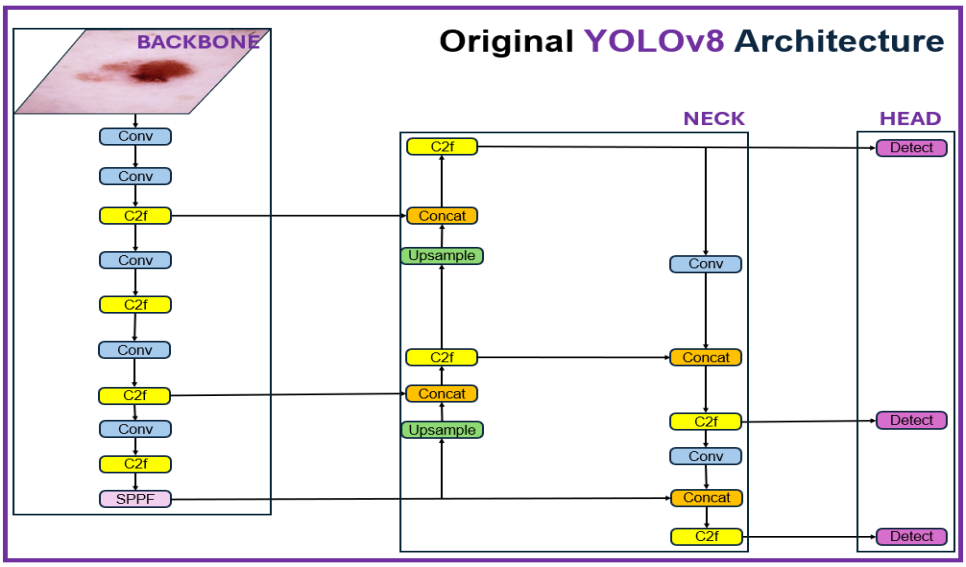


Fig. 1. An overview of the overall architecture.

By integrating these loss functions, YOLOv8 achieves high accuracy and efficiency in object detection tasks. The overall architecture of the basic YOLO model is illustrated in Figure 1.

### B. Attention Modules

**Coordinate Attention** embeds positional information into channel attention to enhance feature representation, especially in mobile networks. It factorizes channel attention into two 1D encoding processes, capturing long-range dependencies along one spatial direction while preserving positional information along the other. This method starts with 1D global pooling operations along vertical and horizontal directions, creating direction-aware feature maps. These maps are concatenated, transformed, and split into tensors that generate attention weights applied to the input feature map. This mechanism captures cross-channel relationships and spatial dependencies, improving object localization and performance in image classification, object detection, and semantic segmentation with minimal computational overhead. These features inspired our integration of Coordinate Attention into the SCSA module.

**Shuffle Attention** combines spatial and channel attention mechanisms efficiently, suitable for environments with limited computational resources. It uses Shuffle Units to process divided input feature maps in parallel, with a unique "channel shuffle" operation facilitating information communication between sub-features. Channel attention uses global average pooling to create channel-wise statistics, while spatial attention employs group normalization. The combined outputs undergo a "channel shuffle" operation, producing an enriched feature map. This method effectively captures spatial and channel dependencies, enhancing performance in image classification, object detection, and semantic segmentation. These attributes led us to integrate Shuffle Attention mechanisms into our SCSA module.

### C. Proposed Method: SCSA

The Spatially Coordinated Shuffling Attention (SCSA) module is designed to integrate the strengths of Shuffle Attention and Coordinate Attention while introducing additional mechanisms to enhance spatial information processing. This module aims to capture both spatial and channel dependencies efficiently and is particularly effective in environments where computational resources are limited.

The SCSA mechanism begins by dividing the input feature map $\mathbf{X}$ of shape $(N, C, H, W)$ into $G$ groups along the channel dimensions:

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_G], \quad \mathbf{X}_k \in \mathbb{R}^{\frac{C}{G} \times H \times W}$$

Each group $\mathbf{X}_k$ is then split into two branches for channel and spatial attention:

$$\mathbf{X}_k = [\mathbf{X}_{k1}, \mathbf{X}_{k2}], \quad \mathbf{X}_{k1}, \mathbf{X}_{k2} \in \mathbb{R}^{\frac{C}{2G} \times H \times W}$$

For channel attention, global average pooling (GAP) is applied to the spatial dimensions of $\mathbf{X}_{k1}$ to generate channel-wise statistics:

$$s = \text{GAP}(\mathbf{X}_{k1}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_{k1}(i, j)$$

These channel-wise statistics are then scaled and shifted using a gating mechanism with learnable parameters, followed by a sigmoid activation function ($\sigma$):

$$\mathbf{X}_{\text{channel}} = \sigma(\mathbf{W}_1 s + \mathbf{b}_1) \cdot \mathbf{X}_{k1}$$

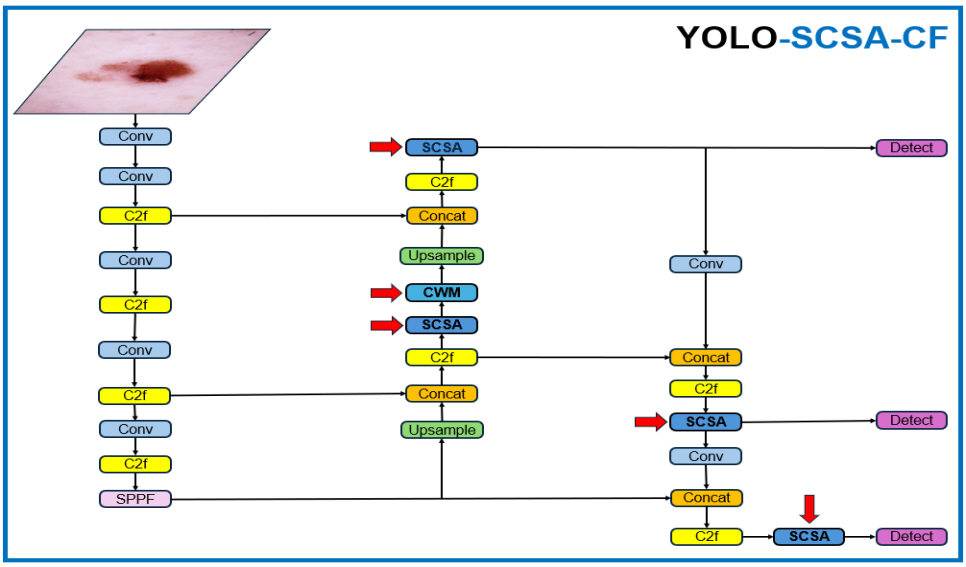


Fig. 2. An overview of the overall architecture for YOLO-SCSA.

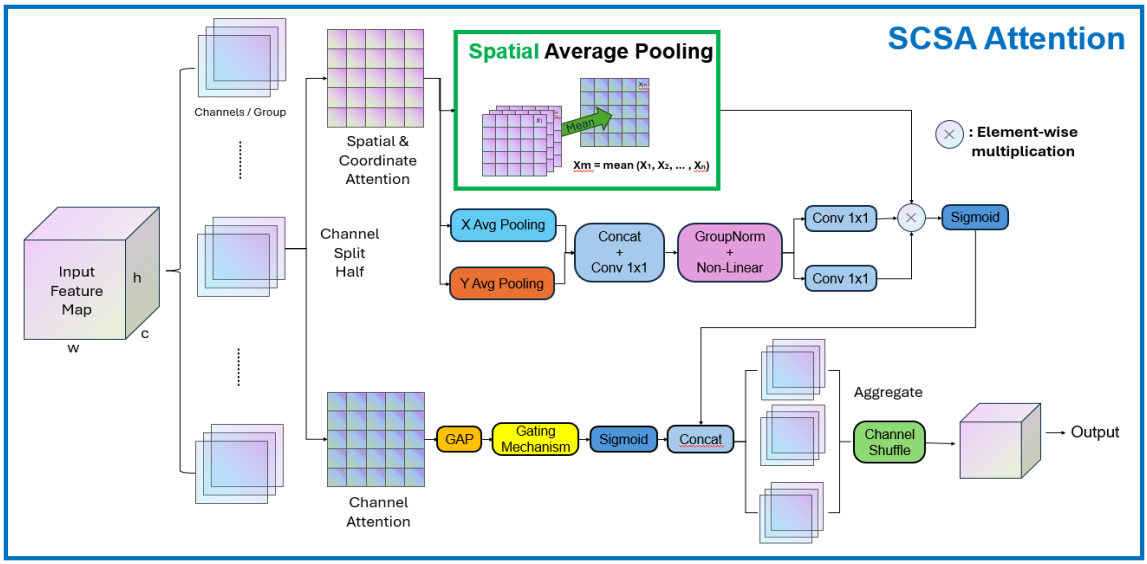


Fig. 3. An overview of SCSA module.

The gating mechanism consists of $\mathbf{W}_1$ and $\mathbf{b}_1$, which are learnable parameters for scaling and shifting.

**Spatial average pooling** is a new branch added in the SCSA module alongside the original mechanism of coordinate attention that only utilizes $x$ and $y$ pooling operations to understand spatial information with channel-encoded information. The model uses the original mechanism of coordinate attention to extract spatially focused information from the feature, enriching spatial information on top of $x$ and $y$ pooling. This spatial average pooling uses the channel dimension of the feature map instead of the width and height dimensions, thus capturing the average value of the channel for each spatial location of the entire feature map. This captures different aspects of spatial dependencies within the entire channels.

To start with the $x$ average pooling step, adaptive average pooling is applied along the horizontal dimension:

$$\mathbf{X}_h = \text{pool}_h(\mathbf{X}_{k2}) = \frac{1}{W}\sum_{j=1}^{W}\mathbf{X}_{k2}(:,:,j)$$

Similarly, adaptive average pooling is applied along the vertical dimension, followed by a permutation to maintain the shape consistency:

$$\mathbf{X}_w = \text{pool}_w(\mathbf{X}_{k2}) = \frac{1}{H}\sum_{i=1}^{H}\mathbf{X}_{k2}(:,i,:)$$

Additionally, spatial average pooling is introduced to enrich spatial information across the entire feature map:

$$\mathbf{X}_s = \text{pool}_s(\mathbf{X}_{k2}) = \frac{1}{C}\sum_{c=1}^{C}\mathbf{X}_{k2}(c,:,:)$$

The pooled features $\mathbf{X}_h$ and $\mathbf{X}_w$ are concatenated along the spatial dimension and processed through a shared $1 \times 1$ convolutional function to combine the spatial information. The transformation through a $1\times1$ convolutional layer, group normalization, and the h-swish activation function is expressed as:

$$\mathbf{Y} = \text{h-swish}(\text{gn}(\text{conv}_{1\times1}(\text{cat}(\mathbf{X}_h, \mathbf{X}_w))))$$

The feature map $\mathbf{Y}$ is then split back into horizontal and vertical components along the spatial dimension:

$$\mathbf{Y} = [\mathbf{Y}_h, \mathbf{Y}_w], \quad \mathbf{Y}_h, \mathbf{Y}_w \in \mathbb{R}^{\frac{C}{2G}\times H\times W}$$

Separate convolutional layers, denoted as $F_h$ and $F_w$, are applied to obtain the attention maps:

$$\mathbf{A}_h = F_h(\mathbf{Y}_h)$$
$$\mathbf{A}_w = F_w(\mathbf{Y}_w)$$

The attention maps are now multiplied element-wise with the spatially pooled feature. The difference between the attention maps are matched with the broadcasting:

$$\mathbf{A}_{hws} = \mathbf{A}_h \cdot \mathbf{A}_w \cdot \mathbf{X}_s$$

The final spatial attention map is obtained by applying a sigmoid activation ($\sigma$) and then multiplied with the original spatial attention branch feature map $\mathbf{X}_{k2}$ to apply the attention weight it had gained during the process:

$$\mathbf{X}_{\text{spatial}} = \mathbf{X}_{k2} \cdot \sigma(\mathbf{A}_{hws})$$

The channel and spatial attention outputs are concatenated along the channel dimension:

$$\mathbf{X}'_k = \text{cat}(\mathbf{X}_{\text{channel}}, \mathbf{X}_{\text{spatial}}) \in \mathbb{R}^{\frac{C}{G}\times H\times W}$$

Finally, all sub-features are aggregated and a channel shuffle operation is applied to enable information communication between different sub-features, producing the final output feature map $\mathbf{Y}$:

$$\mathbf{Y} = \text{ChannelShuffle}([\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_G])$$

This final output maintains the same shape as the input but is enriched with enhanced spatial and channel information. The SCSA module effectively combines spatially coordinated information with additional spatial average pooling and channel dependencies, resulting in improved feature representation and performance across various computer vision tasks. The integration of the module into the YOLOv8 architecture is illustrated in Figure 2, and the inner structure of the SCSA module is illustrated in Figure 3.

### D. Proposed Method: CWM

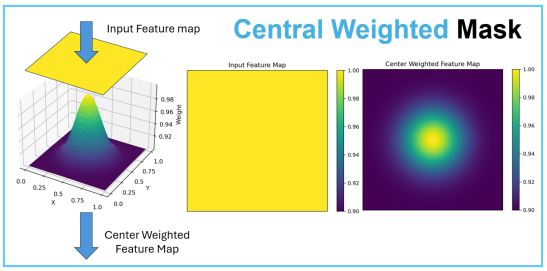


Fig. 4. An overview of CWM module.

The Center Weighted Masking (CWM) module is specifically designed to enhance the processing of skin cancer dermoscopic image data within the architecture. Upon analyzing the dataset, it was observed that the main section of the skin lesion typically appears in the center of the input image, while the bottom and surrounding areas often contain confounding factors such as Dark Corner Artifacts (DCA) [21]. Previous research efforts to remove these areas from the image data directly resulted in negligible improvements in model performance, often considered accidental or even detrimental in our experimental setup.

To address this, we developed the CWM module, which applies a Gaussian-like attention map to focus more on the central region of the input features and reduce the influence of the bottom areas. This attention map is generated using a Gaussian function that smoothly decreases attention from the center to the edges, particularly towards the bottom of the image. By doing so, the CWM module effectively weakens the feature values in the less relevant bottom regions, thus improving the model's focus on the important central lesion area. The construction and application of this Gaussian-based attention map in three dimensions are illustrated in Figure 2. The adjustment of the module into the YOLOv8 architecture is illustrated in Figure 4.

## IV. Experiments

### A. Dataset: HAM10000

For the dataset choice, we selected the HAM10000 dataset [22], which is widely used and well-regarded in dermatoscopic image analysis. The HAM10000 dataset, also known as "Human Against Machine with 10000 training images," comprises 10,015 multi-source dermatoscopic images of common pigmented skin lesions. It includes both melanocytic and non-melanocytic lesions, ensuring comprehensive coverage of conditions encountered in clinical practice. Each image is accompanied by detailed metadata, with over 50% confirmed by pathology and the rest verified through follow-up, expert consensus, or in-vivo confocal microscopy. Additionally, the dataset offers professional-level segmentation labels, making it highly reliable for object detection tasks.

| Model Type | Attention Module | Evaluation Metrics | | | | |
|---|---|---|---|---|---|---|
| | | mAP@50 | mAP@50-95 | Parameters(mil) | GFLOPs | Inference Time(ms) |
| YOLOv8 Small | Basic | 0.783 | 0.585 | 25.8 | 78.7 | 7.2 |
| | Coordinate | 0.805 | 0.583 | 25.9 | 78.8 | 7.7 |
| | ResCBAM | 0.799 | 0.589 | 33.8 | 97.8 | 9.4 |
| | Shuffle | 0.797 | 0.58 | 25.8 | 78.7 | 7.6 |
| | **SCSA** | **0.811** | **0.604** | **25.8** | **78.8** | **8.0** |
| YOLOv8 Medium | Basic | 0.742 | 0.62 | 11.1 | 28.5 | 3.5 |
| | Coordinate | 0.742 | 0.63 | 11.1 | 28.5 | 3.5 |
| | ResCBAM | 0.752 | 0.63 | 16.0 | 38.1 | 4.2 |
| | Shuffle | 0.74 | 0.632 | 11.1 | 28.5 | 3.5 |
| | **SCSA** | **0.78** | **0.639** | **11.1** | **28.5** | **3.6** |

TABLE I
EXPERIMENT RESULTS OF THE ATTENTION MODULES ON YOLOV8 MODEL WITH VARYING SIZES

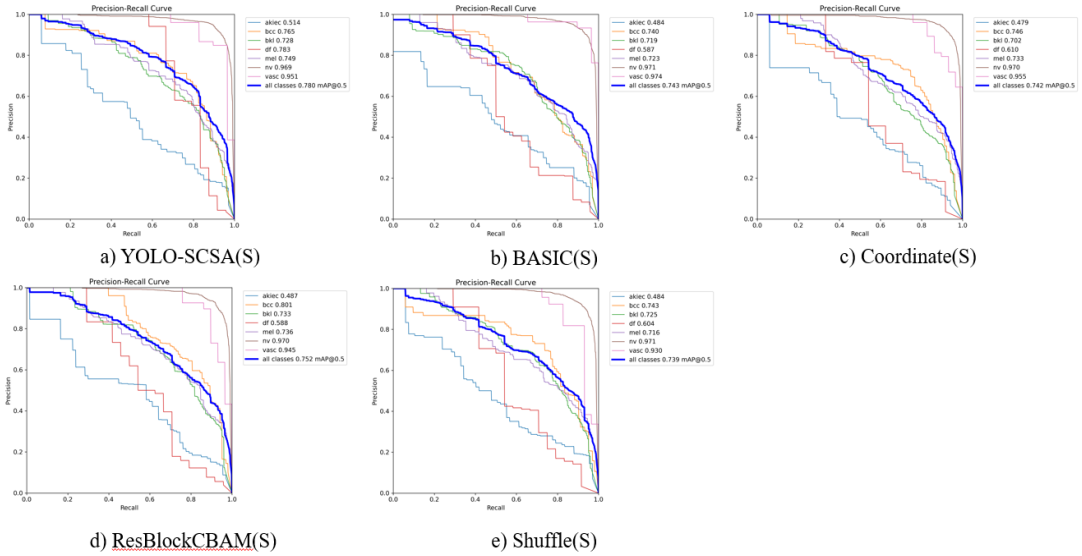


Fig. 5. PR curve metrics for the attention modules in small size YOLOv8.

For our experiments, we utilized mosaic data augmentation, an in-built feature of the YOLOv8 framework, to enhance the robustness and generalization capabilities of our model. We used a fixed input size of 512x512 for all model types to ensure uniformity. The dataset was split into training and validation sets with an 80:20 ratio. This setup provided diverse and varied training samples, allowing us to achieve reliable and accurate results in the automated diagnosis of pigmented skin lesions.

### B. Environmental Setting

For the environmental setting, we used PyCharm IDE and utilized the PyTorch framework for the creation of new modules, since YOLOv8 is built on the same framework. For the graphic card, we utilized the online service Vast.ai, which offers a variety of GPUs for training and running models. Specifically, we chose the RTX 4090 graphic card model for the entire training process of our experiments. This high-performance GPU facilitated efficient processing and accelerated the training times for our models.

In our experiments, we tested five different types of attention module settings: the basic YOLOv8 without any attention module, YOLOv8 with our proposed SCSA and CWM modules, YOLOv8 with Coordinate Attention, YOLOv8 with Shuffle Attention, and YOLOv8 with ResBlockCBAM as suggested in [19]. We conducted the training over 100 epochs and tested both the small and medium sizes of the YOLOv8 object detection model. This approach allowed us to compare the performance of the modules across different model sizes and demonstrate their generalization capabilities.

### C. Evaluation Metrics

The main evaluation metrics of our experiments are categorized into three aspects: performance, efficiency, and speed.

Performance is assessed using the mean Average Precision (mAP) metrics, specifically $mAP_{50}$ and $mAP_{50-95}$. The $mAP_{50}$ measures the precision and recall of the model at an Intersection over Union (IoU) threshold of 0.5, while $mAP_{50-95}$ averages the precision and recall across multiple IoU thresholds (from 0.5 to 0.95 in steps of 0.05). These metrics are widely used in object detection and segmentation tasks within computer vision to evaluate the accuracy and robustness of the model in detecting and classifying objects.

Efficiency is represented by the number of Giga Floating Point Operations per Second (GFLOPs) and the number of parameters in the model. GFLOPs measure the computational complexity and the amount of processing power required for the model to perform a forward pass. The number of parameters, which includes the weights and biases of the model, indicates the model's capacity and complexity. A higher number of parameters can improve the model's ability to learn and generalize from data, but it also increases the computational and memory requirements, potentially leading to overfitting. Therefore, achieving a balance between model complexity and efficiency is crucial.

Speed is represented by the inference time of the model, which measures how quickly the model can make predictions on new data. Lower inference times are desirable for real-time applications, where quick decision-making is critical. In our experiments, we recorded the inference time to evaluate the practical usability of our model in real-world scenarios.

By considering these metrics, we comprehensively evaluated our model's performance, efficiency, and speed to ensure it meets the requirements for practical deployment in clinical settings.

### *D. Experiment Results*

From Table I, it can be observed that the SCSA module outperformed all other attention modules and the basic model in both mAP@50 and mAP@50-95 scores. Specifically, our SCSA module achieved a 4.1% improvement in mAP@50 over the basic YOLOv8 architecture. The Coordinate Attention and Shuffle Attention modules underperformed compared to the original YOLOv8 model in the small size. Although the ResCBAM module closely followed the SCSA module, it was still 2.8% behind in mAP@50 and had significantly more parameters, GFLOPs, and longer inference time. This indicates a lack of computational efficiency and speed compared to the SCSA module. Despite having the highest performance scores, the SCSA module maintained similar GFLOPs, parameters, and inference time as other lightweight modules such as Coordinate Attention, Shuffle Attention, and the basic YOLOv8 model.

In the medium-sized YOLOv8 model, the performance gap between the models decreased, highlighting the efficiency of the SCSA module in small and fast model architectures. The SCSA module still emerged as the best-performing model, with an mAP@50 score of 0.811, 2.8% higher than the basic model. The Shuffle Attention module continued to show the lowest performance among all modules, and the ResCBAM module underperformed compared to the Coordinate Attention module while still having the highest inefficiency in parameters, GFLOPs, and inference time. The Coordinate Attention module followed the SCSA module but remained lower in both mAP@50 and mAP@50-95 scores.

The experimental results demonstrate that the SCSA module achieves the highest performance levels without sacrificing efficiency, avoiding the computational overhead that typically results in high inference times for the architecture.

### *E. PR Curve Metrics*

The YOLOv8 framework provides PR curve graphs for inbuilt evaluation visualization. Figure 5 represents the PR curves for the small model. Each graph displays the PR curves for the basic model and the four types of attention module-integrated models discussed earlier. Each PR curve graph contains the PR curves for each class as well as the average PR curve for all classes in the dataset.

The comparison of the PR curves shows that the curves for each class in the SCSA module drop more slowly and remain higher and more centered than those in other models, especially in the small-sized model. This indicates that the SCSA module maintains a better balance between precision and recall, effectively identifying true positives while minimizing false positives and false negatives. The consistent and higher PR curves suggest that the SCSA module performs uniformly across different classes, providing balanced performance without favoring specific classes. Additionally, the higher curves indicate better generalization to the dataset, making the model robust and reliable.

### *F. Sample Results Comparison*

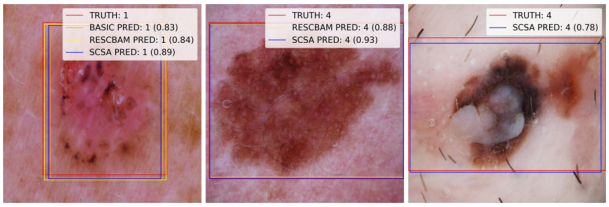


Fig. 6. Sample prediction results of YOLOv8 and the attention modules: Basic, ResBlockCBAM, and SCSA

In Figure 6, there are example prediction results of the basic model, ResBlockCBAM model, and SCSA model. All the sample prediction results came from the small-sized model type. The first sample shows a case where all three types of architecture successfully detected the skin lesion and indicated it with a bounding box based on their prediction results. The second image represents a case where the ResBlockCBAM and SCSA module integrated YOLOv8 architectures could detect the skin lesion, but the basic module could not. The last image represents a case where only the SCSA integrated architecture could detect the skin lesion while the others failed. There was a comparably higher proportion of predictions where the SCSA module could detect the lesion while other types of modules could not. Additionally, the SCSA module more frequently showed higher confidence scores for

its predictions, as depicted in the legend of the sample images.

## V. Conclusion

We proposed the YOLO-SCSA model, which integrates the Spatially Coordinated Shuffling Attention (SCSA) module and the Center Weighted Masking (CWM) module into the YOLOv8 architecture to enhance skin cancer detection. Our comprehensive experiments on the HAM10000 dataset demonstrate that the proposed model outperforms the baseline YOLOv8 model and other state-of-the-art attention modules in both mAP@50 and mAP@50-95 metrics, as well as in various efficiency metrics and speed metrics such as GFLOPs, parameters, and inference time.

There were some limitations in this architecture. While the SCSA integration significantly improves the performance of the architecture, the difference in outperformance decreases as the size of the model increases compared to other attention modules. This is because SCSA utilizes a coordinate attention mechanism that is efficient in small architecture environments that require less computational load. Nonetheless, it still showed better performance than other attention modules integrated into the YOLOv8 framework. Therefore, it may be necessary to experiment with more scenarios where the SCSA module can outperform other attention modules in larger architectures.

Overall, the integration of novel attention mechanisms and specialized modules into the YOLOv8 framework provides a powerful tool for the automated detection of skin cancer, offering a significant advancement over existing models. Future work may explore further optimizations and adaptations of these modules for other medical imaging tasks and datasets. Additionally, enhancements to the module structure could be considered to improve the performance of the proposed attention module in larger architectures, thereby achieving a more significant degree of outperformance compared to other attention modules.

## References


[1] P. Tschandl, G. Argenziano, M. Razmara, and J. Yap, "Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features," *British Journal of Dermatology*, vol. 181, no. 1, pp. 155–165, Oct 2018.

[2] N. Aishwarya, K. M. Prabhakaran, F. T. Debebe, M. S. S. A. Reddy, and P. Pranavee, "Skin cancer diagnosis with yolo deep neural network," *Procedia Computer Science*, vol. 220, pp. 651–658, Jan 2023.

[3] M. Dildar *et al.*, "Skin cancer detection: A review using deep learning techniques," *International Journal of Environmental Research and Public Health*, vol. 18, no. 10, p. 5479, May 2021.

[4] S. Vesal, N. Ravikumar, and A. Maier, "Skinnet: A deep learning framework for skin lesion segmentation," in *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, 2018, pp. 1–3.

[5] H. Ghosh, I. S. Rahat, S. N. Mohanty, J. V. R. Ravindra, and A. Sobur, "A study on the application of machine learning and deep learning techniques for skin cancer detection," *Journal Not Specified*, Jan 2024.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016. [Online]. Available: https://arxiv.org/abs/1506.02640

[7] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with yolov8," 2024. [Online]. Available: https://arxiv.org/abs/2305.09972

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018. [Online]. Available: https://arxiv.org/abs/1807.06521

[9] D. Mandache, E. Dalimier, J. Durkin, C. Boceara, J. Olivo-Marin, and V. Meas-Yedid, "Basal cell carcinoma detection in full field oct images using convolutional neural networks," *HAL*, Apr 2018.

[10] N. A. AlSadhan, S. A. Alamri, M. Maher, and O. Bchir, "Skin cancer recognition using unified deep convolutional neural networks," *Cancers*, vol. 16, no. 7, pp. 1246–1246, Mar 2024.

[11] H. F. Hasya, H. H. Nuha, and M. Abdurohman, "Real time-based skin cancer detection system using convolutional neural network and yolo," *IEEE Xplore*, Sep 2021, available: https://ieeexplore.ieee.org/document/9649224.

[12] H. Hu, F. Wang, J. Su, H. Zhou, Y. Wang, L. Hu, Y. Zhang, and Z. Zhang, "Gam : Gradient attention module of optimization for point clouds analysis," 2023. [Online]. Available: https://arxiv.org/abs/2303.10543

[13] Q.-L. Z. Y.-B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," 2021. [Online]. Available: https://arxiv.org/abs/2102.00240

[14] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," 2021. [Online]. Available: https://arxiv.org/abs/2103.02907

[15] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1910.03151

[16] C. Liu, D. Li, and P. Huang, "Ise-yolo: Improved squeeze-and-excitation attention module based yolo for blood cells detection," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 3911–3916.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2019. [Online]. Available: https://arxiv.org/abs/1709.01507

[18] Y. Shi, X. Li, and M. Chen, "Sc-yolo: A object detection model for small traffic signs," *IEEE Access*, vol. 11, pp. 11 500–11 510, 2023.

[19] C.-T. Chien, R.-Y. Ju, K.-Y. Chou, E. Xieerke, and J.-S. Chiang, "Yolov8-am: Yolov8 with attention mechanisms for pediatric wrist fracture detection," 2024. [Online]. Available: https://arxiv.org/abs/2402.09329

[20] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," 2020. [Online]. Available: https://arxiv.org/abs/2006.04388

[21] S. W. Pewton, B. Cassidy, C. Kendrick, and M. H. Yap, "Dermoscopic dark corner artifacts removal: Friend or foe?" *Computer Methods and Programs in Biomedicine*, vol. 244, p. 107986, Feb. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2023.107986

[22] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, Aug. 2018. [Online]. Available: http://dx.doi.org/10.1038/sdata.2018.161