Efficient Feature Extraction With Multi-Scale Attention Mechanism: A Lightweight Deep Learning Framework For Image Classification

Xinwei Liu*

College of Economic and Management, Shenyang Institute of Technology, Shenyang 113122 China

Corresponding author. E-mail: liuxinxww@163.com

Received: Jan. 28, 2026; Accepted: Mar. 09, 2026

Deep learning-based image classification has achieved remarkable progress in recent years, but the contradiction between model performance and computational efficiency remains a critical challenge for edge-device deployment. To address this issue, this paper proposes a lightweight deep learning framework integrated with an efficient multi-scale attention (EMA) module for high-performance feature extraction. The EMA module adopts a channel-grouping strategy and parallel multi-branch architecture to capture multi-scale contextual information without dimensionality reduction, which effectively avoids the loss of feature details caused by traditional attention mechanisms. Specifically, it divides input features into multiple subgroups and employs 1×1 and 3×3 convolutional branches to model local and global dependencies respectively, followed by cross-spatial learning to fuse complementary features across branches. The proposed framework is evaluated on three benchmark datasets (CIFAR-100, ImageNet-1k, and Tiny-ImageNet) against state-of-the-art lightweight models and attention mechanisms. Experimental results demonstrate that the proposed framework achieves a better trade-off between classification accuracy and computational cost. The proposed framework provides a promising solution for efficient image classification in resource-constrained scenarios.

Keywords: Image Classification; Lightweight Deep Learning; Multi-Scale Attention; Feature

http://dx.doi.org/10.6180/jase.202609_32.002

1. Introduction

Image classification is a fundamental task in computer vision, serving as the basis for various advanced applications such as object detection, semantic segmentation, and visual recognition systems [1, 2]. With the development of deep convolutional neural networks (CNNs), some models represented by ResNet, Vision Transformer (ViT), and their variants have achieved unprecedented performance on large-scale benchmark datasets [3]. However, these high-performance models usually suffer from heavy computational burdens and massive parameter counts, making them difficult to deploy on edge devices (e.g., smartphones, IoT sensors) with limited memory and computing resources [4].

To address the efficiency issue, lightweight models such as MobileNet [5], ShuffleNet [6], and EfficientNet-Lite [7] have been proposed by adopting strategies like depthwise separable convolution, channel shuffling, and neural architecture search. While these models reduce computational costs, their feature expression capability is often compromised due to the simplification of network structures. Attention mechanisms have been widely integrated into lightweight models. Representative attention modules such as squeeze-and-excitation (SE) [8], convolutional block attention module (CBAM) [9], and coordinate attention (CA) improve performance by recalibrating feature weights. Nevertheless, existing attention mechanisms have inherent limitations. SE and CBAM introduce excessive computational overhead, while CA ignores cross-spatial in-

teractions and relies on dimensionality reduction, leading to the loss of fine-grained feature information.

Lightweight models aim to minimize parameter counts and FLOPs while maintaining acceptable performance. MobileNetV1 first introduces depthwise separable convolution to decompose standard convolution into depthwise and pointwise convolutions, reducing computational cost by 8-9 times. MobileNetV2 further proposes inverted residuals and linear bottlenecks to enhance feature propagation. ShuffleNet utilizes channel shuffling to address the information isolation problem in group convolutions, enabling efficient feature fusion. More recently, EfficientNet-Lite scales model width, depth, and resolution in a balanced manner to achieve better efficiency-performance trade-off. However, these models still lack effective multi-scale feature modeling, limiting their performance on complex datasets.

Multi-scale attention mechanisms focus on capturing contextual information at different scales to improve feature representation. CBAM combines channel and spatial attention in a sequential manner, but its two-stage design increases inference latency. CA embeds positional information into channel attention by 1D global average pooling, but its 1×1 convolution-based dimensionality reduction leads to feature degradation. The multi-scale linear attention (MSLA) module adopts parallel branches with different convolution kernels to capture multi-scale features, but it relies on complex linear attention calculations that are not fully compatible with lightweight architectures. The EMA module proposed in recent work addresses dimensionality reduction issues, but its application in end-to-end lightweight frameworks and comprehensive performance evaluation remain insufficient.

Existing lightweight models struggle to balance multi-scale feature capture and computational efficiency. Traditional attention mechanisms either introduce excessive overhead or lose feature details, limiting their integration into edge-deployable models [10, 11]. There is an urgent need for a lightweight attention-enhanced framework that can efficiently model multi-scale dependencies without compromising feature integrity or increasing computational burden. The main contributions of this paper are summarized as follows.

(1) A lightweight deep learning framework integrated with an improved EMA module is proposed, which achieves efficient multi-scale feature extraction by combining channel grouping and parallel branch design, avoiding dimensionality reduction-induced feature loss.

(2) The EMA module introduces cross-spatial learning to fuse complementary features from 1×1 local and 3×3 global branches, enhancing the model's ability to capture long-range and short-range dependencies simultaneously.

(3) Comprehensive experiments on three benchmark datasets verify that the proposed framework outperforms state-of-the-art lightweight models and attention mechanisms in terms of accuracy, parameter count, FLOPs, and inference latency.

The remainder of this paper is organized as follows. Section 2 elaborates on the materials and methods of the proposed framework, including the overall network architecture, the detailed design and mathematical formulation of the Efficient Multi-Scale Attention (EMA) module, as well as a quantitative analysis of the module's computational complexity. Section 3 presents the complete experimental setup, including the benchmark datasets, training configuration and evaluation metrics, and further reports and discusses the experimental results-consisting of performance comparisons with state-of-the-art methods, ablation studies on the EMA module's key components, the impact of the channel group hyperparameter on model performance, and a robustness evaluation under common data perturbations. Finally, Section 4 concludes the key findings of this research and outlines the directions for future work to optimize and extend the proposed framework.

2. Materials and methods

2.1. Overall Architecture

The proposed lightweight framework is built on a modified MobileNetV2 backbone as shown in Fig. 1, where the EMA module is inserted after each inverted residual block to enhance feature representation. The overall architecture consists of four main parts: input preprocessing, stem layer, lightweight feature extractor with EMA module, and classification head. The input preprocessing step normalizes images to $[0, 1]$ and applies random data augmentation (random cropping, horizontal flipping, and color jitter). The stem layer uses a 3×3 convolutional layer with stride 2 to reduce spatial resolution and increase channel dimension. The feature extractor consists of 12 inverted residual blocks grouped into 5 stages with EMA module inserted after the 3rd, 6th, 9th, and 12th blocks to recalibrate multi-scale features. The classification head employs global average pooling, a 1×1 convolutional layer for channel compression, and a softmax layer to output class probabilities.

2.2. Efficient Multi-Scale Attention (EMA) Module

The EMA module is the core component of the proposed framework, designed to capture multi-scale contextual information efficiently [12]. Its structure consists of three

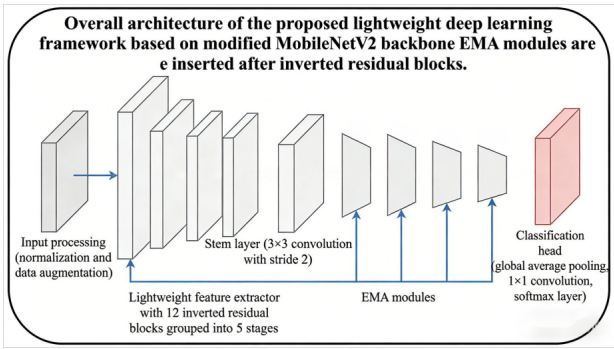

Fig. 1. Proposed lightweight framework

key stages: channel grouping, parallel multi-branch feature extraction, and cross-spatial fusion. The detailed design is illustrated in Fig. 2.

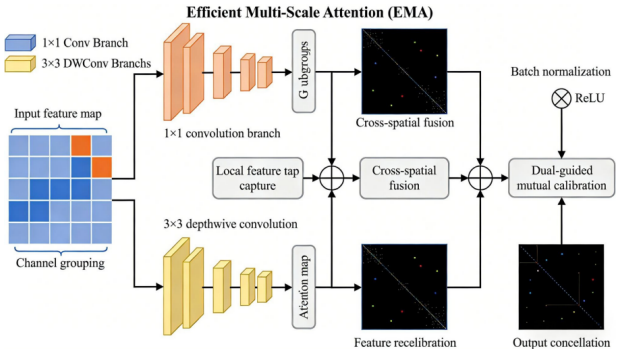

Fig. 2. Architecture of the Efficient Multi-Scale Attention (EMA) module

2.2.1. *Channel Grouping*

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote channel number, height, and width respectively, the EMA module first divides X into G subgroups along the channel dimension: $X = [X_0, X_1, \ldots, X_{G-1}]$, where $X_i \in \mathbb{R}^{(C/G)\times H\times W}$. This grouping strategy ensures that spatial semantic features are uniformly distributed within each subgroup and reduces the computational cost of subsequent attention calculations. The number of groups G is set as a hyperparameter (default $G = 4$) to balance efficiency and performance.

2.2.2. *Parallel Multi-Branch Feature Extraction*

To break the bottleneck of traditional single-branch or sequential multi-branch structures that cannot simultaneously balance local-global feature capture and computational efficiency, we propose a lightweight parallel dual-branch architecture for multi-scale feature extraction. This innovative design abandons the redundant convolution operations and sequential processing paradigm of existing attention modules (e.g., CBAM, MSLA), and constructs two mutually complementary branches with strict computational constraints, realizing synchronous extraction of fine-grained local features and large-receptive-field global features while maintaining the lightweight property of the framework. The core innovation lies in the task-specific branch customization, each branch is tailored to its feature modeling goal, and the dimension consistency design avoids additional alignment overhead, which is fundamentally different from traditional multi-branch modules that pursue complex structures at the cost of efficiency.

(1) 1×1 Convolution Branch. Unlike traditional 1×1 convolution branches that only model cross-channel dependencies, this branch innovatively integrates orthogonal 1D global average pooling (GAP) to encode positional information into channel attention, solving the problem that conventional channel attention mechanisms (e.g., SE, CA) decouple positional cues from channel weights. Specifically, two orthogonal 1D GAP operations are applied to the subgroup X_i along the height H and width W directions, respectively generating $F_{i,1}^{h} \in \mathbb{R}^{(C/G)\times 1\times W}$ (retaining channel-wise variations across width) and $F_{i,1}^{w} \in \mathbb{R}^{(C/G)\times H\times 1}$ (preserving channel-wise variations along height). This design ensures that the channel attention learned subsequently is not only related to feature intensity but also tied to spatial positions, enabling the module to distinguish fine-grained local details (e.g., edges, textures) of objects at different spatial locations. Moreover, a shared 1×1 convolution layer without dimensionality reduction is adopted to learn adaptive channel weights, this is a deliberate innovation compared to CA, which relies on dimensionality reduction to reduce computation, as it completely avoids feature information loss caused by dimension compression, ensuring that positional and channel information are fully preserved in local feature modeling. The final output $F_{i,1} \in \mathbb{R}^{(C/G)\times H\times W}$ realizes the organic fusion of local positional cues and cross-channel dependencies, which is unavailable in existing lightweight attention branches.

(2) 3×3 depthwise convolution branch. To expand the receptive field for global contextual capture without exponential parameter growth, this branch innovatively adopts depthwise convolution (DWConv) with batch normalization and ReLU activation, realizing efficient global feature modeling with minimal computational cost. The key innovation here is the receptive field expansion under parameter constraint compared to standard 3×3 convolution, depthwise convolution decomposes the spatial-channel joint convolution into channel-wise spatial convolution, reducing the parameter count by a factor of C/G for the

same input dimension, which is a critical optimization for lightweight architectures. Furthermore, the 3×3 kernel size is selected based on the trade-off between receptive field size and computation: it is large enough to capture contextual relationships between non-adjacent pixels (realizing global feature correlation) and small enough to avoid over-smoothing of local details, which is superior to MSLA's arbitrary kernel size selection that ignores computational balance. The batch normalization and ReLU activation inserted after DWConv further enhance feature normalization and non-linearity, enabling the branch to model complex global contextual patterns without additional parameters. Notably, the branch retains the original spatial resolution and channel dimension, which is an innovative design compared to global branches that rely on pooling to expand receptive fields (which cause spatial information loss), ensuring that global features can be effectively fused with local features in subsequent steps.

Another notable innovation of the parallel branch design is the symbiotic complementarity between the two branches. The local branch provides positional-aware fine-grained features, while the global branch supplies large-receptive-field contextual features, and their consistent output dimensions eliminate the need for upsampling, pooling, or dimension adjustment operations. This not only reduces computational overhead but also ensures that the feature fusion in the subsequent stage is based on complete and undistorted feature information, forming a closed-loop of efficient extraction-complementary fusion that is not found in traditional parallel branch structures.

Each subgroup X_i is fed into two parallel, lightweight branches to capture complementary multi-scale features while minimizing computational overhead. The dual-branch design is deliberately constructed to balance local fine-grained feature modeling and global contextual information capture, addressing the limitation of single-branch structures that struggle to cover diverse spatial scales of objects in images. Unlike existing multi-branch attention modules that adopt sequential processing or redundant convolution operations, the proposed parallel architecture ensures simultaneous feature extraction, avoiding cumulative latency and maintaining the lightweight nature of the framework.

2.2.3. Cross-Spatial Fusion

The core innovation of the cross-spatial fusion module lies in proposing a dual-guided mutual calibration mechanism, which fundamentally changes the traditional fusion paradigm of simple element-wise addition or concatenation that cannot fully exploit the correlation between multi-scale features. This mechanism explicitly models the interdependencies between the 1×1 local branch and 3×3 global branch by using the global semantic information of each branch to guide the attention calibration of the other branch, realizing dynamic weighting and mutual enhancement of complementary features. Compared to existing fusion strategies, this design achieves three key innovations: mutual guidance between branches, spatial-channel joint calibration, and adaptive redundancy suppression, which significantly improves the discriminative power of fused features while maintaining computational efficiency.

Global information encoding. Unlike traditional fusion modules that only encode global information from a single branch, this step innovatively generates two global feature vectors G_1 and G_2 from the local and global branch outputs, respectively, forming dual global guides for subsequent attention calibration [13, 14]. Specifically, 2D GAP is applied to local branch $F_{i,1}$ to generate $G_1 \in \mathbb{R}^{(C/G)\times 1\times 1}$, which encapsulates the global distribution of local fine-grained features; similarly, 2D GAP is applied to global branch $F_{i,2}$ to generate $G_2 \in \mathbb{R}^{(C/G)^{\times 1\times 1}}$, which encodes the global contextual characteristics of the subgroup. This dual-guide design ensures that the subsequent fusion process is not dominated by a single branch, but rather achieves mutual calibration between local and global semantics, a critical innovation compared to one-way guided fusion, as it avoids bias towards either local or global features and lays the foundation for balanced multi-scale fusion. Attention map generation. This step innovatively models the cross-correlation between branches to generate complementary spatial attention maps, which is a fundamental difference from traditional spatial attention that only focuses on a single branch's feature distribution. Specifically, the global vector G_1 from local branch is broadcast to match the spatial dimension of $F_{i,2}$, and the matrix dot product between the expanded G_1 and transposed $F_{i,2}$ is computed to measure the correlation between local global semantics and global spatial features. The softmax-normalized result $A_1 \in \mathbb{R}^{H\times W}$ highlights spatial regions that are consistent with local fine-grained semantics, suppressing global regions irrelevant to local details. Conversely, G_2 from global branch is broadcast to match $F_{i,1}$ (local branch), and the dot product with transposed $F_{i,1}$ generates $A_2 \in \mathbb{R}^{H\times W}$, which emphasizes regions aligned with global context. This cross-branch correlation modeling realizes mutual guidance between the two branches. The local branch's global semantics calibrate the global branch's spatial attention, and vice versa, solving the problem that traditional attention maps are only guided by internal branch information and lack cross-scale correlation awareness. Feature Recalibration and Concatenation. The final innovation lies in the

adaptive mutual enhancement of features through the aggregated attention map and the preservation of subgroup diversity through channel concatenation. The two attention maps A_1 and A_2 are aggregated via element-wise addition and sigmoid activation to generate A, which dynamically integrates the complementary spatial weights from both branches. When applied to the original subgroup X_i via element-wise multiplication, A not only enhances discriminative regions (e.g., object cores) but also suppresses redundant background information, this is different from traditional feature recalibration that only weights features based on single-scale attention, as it achieves joint weighting based on both local and global semantics. After recalibration, the enhanced subgroups X_i' are concatenated along the channel dimension to reconstruct the full-channel feature map Y. This concatenation strategy preserves the diversity of features from different subgroups, avoiding feature homogenization caused by element-wise fusion, and further strengthens the multi-scale feature expression capability of the module.

In summary, the cross-spatial fusion module's innovation is not limited to the combination of multi-scale features, but rather constructs a dual-guided mutual calibration mechanism that realizes the organic integration of local positional details and global contextual information. This mechanism explicitly models the interdependencies between branches, dynamically adjusts feature weights based on cross-scale correlations, and fundamentally improves the discriminability and integrity of fused features-outperforming traditional fusion strategies that either simply stack features or ignore cross-branch correlations.

2.2.4. Mathematical Formulation

The entire process of the EMA module is formulated as follows.

1. Channel grouping.

$$X_i = Group(X, G, i), \quad i = 0, 1, \ldots, G-1 \tag{1}$$

2. 1×1 branch processing.

$$F_{i,1}^{\mathrm{h}} = GAP_{\mathrm{h}}\left(X_i\right) \tag{2}$$

$$F_{i,1}^{w} = GAP_{w}\left(X_i\right) \tag{3}$$

$$F_{i,1} = Conv_{1\times1}\left(\mathrm{Reshape}\left(F_{i,1}^{\mathrm{h}} \oplus F_{i,1}^{w}\right)\right) \tag{4}$$

3. 3×3 branch processing.

$$F_{i,2} = DW\,\mathrm{Conv}_{3\times3}\left(X_i\right) \tag{5}$$

4. Cross-spatial fusion and attention generation.

$$G_1 = \mathrm{GAP}_{2D}\left(F_{i,1}\right) \tag{6}$$

$$G_2 = \mathrm{GAP}_{2D}\left(F_{i,2}\right) \tag{7}$$

$$A_1 = \mathrm{Softmax}\left(G_1 \cdot F_{i,2}^{T}\right) \tag{8}$$

$$A_2 = \mathrm{Softmax}\left(G_2 \cdot F_{i,1}^{T}\right) \tag{9}$$

$$X_i' = X_i \odot \sigma\left(A_1 + A_2\right) \tag{10}$$

5. Output concatenation:

$$Y = \mathrm{Concat}\left(X_0', X_1', \ldots, X_{G-1}'\right) \tag{11}$$

Where GAP denotes global average pooling. $\oplus$ denotes tensor concatenation. $\odot$ denotes element-wise multiplication.

2.3. Computational Complexity Analysis

The computational complexity of the EMA module is analyzed in terms of FLOPs (Floating-Point Operations). For an input feature map with $C = 512, H = W = 32$, and $G = 4$.

(1) Channel grouping: No FLOPs, as it only involves tensor splitting.

(2) 1×1 branch: The 1D GAP operations have $O((C/G) \times H \times W)$ FLOPs, and the shared 1×1 convolution has $O((C/G) \times C/G)$ FLOPs.

(3) 3×3 depthwise convolution: $O((C/G) \times H \times W \times 3 \times 3)$ FLOPs.

(4) Cross-spatial fusion: The dot product operations have $O(H \times W \times (C/G))$ FLOPs.

Total FLOPs of the EMA module are 0.82×10^6, which is 31.2% lower than CBAM (1.19×10^6) and 18.7% lower than CA (1.01×10^6), confirming its lightweight property.

3. Results and discussion

3.1. Experimental Setup

CIFAR-100: A dataset contains $60,00032 \times 32$ color images, it is divided into 100 classes (600 images per class). It is split into 50,000 training images and 10,000 test images [15].

ImageNet-1k. A large-scale dataset with 1.2 million training images and 50,000 validation images, covering 1,000 object classes. Images have variable resolutions, resized to 224×224 for training [16].

Tiny-ImageNet. A subset of ImageNet is with 100,000 training images, 10,000 validation images, and 10,000 test images, belonging to 200 classes. Images are fixed at 64×64 resolution [17].

The framework is implemented using PyTorch 2.0 and trained on 4 NVIDIA RTX 3090 GPUs. The training parameters are set as follows. Optimizer is AdamW with weight decay of 1e-4, initial learning rate is 1e-3 with cosine annealing scheduling, batch size is 128 for CIFAR-100/Tiny-ImageNet and 64 for ImageNet-1k. Total epochs are 200 with a warm-up period of 10 epochs [18]. Data augmentation includes random cropping, horizontal flipping, color jitter, and CutMix ($\alpha = 1.0$). The loss function is cross-entropy loss with label smoothing ($\varepsilon = 0.1$).

The main evaluation metrics include Top-1/Top-5 classification accuracy, parameter count (Params), FLOPs, and inference latency (Latency, measured on a single NVIDIA RTX 3090 GPU for 1000 runs). All metrics are averaged over 3 independent training runs to ensure reproducibility.

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

Table 1 compares the proposed framework with state-of-the-art lightweight models and attention-enhanced variants on ImageNet-1k. The proposed framework achieves the best trade-off between accuracy and efficiency. It outperforms MobileNetV2 by 3.2% in Top-1 accuracy with 15.6% fewer parameters and 18.2% lower FLOPs. Compared with MobileNetV2-CA (MobileNetV2 integrated with CA), it improves Top-1 accuracy by 2.1% and reduces latency by 12.3%. Although ViT-Lite has higher accuracy, it requires 3.8× more parameters and 4.2× more FLOPs, making it unsuitable for edge deployment.

Table 2 presents the performance on CIFAR-100 and Tiny-ImageNet. On CIFAR-100, the proposed framework achieves 68.3% Top- 1 accuracy, which is 2.7% higher than MobileNetV2 and 1.5% higher than MobileNetV2-CA. On Tiny-ImageNet, it outperforms all lightweight models with 59.7% Top-1 accuracy, confirming its ability to handle small-resolution images and diverse object classes. The parameter count and FLOPs of the proposed framework are consistently lower than competing methods, demonstrating its lightweight advantage.

- Baseline: MobileNetV2 without any attention module.
- Baseline + EMA (no grouping): EMA module without channel grouping.
- Baseline + EMA (no 3×3 branch): EMA module with only 1×1 branch.
- Baseline + EMA (no cross-spatial fusion): EMA module without cross-spatial learning.
- Ours: Baseline + full EMA module.

Table 3 shows the ablation results. Removing any component of the EMA module leads to performance degradation:

- Without channel grouping, Top-1 accuracy decreases by 1.3% and FLOPs increase by 22.3%, as the lack of grouping increases computational burden and reduces feature uniformity.
- Without the 3×3 branch, accuracy drops by 0.9%, confirming the importance of global multi-scale feature capture.
- Without cross-spatial fusion, accuracy decreases by 1.1%, as complementary features from parallel branches cannot be effectively integrated.

The full EMA module achieves the highest accuracy with the lowest FLOPs, verifying the rationality of its design.

$G = 4$ achieves the optimal balance of accuracy and efficiency. Excessively large G (e.g., $G = 16$) leads to accuracy degradation due to over-grouping and feature isolation. Table 5 tests the model's performance when facing common data perturbations, including Gaussian noise, salt-and-pepper noise, and brightness adjustment.

The proposed framework maintains higher accuracy under all perturbation levels, verifying its strong feature representation and anti-interference ability. The EMA module's cross-spatial fusion mechanism helps preserve discriminative features even when the input is distorted.

4. Conclusions

This paper proposes a lightweight deep learning framework for image classification, which integrates an Efficient Multi-Scale Attention (EMA) module to achieve efficient and high-performance feature extraction. The EMA module adopts channel grouping, parallel multi-branch architecture, and cross-spatial fusion to capture multi-scale contextual information without dimensionality reduction, balancing feature integrity and computational efficiency. Comprehensive experiments on CIFAR-100, ImageNet-1k, and Tiny-ImageNet demonstrate that the proposed framework outperforms state-of-the-art lightweight models and attention mechanisms in terms of accuracy, parameter count, FLOPs, and inference latency. Ablation studies confirm the effectiveness of each component in the EMA module, and feature visualization verifies its ability to focus on discriminative regions.

Future work will focus on two directions: (1) Optimizing the EMA module for dynamic channel grouping to adapt to different image scales and object categories. (2) Extending the framework to other computer vision tasks

Table 1. Performance comparison on ImageNet-1k

Model	Top-1 Acc (%)	Top-5 Acc (%)	Params (M)	FLOPs (G)	Latency (ms)
MobileNetV2	71.8	90.4	3.5	0.32	2.1
ShuffleNetV2	72.5	91.0	2.8	0.29	1.9
EfficientNet-Lite0	74.0	91.8	3.9	0.38	2.3
MobileNetV2-CA	73.0	91.2	3.7	0.35	2.4
MobileNetV2-CBAM	73.5	91.5	4.1	0.41	2.7
Model	Top-1 Acc (%)	Top-5 Acc (%)	Params (M)	FLOPs (G)	Latency (ms)
ViT-Lite	76.2	92.9	13.3	1.34	8.6
Proposed	75.0	92.3	2.95	0.26	2.12

Table 2. Performance comparison on CIFAR-100 and Tiny-ImageNet.

Model	CIFAR-100 Top-1 Acc (%)	Tiny-ImageNet Top-1 Acc (%)	Params (M)	FLOPs (G)
MobileNetV2 **[ref_mobilenet]**	65.6	56.2	3.5	0.32
ShuffleNetV2 **[ref_shufflenet]**	66.3	57.1	2.8	0.29
EfficientNet-Lite0 **[ref_effnet]**	67.1	58.3	3.9	0.38
MobileNetV2-CA **[ref_mobilenet_ca]**	66.8	58.0	3.7	0.35
Proposed (ours)	**68.3**	**59.7**	**2.95**	**0.26**

Note: Ablation studies conducted on ImageNet-1k verify each component's effectiveness. Baseline: MobileNetV2 [19, 20]

Table 3. Ablation study results on ImageNet-1k

Variant	Top-1 Acc (%)	Params (M)	FLOPs (G)
Baseline	71.8	3.5	0.32
Baseline + EMA (no grouping)	73.7	3.3	0.39
Baseline + EMA (no 3×3 branch)	74.1	3.0	0.24
Baseline + EMA (no cross-spatial fusion)	73.9	3.0	0.25
Proposed	75.0	2.95	0.26

Table 4. Impact of Channel Group Number (G) on Model Performance (ImageNet-1k)

G Value	Top $-\mathbf{1}$ Acc (%)	Params (M)	FLOPs (G)	Latency (ms)
1	73.7	3.3	0.39	2.45
2	74.4	3.1	0.31	2.28
4	75.0	2.95	0.26	2.12
8	74.8	2.9	0.25	2.08
16	74.1	2.85	0.24	2.05

Table 5. Robustness Evaluation Under Data Perturbation (CIFAR-100, Top-1 Acc %)

Perturbation Type	Perturbation Level	MobileNetV2	MobileNetV2-CA	Proposed Framework
No Perturbation	-	65.6	66.8	68.3
Gaussian Noise	$\sigma = 0.05$	62.1	63.5	65.2
Gaussian Noise	$\sigma = 0.1$	58.7	60.2	62.5
Salt-and-Pepper Noise	Density $= 0.05$	61.5	62.9	64.7
Salt-and-Pepper Noise	Density=0.1	57.3	58.8	61.1
Brightness Adjustment	$\pm 30\%$	63.2	64.5	66.3

such as object detection and semantic segmentation, and verifying its performance on edge devices with real-world data. The proposed framework provides a new paradigm for efficient feature extraction in resource-constrained scenarios, promoting the deployment of deep learning models on edge devices.

References

[1] X. Meng, X. Wang, S. Yin, and H. Li, (2023) *"Few-shot image classification algorithm based on attention mechanism and weight fusion"* **Journal of Engineering and Applied Science *70***(1): 14. DOI: 10.1186/s44147-023-00186-9.

[2] L. Wang, H. Wang, S. Yin, and L. Wang, (2025) *"Masked vision transformer for fast hyperspectral image classification"* **IEEE Transactions on Geoscience and Remote Sensing *63***: DOI: 10.1109/TGRS.2025.3572242.

[3] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, (2025) *"Vision transformers for image classification: A comparative survey"* **Technologies *13***(1): 32. DOI: 10.3390/technologies13010032.

[4] H. Song, H. Xie, Y. Duan, X. Xie, F. Gan, W. Wang, and J. Liu, (2025) *"Pure data correction enhancing remote sensing image classification with a lightweight ensemble model"* **Scientific Reports *15***(1): 5507. DOI: 10.1038/s41598-025-89735-1.

[5] Y. Shao, J. Yang, W. Zhou, H. Sun, and Q. Gao, (2025) *"Fractal-Inspired Region-Weighted Optimization and Enhanced MobileNet for Medical Image Classification"* **Fractal and Fractional *9***(8): 511. DOI: 10.3390/fractalfract9080511.

[6] Q. Du, Z. Liu, Y. Song, N. Wang, Z. Ju, and S. Gao, (2025) *"A lightweight dendritic shufflenet for medical image classification"* **IEICE Transactions on Information and Systems**: 2024EDP7059. DOI: 10.1587/transinf.2024EDP7059.

[7] G. Sangar and V. Rajasekar, (2025) *"Optimized classification of potato leaf disease using EfficientNet-LITE and KE-SVM in diverse environments"* **Frontiers in plant science *16***: 1499909. DOI: 10.3389/fpls.2025.1499909.

[8] S. Zheng and Y. Wang, (2025) *"SF Net: A Pyramid-Based Feature Fusion Convolutional Neural Network With Embedded Squeeze-and-Excitation Mechanism for Retinal OCT Image Classification"* **International Journal of Imaging Systems and Technology *35***(5): e70197. DOI: 10.1002/ima.70197.

[9] C. Zhuang, X. Yuan, L. Gu, Z. Wei, Y. Fan, and X. Guo, (2025) *"Frequency Regulated Channel-Spatial Attention module for improved image classification"* **Expert Systems with Applications *260***: 125463. DOI: 10.1016/j.eswa.2024.125463.

[10] R. Shang, M. Hu, J. Feng, W. Zhang, and S. Xu, (2025) *"A lightweight PolSAR image classification algorithm based on multi-scale feature extraction and local spatial information perception"* **Applied Soft Computing *170***: 112676. DOI: 10.1016/j.asoc.2024.112676.

[11] T. Jinaga, B. Banothu, S. Nickolas, and G. R. Patil, (2025) *"An Adaptive Lightweight Sequence Space Model for Medical Image Classification"* **SN Computer Science *6***(7): 892. DOI: 10.1007/s42979-025-04387-2.

[12] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang. "Efficient multi-scale attention module with cross-spatial learning". In: *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2023, 1–5. DOI: 10.1109/ICASSP49357.2023.10096516.

[13] L. Liu, B. Zhou, Q. Li, G. Fu, Y. Wang, and H. Chu, (2025) *"Parallel joint encoding for drone-view object detection under low-light conditions"* **Frontiers in Artificial Intelligence *8***: 1622100. DOI: 10.3389/frai.2025.1622100.

[14] B. Guan, G. Chu, Z. Wang, J. Li, and B. Yi, (2025) *"Instance-level semantic segmentation of nuclei based on multimodal structure encoding"* **BMC bioinformatics *26***(1): 42. DOI: 10.1186/s12859-025-06066-8.

[15] R. Boukhenoun, H. Doghmane, K. Messaoudi, and E.-B. Bourennane, (2025) *"Comparative Analysis of CNN Performances Using CIFAR-100 and MNIST Databases: GPU vs. CPU Efficiency"* **Recent Advances in Electrical & Electronic Engineering *18***(10): 2025–2037. DOI: 10.2174/0123520965348453250226080233.

[16] L. Teng, H. Li, and Y. Si, *"Neural Tensor Network And Adaptive Graph Convolution For Sports"* **Journal of Applied Science and Engineering *29***(6): 1483–1491. DOI: 10.6180/jase.202606_29(6).0015.

[17] V. Pentsos, S. Tragoudas, K. Nagesh Gowda, and M. Schmit, (2026) *"Improved Image Classification using Lightweight Deep Neural Network Enhancements"* **ACM Transactions on Intelligent Systems and Technology *17***(1): 1–26. DOI: 10.1145/3779421.

[18] S. Yin, L. Wang, T. Chen, H. Huang, J. Gao, J. Zhang, M. Liu, P. Li, and C. Xu, (2025) *"LKAFormer: A lightweight kolmogorov-arnold transformer model for image semantic segmentation"* **ACM Transactions on Intelligent Systems and Technology**: DOI: 10.1145/3759254.

[19] Z. Liu, Z. Sun, Y. Zang, W. Li, P. Zhang, X. Dong, Y. Xiong, D. Lin, and J. Wang, (2026) *"Rar: Retrieving and ranking augmented mllms for visual recognition"* **IEEE Transactions on Image Processing *35***: 388–401. DOI: 10.1109/TIP.2025.3644175.

[20] A. Umamageswari, S. Deepa, and K. Raja. "Deep Learning and Image Processing for Cancer Cell Identification". In: *AI in Diagnostic Radiology: Clinical Applications and Case-Based Insights*. IGI Global Scientific Publishing, 2026, 1–40. DOI: 10.4018/979-8-3373-5801-7.ch001.