Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AMS-Net: Lightweight Self-Attention Enhanced MobileNetV3 for Real-Time Multimedia Classification on Edge Devices

Shitong Wang
Anyang University
Jiaquan Shen
shenjiaquan_cv@163.com

Luoyang Normal University
Jiaxi Yang
Zhengzhou Business University

Article

Keywords: Lightweight neural networks, Self-attention mechanism, Edge device deployment, Real-time multimedia processing, Fine-grained image classification, MobileNetV3

Posted Date: August 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10006104/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AMS-Net: Lightweight Self-Attention Enhanced MobileNetV3 for Real-Time Multimedia Classification on Edge Devices

Shitong Wang[1], Jiaquan Shen[2*], Jiaxi Yang[3]

[1]School of Computer Science, Anyang University, Xinxiang, 453506, China.

[2*]School of Information and Technology, Luoyang Normal University, Luoyang, 471000, China.

[3]School of Computer Engineering and Big Data, Zhengzhou Business University, Zhengzhou, 451200, China.

*Corresponding author(s). E-mail(s): shenjiaquan_cv@163.com;
Contributing authors: m15093266334@163.com;
jiaxi.yang@mail.concordia.ca;

Abstract

Purpose: Real-time visual content classification on resource-constrained edge devices presents a fundamental challenge in multimedia systems, requiring models that balance computational efficiency with discriminative accuracy. Existing lightweight architectures sacrifice fine-grained recognition capability for deployment feasibility.

Methods: This paper proposes AMS-Net (Attention MobileNet Small), a lightweight architecture that integrates a novel Lightweight Spatial Self-Attention Module (LSSAM) into MobileNetV3-Small. LSSAM achieves 1/4096 computational complexity reduction relative to standard self-attention through channel reduction ($\boldsymbol{r}$=16), spatial downsampling ($\boldsymbol{s}$=2), and grouped attention ($\boldsymbol{G}$=8). A dynamic transfer learning strategy with progressive layer unfreezing and differentiated learning rates further optimizes training efficiency.

Results: On the Flowers-WTT dataset, AMS-Net attains 97.58% classification accuracy with 1.53M parameters, 0.124 GFLOPs, and 0.0168s per-image inference latency (≈60 FPS). Ablation studies confirm LSSAM contributes 3.05 percentage points accuracy improvement with merely 0.39% parameter overhead.

The dynamic transfer learning strategy reduces training time by 35.4%. Cross-dataset validation on Oxford-17 achieves 94.0% accuracy, surpassing VGG-16 with 90× fewer parameters.
Conclusion: AMS-Net establishes a superior accuracy-efficiency trade-off for real-time multimedia classification on edge devices. The proposed techniques generalize to other fine-grained visual recognition tasks, enabling practical multimedia applications in smart agriculture, ecological monitoring, and botanical education.

Keywords: Lightweight neural networks, Self-attention mechanism, Edge device deployment, Real-time multimedia processing, Fine-grained image classification, MobileNetV3

1 Introduction

Real-time visual content classification on edge devices constitutes a critical requirement for practical multimedia systems. Mobile phones, embedded sensors, and IoT devices increasingly demand high-accuracy visual recognition under stringent computational constraints [1]. This demand extends across diverse multimedia application domains, including aerial traffic monitoring [2, 3], biometric verification [4], and fine-grained botanical recognition [5]. Among these, flower classification exemplifies challenging fine-grained recognition scenarios where subtle morphological differences between species necessitate sophisticated feature extraction while deployment contexts require real-time inference capability.

Conventional deep architectures achieve high classification accuracy at prohibitive computational cost. VGG-16 requires 138M parameters and 30.9 GFLOPs per inference [6], rendering edge deployment impractical. Lightweight architectures including MobileNet [1], ShuffleNet [7], and EfficientNet [8] address efficiency concerns through depthwise separable convolutions, channel shuffling, and compound scaling, respectively. However, parameter reduction typically compromises fine-grained discrimination, yielding 5–8% accuracy degradation on challenging classification tasks [9].

Self-attention mechanisms capture long-range spatial dependencies critical for fine-grained recognition [10]. Vision Transformers demonstrate superior performance through global context modeling [11], yet standard self-attention exhibits $O(n^2)$ complexity with respect to spatial resolution, precluding direct integration into lightweight architectures for real-time multimedia applications [12]. Recent efficient attention variants such as ECA-Net [13] and CBAM [14] provide partial solutions, but systematic integration of spatial self-attention into MobileNet architectures for real-time deployment remains underexplored. Prior work on lightweight attention for domain-specific tasks [2, 4, 15] further motivates efficient attention design for resource-constrained multimedia processing.

This paper addresses the fundamental challenge of integrating attention mechanisms into lightweight networks for real-time edge-based multimedia classification. We propose AMS-Net incorporating three principal contributions:

(1) Lightweight Spatial Self-Attention Module (LSSAM): A computationally efficient attention mechanism achieving 1/4096 complexity reduction through channel compression, spatial downsampling, and grouped attention, enabling real-time inference without sacrificing discriminative capability.

(2) Strategic Hierarchical Attention Integration: Systematic placement of LSSAM modules at blocks 5, 8, and 11 to capture multi-scale spatial dependencies across network depth, validated through comprehensive ablation studies.

(3) Dynamic Transfer Learning Strategy: Progressive layer unfreezing with differentiated learning rates, reducing training time by 35.4% while maintaining classification accuracy.

Experimental validation demonstrates AMS-Net achieves 97.58% accuracy with 0.0168s inference latency (≈60 FPS), 1.53M parameters, and 0.124 GFLOPs, establishing a state-of-the-art accuracy-efficiency trade-off for real-time fine-grained multimedia classification on edge devices.

2 Related Work

2.1 Lightweight Network Architectures

MobileNetV1 [1] introduced depthwise separable convolutions, reducing computational cost by 8–9× for 3×3 kernels. MobileNetV2 [16] extended this through inverted residual structures with linear bottlenecks. MobileNetV3 [17] combined neural architecture search with hardware-aware optimization, incorporating Squeeze-and-Excitation modules and h-swish activation. ShuffleNet [7] introduced channel shuffle operations for cross-group information flow, while EfficientNet [8] proposed compound scaling for balancing depth, width, and resolution. Recent advances include RepViT [18] employing structural reparameterization and MobileNetV4 [19] introducing universal inverted bottlenecks with Mobile MQA attention achieving 87% ImageNet accuracy at 3.8ms mobile inference. Beyond general-purpose backbones, specialized lightweight architectures have demonstrated efficacy in domain-specific multimedia tasks including aerial vehicle detection [3] and biometric verification [15].

2.2 Efficient Attention Mechanisms

Standard self-attention [10] computes pairwise token relationships with $O(n^2)$ complexity, which is prohibitive for high-resolution visual features. Linear attention approximations [20] reduce complexity to $O(n)$ through kernel feature maps, while sparse attention patterns [21] restrict computation to local windows. For lightweight networks, ECA-Net [13] proposes efficient channel attention through 1D convolution, and CBAM [14] combines channel and spatial attention with minimal overhead. Recent work further explores integrating transformer-style attention into CNNs while maintaining efficiency [22]. Attention mechanisms have also been applied in multimedia-specific contexts, including pedestrian detection in aerial imagery [23] and fine-grained instrument recognition [4]. However, systematic integration of spatial self-attention into MobileNet architectures specifically optimized for real-time multimedia applications remains underexplored.

2.3 Real-Time Multimedia Classification

Real-time classification requires balancing accuracy with strict latency constraints, typically ≤30ms for 30 FPS multimedia applications [24]. Edge deployment introduces additional constraints on model size, memory footprint, and power consumption. Prior lightweight flower classification systems achieve 0.020s per-image inference [9], establishing baseline latency requirements. Direction-aware semantic feature extraction [2] has further demonstrated the importance of efficient feature design for real-time multimedia processing in complex environments. This work targets sub-20ms inference latency while maximizing classification accuracy for practical edge-based multimedia deployment.

3 Methodology

3.1 Architecture Overview

AMS-Net extends MobileNetV3-Small through strategic integration of LSSAM modules at three hierarchical positions within the inverted residual block sequence. Fig. 1 illustrates the MobileNetV3-Small backbone architecture serving as the foundation for AMS-Net. The base architecture contains 11 inverted residual blocks with depthwise separable convolutions, SE modules, and h-swish activations. Input images of size 224×224×3 undergo progressive spatial reduction through stride-2 convolutions while channel dimensions increase from 16 to 960.

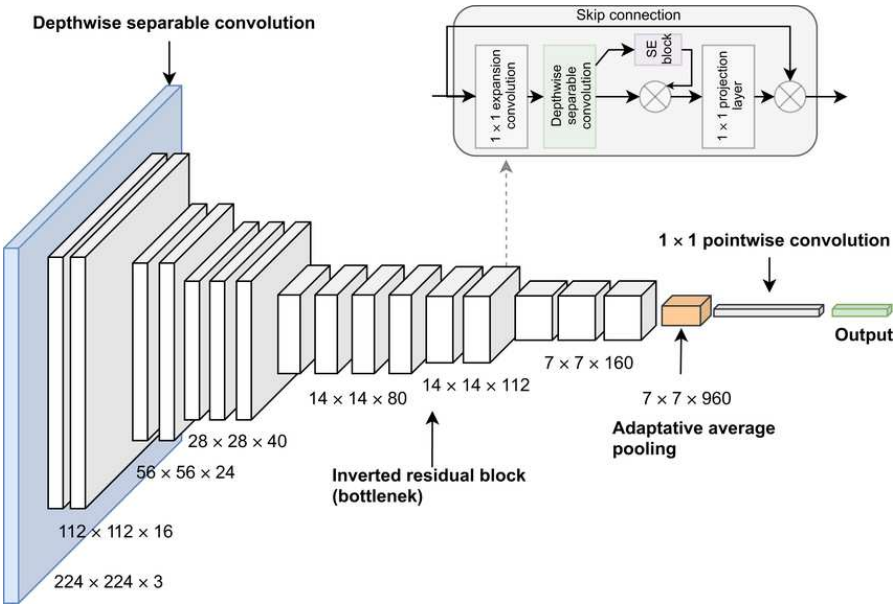


Fig. 1 MobileNetV3-Small base architecture showing depthwise separable convolutions, inverted residual blocks with SE modules, and progressive spatial reduction from 224×224×3 input to 7×7×960 feature maps [25].

LSSAM modules are strategically inserted after blocks 5, 8, and 11, operating on feature maps of dimensions 28×28 (block 5), 14×14 (block 8), and 7×7 (block 11), capturing hierarchical spatial dependencies at multiple semantic levels. This distributed placement strategy ensures comprehensive multi-scale feature enhancement while the progressively decreasing spatial resolution minimizes computational overhead.

3.2 Lightweight Spatial Self-Attention Module (LSSAM)

LSSAM establishes global spatial dependencies through attention computation while minimizing computational overhead via three synergistic optimization strategies. Fig. 2 presents the complete LSSAM architecture.

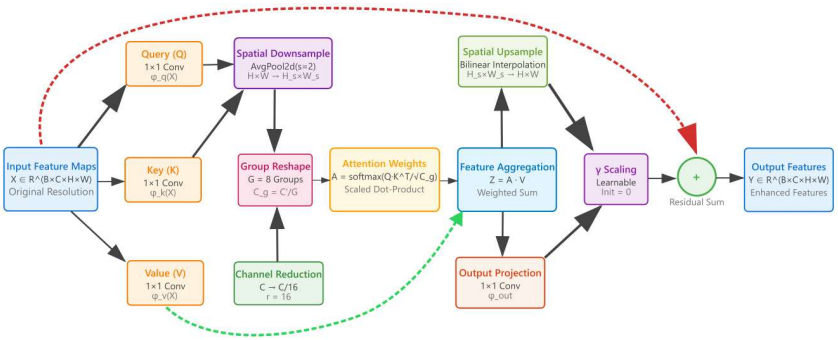


Fig. 2 Lightweight Spatial Self-Attention Module (LSSAM) architecture with channel reduction (r=16), spatial downsampling (s=2), grouped attention (G=8), and residual fusion. Combined optimizations achieve 1/4096 complexity reduction versus standard self-attention.

3.2.1 Mathematical Formulation

Given input features $X \in \mathbb{R}^{B\times C\times H\times W}$ where B denotes batch size, C channels, and $H \times W$ spatial dimensions, LSSAM performs the following transformations.

Feature Projection: Three 1×1 convolutions generate query, key, and value representations:

$$Q = \phi_q(X) \in \mathbb{R}^{B\times C'\times H_s\times W_s} \tag{1}$$

$$K = \phi_k(X) \in \mathbb{R}^{B\times C'\times H_s\times W_s} \tag{2}$$

$$V = \phi_v(X) \in \mathbb{R}^{B\times C'\times H_s\times W_s} \tag{3}$$

where φ_q, φ_k, and φ_v are 1×1 convolutions, $C' = C/r$ (channel reduction ratio r), and $H_s = H/s$, $W_s = W/s$ (spatial downsampling factor s).

Grouped Attention Computation. Channels are partitioned into G groups with $C_g = C'/G$ channels each. Group-wise attention:

$$A_g = \text{softmax}\left(\frac{Q_g \cdot K_g^T}{\sqrt{C_g}}\right) \in \mathbb{R}^{B\times(H_s\cdot W_s)\times(H_s\cdot W_s)} \tag{4}$$

Feature Aggregation and Output. Group outputs $Z_g = A_g \cdot V_g$ are concatenated, upsampled via bilinear interpolation to restore spatial dimensions, and projected through 1×1 convolution with learnable scaling parameter γ (initialized to 0):

$$Y = X + \gamma \cdot \varphi_{\text{out}}(\text{Upsample}(Z)) \tag{5}$$

3.2.2 Computational Complexity Analysis

Standard self-attention exhibits complexity $O(C^2HW)$ for feature maps of size $C \times H \times W$. LSSAM achieves dramatic complexity reduction through three synergistic mechanisms: (1) channel reduction (r=16) reduces complexity by 256×; (2) spatial downsampling (s=2) contributes 16× reduction; and (3) grouped attention (G=8) provides additional reduction. The combined complexity reduction factor is:

$$\frac{\text{Standard Attention}}{\text{LSSAM}} = r^2 \times s^4 \times G = 16^2 \times 2^4 \times 1 = 4096\times \tag{6}$$

This 4096× reduction enables practical integration without compromising real-time performance. LSSAM introduces minimal parameters through four 1×1 convolution layers plus the scaling parameter γ. Total parameter overhead across three modules equals 5,891 parameters, representing merely 0.39% increase over the 1.526M baseline parameters.

3.3 Hierarchical Attention Placement Strategy

Strategic LSSAM placement captures multi-scale spatial dependencies essential for fine-grained discrimination. Systematic ablation studies (Section 4.3) evaluated multiple three-position configurations across the 11-block architecture. The optimal configuration (5, 8, 11) positions LSSAM modules to capture: (a) low-level texture patterns, color gradients, and edge relationships critical for variety discrimination at Block 5 (28×28 features); (b) mid-level structural components including petal arrangements and local geometric patterns at Block 8 (14×14 features); and (c) high-level semantic understanding and global morphological patterns for species-level discrimination at Block 11 (7×7 features). This distributed placement outperforms concentrated configurations by enabling comprehensive multi-scale feature integration while progressively decreasing spatial resolution minimizes computational overhead.

Table 1 Computational Efficiency Metrics of AMS-Net

Metric	Baseline	+LSSAM	AMS-Net	Δ	Δ%
Parameters (M)	1.526	1.532	1.532	+0.006	+0.39%
FLOPs (G)	0.123	0.124	0.124	+0.001	+0.49%
Model Size (MB)	17.71	17.80	17.80	+0.09	+0.51%
Inference (s)	0.0156	0.0167	0.0168	+0.0012	+7.7%
Throughput (FPS)	64.1	59.9	59.5	−4.6	−7.2%

3.4 Dynamic Transfer Learning Strategy

Traditional static transfer learning employs fixed unfreezing schedules and uniform learning rates, ignoring stage-specific optimization requirements. The proposed dynamic strategy implements three-phase progressive training:

Phase 1 (Epochs 1–5): Unfreeze late-stage blocks (9–11), LSSAM modules, and classifier head. Base learning rate $\eta_0 = 0.001$.

Phase 2 (Epochs 6–10): Additionally unfreeze mid-level blocks (5–8). Learning rate $\eta_1 = 0.5\eta_0$.

Phase 3 (Epochs 11+): Full network fine-tuning. Learning rate $\eta_2 = 0.1\eta_0$.

Differentiated learning rates preserve pre-trained features while enabling task adaptation: backbone layers at 0.01η, LSSAM modules at 0.5η, and classifier head at 1.0η. This strategy reduces training time by 35.4% while maintaining classification accuracy through efficient utilization of pre-trained ImageNet features.

4 Experimental Evaluation

4.1 Experimental Setup

Datasets. Primary evaluation employs the Flowers-WTT dataset containing 8,000 images across 8 categories (1,000 images/category), including four rose varieties with subtle color differences for fine-grained evaluation. Data partitioning follows 68%/16%/16% training/validation/test splits. Secondary validation utilizes Oxford-17 [5] with the standard 40/20/20 per-category protocol.

Implementation. All experiments employ PyTorch 2.1.0 with CUDA 12.2 on NVIDIA GeForce GTX 1660Ti (6GB). Training uses Adam optimizer (β_1=0.9, β_2=0.999) with cosine annealing scheduler, batch size 32, and early stopping with patience 10. Data augmentation includes RandomResizedCrop(224), RandomHorizontalFlip(0.5), and ColorJitter.

Evaluation Metrics. Classification accuracy, parameters (M), FLOPs (G), model size (MB), and inference time (s) quantify accuracy-efficiency trade-offs. Per-class precision, recall, and F1-scores evaluate fine-grained discrimination capability.

4.2 Real-Time Performance Analysis

Table 1 presents comprehensive efficiency metrics establishing AMS-Net's real-time multimedia processing capability.

AMS-Net achieves 0.0168s per-image inference (≈60 FPS), satisfying real-time requirements for mobile multimedia applications. The 7.7% latency increase from LSSAM integration remains imperceptible in practical deployment. Model size of 17.8MB enables storage on resource-constrained devices, representing 29.7× reduction compared to VGG-16 (528MB). FLOPs of 0.124G confirm minimal computational overhead, enabling deployment on mobile CPUs and low-power embedded processors. Peak memory consumption of ∼45MB at batch size 1 ensures stable operation across diverse mobile platforms.

4.3 Ablation Studies

4.3.1 Component Contribution Analysis

Table 2 quantifies individual component contributions through systematic ablation.

Table 2 Ablation Study Results on Flowers-WTT Dataset

Configuration	Acc. (%)	Δ Acc.	Train (s)	Params (M)
MobileNetV3-Small	95.00	—	5,681	1.526
+ LSSAM	98.05	+3.05	10,604	1.532
+ Dynamic Transfer (AMS-Net)	**97.58**	**+2.58**	**6,853**	**1.532**

LSSAM integration contributes 3.05 percentage points accuracy improvement with merely 0.39% parameter overhead, demonstrating exceptional efficiency in enhancing discriminative capability. Dynamic transfer learning reduces training time by 35.4% (10,604s → 6,853s) while maintaining 97.58% accuracy, validating the progressive unfreezing strategy's effectiveness.

Fig. 3 illustrates the training dynamics comparison across configurations, revealing distinct convergence patterns.

Fig. 4 quantifies training efficiency gains achieved through the dynamic transfer learning strategy. AMS-Net reaches 97% validation accuracy at epoch 10 versus epoch 16 without dynamic transfer learning, demonstrating a 37.5% reduction in convergence time.

4.3.2 LSSAM Position Optimization

Table 3 presents systematic evaluation of LSSAM placement configurations across the 11-block architecture.

The optimal (5, 8, 11) configuration achieves 98.05% accuracy through distributed placement enabling multi-scale feature integration. Concentrated late-stage configurations (9, 10, 11) yield inferior performance (97.11%), confirming that attention mechanisms benefit from hierarchical rather than concentrated deployment. The 0.94 percentage point gap between optimal and worst configurations represents significant practical impact at high baseline accuracies.

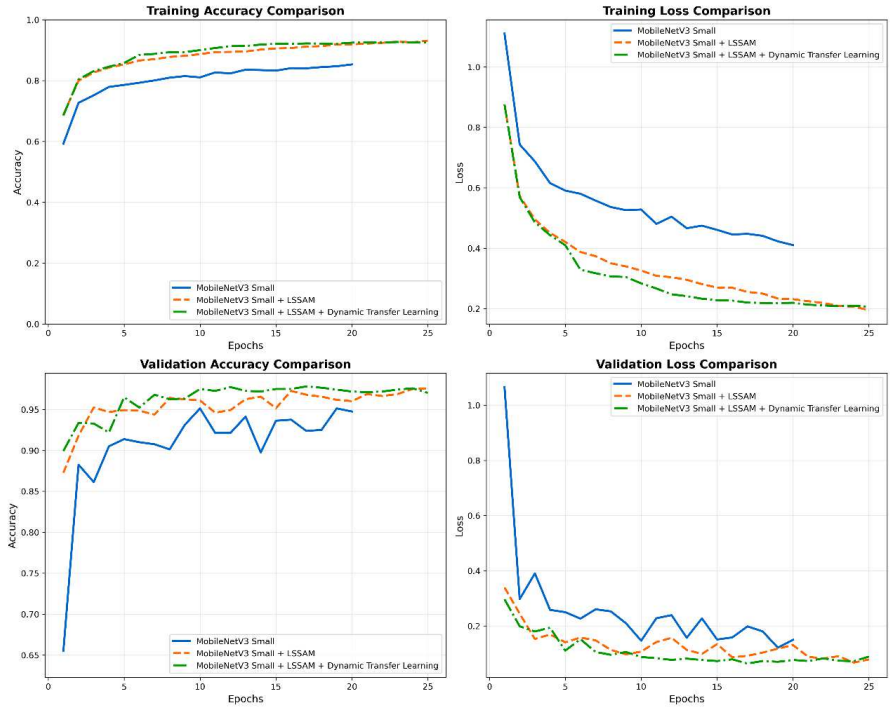


Fig. 3 Training and validation curves comparison. Top: Training accuracy and loss. Bottom: Validation accuracy and loss across baseline (blue), +LSSAM (orange), and AMS-Net with dynamic transfer learning (green). LSSAM enhances generalization (reduced train-validation gap), while dynamic transfer learning accelerates convergence.

Table 3 Performance Comparison of LSSAM Position Configurations

Positions	Acc. (%)	Feature Map Sizes	Rel. Perf.
(5, 8, 11)	**98.05**	28×28, 14×14, 7×7	**Optimal**
(3, 6, 11)	97.73	56×56, 28×28, 7×7	−0.32%
(4, 7, 10)	97.66	28×28, 14×14, 7×7	−0.39%
(9, 10, 11)	97.11	7×7 concentrated	−0.94%

4.4 Comparison with State-of-the-Art Methods

Table 4 presents comprehensive comparison against representative architectures spanning traditional deep networks and modern lightweight designs.

AMS-Net achieves the highest accuracy (97.58%) among all evaluated architectures while maintaining the lowest computational complexity (0.12 GFLOPs). Compared to VGG-16, AMS-Net achieves 0.33% higher accuracy with 90.4× fewer parameters, 250× lower FLOPs, 5.3× faster inference, and 29.7× smaller model size. Among lightweight

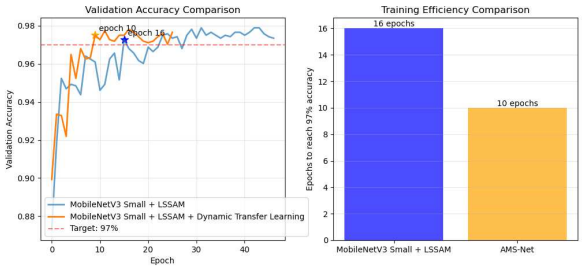


Fig. 4 Training efficiency comparison. Left: Validation accuracy trajectories showing AMS-Net with dynamic transfer learning reaches 97% accuracy 6 epochs earlier. Right: Bar chart comparing epochs required to reach 97% accuracy threshold (16 vs. 10 epochs), demonstrating 37.5% convergence acceleration.

Table 4 Performance Comparison on Flowers-WTT Dataset

Architecture	Acc. (%)	Params (M)	FLOPs (G)	Infer. (s)	Size (MB)
VGG-16	97.25	138.36	30.94	0.0892	528.5
EfficientNet-V2-S	94.88	20.19	5.79	0.1678	232.3
MobileNetV3-Large	95.75	4.21	0.47	0.0437	48.5
MobileNetV2	95.25	2.23	0.65	0.0553	25.9
ShuffleNetV2-x1.0	96.00	1.26	0.30	0.0237	14.7
AMS-Net (Ours)	**97.58**	**1.53**	**0.12**	**0.0168**	**17.8**

architectures, AMS-Net surpasses ShuffleNetV2-x1.0 by 1.58 percentage points accuracy while achieving 2.4× lower FLOPs and maintaining comparable inference speed suitable for real-time multimedia applications.

4.5 Classification Performance Visualization

Fig. 5 presents the confusion matrix for the complete AMS-Net model, revealing classification patterns across flower categories.

The confusion matrix demonstrates AMS-Net's strong discriminative capability, with perfect or near-perfect classification for morphologically distinct categories (Blue Iris: 160/160, White Water Lily: 160/160, Yellow Iris: 160/160). Residual confusion concentrates between visually similar rose varieties, where subtle color gradient differences represent the primary classification challenge.

4.6 Fine-Grained Discrimination Performance

Fig. 6 presents per-category performance metrics, and Table 5 provides detailed F1-scores for rose variety classification demonstrating LSSAM's impact on fine-grained discrimination.

LSSAM significantly enhances fine-grained variety discrimination, with all rose categories achieving F1-scores exceeding 0.97. The challenging Pink-Red Rose category

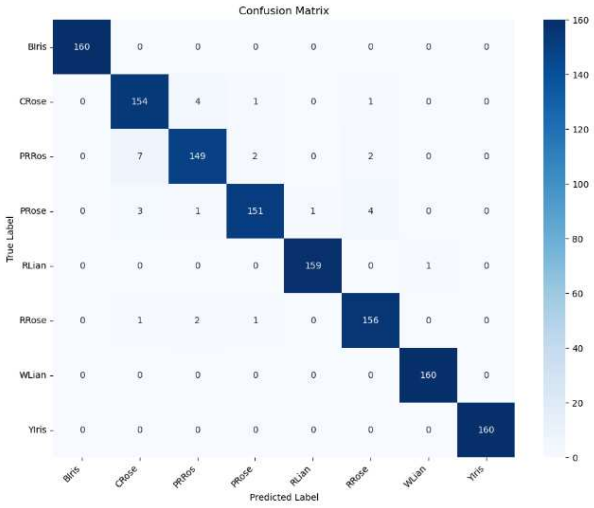


Fig. 5 Confusion matrix for AMS-Net on Flowers-WTT test set. The model achieves near-perfect classification for morphologically distinct categories (water lilies, irises) and high accuracy for fine-grained rose varieties. Minor confusion occurs primarily between visually similar rose categories (Champagne Rose vs. Pink-Red Rose).

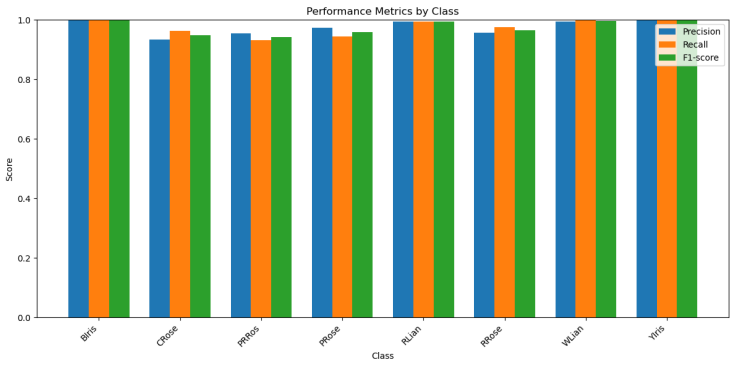


Fig. 6 Per-category precision, recall, and F1-score for AMS-Net. All categories achieve F1-scores exceeding 0.94, with morphologically distinct species (Blue Iris, Yellow Iris, Water Lilies) reaching near-perfect scores. Rose varieties demonstrate strong discrimination despite subtle color differences.

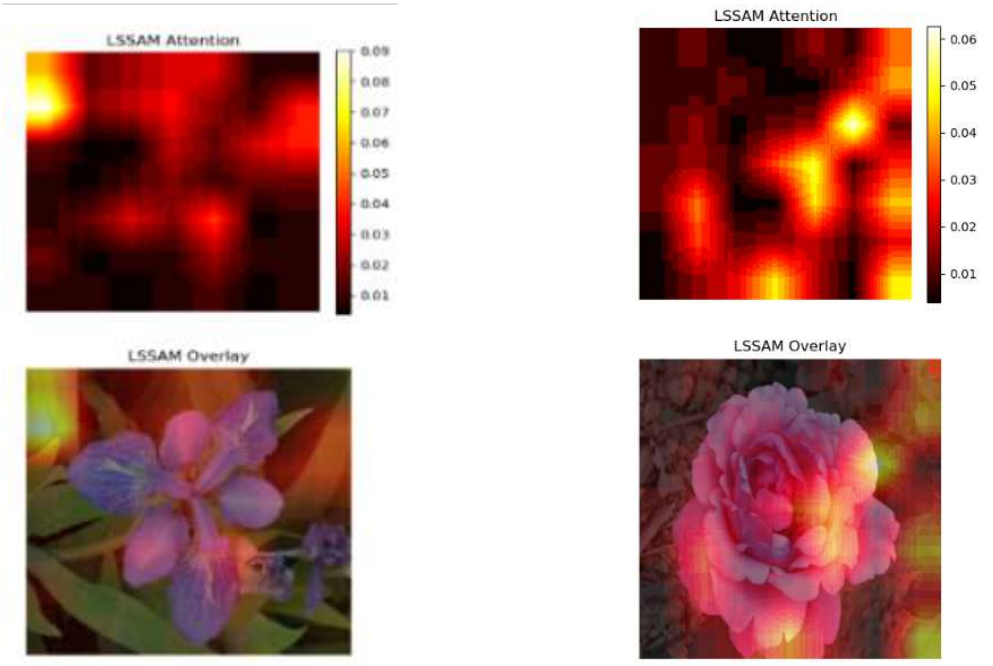
Table 5 Per-Category F1-Scores for Rose Variety Classification

Rose Variety	Baseline F1	AMS-Net F1	Improvement
Red Rose	0.9014	0.9688	+0.0674
Pink Rose	0.9067	0.9750	+0.0683
Champagne Rose	0.9231	0.9811	+0.0580
Pink-Red Rose	0.8824	0.9753	+0.0929

shows the greatest improvement (+0.0929), demonstrating LSSAM's effectiveness in capturing subtle color gradient differences essential for variety-level classification.

4.7 Attention Visualization Analysis

To validate LSSAM's discriminative feature capture mechanisms, we visualize attention patterns using Grad-CAM-style analysis at LSSAM layer 5 (28×28 feature maps). Fig. 4.7 presents representative attention visualizations for morphologically distinct and fine-grained categories.



(a) Blue Iris (b) Pink Rose

Fig. 7 LSSAM attention visualization. (a) Blue Iris: Attention heatmap (top) showing concentrated weights on petal structures and distinctive patterns; overlay (bottom) demonstrating precise localization of discriminative regions while suppressing background foliage. (b) Pink Rose: Attention heatmap (top) revealing distributed focus across petal layers, color gradients, and structural details; overlay (bottom) highlighting fine-grained features essential for variety discrimination including petal edges and color transitions.

The visualizations reveal LSSAM's adaptive attention allocation strategy. For morphologically distinct categories (irises, water lilies), attention concentrates on characteristic structural features. For fine-grained rose varieties, attention distributes across color gradients, texture patterns, and subtle morphological differences—capabilities absent in baseline models. Background regions consistently receive low

attention weights, demonstrating effective irrelevant feature suppression. These attention patterns provide interpretable evidence of LSSAM's contribution to fine-grained multimedia content discrimination, which is particularly relevant for practical multimedia classification systems where model interpretability supports user trust and adoption.

4.8 Generalization Validation on Oxford-17

Table 6 validates AMS-Net's generalization capability on the standard Oxford-17 benchmark with limited training data (40 images/category).

Table 6 Comparison on Oxford-17 Dataset

Method	Accuracy (%)	Parameters (M)
VGG-16	93.5	138.36
MobileNetV2	91.2	2.23
MobileNetV3-Small	92.1	1.53
AMS-Net (Ours)	**94.0**	**1.53**

AMS-Net achieves 94.0% accuracy on Oxford-17, surpassing VGG-16 (93.5%) with 90× fewer parameters. This validates the architecture's generalization capability and effectiveness of transfer learning under data-constrained scenarios typical of edge multimedia deployment applications.

5 Discussion

5.1 Real-Time Multimedia Deployment Considerations

AMS-Net's computational characteristics establish practical viability for diverse multimedia processing scenarios on edge devices. The 17.8MB model size and 0.124 GFLOPs computation enable deployment on contemporary smartphones via iOS Core ML and Android NNAPI frameworks, supporting real-time camera-based classification at ≈60 FPS for botanical identification, smart agriculture, and ecological monitoring applications. For embedded platforms including Raspberry Pi and NVIDIA Jetson Nano, the ∼45MB peak memory footprint accommodates devices with limited RAM. Quantization to INT8 precision can further reduce model size to ∼4.5MB and accelerate inference on specialized edge hardware with minimal accuracy degradation. For IoT sensor networks in agricultural monitoring and biodiversity assessment, AMS-Net's minimal FLOPs translate to reduced energy consumption, extending operational lifetime for remote sensing deployments.

5.2 Attention Mechanism Effectiveness

The 3.05 percentage point accuracy improvement from merely 0.39% parameter overhead demonstrates that strategic attention integration can significantly enhance

lightweight network discriminative capability without compromising real-time performance. The attention visualization analysis (Fig. 7) provides interpretable evidence: for morphologically distinct categories, attention concentrates on characteristic structural features; for fine-grained rose varieties, attention distributes across color gradients, texture patterns, and subtle morphological differences. This adaptive attention behavior—absent in baseline models—directly explains LSSAM's contribution to the substantial F1-score improvements observed in Table 5, particularly the +0.0929 gain for the challenging Pink-Red Rose category. This finding generalizes beyond flower classification to other fine-grained multimedia recognition tasks requiring edge deployment.

5.3 Limitations and Future Directions

Despite strong performance, several limitations warrant future investigation. Current evaluation employs 8-category classification; expansion to larger-scale datasets (e.g., 102+ categories) and evaluation on additional fine-grained benchmarks would strengthen generalization claims. Performance under severe lighting variation, heavy occlusion, and unusual viewpoints requires systematic evaluation for robust outdoor multimedia deployment. Hardware-specific optimizations through TensorRT, ONNX Runtime, and platform-specific quantization can further optimize inference for specific edge hardware targets. Extension to video-based classification incorporating temporal attention for dynamic multimedia monitoring represents a promising future direction, as does multimodal fusion integrating visual features with geographic and temporal metadata for enhanced recognition accuracy in practical multimedia systems.

6 Conclusion

This paper presented AMS-Net, a lightweight architecture achieving real-time multimedia classification through efficient integration of self-attention mechanisms into MobileNetV3-Small. The proposed LSSAM module reduces computational complexity to 1/4096 of standard self-attention through channel reduction (r=16), spatial downsampling (s=2), and grouped attention (G=8), enabling attention benefits without compromising real-time inference capability.

Comprehensive experimental evaluation demonstrates AMS-Net achieves 97.58% classification accuracy with 0.0168s inference latency (≈60 FPS), 1.53M parameters, and 0.124 GFLOPs computational cost. Ablation studies confirm LSSAM contributes 3.05 percentage points accuracy improvement with merely 0.39% parameter overhead. The dynamic transfer learning strategy reduces training time by 35.4% while maintaining classification performance. Comparative evaluation against state-of-the-art architectures establishes AMS-Net's superior accuracy-efficiency trade-off, achieving the highest accuracy with the lowest computational complexity among evaluated models.

These results demonstrate practical viability for deployment on mobile devices and embedded edge platforms, enabling real-time multimedia recognition applications in smart agriculture, ecological monitoring, and botanical education. The proposed techniques generalize beyond flower classification to other fine-grained visual recognition

tasks, advancing real-time multimedia processing capabilities in resource-constrained environments.

Data Availability

The datasets analysed during the current study are available in the Kaggle repository, https://www.kaggle.com/datasets/tommeyshelby208/flowers-wtt.

Author contributions statement

SW conducted experiments, collected and analyzed experimental data, prepared Figures and wrote the first draft of main manuscript text. JS, SW and JY designed the experimental scheme, reviewed the experimental data and polished the first draft of the manuscript. JY, SW and JY reviewed the manuscript and agreed to publish the article.

Funding

This research was supported by the Key Scientific Research Project of Higher Education of Henan Province (No.24A520025), and Henan Province Science and Technology Research Project (No. 252102320286, No.262102240133), Henan Natural Science Foundation (No.262300420683).

References

[1] Howard, A.G.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[2] Shen, J., Liu, N., Sun, H., Wu, S., Liang, Z., Han, L., Zhang, Y., Li, D.: Lightweight semantic feature extraction model with direction awareness for aerial traffic object detection. IEEE Transactions on Intelligent Transportation Systems (2025)

[3] Shen, J., Zhou, W., Liu, N., Sun, H., Li, D., Zhang, Y.: An anchor-free lightweight deep convolutional network for vehicle detection in aerial images. IEEE Transactions on Intelligent Transportation Systems **23**(12), 24330–24342 (2022)

[4] Shen, J., Liu, N., Sun, H., Li, D., Zhang, Y.: An instrument indication acquisition algorithm based on lightweight deep convolutional neural network and hybrid attention fine-grained features. IEEE Transactions on Instrumentation and Measurement **73**, 1–16 (2024)

[5] Nilsback, M.-E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 722–729 (2008). IEEE

[6] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[7] Ma, N., Zhang, X., Zheng, H.-T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 116–131 (2018)

[8] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR

[9] Fei, Y., Li, Z., Zhu, T., Ni, C.: A lightweight attention-based convolutional neural networks for fresh-cut flower classification. IEEE Access **11**, 17283–17293 (2023)

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[11] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[12] Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al.: Rethinking attention with performers. arXiv preprint arXiv:2009.14794 (2020)

[13] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)

[14] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

[15] Shen, J., Liu, N., Xu, C., Sun, H., Xiao, Y., Li, D., Zhang, Y.: Finger vein recognition algorithm based on lightweight deep convolutional neural network. IEEE Transactions on Instrumentation and Measurement **71**, 1–13 (2021)

[16] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

[17] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., *et al.*: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)

[18] Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: Repvit: Revisiting mobile cnn from vit perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15909–15920 (2024)

[19] Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., *et al.*: Mobilenetv4: Universal models for the mobile ecosystem. In: European Conference on Computer Vision, pp. 78–96 (2024). Springer

[20] Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International Conference on Machine Learning, pp. 5156–5165 (2020). PMLR

[21] Child, R.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)

[22] Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S.W., Anwer, R.M., Shahbaz Khan, F.: Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In: European Conference on Computer Vision, pp. 3–20 (2022). Springer

[23] Yang, J., Shen, J., Wang, S.: Pedestrian detection in aerial image based on convolutional neural network with attention mechanism and multi-scale prediction. Scientific Reports **15**(1), 43677 (2025)

[24] Bianco, S., Cadene, R., Celona, L., Napoletano, P.: Benchmark analysis of representative deep neural network architectures. IEEE access **6**, 64270–64277 (2018)

[25] Abd Elaziz, M., Al-qaness, M.A.A., Dahou, A., Alsamhi, S.H., Abualigah, L., Ibrahim, R.A., Ewees, A.A.: Evolution toward intelligent communications: Impact of deep learning applications on the future of 6G technology. WIREs Data Min. Knowl. Discov. **14**(1), 1521 (2024)