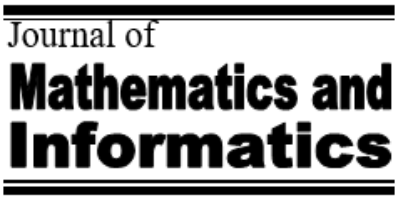
Journal of Mathematics and Informatics
Vol. 31, 2026, 19-31
ISSN: 2349-0632 (P), 2349-0640 (online)
Published 19 September 2026
www.researchmathsci.org
DOI: http://doi.org/10.22457/jmi.v31a03273



A Multi-scale Fuzzy Hybrid Attention Method for Fine-grained Image Classification

Lianwei Lv[1] **and** ***Na Qin***[2*]

[1,2]College of Artificial Intelligence and Computer Science, Northwest Normal University, Lanzhou, 730070, Lanzhou, China. E-mail: 202521162199@nwnu.edu.cn
*Corresponding author. E-mail: qinna@nwnu.edu.cn

Received 31 July 2026; accepted 12 September 2026

Abstract. Fine-grained image classification is vulnerable to attention drift when complex background texture and high-frequency noise compete with subtle inter-class cues. This paper proposes a multi-scale fuzzy hybrid attention network (MSFHA-Net). Its learnable, channel-wise Gaussian membership kernels generate smoothed guidance features for channel and spatial attention, while a residual modulation path preserves discriminative high-frequency details. A multi-scale fuzzy aggregation module combines four spatial scales and three kernel sizes, followed by stacked fuzzy refinement before classification. On Stanford Dogs and Oxford 102 Flowers, MSFHA-Net achieves accuracies of 91.06% and 92.27%, improving on the ResNet-50 baseline by 4.19 and 6.74 percentage points, respectively. Component-wise ablation and interaction analysis show that the fuzzy guidance, multi-scale aggregation, and stacked refinement are complementary. Grad-CAM visualizations further indicate that the model suppresses background responses and concentrates on class-relevant object parts. These results support learnable Gaussian membership weighting as a practical way to couple uncertainty-aware guidance with detail-preserving fine-grained recognition.

Keywords: Fine-grained image classification; attention mechanism; learnable Gaussian membership kernel; multi-scale features; feature decoupling

AMS Mathematics Subject Classification (2010): 68T45, 68U10, 03E72, 68T30

1. Introduction

Fine-grained image classification (FGIC) distinguishes subordinate categories within the same basic-level class, such as dog breeds or flower species. Because inter-class differences are subtle and intra-class appearance varies with pose, scale, and background, an effective model must localize small discriminative regions without discarding context [1].

Existing FGIC methods can be grouped into strongly supervised methods, weakly supervised or unsupervised methods, and methods that use external knowledge. A representative strongly supervised method is Part-based R-CNN [2], which extends the R-CNN detection framework [3] to estimate object and part locations before classification.

Although part annotations can improve localization, their acquisition is expensive and limits scalability. Consequently, recent work has emphasized image-level supervision and end-to-end feature learning.

Weakly supervised approaches use attention or multi-scale learning to discover discriminative regions from class labels alone. Ji et al. [4] combined attention with multi-scale ensemble learning, Zheng et al. [5] introduced multi-attention convolutional learning, and Rao et al. [6] used counterfactual attention to reduce reliance on misleading regions.

Attention estimated directly from noisy features can nevertheless drift toward background texture. Wang et al. [7] addressed scale variation through multi-granularity part sampling attention, but the separation of stable object evidence from high-frequency background responses remains an open problem.

MSFHA-Net addresses this problem by interpreting normalized Gaussian responses as differentiable membership weights. The smoothed feature is used to estimate channel and spatial attention, whereas the original feature remains in the residual modulation path. This separation is the central distinction from conventional attention and from decision-level fuzzy inference [1].

The main contributions are as follows:

(1) A learnable fuzzy attention core (FAC) is formulated from channel-wise Gaussian membership kernels, with complete parameter constraints and an analytical frequency-response interpretation.

(2) A fuzzy hybrid attention mechanism (FHAM) uses the smoothed membership-guidance branch to estimate channel and spatial masks while preserving fine details through residual modulation.

(3) A multi-scale fuzzy aggregation (MSFA) module combines four spatial scales and three kernel sizes; ablation, interaction, computational, and Grad-CAM analyses clarify its contribution and limitations.

2. Related work and mathematical preliminaries

Attention mechanisms are widely used to reweight informative representations. Their roles include improving transformer explanations [8], medical-image segmentation [9], multimodal brain-computer interfaces [10], and long-sequence modeling [11]. Related studies in this journal have also shown that convolutional classification is sensitive to low-quality retinal imagery [26] and to variation in microscopic malaria images [28], motivating explicit control of nuisance responses in visual features.

Multi-scale representations provide complementary local and contextual evidence [12, 13]. They have improved image super-resolution [14], semantic segmentation [15], efficient convolutional learning [16], and feature-pyramid detection [17]. Fuzzy C-means segmentation has likewise demonstrated how graded memberships can stabilize image partitioning under ambiguous boundaries [27]. The present work transfers this graded-membership idea from image-level clustering to differentiable channel-wise feature filtering.

For a feature activation at spatial location (u, v) in channel c, the Gaussian response $\mu_c(u, v; \sigma_c)$lies in [0, 1] and is treated as a membership grade expressing how strongly that location belongs to the channel's locally coherent structure. Normalizing these grades over a k by k neighborhood produces a depthwise convolution kernel. Thus, fuzzification in

MSFHA-Net is differentiable membership weighting, not a rule-based fuzzy controller or a separate defuzzification stage.

The terms used below have precise operational meanings. A fuzzy kernel is the normalized Gaussian membership field K_c. A fuzzy feature is the depthwise-smoothed tensor X_{blur} produced with K_c. Fuzzy attention denotes channel and spatial masks estimated fromX_{blur} and applied to the unsmoothed feature. Fuzzy aggregation denotes averaging FAC outputs over kernel sizes and concatenating them over spatial scales. Feature decoupling refers to separating the guidance path from the detail-preserving modulation path; it does not claim statistical independence.

Compared with prior fuzzy inference that reweights multi-scale decisions [1], MSFHA-Net embeds learnable membership functions inside feature extraction. The novelty is therefore architectural and differentiable: uncertainty is represented by channel-specific smoothing strengths, and these strengths guide attention before classification.

3. Method

3.1 MSFHA-net model

MSFHA-Net is designed to suppress high-frequency background responses without erasing the local texture and shape cues required by fine-grained classification.

For an input batch X_0 in $R^{B\times3\times448\times448}$, the ResNet-50 backbone excludes its global pooling and fully connected layers and returns the Stage-4 feature F_0 in $R^{B\times2048\times14\times14}$. Here B is the batch size, C is the channel count, and H and W are spatial dimensions.

MSFA first projects F_0 to $C' = 512$ channels. It then evaluates the scale set S = {1, 2, 4, 8} and the kernel-size set K = {3, 5, 7}. Each scale branch produces a tensor in $R^{B\times C'\times H_s\times W_s}$; its three FAC outputs are averaged, resized to $H \times W$, concatenated, and fused into F_{MSFA} in $R^{B\times C'\times H\times W}$.

The stacked fuzzy refinement head applies two single-scale FAC blocks to F_{MSFA}. Global average pooling and a linear classifier then map the refined representation to logits z in $R^{B\times N}$, where N is the number of fine-grained classes.

3.2. Parameterized depthwise gaussian fuzzy operator

A fixed Gaussian filter cannot adapt to channels that encode different semantic frequencies. Figure 1 therefore parameterizes a separate Gaussian membership field for each channel and uses it as a differentiable depthwise convolution kernel.

For channel c in $\{1,\ldots,C\}$, let k be an odd kernel size,$r = (k-1)/2$, and (u, v) be an integer coordinate in $\{0,\ldots,k-1^2\}$. The center is (r, r), and $\sigma_c > 0$ controls the membership spread.

The normalized kernel weight at position (u, v) is

$$K_c(u,v) = \frac{1}{Z_c}\exp\left(-\frac{(u-\mu)^2+(v-\mu)^2}{2\sigma_c^2}\right)$$

where $Z_c$is the sum of the unnormalized Gaussian memberships over all k^2 positions. The constant epsilon $= 10^{-8}$ is added only for numerical stability, so the resulting kernel is nonnegative and approximately sums to one.

The channel-wise standard deviation θ_c is optimized jointly with the network. The implementation enforces the interval $[\sigma_{min}, \sigma_{max}]$ through

$$\sigma_c = \mathrm{Clamp}(\exp(\theta_c), \sigma_{min}, \sigma_{max})$$

where θ_c is an unconstrained real-valued parameter, $\sigma_{min} = 0.1$, and $\sigma_{max} = 5.0$. Initializing $\theta_c = 0$ gives $\sigma_c = 1.0$ before clipping, matching the implementation.

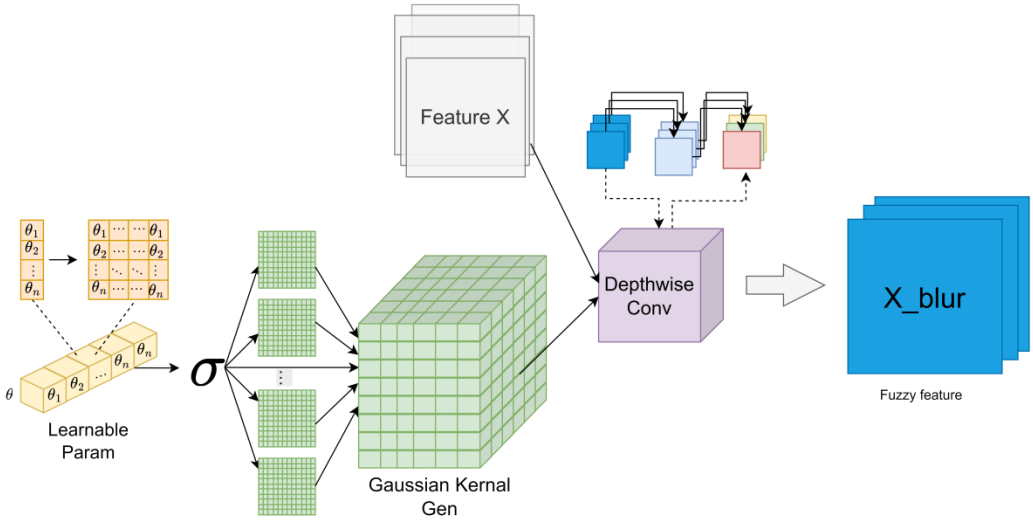


Figure 1: Learnable depthwise Gaussian membership operator

The role of σ_c follows from the frequency response of a Gaussian. In the continuous approximation, the response magnitude is

$$|H_c(\omega)| = \exp[-\frac{\sigma_c^2}{2}(\omega_x^2 + \omega_y^2)]$$

For every nonzero spatial frequency, differentiating this expression with respect to σ_c gives a negative value. Increasing σ_c therefore attenuates high-frequency responses more strongly and emphasizes slowly varying shape, whereas decreasing σ_c retains more edge and texture energy. The finite normalized kernel used in the network follows the same qualitative relation within the imposed bounds.

3.3 Fuzzy Hybrid Attention Mechanism

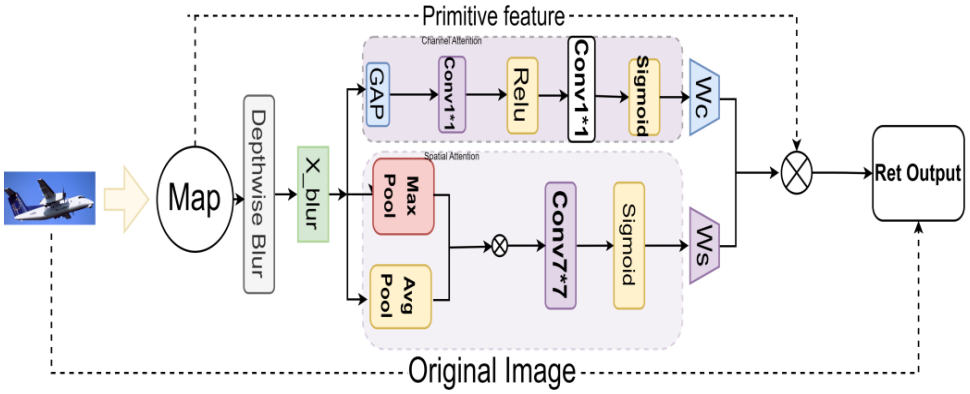


Figure 2: Fuzzy hybrid attention mechanism

Figure 2 shows one fuzzy attention core (FAC). The upper and lower mask paths estimate channel and spatial importance from the smoothed guidance feature X_{blur}. Their product modulates the original preprocessed feature X, and a scaled residual connection return Y. The guidance path is therefore noise-resistant, while the modulation path preserves high-frequency evidence.

(1) Structural Semantics Extraction

Let X in $R^{B\times C\times H\times W}$ be the preprocessed input. The depthwise Gaussian operator acts independently on each channel:

$$X_{blur}^{(c)} = X^{(c)} * K_c, c = 1, \dots, C$$

(2) Blur-Guided Channel Attention

Channel importance is estimated only from X_{blur}, which reduces the chance that isolated background responses dominate the descriptor.

Global average pooling maps X_{blur} to $R^{B\times C\times 1\times 1}$. A two-layer bottleneck with reduction ratio $r = 4$ then produces the channel mask

$$w_c = \text{sigmoid}(W_2\delta(W_1 GAP(X_{blur}))) \in R^{B\times C\times 1\times 1}$$

Here delta is ReLU, sigmoid is the logistic function, W_1 is in $R^{((C/r)\times C)}$, and W_2 is in $R^{(C\times(C/r))}$. Broadcasting applies w_c across H and W.

(3) Blur-Guided Spatial Attention

Spatial importance is computed by average and maximum pooling over the channel dimension of X_{blur}. Concatenating the two $B \times 1 \times H \times W$ maps gives F_{cat} in $R^{B\times 2\times H\times W}$; a 7×7 convolution followed by a sigmoid produces w_s in $R^{B\times 1\times H\times W}$.

$$F_{cat} = [\text{AvgPool}(X_{blur}); \text{MaxPool}(X_{blur})] \in R^{B\times 2\times H\times W}$$

$$w_s = \sigma(\text{Cov}_{7\times 7}(F_{cat})) \in R^{B\times 1\times H\times W}$$

(4) Origin Modulation and Residual

The channel and spatial masks are broadcast and multiplied with the unsmoothed feature:

$$X_{refined} = X \odot w_c \odot w_s$$

The learnable scale gamma in $R^{1\times C\times 1\times 1}$ is initialized to 10^{-2}. DropPath applies stochastic depth during training and is the identity at inference. The residual output Y therefore has the same $B \times C \times H \times W$ shape as X.

$$Y = X + \text{DropPath}(\gamma \cdot X_{refined})$$

This two-path construction explains the term feature decoupling: X_{blur} provides stable guidance for localization, whereas X retains the high-frequency details needed to distinguish subtle structures such as facial markings or stamens. The residual connection prevents the Gaussian branch from becoming an irreversible low-pass bottleneck.

3.4 Multi-scale fuzzy aggregation

Figure 3 illustrates MSFA as a grid of spatial scales and blur extents. The horizontal dimension changes spatial resolution, the within-branch FAC units change Gaussian

support, and the final fusion aligns all outputs to a common $B \times C' \times H \times W$ representation.

(1) Parallel Spatial Pyramid

Let S={1,2,4,8}. For each s in S, adaptive average pooling maps the input F_{in} in $R^{B\times C\times H\times W}$ to $H_s=max(1,floor(H/s))$ and $W_s=max(1,floor(W/s))$; s=1 is the identity branch.

$$F_{down}^{(s)}=AvgPool(F_{in},kernel=s,stride=s)$$

(2) Multi-Kernel Fuzzy Ensemble

For each scale, the kernel-size set K={3,5,7} supplies three FAC units. Their outputs have identical shape and are averaged, so each scale contributes equally without increasing the channel count before resizing.

$$F_{feat}^{(s)}=\frac{1}{|K|}\sum_{k\in K} FAC(F_{down}^{(s)};k,\sigma_{learnable})$$

(3) Global Fusion

Each $F_{feat}^{(s)}$ is bilinearly resized to $H\times W$. Concatenation over $|S|$=4 scales producesF_{concat} in $R^{B\times(4C')\times H\times W}$.

$$F_{concat}=Concat(Up(F_{feat}^{(1)}),\ldots,Up(F_{feat}^{(8)}))$$

$$F_{concat}\in R^{(|S|\cdot C)\times H\times W}$$

A 1 × 1 convolution, batch normalization, and ReLU fuse F_{concat} into F_{MSFA} in $R^{B\times C'\times H\times W}$. This operation learns cross-scale mixtures while keeping the spatial resolution unchanged for the refinement head.

3.5 Stacked fuzzy refinement head

The stacked fuzzy refinement head progressively recalibrates the aggregated feature immediately before classification.

As shown in Figure 4, the pipeline is feature extraction, channel reduction, MSFA, two serial FAC blocks, global average pooling, and a linear classifier. The two FAC blocks keep the $B\times C'\times H\times W$ shape fixed and successively refine channel and spatial masks.

For L serial refinement blocks, stochastic-depth probability increases linearly with block index l to regularize deeper residual paths:

$$p_l=p_{max}\cdot\frac{l-1}{L-1}$$

where, $l=\{0,\ldots,L-1\}$ and p_{max} = 0.1. At inference, all residual paths are active. Global average pooling yields h in $R^{B\times C'}$, and the classifier produces logits z in $R^{B\times N}$

Training minimizes the mean cross-entropy between softmax(z) and the ground-truth labels

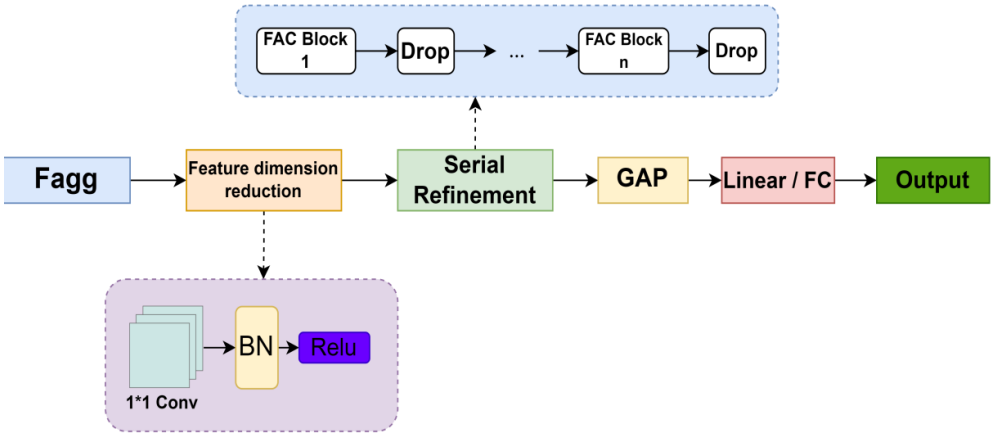


Figure 3: Multi-Scale Fuzzy Aggregation

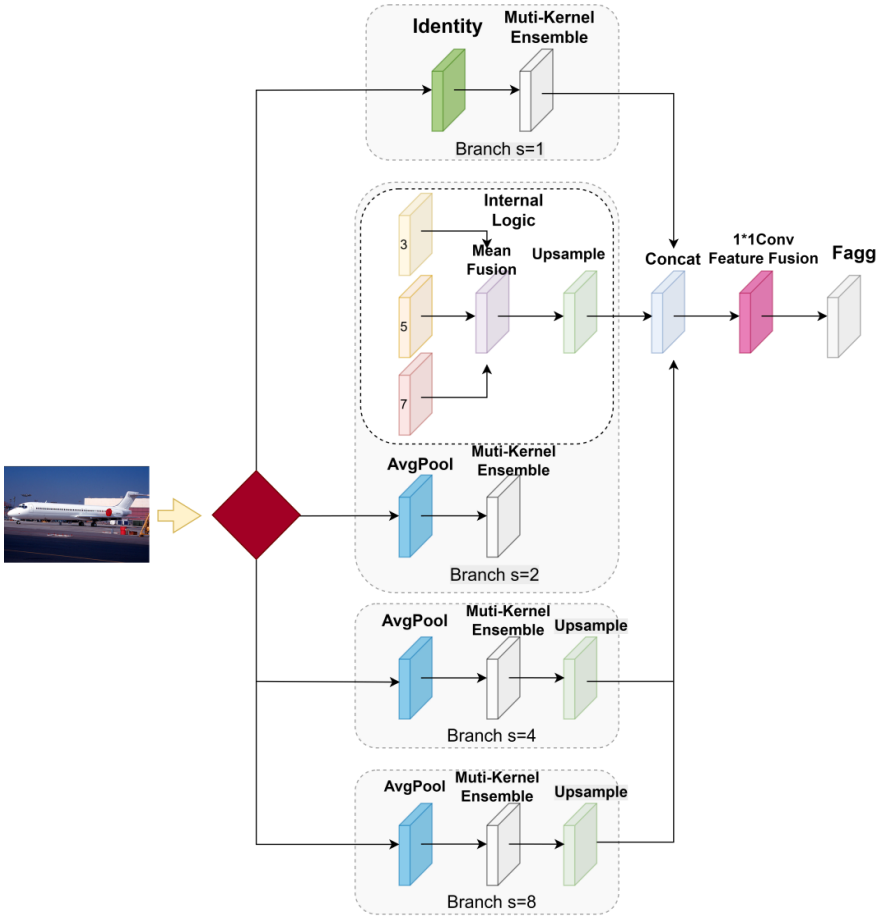


Figure 4: Stacked Fuzzy Refinement

4. Experimental results

4.1. Configuration and reproducibility

Experiments use Stanford Dogs [18] and Oxford 102 Flowers [19]. Stanford Dogs contains 20,580 images from 120 breeds, with 12,000 training and 8,580 test images. Oxford 102 Flowers contains 8,189 images from 102 classes and follows its standard train, validation, and test partitions. All images are resized and center-cropped to 448 × 448 pixels. The remaining settings are summarized in Table 1. The random seed is fixed at 42 for Python, NumPy, and PyTorch. The reported accuracies are the retained outcomes of this fixed-seed protocol; repeated-run confidence intervals cannot be reconstructed from the available checkpoints and are therefore not claimed. This limitation is considered in Section 4.4.

Table 1: Experimental configuration and training hyperparameters

Category	Item	Parameters / Settings
Hardware Environment	Framework	PyTorch
	GPU Model	NVIDIA Tesla A30 × 2
Backbone Network	Architecture	ResNet-50
	Pre-trained Weights	ImageNet-1k
Data Preprocessing	Method	Resize + Center Crop
	Input Resolution	448 × 448
Model Architecture	Blur Blocks	2 (Default)
	Multi-scale Pyramid	{1, 2, 4, 8}
	Parallel Blur Kernels	{3, 5, 7}
Training Strategy	Phase 1: Warm-up	80 Epochs, LR: 5×10^{-4}, Frozen Backbone
	Phase 2: Fine tuning	20 Epochs, LR: 5×10^{-5}, End-to-End
Optimizer Settings	Type	AdamW
	Weight Decay	1×10^{-4}
	LR Scheduler	Cosine Annealing
Reproducibility	Random Seed	42

4.2. Ablation study

Table 2 isolates the two principal paths of MSFHA-Net. On Oxford 102 Flowers, ResNet-50 achieves 85.53%. Adding the single-scale FAC increases accuracy to 87.87% , correcting the inconsistency in the previous text. Using fuzzification with MSFA but without stacked single-scale refinement reaches 90.19%. The complete model attains 92.27%, confirming that the final refinement remains useful after multi-scale aggregation.

On Stanford Dogs, the single-scale FAC improves accuracy from 86.87% to 87.46%, while the MSFA configuration reaches 88.16%. Combining MSFA with stacked refinement raises accuracy to 91.06%, 4.19 percentage points above the baseline and 2.90 points above the stronger isolated branch.

Module interaction can be quantified as $\Delta_{int} = A_{full} - A_{FAC} - A_{MSFA} + A_{base}$. Δ_{int} is +2.31 points on Stanford Dogs, indicating super-additive complementarity, and -0.26 points on Oxford 102 Flowers, indicating mild redundancy even though the full model

remains 2.08 points better than MSFA alone. Thus, fuzzy guidance supplies noise-resistant localization, MSFA expands scale coverage, and stacked refinement is most beneficial when the two cues are complementary.

Table 2: Ablation study on Stanford Dogs and Oxford 102 Flowers

ResNet-50	Stacked FAC	Fuzzification	MSFA	Dogs (%)	Flowers (%)
√				86.87	85.53
√	√	√		87.46	87.87
√		√	√	88.16	90.19
√	√	√	√	91.06	92.27

4.3. Comparative results and analysis

Table 3 compares MSFHA-Net with MSA [20], complemented spatial enhancement (CSE) [21], a hybrid attention module (HAM) [22], feature reconstruction bias (FRB) [23], and ANI-CNN [24]. MSFHA-Net reaches 91.06% on Stanford Dogs and 92.27% on Oxford 102 Flowers. The gain is attributable to channel-wise learnable smoothing, parallel scale processing, and dual channel-spatial calibration. Because the cited methods use their published training protocols, the table is a literature comparison rather than a controlled reimplementation; the ResNet-50 ablation in Table 2 provides the controlled baseline.

Table 3: Comparative classification accuracy (%)

Method	Stanford Dogs	Oxford 102 Flowers
MSA	89.62	88.89
CSE	89.84	84.29
HAM	89.11	89.57
FRB	-	87.60
ANI-CNN	-	91.90
Ours	91.06	92.27

4.4. Computational cost and limitations

Computational cost was measured from the implementation described in Section 3 using one 448 x 448 input and a 120-class output layer. PyTorch operation counting gives 31.21 million parameters and 34.16 GFLOPs for MSFHA-Net, compared with 23.75 million parameters and 32.70 GFLOPs for ResNet-50. The proposed head therefore adds 7.46 million parameters (31.4%) but only 1.46 GFLOPs (4.5%), because Gaussian filtering is depthwise and most multi-scale branches operate at reduced resolution.

Table 4: Computational cost at 448 x 448 input resolution

Model	Parameters (M)	FLOPs (G)	Accuracy (%)
ResNet-50	23.75	32.70	86.87 / 85.53
MSFHA-Net	31.21	34.16	91.06 / 92.27

The two accuracy values are reported in the order Stanford Dogs / Oxford 102 Flowers. FLOPs count multiplication and addition as separate operations. Changing the output layer from 120 to 102 classes changes the parameter total by less than 0.01 million.

The evaluation has three boundaries. First, both benchmarks are curated and do not represent severe domain shift, label noise, or occlusion. Second, only a fixed-seed run is available, so run-to-run variance is unknown. Third, Grad-CAM provides qualitative localization evidence but does not prove causal reliance on the highlighted regions. Future work will evaluate repeated seeds, corrupted labels, cross-domain transfer, severe occlusion, and larger fine-grained datasets while reporting latency on deployment hardware.

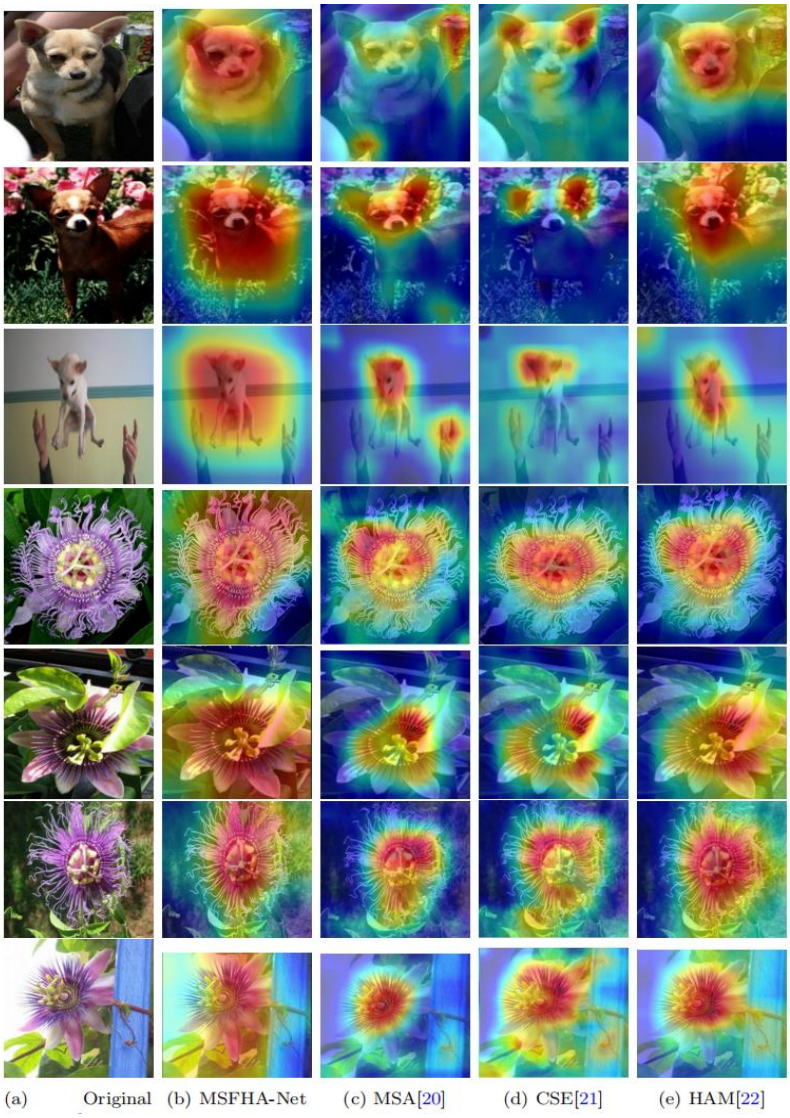


Figure 5: Visual analysis of comparative experiments

4.5. Visualization analysis

Grad-CAM [25] is used to inspect the class-discriminative regions of MSFHA-Net, MSA [20], CSE [21], and HAM [22]. Figure 5 arranges the original image in the first column

and the four model heatmaps in the remaining columns, allowing foreground coverage and background activation to be compared row by row.

For dog images, competing methods sometimes activate cans, grass, or fingers. MSFHA-Net concentrates more consistently on the dog and covers complementary parts such as the face, ears, chest, and trunk. This pattern is consistent with blur-guided masks suppressing isolated background responses.

For flowers with dense textures, MSFHA-Net covers the central stamens and inner petals more continuously than the comparison methods. These visualizations support the localization interpretation of the method, although, as noted in Section 4.4, they remain qualitative explanations rather than causal tests.

5. Conclusion

This paper presented MSFHA-Net, which embeds channel-wise learnable Gaussian membership kernels into a hybrid attention architecture. The smoothed guidance path estimates stable channel and spatial masks, the original path preserves fine details, and MSFA combines four spatial scales with three kernel sizes before stacked refinement. The model achieves 91.06% accuracy on Stanford Dogs and 92.27% on Oxford 102 Flowers, improving on ResNet-50 by 4.19 and 6.74 percentage points. Ablation and interaction analyses show that the modules are complementary, while the cost analysis shows a 4.5% FLOP increase over ResNet-50 at 448 × 448 resolution. The approach is useful when background texture competes with subtle object cues, but its generalization to noisy labels, domain shift, severe occlusion, and more diverse datasets remains to be established through repeated-run and cross-domain evaluation.

Acknowledgements. The authors thank the anonymous reviewer for constructive comments that improved the mathematical presentation, reproducibility, and clarity of this paper.

Conflicts of Interest. The authors confirm that there are no conflicts of interest associated with the research, authorship, or publication of this manuscript.

Author Contributions. Both authors contributed to the conceptualization of the study, conducted the research, contributed to the preparation of the manuscript, and reviewed and approved the final version for publication.

Declarations funding. This work was supported by the National Natural Science Foundation of China (12471444) and the Northwest Normal University Research Capacity Enhancement Program for Young Teachers (NWNU-LKQN2022-03).

REFERENCES

1. Q. Chen, F. He, G. Wang, X. Bai, L. Cheng and X. Ning, Dual guidance enabled fuzzy inference for enhanced fine-grained recognition, IEEE Transactions on Fuzzy Systems 33(1) (2025) 418–430. https://doi.org/10.1109/TFUZZ.2024.3427654.
2. N. Zhang, J. Donahue, R. Girshick and T. Darrell, Part-based R-CNNs for fine-grained category detection, in: European Conference on Computer Vision (2014) 834–849. https://doi.org/10.1007/978-3-319-10599-4_54.
3. R. Girshick, J. Donahue, T. Darrell and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (2014) 580–587. https://doi.org/10.1109/CVPR.2014.81.

4. S. Ji, Z. Jiang, X. Ma and L. Yang, Attention and multi-scale ensemble learning based network for fine-grained image recognition, Journal of Data Acquisition and Processing 40(2) (2025) 384–400.
5. H. Zheng, J. Fu, T. Mei and J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in: IEEE International Conference on Computer Vision (2017) 5209–5217.
6. Y. Rao, G. Chen, J. Lu and J. Zhou, Counterfactual attention learning for fine-grained visual categorization and re-identification, in: IEEE/CVF International Conference on Computer Vision (2021) 1025–1034.
7. J. Wang, Q. Xu, B. Jiang, B. Luo and J. Tang, Multi-granularity part sampling attention for fine-grained visual classification, IEEE Transactions on Image Processing 33 (2024) 4529–4542.
8. M. Eggen, J. Lysnaes-Larsen and I. Struemke, Integrating attention into explanation frameworks for language and vision transformers, arXiv:2508.08966 (2025). https://doi.org/10.48550/arXiv.2508.08966.
9. I. U. Haq, A. Iqbal, M. Anas, F. Masood, A. S. Alzahrani and M. AlNaeem, GAME-Net: An ensemble deep learning framework integrating generative autoencoders and attention mechanisms for automated brain tumor segmentation in MRI, Frontiers in Computational Neuroscience 19 (2025) 1702902. https://doi.org/10.3389/fncom.2025.1702902.
10. J. Wang, W. Ye, J. He, L. Zhang, G. Huang, Z. Yu and Z. Liang, Integrating biological and machine intelligence: Attention mechanisms in brain-computer interfaces, Information Fusion 125 (2026) 103417. https://doi.org/10.1016/j.inffus.2025.103417.
11. D. Bai, T. Zhang, Q. Zhang, J. Zhang, Z. Jin and L. Li, Investigating the effect of attention mechanisms on deep neural network engineering performance and brain-like properties, Neurocomputing 664 (2026) 132094.
12. R. F. Betzel and D. S. Bassett, Multi-scale brain networks, NeuroImage 160 (2017) 73–83.
13. H. Shen, X. Cheng and B. Fang, Covariance, correlation matrix and the multi-scale community structure of networks, Physical Review E 82(1) (2010) 016114. https://doi.org/10.1103/PhysRevE.82.016114.
14. F. Liu, A. Jiang, B. Wang and L. Chen, A multi-scale and multi-dense network for image super-resolution, Circuits, Systems, and Signal Processing 44(12) (2025).
15. T. Li, Z. Cui and D. Wei, Enhanced multi-scale networks for semantic segmentation, Complex & Intelligent Systems 10(2) (2024) 2557–2568.
16. G. Wang, G. Yuan, T. Li and L. Meng, A multi-scale learning network with depthwise separable convolutions, IPSJ Transactions on Computer Vision and Applications 10(11) (2018) 1–8.
17. X. Zhao, J. Chen, M. Liu, K. Ye and L. Shen, Multi-scale attention-based feature pyramid networks for object detection, in: 11th International Conference on Image and Graphics (2021) 405–417.
18. A. Khosla, N. Jayadevaprakash, B. Yao and L. Fei-Fei, Novel dataset for fine-grained image categorization: Stanford Dogs, in: CVPR Workshop on Fine-Grained Visual Categorization (2011). http://vision.stanford.edu/aditya86/ImageNetDogs/.

19. M.-E. Nilsback and A. Zisserman, Automated flower classification over a large number of classes, in: Indian Conference on Computer Vision, Graphics and Image Processing (2008) 722–729. https://doi.org/10.1109/ICVGIP.2008.47.
20. P. Zhao, S. Yang, W. Ding, R. Liu, W. Xin, X. Liu and Q. Miao, Learning multi-scale attention network for fine-grained visual classification, Journal of Information and Intelligence 3(6) (2025) 492–503. https://doi.org/10.1016/j.jiixd.2025.04.005.
21. X. Zhao, J. Wang, Y. Li, Y. Wang and Z. Miao, Complemented attention method for fine-grained image classification, Journal of Image and Graphics 26(12) (2021) 2860–2869.
22. W. Lu, Y. Yang and L. Yang, Fine-grained image classification method based on a hybrid attention module, Frontiers in Neurorobotics 18 (2024) 1391791. https://doi.org/10.3389/fnbot.2024.1391791.
23. Z. Li, L. Wang, S. Ding, X. Yang and X. Li, Few-shot classification with feature reconstruction bias, in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (2022) 526–532. https://doi.org/10.23919/APSIPAASC55919.2022.9980086.
24. M. Qin, Y. Xi and F. Jiang, A new improved convolutional neural network flower image recognition model, in: IEEE Symposium Series on Computational Intelligence (2019) 3110–3117.
25. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision (2017) 618–626. https://doi.org/10.1109/ICCV.2017.74.
26. Y. Ge, G. Zhang and J. Wang, Low quality retinal image recognition based on convolutional neural network, Journal of Mathematics and Informatics 19 (2020) 37–46. https://doi.org/10.22457/jmi.v19a05178.
27. Y. Cheng, G. Zhang, G. Han and Y. Song, Based on improved level set algorithm to enhanced fuzzy C-means clustering for controlled knife image segmentation, Journal of Mathematics and Informatics 20 (2020) 51–60. https://doi.org/10.22457/jmi.v20a06193.
28. H. Wapalila, S. Kaijage and J. Leo, Evaluation of the performance of deep learning classification models in microscopic image results for malaria disease, Journal of Mathematics and Informatics 26 (2024) 49–63. https://doi.org/10.22457/jmi.v26a05241.