Review of Computer Engineering Research
2025 Vol. 12, No. 3, pp. 195-205
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v12i3.4439
© 2025 Conscientia Beam. All Rights Reserved.

Spatial multi-scale feature transformer network for fine-grained few-shot image classification

Liyong Guo[1]
Erzam Marlisah[2+]

[1,2]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia.*
[1]*Email:* guoliyong685@gmail.com
[2]*Email:* erzam@upm.edu.my

(+ Corresponding author)

Article History
Received: 5 June 2025
Revised: 26 August 2025
Accepted: 18 September 2025
Published: 2 October 2025

Keywords
Few-shot learning
Fine-grained few-shot image classification
Fine-grained image classification
Multi-scale features
Spatial transformer network.

ABSTRACT

This year has seen significant advancements in deep learning, and fine-grained few-shot image classification (FGFSIC) has also made substantial progress. FGFSIC faces two key challenges: high intra-class variance and low inter-class variance, which hinder accurate classification with limited data. Despite considerable efforts to extract more discriminative features using powerful networks, few studies have specifically addressed these challenges. This paper proposes a Spatial Multi-Scale Feature Transformer Network to overcome these issues. The approach first modifies the backbone network to extract multi-scale features, with classification results derived from comparing these multi-scale representations. Additionally, a Spatial Feature Transformer network is introduced to adjust the spatial positions of multi-scale features, which helps to reduce intra-class variance. Experiments were conducted on three widely used datasets—CUB-200-2011, Stanford Cars, and Stanford Dogs. The results demonstrate that both components of the proposed model significantly enhance FGFSIC performance, with final accuracies surpassing those of most existing methods. The findings emphasize the effectiveness of the proposed approach in tackling the critical issues of high intra-class variance and low inter-class variance, making it a promising solution for fine-grained image classification tasks, particularly in situations where only limited data is available. This work paves the way for improved performance in real-world applications requiring precise, few-shot learning in fine-grained domains.

Contribution/Originality: This study introduces a Spatial Multi-Scale Feature Transformer Network designed to address the primary challenges of Fine-Grained Forest Species Identification and Classification (FGFSIC): high intra-class variance and low inter-class variance. The utilization of multi-scale features enhances inter-class variance, enabling better differentiation among species. Simultaneously, the spatial transformer reduces intra-class variance by aligning features within the same class, leading to more consistent representations.

1. INTRODUCTION

With the rapid development of deep learning in recent years, significant breakthroughs have also been made in the field of computer vision. As an essential area within computer vision, fine-grained few-shot image classification (FGFSIC) has presented numerous development opportunities. Fine-grained image classification (FGIC) involves distinguishing between categories within the same super-class but belonging to different subordinate classes. FGIC aims to differentiate closely related categories, such as various dog breeds, bird species, or car models. For example, while both the Alaskan Malamute and the Alaskan Husky belong to the same super-class dogs they represent different subordinate classes. As illustrated in Figure 1, FGIC focuses on classifying different bird species, all of which belong to the same super-class, bird.

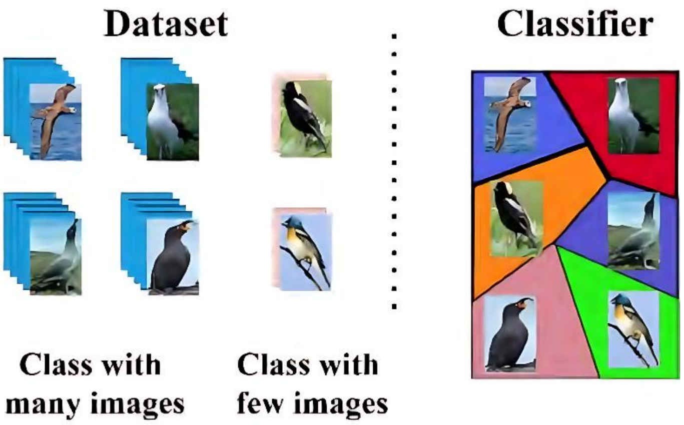

Figure 1. The difference between traditional fine-grained image classification and fine-grained few-shot image classification.

FGFSIC is a type of FGIC that operates under more constrained conditions, where only a few images per class are available for training. The difference between FGIC and FGFSIC is illustrated in Figure 1: While FGIC typically has abundant samples for each class, FGFSIC must learn to classify categories with minimal examples. This presents greater challenges, as it requires overcoming the inherent difficulties of fine-grained classification with sparse data.

The main challenge of FGFSIC is addressing the problem of high intra-class variance and low inter-class variance with limited data. Low inter-class variance is caused by fine-grained images being too similar to each other, while high intra-class variance results from differences in pose, background, viewpoint, and other factors. These issues significantly affect the final classification performance. To address the low inter-class variance problem, some works [1, 2] employ complex networks to extract more discriminative features and better distinguish fine-grained images. However, these methods fail to resolve the high intra-class variance issue. Other studies, Huang et al. [3] and Huang et al. [4], have attempted to address high intra-class variance by aligning the features of query and support images. Nevertheless, neither of these two types of methods addresses both high intra-class variance and low inter-class variance comprehensively. Based on these observations, this paper proposes a Spatial Multi-Scale Feature Transformer Network, which exploits multi-scale features to enhance feature discriminativeness and learns a Spatial Transformer Network to adjust the spatial positions of multi-scale features. The use of multi-scale features increases inter-class variance, while the Spatial Transformer Network refines the structure of extracted features, thereby reducing intra-class variance. The main contributions of this paper are summarized as follows:

1) This paper proposes the use of multi-scale features to enhance the discriminativeness of extracted features and increase inter-class variance, thereby improving classification accuracy.
2) A Spatial Transformer Network is employed to reduce high intra-class variance, making images from the same class easier to recognize.
3) The proposed method is evaluated on three widely used fine-grained few-shot image classification datasets. Experimental results demonstrate that our approach effectively improves classification performance and outperforms several recent state-of-the-art methods.

2. RELATED WORK

2.1. Few-Shot Learning

Few-shot learning involves learning models that can efficiently learn and make accurate predictions using only a small number of samples. There are several types of few-shot learning methods: metric-based methods, optimization-based methods, and augmentation-based methods. There are several famous metric-based methods, including prototypical networks [5], relation networks [6], and matching networks [7]. They all aim to learn a metric space where image features from the same class cluster together while image features from different classes are separated. A prototypical network [5] is proposed to calculate the Euclidean distance between the query features and support prototypes, which are the means of the features from the same class of the support images. A relation network [6] compares query features with each support feature to learn a Cosine distance, which is used to form the final prediction. The matching network [7] is based on concepts from metric learning, utilizing external memory to enhance the network and improve its learning capabilities.

MAML [8] is an optimization-based method, and its main idea is to train a set of initial parameters, enabling the model to achieve fast convergence with only a small amount of data by relying on these initial parameters. Latent Embedding Optimization [9] is a meta-learning framework that, instead of directly optimizing model parameters for each task, optimizes a latent embedding space that indirectly generates those parameters via a decoder or generator. It adapts to new tasks by optimizing in a latent space rather than a parameter space, thereby reducing overfitting when data is scarce.

Augmentation-based methods can generate additional samples to enhance the accuracy of few-shot learning. MetaGAN [10] integrates a generative adversarial network (GAN) into the meta-learning pipeline to synthesize additional examples in feature space for novel classes, improving classification with very few samples. Hallucinating new examples by using a learned generator to produce plausible features for unseen classes is another augmentation-based method, Hariharan and Girshick [11]. Wang et al. [12] extend the feature hallucination concept by introducing an adversarial learning framework, where a feature hallucinator generates new features, and the discriminator ensures their plausibility [13].

Among the three categories of methods, metric-based approaches are particularly popular and serve as the primary focus of this study. This preference is attributed to their simplicity, ease of implementation, and interpretability. Metric-based methods perform classification by measuring the distances between support and query embeddings. Furthermore, they eliminate the need for complex fine-tuning of model parameters during the testing phase—requiring only feature embedding extraction and similarity computations. These methods also enable rapid inference on new tasks without the necessity for retraining or adaptation loops. By directly computing support set prototypes and comparing them to query samples, metric-based approaches achieve high efficiency, making them especially suitable for real-time or online few-shot learning applications.

2.2. Fine-Grained Image Classification

Fine-grained image classification (FGIC) aims to classify images into subordinate categories that belong to the same super-class. Since these categories share the same super-class, they are often very similar and difficult to distinguish. Early works [14, 15] typically relied on part annotations or bounding boxes to extract more discriminative features for FGIC. However, these methods are labor-intensive, time-consuming, and require additional annotations.

With the advancement of deep learning, alternative approaches that do not depend on part annotations have been proposed. These methods are generally referred to as deep learning-based approaches [1, 7, 16]. They primarily rely on specially designed networks to extract more discriminative features. These approaches can be further divided into two subfields: feature encoding-based methods and part localization-based methods.

Feature encoding-based methods [1, 2, 8] focus on designing effective network architectures or modules to capture more discriminative and fine-grained visual features from the entire image. For example, bilinear CNN [1], compact bilinear pooling [2], and Attentional Kernel Encoding Networks [17] have been proposed to enhance feature representation at the global or local level. These methods improve the model's ability to distinguish between visually similar categories without requiring explicit part annotations.

Part localization-based methods [18-20] aim to identify and localize informative object parts or regions that are crucial for distinguishing between fine-grained categories. These approaches typically involve learning to detect key parts of an object (e.g., the beak, wings, or tail of a bird) and then leveraging features extracted from these localized regions to enhance classification performance.

Feature encoding-based methods and part localization-based methods can address the key challenges of FGIC and achieve relatively higher accuracy. However, these methods typically rely on large, labeled datasets, which are difficult to obtain in real-world scenarios. Effectively utilizing existing data and learning from a limited number of samples remain significant challenges.

2.3. Fine-Grained Few-Shot Image Classification

Fine-grained few-shot image classification (FGFSIC) initially adopted a bilinear CNN to extract more discriminative features, while sub-classifiers were employed to reduce the number of parameters [21]. To further improve FGFSIC performance, LRPABN [3] was proposed to align corresponding regions between query and support images spatially. CPSN [22] compares images at the patch level, focusing on local discriminative regions—a crucial approach for fine-grained tasks where subtle visual cues are essential. Performing patch-wise similarity comparisons effectively reduces the influence of irrelevant background areas by emphasizing regions with high similarity and down-weighting uninformative patches. TOAN [4] performs target-oriented alignment between query and support images, ensuring that corresponding semantic or spatial regions are accurately matched. Meanwhile, long-short-range alignment [23] enables the model to handle part misalignments, occlusions, and viewpoint changes by flexibly aligning semantically similar regions, regardless of their spatial position in the image.

All these models focus on reducing misalignment between query and support images, which helps decrease high intra-class variance. However, they often overlook the importance of increasing low inter-class variance. MattML [24] addresses this by capturing multi-level discriminative features that help distinguish subtle inter-class differences in fine-grained images. Results show that addressing only one of these challenges is insufficient to achieve very high accuracy. To this end, a spatial multi-scale feature transformer network is introduced to tackle both problems simultaneously and further boost performance.

3. METHODOLOGY

This section primarily introduces the proposed method of the spatial multi-scale feature transformer network. It begins with an overview of the problem. The multi-scale feature extraction and the spatial multi-scale feature transformer are explained to demonstrate how to address the issues of high intra-class variance and low inter-class variance. The entire process is illustrated in Figure 2.

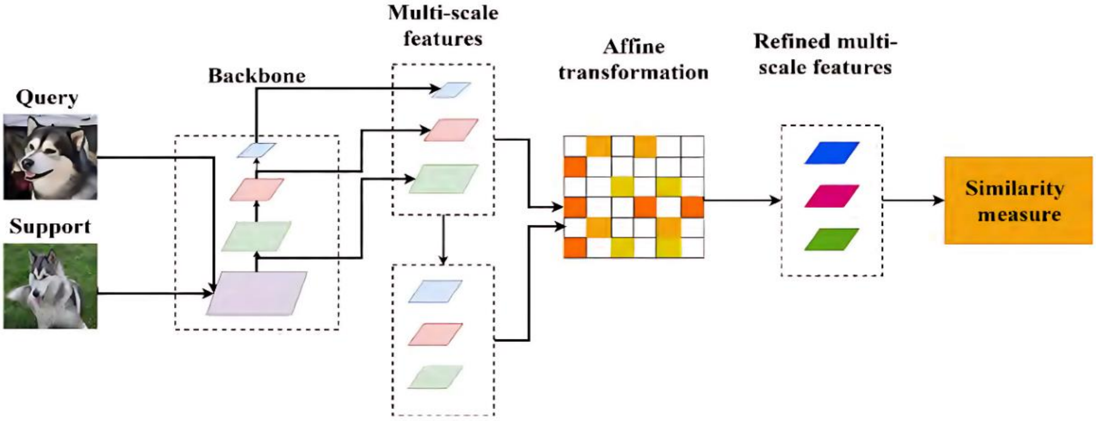

Figure 2. The framework of the proposed method. The original images are input into the backbone to extract multi-scale features, which are then used to learn an affine transformation matrix that adjusts the spatial positions of the multi-scale features. After the transformation, the refined multi-scale features are used to compute similarity scores across three scales.

3.1. Problem Definition

This research primarily focuses on N-way K-shot learning. This is a type of few-shot learning where the data is limited to N classes, and each class has K images.

The dataset in few-shot learning is typically divided into a training set, a validation set, and a test set, which are disjoint from each other. Usually, the test set is divided into a support set and a query set. The query images are used to compare with the support images to obtain the predicted labels. To minimize the loss and enable rapid model adaptation (As in Vinyals et al. [7]), the training set and validation set are also divided into support and query sets. The training, validation, and test procedures are implemented in multiple episodes. Each episode randomly selects N×K images as support images and some query images to update the model parameters. After many episodes, the model parameters can stabilize. This is referred to as the episodic training mechanism [7].

3.2. Multi-Scale Feature Extraction

Most recent works [3, 4] rely solely on global features for classification. However, this approach is insufficient for fine-grained image classification, which requires extracting more discriminative information from the original images. When the backbone network is fixed, the global feature extraction is also limited. Therefore, exploring how to extract richer information based on the same backbone is a crucial direction for enhancing FGFSIC performance. By analyzing the main backbones used in FGFSIC, it can be observed that the intermediate outputs of the backbone contain valuable local features that are beneficial for improving classification accuracy. Retaining these intermediate (multi-scale) features can enhance the discriminative power of the extracted representations. One possible approach to leveraging these multi-scale features is to compare and combine information from different scales to produce the final prediction.

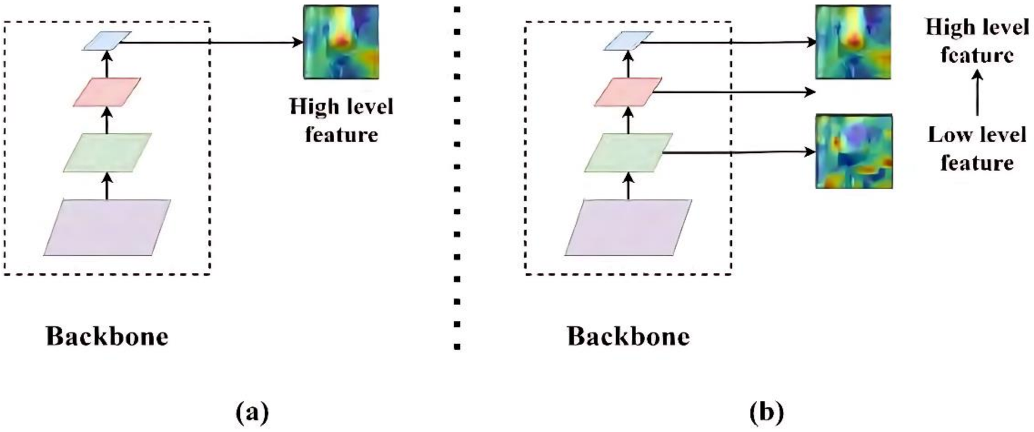

Figure 3. The difference between the original backbone (a) and the multi-scale feature extracted backbone (b).

As illustrated in Figure 3, the multi-scale feature extraction backbone generates additional outputs by retaining the intermediate results of the original backbone. As a result, the final production of the backbone consists of multiple feature maps at different scales, denoted as $\{F_1, F_2 \ldots F_n\}$. In this paper, Conv64 is used as the backbone, which produces four intermediate feature maps. Among these, three are selected for use, referred to as $\{F_1, F_2, F_3\}$, where F_1 (Blue) represents the minor scale and F_3 (Green) represents the most enormous scale.

3.3. Spatial Multi-Scale Feature Transformer

The extracted features from different scales share a common problem: high intra-class variance. Different poses and viewpoints, among other factors, primarily cause this issue. If we can learn a network to automatically transform the features' spatial positions and reduce the effects of different poses, this can effectively improve the performance of FGFSIC.

Our activation is based on the Spatial Transformer Network [25], which enables a network to learn how to manipulate the spatial positioning of features within itself. The core idea is to introduce a learnable module that predicts an affine transformation matrix, which is then used to establish a spatial mapping between the input feature map and an adjusted, transformed feature map. An affine transformation is defined by six parameters that control scaling, translation, and rotation, enabling spatial adjustments that conventional CNNs typically find difficult to manage effectively [25].

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} \quad (1)$$

As is illustrated in function 1, (x, y) represents the original coordinates of a pixel, while (xㅇㅅ, y′) denotes the new coordinates after an affine transformation, which is parameterized by six values: a to f. The entire process is shown in Figure 4. The input feature map is first used to predict the affine transformation parameters, and these parameters, along with the input feature map, are then used to generate an output feature map with spatial positions adjusted accordingly.

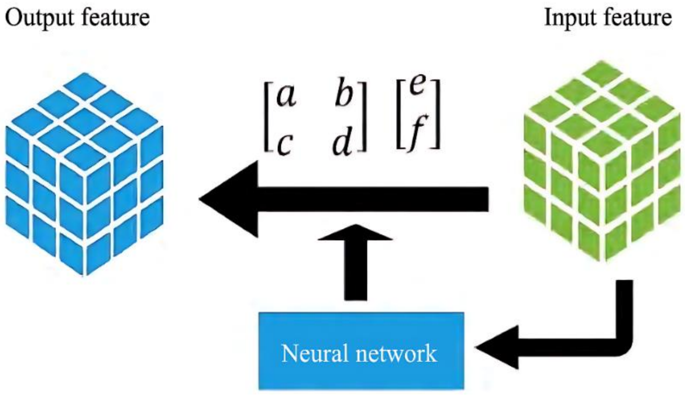
Output feature
Input feature
$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e \\ f \end{bmatrix}$
Neural network

Figure 4. Feature affine transformation process.

Since there are multiple scales of features, three different affine transformations would traditionally be required to transform these features. However, to avoid conflicts among different transformations and further improve accuracy, this paper proposes a Spatial Multi-Scale Feature Transformer to learn a unified transformation scheme for multi-scale features. In other words, for the three scales of features, only a single affine transformation is computed. As shown in Figure 5, the affine transformation matrix is predicted based on the middle-scale feature (Indicated by the red one). Subsequently, the three feature maps are resized to a consistent spatial size using two convolutional layers. Finally, the shared affine transformation is applied, aligning the original multi-scale features into a new, spatially consistent set of multi-scale features.

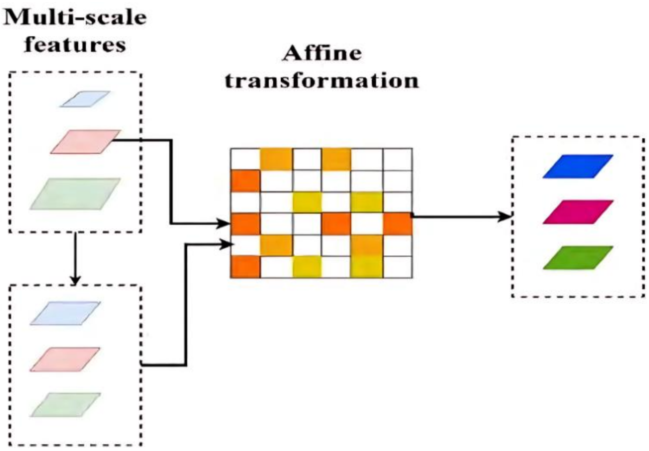

Figure 5. Spatial multi-scale feature transformer.

The spatial multi-scale feature transformer reduces intra-class variance by addressing spatial differences and prevents conflicts from multiple independent transformations that could negatively impact accuracy. To further improve FGFSIC performance, channel and spatial attention mechanisms are introduced to emphasize important feature regions and alleviate the problem of high intra-class variance.

3.4. Loss Function

This paper adopts the cross-entropy loss function. The loss function for each layer in the N-way K-shot setting is formulated as follows:

$$\mathcal{L}_a = -\sum_{i=1}^{M}\sum_{j=1}^{N} y_{ic} log\,(h_\theta(x_i)_c) \quad (2)$$

Where a = 1,2, 3 means the number of different feature scales, N is the number of categories, and M is the number of samples. y_{ic} is the one-hot encoded target value for a sample. And $h\theta(xi)_C$ is the predicted probability that the observed sample x_i belongs to category c. The predicted probability of this paper can be calculated by the similarity score between query features and support prototypes, measured by Euclidean distance.

The total loss function can be formulated as:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + (1 - \alpha - \beta)\mathcal{L}_3 \quad (3)$$

Where α, β are parameters within a range of [0,1] and $\alpha + \beta \leq 1$.

4. EXPERIMENT

4.1. Dataset

Three popular datasets CUB-200-2011, Stanford Cars, and Stanford Dogs are used in the experiment to validate the effectiveness of the proposed method.

The CUB-200-2011 dataset comprises 11,788 images from 200 bird species, featuring detailed annotations that include species labels, part locations, and bounding boxes. It is commonly used in both few-shot learning and fine-grained image classification tasks. Figure 6 presents example images from the CUB-200-2011 dataset.

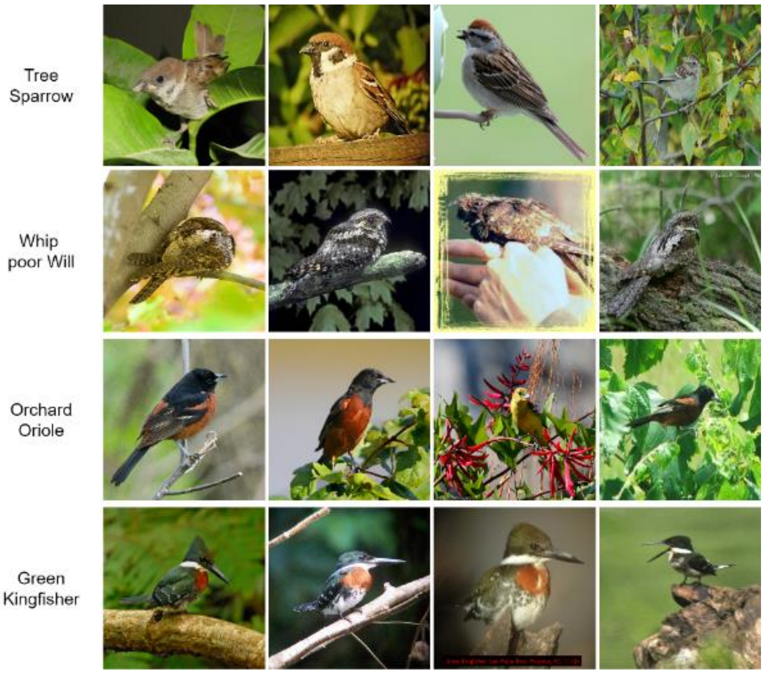

Figure 6. Example images from CUB-200-2011.

The Stanford Cars dataset comprises 16,185 images categorized into 196 distinct car makes and models. This dataset presents a significant challenge due to the high visual similarity between different car models, making it a suitable benchmark for evaluating fine-grained classification methods. Figure 7 shows several examples from the Stanford Cars dataset.

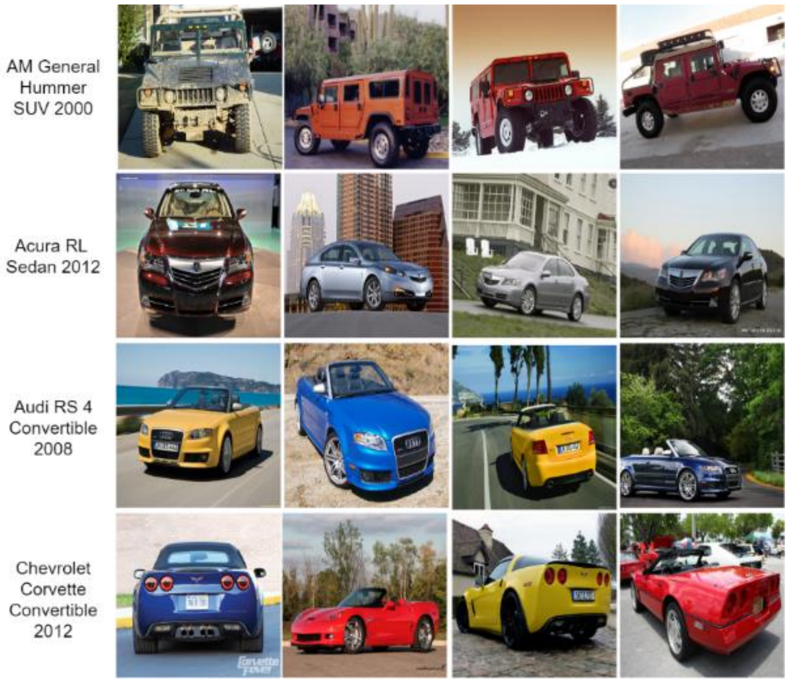
AM General Hummer SUV 2000
Acura RL Sedan 2012
Audi RS 4 Convertible 2008
Chevrolet Corvette Convertible 2012

Figure 7. Example images from Stanford-Cars.

The Stanford Dogs dataset contains 20,580 images spanning 120 dog breeds, with significant variations in pose, background, and lighting conditions. This makes it a challenging benchmark for fine-grained image classification algorithms. Example images from the dataset are shown in Figure 8.

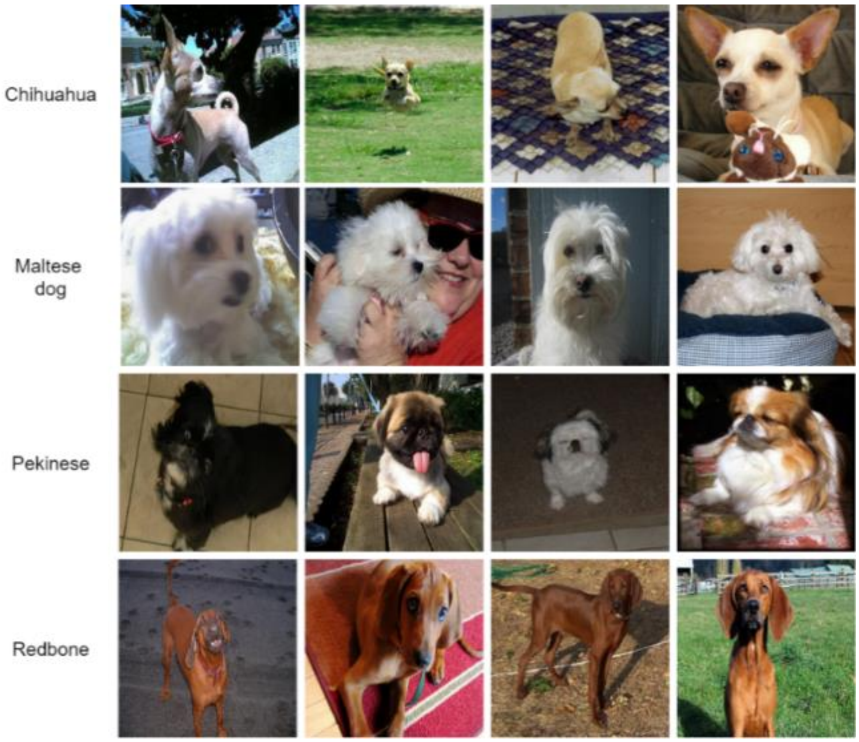

Figure 8. Example images from Stanford-Dogs.

4.2. Experiment Setting

In our experiments, we use Conv64 as the backbone network, which is a commonly used feature extractor in few-shot learning benchmarks [5, 7]. All input images are resized to 84 × 84 pixels. We apply the same standard data augmentation techniques as used in prior works, including random cropping, horizontal flipping, and normalization.

During training, we follow an episodic training strategy, constructing 60,000 episodes. In each episode, $N \times K$ images are randomly selected as the support set, where N represents the number of categories (or ways), and K represents the number of examples per category (or shots).

In addition, each query set contains 16 × N images. The experiment uses Adam as the optimizer with an initial learning rate of 0.001. To ensure the statistical reliability of the results, the number of test episodes is set to 2,000, and the final result is the average classification accuracy from these 2,000 episodes, with 95% confidence intervals.

4.3. Results and Discussion

The experiment results are illustrated as follows:

Table 1. Average classification accuracy (%) compared to both traditional few-shot learning methods and FGFSIC methods across three different datasets.

Methods	Backbone	CUB-200-2011		Stanford-Cars		Stanford-Dogs	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchNet [7]	Conv64	60.52±0.88	75.29±0.75	34.80±0.98	44.70±1.03	35.80±0.99	47.50±1.03
ProtoNet [5]	Conv64	50.46±0.88	76.39±0.64	40.90±1.01	52.93±1.03	37.59±1.00	48.19±1.03
RelationNet [6]	Conv64	62.34±0.94	77.84±0.68	47.79±0.49	60.60±0.41	43.29±0.46	55.15±0.39
MAML [8]	Conv64	54.73±0.97	75.75±0.76	47.25±0.30	61.11±0.29	44.84±0.31	58.61±0.30
PCM [21]	Conv64	42.10±1.96	62.48±1.21	29.63±2.38	52.28±1.46	28.78±2.33	46.92±2.00
LRPABN [3]	Conv64	63.63±0.77	76.06±0.58	60.28±0.76	73.29±0.58	45.72±0.75	60.94±0.66
MattML [24]	Conv64	66.29±0.56	80.34±0.30	66.11±0.54	82.80±0.28	54.84±0.53	71.34±0.38
Ours	Conv64	65.90±0.52	85.59±0.29	68.72±0.49	84.76±0.30	53.29±0.52	75.98±0.36

Note: Results are reported under two settings: 5-way 1-shot and 5-way 5-shot, using Conv64 as the backbone network. The best performance in each setting is highlighted in bold.

Table 1 presents the comparative performance of several few-shot learning and FGFSIC methods on three fine-grained image classification datasets: CUB-200-2011, Stanford Cars, and Stanford Dogs, under both 1-shot and 5-shot settings, using Conv64 as the backbone.

Overall, our proposed method consistently outperforms other approaches across most settings and datasets. On the CUB-200-2011 dataset, our method achieves 65.90% ± 0.52% in the 1-shot setting and 85.59% ± 0.29% in the 5-shot setting, which is competitive with the best-performing methods. Notably, it surpasses MattML, the previous top method, in the 5-shot scenario by a considerable margin (85.59% vs. 80.34%).

On the Stanford-Cars dataset, which is particularly challenging due to high intra-class similarity, our method achieves the highest performance among all methods, with 68.72% ± 0.49% accuracy in the 1-shot setting and 84.76% ± 0.30% accuracy in the 5-shot setting. This represents a significant improvement over MattML (66.11% and 82.80%) and other baselines such as RelationNet.

For the Stanford-Dogs dataset, known for its significant variations in pose, background, and illumination, our method also demonstrates strong performance, achieving 53.29% ± 0.52% in the 1-shot setting and 75.98% ± 0.36% in the 5-shot setting. While MattML slightly outperforms our model in the 1-shot setting (54.84% vs. 53.29%), our method delivers a notable improvement in the 5-shot setting (75.98% vs. 71.34%).

These results indicate that our model's combination of multi-scale feature extraction and spatial feature transformer is effective in addressing the challenges of fine-grained few-shot image classification. Notably, the consistent improvement in 5-shot scenarios suggests that the proposed method benefits from increased sample availability, effectively leveraging multiple support images to enhance classification accuracy.

5. CONCLUSION

In this paper, we propose a Spatial Multi-Scale Feature Transformer Network to address the challenges of high intra-class variance and low inter-class variance. First, to enhance the discriminative power of the extracted features, we utilize intermediate outputs from the backbone and compare them with more discriminative aspects to improve classification accuracy. Then, a Spatial Feature Transformer Network is applied to adjust the spatial positions of features, effectively reducing high intra-class variance. To avoid conflicts between transformations at different feature scales, we adopt a unified transformation approach for multi-scale feature processing. By simultaneously reducing intra-class variance and increasing inter-class variance, our method improves the performance of FGFSIC. Experimental results validate the effectiveness of our hypotheses and proposed methods. However, the current Spatial Multi-Scale Feature Transformer cannot handle variations caused by different viewpoints and backgrounds within the same class, which also contribute to high intra-class variance. Addressing this limitation may be a valuable direction for future research.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] T. Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449-1457.

[2] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 317-326.

[3] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Transactions on Multimedia*, vol. 23, pp. 1666-1680, 2021. https://doi.org/10.1109/TMM.2020.3001510

[4] H. Huang, J. Zhang, L. Yu, J. Zhang, Q. Wu, and C. Xu, "TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 853-866, 2022. https://doi.org/10.1109/TCSVT.2021.3065693

[5] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems," vol. 30. Red Hook, NY, USA: Curran Associates, Inc, 2017, pp. 4077–4087.

[6] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199-1208.

[7] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), Advances in Neural Information Processing Systems," vol. 29. Red Hook, NY, USA: Curran Associates, Inc, 2016, pp. 3630–3638.

[8] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," presented at the International Conference on Machine Learning. PMLR, 2017.

[9] A. A. Rusu *et al.*, "Meta-learning with latent embedding optimization," *arXiv preprint arXiv:1807.05960*, 2018. https://doi.org/10.48550/arXiv.1807.05960

[10] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, "MetaGAN: An adversarial approach to few-shot learning. In Advances in Neural Information Processing Systems," vol. 31. Red Hook, NY: Curran Associates, Inc, 2018, pp. 2365–2374.

[11] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proceedings of the IEEE international Conference on Computer Vision*, 2017, pp. 3018-3027.

[12] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan,, "Low-shot learning from imaginary data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City, UT, USA: IEEE.* https://doi.org/10.1109/CVPR.2018.00760, 2018, pp. 7278–7286.

[13] K. Li, Y. Zhang, K. Li, and Y. Fu, "Adversarial feature hallucination networks for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13470-13479.

[14] M. Lam, B. Mahasseni, and S. Todorovic, "Fine-grained recognition as hsnet search for informative image parts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2520-2529.

[15] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer International Publishing*, 2014, pp. 834-849.

[16] H. Zheng, J. Fu, Z. J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5012-5021.

[17] Y. Hu, Y. Yang, J. Zhang, X. Cao, and X. Zhen, "Attentional kernel encoding networks for fine-grained visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 301-314, 2020. https://doi.org/10.1109/TCSVT.2020.2978115

[18] X. S. Wei, C. W. Xie, and J. Wu, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained image recognition," *arXiv preprint arXiv:1605.06878*, 2016. https://doi.org/10.48550/arXiv.1605.06878

[19] X. He, Y. Peng, and J. Zhao, "Fast fine-grained image classification via weakly supervised discriminative localization " *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1394-1407, 2019. https://doi.org/10.1109/TCSVT.2018.2834480

[20] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4438-4446.

[21] X. S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6116-6125, 2019. https://doi.org/10.1109/TIP.2019.2924811

[22] S. Tian, H. Tang, and L. Dai, "Coupled patch similarity network for one-shot fine-grained image recognition," presented at the 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021.

[23] Y. Wu *et al.*, "Object-aware long-short-range spatial alignment for few-shot fine-grained image classification," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 107-115.

[24] Y. Zhu, C. Liu, and S. Jiang, "Multi-attention meta learning for few-shot fine-grained image recognition," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20). Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization*, 2020, pp. 1052–1058.

[25] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS 2015)*, 2015, pp. 2017–2025.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.