iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73344

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Spatial Gated Fusion Attention in ConvNeXt: A Novel Dual-Branch Architecture for Lightweight Image Classification

A. Haq[1], Liu Haoran[2], Imdadul Haque Enan[3], Rana A.[4], Hui Zong[5]
Faculty of Computer and Software Engineering, (HYIT) Huaiyin institute of technology

Abstract: ***To enhance the feature extraction capability and computational efficiency of Convolutional Neural Networks (CNNs) in image classification tasks, this paper proposes a novel attention-augmented architecture, SGFA-ConvNeXt, based on the ConvNeXt backbone. The model embeds Spatial Gated Fusion Attention (SGFA) modules at critical transition points of each stage. These modules adopt a dual-branch parallel structure to model salient features along spatial and channel dimensions. The spatial branch combines multi-scale pooling with depthwise convolution to effectively capture long-range dependencies, while the channel branch utilizes global average pooling to recalibrate channel weights for feature refinement. Ultimately, the two branches are fused via a gating mechanism and residual connection, enhancing representational capacity while preserving gradient stability. Experimental results on the CIFAR-10 dataset demonstrate that SGFA-ConvNeXt improves classification accuracy by over 2% compared to the ConvNeXt-Tiny baseline, with only a marginal increase in FLOPs. Moreover, it shows competitive performance among various advanced CNN architectures. Ablation studies further validate the complementary nature of the spatial and channel attention paths in SGFA, underscoring its effectiveness in performance enhancement under low computational cost. This method offers a novel design strategy for efficient image classification in resource-constrained scenarios.***

Keywords: Image Classification; SGFA; Dual-Branch Modelling; Spatial Gated Fusion; Attention Mechanism.

I. INTRODUCTION

Image classification is a fundamental task in computer vision, widely applied in autonomous driving, medical imaging, security surveillance, and more. In recent years, deep learning—particularly Convolutional Neural Networks (CNNs) [1]—has achieved breakthroughs in image classification by learning hierarchical features and automatically extracting effective representations from raw images, significantly boosting accuracy.

However, despite many efficient CNN models and their impressive performance, current classification methods still face several challenges:

1) Resource Intensity: With increasing image size and complexity, traditional CNNs often demand substantial computation and storage. Reducing computation without sacrificing accuracy remains a key problem.

2) Feature Representation Limitations: Existing models struggle to balance global semantics and local details, limiting robustness under complex backgrounds, lighting variations, or multi-scale objects.

3) Diverse Feature Sensitivity: As datasets become more varied, models show inconsistent sensitivity to features like texture, colour, and shape. Enhancing feature capture comprehensively is still a major research direction.

To address these issues, this study introduces a new image classification method by enhancing the ConvNeXt architecture [2] with a Spatial Gated Fusion Attention (SGFA) module. This module introduces both spatial and channel attention mechanisms at various network levels to improve feature representation and classification accuracy. SGFA employs efficient depth wise separable convolutions and layer normalization to minimize computational cost while maintaining high precision.

A. Main Contributions

1) Proposes the SGFA module combining spatial and channel attention to boost feature expressiveness.

2) Designs a scheme to embed SGFA into ConvNeXt, significantly improving classification performance with minimal extra computation.

3) Conducts extensive experiments on CIFAR-10 and benchmarks SGFA-ConvNeXt against popular models.

4) Performs ablation analysis to assess each component's contribution, offering valuable guidance for future research.

This work not only enhances classification performance but also proposes an optimization pathway for deep models in limited-resource environments.

II. RELATED WORK

A. Development of CNNs

CNNs have become central to image classification. Early models like VGGNet [3] stacked small kernels to learn deep features but suffered from parameter redundancy. ResNet [4] introduced residual connections to mitigate gradient vanishing, enabling deeper networks. Lightweight architectures like MobileNet [5] and ShuffleNet [6] used depthwise separable and grouped convolutions to reduce computation while maintaining performance.

Despite success, CNNs mainly rely on local receptive fields and inherently struggle to model global semantics and long-range dependencies.

B. Attention Mechanisms

To address CNN limitations, attention mechanisms have been introduced. SE-Net [7] pioneered global average pooling for channel-wise feature recalibration. CBAM [8] added spatial attention, allowing models to focus on salient regions. However, most combine spatial and channel features via concatenation or summation, leading to redundancy or inefficient fusion.

Recent strategies explore parallel branches, gating, and complementary attention to further improve feature selection.

C. Transformer Architectures

Transformers [9] have gained traction in vision tasks for their global modeling capabilities. Vision Transformer [10] and Swin Transformer [11] set new benchmarks, yet their self-attention incurs high computational cost and needs large datasets—limiting performance on small datasets like CIFAR-10.

D. ConvNeXt Development

ConvNeXt bridges the gap between Transformers and CNNs by integrating design elements like LayerNorm, GELU activation, and scaling factors, while maintaining convolutional efficiency. Though ConvNeXt achieves competitive performance, it lacks explicit attention mechanisms, limiting feature selectivity.

E. Our Contribution

In lightweight model design, improving perception of key regions and semantic channels without excessive computation has become a focus. Spatial attention captures local structure and multi-scale objects; channel attention enhances abstract, class-relevant features.

The proposed SGFA module adopts parallel branches for spatial and channel modeling and fuses them using residual connections and depthwise convolution. Integrated at ConvNeXt's key stages, this design boosts representational diversity and offers a new lightweight classification paradigm.

III.METHODOLOGY

This section details the SGFA-ConvNeXt model, an enhanced ConvNeXt integrating a Spatial Gated Fusion Attention (SGFA) module to jointly strengthen spatial and channel features while maintaining low computational cost.

A. Overall Network Architecture

As shown in the diagram, the model accepts standard RGB images and applies an initial downsampling module (Conv 4×4 + LayerNorm). This feeds into the ConvNeXt backbone consisting of multiple ConvNeXt Blocks. At the end of each stage, an SGFA module is inserted for attention enhancement, followed by downsampling to reduce spatial dimensions and increase channels.

The final stage applies Global Average Pooling, followed by LayerNorm and a fully connected classification head to output probabilities. By embedding SGFA modules at key nodes, the model balances efficiency with strong discriminative capability and robustness.

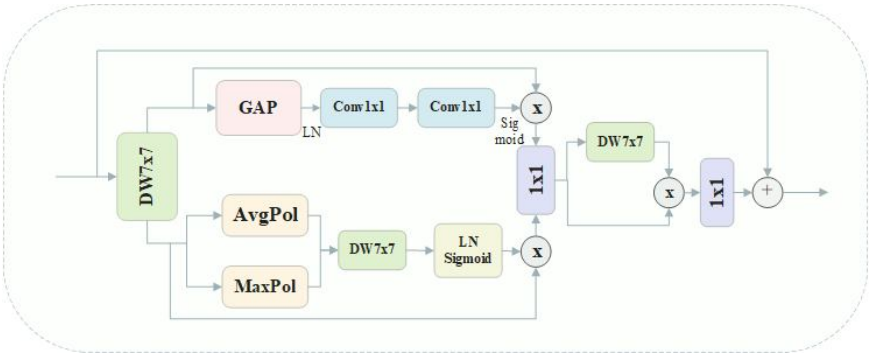


Fig. 1 SGFA Module Architecture Overview

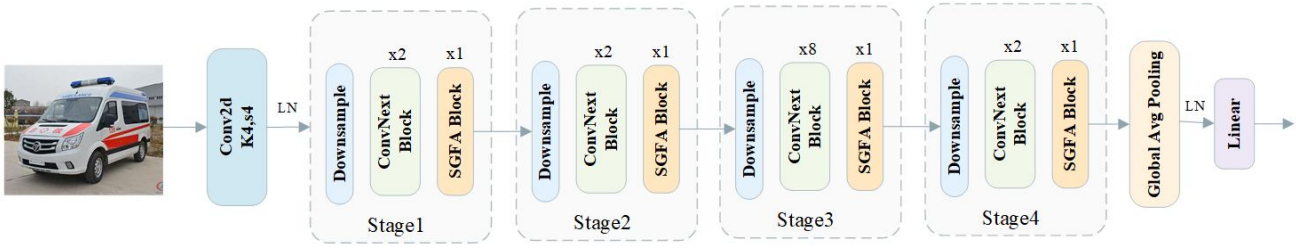


Fig. 2 SGFA-ConvNeXt Network Architecture

B. ConvNeXt Backbone

ConvNeXt is a high-efficiency hierarchical CNN inspired by Transformer design. It uses:

- 7×7 depthwise separable convolutions (DwConv)
- A [3, 3, 9, 3] block structure for pyramid features
- Downsampling strategy reducing spatial resolution while doubling channels:

Let X be the input feature map and Y the downsampled output. ConvNeXt achieves strong baseline performance through unified blocks and convolutions, serving as an excellent base for classification.

C. SGFA Module Design

The SGFA module has two parallel branches: spatial and channel attention, which are fused with residual connections.

1) Spatial Attention Branch

- Uses multi-scale pooling (MaxPool + AvgPool) to capture local/global spatial responses:
  $M_s = \text{Sigmoid}(\text{Normalize}(\text{Conv7x7}(\text{Pool}(\text{Max}+\text{Avg}))))$ （1）
- Produces spatial attention mask M_s, which is elementwise multiplied with input features to yield enhanced spatial features.

2) Channel Attention Branch

- Applies Global Average Pooling (GAP) to compress spatial dimensions into channel descriptors.
- Uses a 1×1 convolution bottleneck (reduce-then-expand) and activation to generate attention mask M_c:
  $M_c = \text{Sigmoid}(\text{FC}(\text{Activation}(\text{FC}(\text{GAP}(X)))))$ （2）
- Multiplied elementwise with input to produce channel-enhanced features.

3) Fusion and Residual Connection

- Outputs from both branches are concatenated, passed through a 1×1 convolution to reduce channels, and added to the original input via residual connection:

$\text{Output} = \text{Input} + \text{Conv1x1}([M_s * X, M_c * X])$ （3）

This fusion preserves shallow features and enriches high-level semantic representations, improving gradient flow and model stability.

D. Efficient Spatial Feature Fusion

To avoid high computational cost, SGFA modules are only inserted at the last block of each stage, targeting key feature transition points. The use of 7×7 DwConv ensures low overhead. This selective strategy significantly improves classification performance while maintaining efficiency.

IV. EXPERIMENT

A. Dataset And Metrics

Experiments are conducted on the CIFAR-10 benchmark, comprising 60,000 32×32 color images in 10 categories (6,000 per class), with a 50,000/10,000 train/test split.

We evaluate using:

1) Accuracy
2) Precision
3) Recall
4) F1-score
5) Model complexity (Params, FLOPs)

B. Implementation Details

Training is performed using an NVIDIA RTX 4060 GPU with Automatic Mixed Precision (AMP). Optimizer: SGD with momentum 0.9, weight decay 5e-2, initial LR 5e-4, cosine annealing for decay, and early stopping to prevent overfitting. Batch size: 64, Epochs: 100. Data augmentations: random flip, crop, and colour jitter.

C. Ablation Story

To verify the contribution of each component within the proposed SGFA module, we conducted ablation experiments on the CIFAR-10 dataset. We compared five configurations:

Baseline: The original ConvNeXt-Tiny block.

1) +CA: Baseline with added Channel Attention.
2) +SCPA: Baseline with Spatial and Channel Parallel Attention.
3) +SGU: Baseline with added Spatial Gated Unit.
4) SGFA (Full): Baseline with both SCPA and SGU integrated.

All models were trained using the same hyperparameters: 100 training epochs, a batch size of 64, the AdamW optimizer, a learning rate of 5e-4, and a weight decay of 5e-2.

TABLE I
ABLATION RESULTS OF SGFA COMPONENTS.

Method	FLOPs	Acc
Baseline	0.091	0.8850
+CA	0.091	0.8861
+SCPA	0.094	0.8971
SGFA	0.109	0.9088

D. Comparison with Classical Classification Network Models

To thoroughly evaluate the performance advantages of the proposed SGFA-ConvNeXt model, we conducted a detailed comparative analysis against six mainstream image classification models, including:

1) ResNet50 (a representative deep residual network),
2) MobileNetV3 (a benchmark lightweight CNN),
3) EfficientViT [13] (an efficient Vision Transformer model),
4) ShuffleNetV2 the original ConvNeXt-Tiny (used as the baseline for this study).

Experiments were conducted on the CIFAR-10 dataset [14], using the same experimental environment and hyperparameter settings. Key performance metrics evaluated include Accuracy, Precision, Recall, F1-score, and Parameter count (Params).

TABLE II
COMPARISON OF CLASSIFICATION MODEL PERFORMANCE.

Method	Loss	Acc	P	R	F1	Params
ResNet50	0.5690	0.8231	0.8233	0.8227	0.8229	23.5
MobileNetv3	0.015	0.8441	0.8440	0.8413	0.8421	1.24
EfficientViT	0.3413	0.8825	0.8823	0.8825	0.8820	7.5
Swin-Transformer	0.4487	0.8798	0.8796	0.8798	0.8796	27.5
ShufflenetV2	0.3997	0.8665	0.8658	0.8665	0.8659	1.26
ConvNext	0.3737	08807	0.8816	0.8807	0.8802	29

In a comprehensive comparison, the proposed SGFA-ConvNeXt model achieved higher accuracy than the other network models, with a 2% improvement over the original baseline model. It strikes a good balance between classification accuracy and model parameter size.

E. Visualization of Experimental Results

In this section, we provide a detailed visualization and analysis of the experimental results. The experiments were conducted to evaluate and compare the performance of seven representative models—ResNet50, MobileNetV3, ShuffleNetV2, Swin Transformer, EfficientNet, ConvNeXt, and the proposed SGFA-ConvNeXt—on the widely used CIFAR-10 dataset

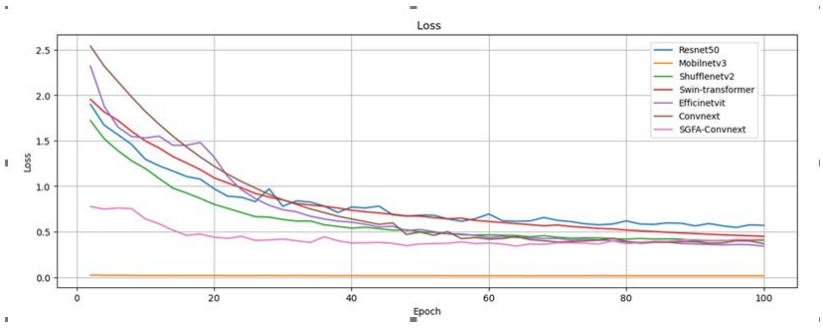


Fig. 3 Training loss comparison across models.

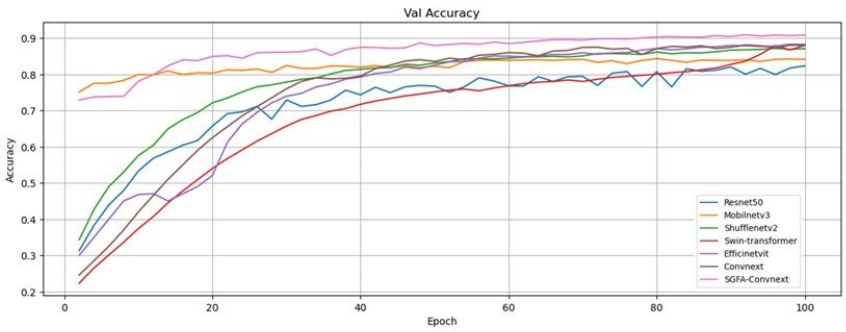


Fig. 4 Validation accuracy trends across models.

V. CONCLUSIONS

To address the issues of insufficient recognition accuracy and incomplete feature extraction in current image classification tasks, this study proposes an improved model—SGFA-ConvNeXt—based on the ConvNeXt-Tiny architecture, integrating both spatial and channel attention. By incorporating the Spatial Gated Fusion Attention (SGFA) module at the end of each stage, the model performs multiscale and multidimensional key region modeling and channel feature enhancement on vehicle images, thereby improving both its discriminative capability and feature representation.

Experimental results on the CIFAR-10 dataset demonstrate that SGFA-ConvNeXt outperforms the original ConvNeXt and other mainstream models while maintaining low computational overhead, achieving a 2% increase in classification accuracy on CIFAR-10.

The primary contributions of this research are twofold:

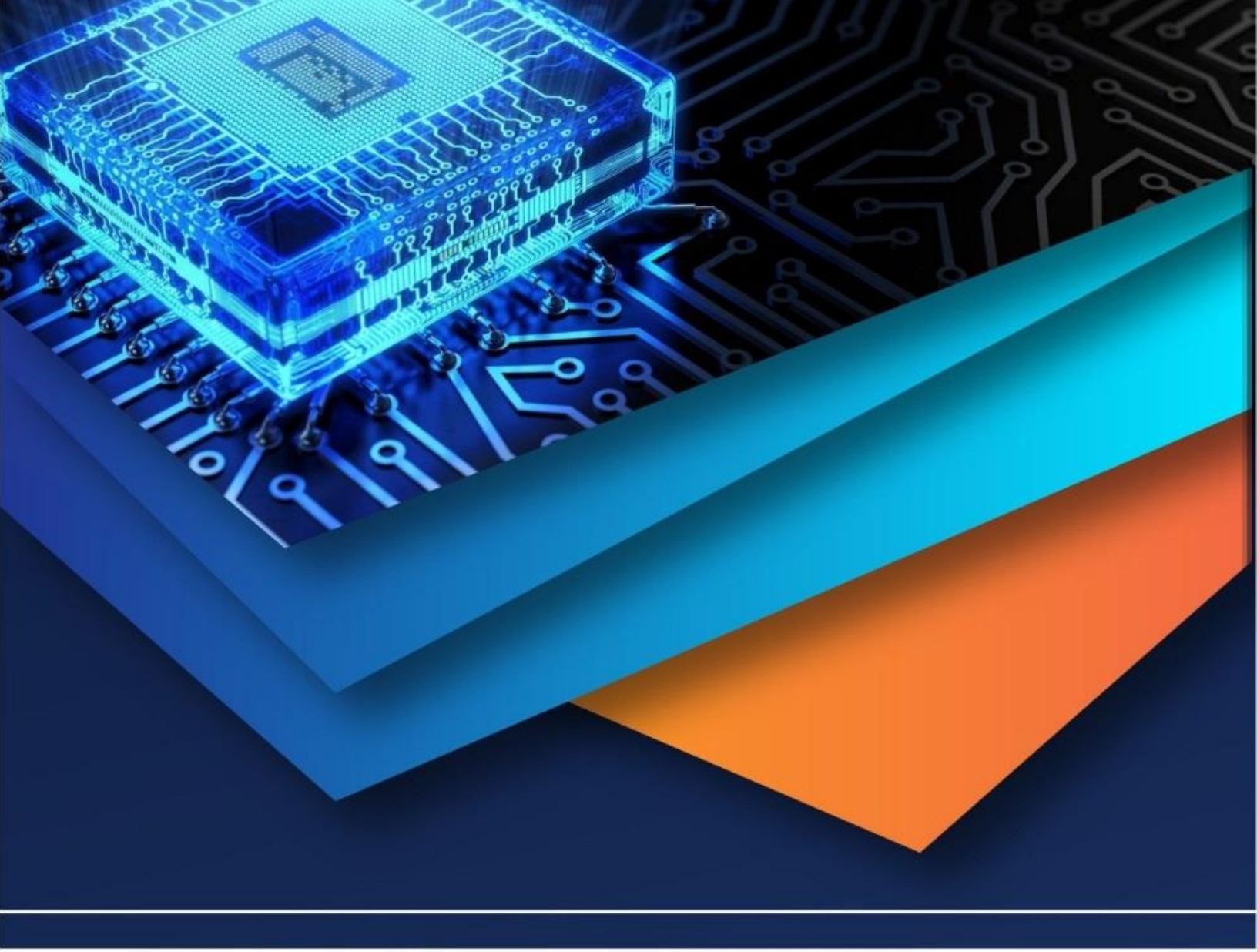
1) The design of the SGFA module, which fuses spatial and channel attention, effectively addresses the difficulty traditional networks face in simultaneously modeling local details and global semantics.

2) A significant improvement in classification performance is achieved without compromising efficiency, making the method well-suited for resource-constrained environments such as edge computing.

Future research can focus on further optimizing the SGFA module to enhance its performance on more complex tasks and exploring its application to other image classification domains, such as medical image analysis and object detection. Additionally, how to further reduce computational cost while maintaining performance is also a promising direction for in-depth investigation.

REFERENCES

[1] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 2002, 86(11): 2278-2324.
[2] Liu Z, Mao H, Wu C Y, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, 2022: 11966-11976.
[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[5] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision
[6] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.
[7] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[8] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[11] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
[12] Jamali A, Roy S K, Hong D, et al. Spatial-Gated Multilayer Perceptron for Land Use and Land Cover Mapping[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 1-5.
[13] Liu X, Peng H, Zheng N, et al. Efficientvit: Memory efficient vision transformer with cascaded group attention[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 14420-14430.
[14] Krizhevsky A. Learning multiple layers of features from tiny images[Z]. Toronto: University of Toronto, 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY