ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Multi-granularity Self-attention Mechanisms for Few-shot Learning

Wang Jing[1] and Chia S. Lim[2]

ABSTRACT

Few-shot learning aims to classify novel data categories with limited labeled samples. Although metric-based meta-learning has shown better generalization ability as a few-shot classification method, it still faces challenges in handling data noise and maintaining inter-sample distance stability. To address these issues, our study proposes an innovative few-shot learning approach to enhance image features' global and local semantic representation. Initially, our method employs a multiscale residual module to facilitate extracting multi-granularity features within images. Subsequently, it optimizes the fusion of local and global features using the self-attention mechanism inherent in the Transformer module. Additionally, a weighted metric module is integrated to improve the model's resilience against noise interference. Empirical evaluations on CIFAR-FS and Mini-ImageNet few-shot datasets using 5-way 1-shot and 5-way 5-shot scenarios demonstrate the effectiveness of our approach in capturing multi-level and multi-granularity image representations. Compared to other methods, our method improves accuracy by 2.63% and 1.27% for 5-shot scenes on these two datasets. The experimental results validate the efficacy of our model in significantly enhancing few-shot image classification performance.

Article information:

Keywords: Few-shot Learning, Multi-granularity Self-attention, Visual Transformer, Features Fusion, Metric-Learning

Article history:
Received: April 19, 2024
Revised: July 2, 2024
Accepted: September 19, 2024
Published: October 5, 2024 (Online)

DOI: 10.37936/ecti-cit.2024184.256469

1. INTRODUCTION

Few-shot learning poses a significant challenge pervasive in image recognition [1], target detection [2], and image segmentation [3]. Despite the impressive performance of deep learning models when trained on ample labeled data, they often grapple with overfitting issues when confronted with a scarcity of samples [4].

In contrast, humans exhibit rapid learning with minimal data. For instance, after grasping the "horse" concept, humans can often assimilate the idea of "zebra" with merely one or a few images with "***Fig. 1:*** ". Informed by this phenomenon, few-shot image classification seeks to train a limited number of labeled samples for each category, thereby facilitating the recognition of new images of these categories [5].

Methods employed in few-shot learning encompass model fine-tuning [6], data augmentation [7], and transfer learning [8]. Model fine-tuning involves pre-training the model with large-scale datasets and refining it for a specific task with limited training samples.

Fig.1: *Examples of few-shot image concept.*

Nonetheless, such models may still succumb to overfitting when confronted with inadequate instances [9]. Data augmentation approaches can mitigate the issue of overfitting by enlarging the training set. However, it is essential to note that these strategies may add noisy data, negatively influencing the model results [10]. Transfer learning for small samples incorporates metric learning [11] and meta-learning [12]. Metric learning determines the category of an unknown sample based on labeled samples by calculating the distance between samples [13]. Optimization-based approaches allow models to adapt quickly to new tasks but still have difficulty handling domain transitions

[1,2]The authors are with the Graduate School of Technology, Asia Pacific University of Technology and Innovation, 57000, Kuala Lumpur, Malaysia, E-mail: lim.chiasien@apu.edu.my and wjing985@163.com
[1]Corresponding author: wjing985@163.com

between base and new classes [14]. In contrast, meta-learning methods are closer to human learning patterns but tend to neglect exploring relationships between samples [15].

However, with a small sample size, improving model effectiveness encounters bottlenecks. Few-shot learning methods are susceptible to overfitting with very few or imbalanced pieces. Some approaches focus on single-scale feature extraction, limiting their applicability in the domain of few-shot images.

Images encapsulate semantic information with varying granularity features, encompassing low-level features with high resolution and robust local information yet limited global semantic information. On the other hand, high-level features are abundant in semantic information but possess low resolution and weak perception of local details [16]. Thus, features with diverse scales of multi-granularity can more comprehensively extract sample features with "***Fig. 2:***".

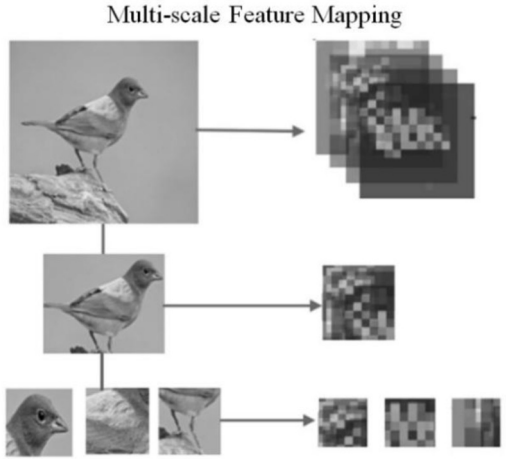

Fig.2: *Extraction of multi-grained features.*

A feature extractor with good generalizability is essential for few-shot learning. Feature enhancement can increase the diversity of samples and achieve more accurate classification [17]. Improved residual network Res2net [18] obtains multi-granularity features by performing multiscale convolution inside each residual block to form different sensory fields (as shown in "***Fig. 3:***"). Building upon this premise, a multi-granularity residual attention network, as delineated in [19], exhibits heightened efficacy in capturing features across varying levels of granularity within images. This augmentation leads to a discernible enhancement in the classification prowess of deep networks.

For small samples, Dong *et al.* (2021) introduced the multiscale feature network (MSFN) for feature enhancement, followed by the computation of distances between improved prototypes using enhanced labeled features for classification [20]. In a similar vein, Chen *et al.* (2022) presented the multi-scale adaptive task attention network (MTAN) [21].

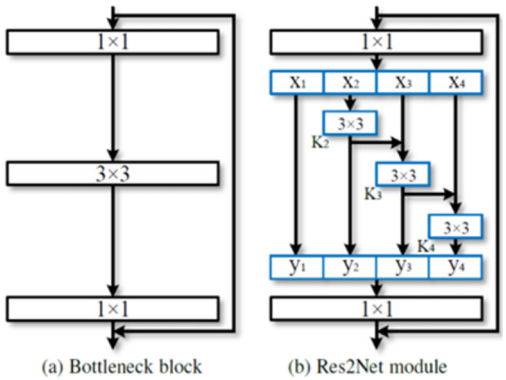

Fig.3: *Multi-granularity residual network architecture(Res2net).*

MTAN can dynamically allocate attention across multiple feature scales within input contexts, particularly in Few-shot learning tasks. This adaptive attention allocation allows for improved capture of feature information across different scales.

Self-attentive mechanisms can better capture global and local features in various computer vision tasks, and it has emerged as a compelling alternative to convolutional neural networks (CNNs) [22]. However, the transformer exhibits sub-optimal efficiency in learning from sparsely annotated tasks [23]. In this study, we aim to explore the efficacy of multi-granularity visual transformers to improve performance under conditions of limited data availability.

We aim to acquire and fuse information from small sample data at multiple granularities to overcome the limitation of extracting feature information from a single granularity. This paper uses multi-granularity feature enhancement and Transformer self-attention aggregation to improve few-shot image classification performance. The main contributions are as follows:

(1) A novel multi-granularity self-attention mechanism-based few-shot learning technique is presented that effectively extracts and improves multi-granularity feature representations from small-sample pictures.

(2) A multi-branch self-attention fusion module has been devised to mitigate the impact of partially irrelevant information within image samples on the classification task. Operating through the Transformer mechanism, this module enhances multi-granularity local features containing global information to yield a more precise representation of sample features.

(3) A weighted distance metric module is proposed to address potential issues stemming from sample diversity within the same class. The similarity between the query set and each sample in the class is calculated by this module using the weighted average distance method. Addressing sample diversity seeks to mitigate potential class centroid bias.

The paper is structured as follows: Section 1 pro-

vides an overview of the background of few-shot learning; Section 2 elaborates on the proposed method; Section 3 presents the experimental results and analysis; and Section 4 concludes the paper and outlines future directions. By delving into this research, we anticipate offering novel insights and solutions to advance the field of few-shot image classification.

2. MATERIALS AND METHODS

2.1 Few-shot learning

Few-shot learning trains the model to adapt to the new classes, allowing for effective classification performance with small samples [24]. It is assumed that the support set S consists of N data categories, each comprising K-labeled samples, in the context of a few-shot target classification issue. The query set Q comprises N classes corresponding to the support set S and the same images as q unlabeled samples. Every class is included in query set Q. Thus; the N-way K-way few-shot classification task is indicated.

On the other hand, it becomes difficult to train a model directly to categorize the unlabeled samples in set Q when set S contains few labeled examples per class. Because of this, few-shot learning usually uses a meta-learning paradigm, gaining transferable information on an auxiliary set A to enhance classification on set Q. Set A is not connected to set S, even if it contains several classes and labeled samples. As a general term, "base class" refers to Set A. Name set S as the new class, or target class, and set A as the base class, or source class, according to convention. Few-shot learning uses a set of base courses with labeled data to build a classifier and then adapts it to new types with fewer labeled samples.

2.2 Multi-Granularity Feature Generation

The multi-granularity feature module, depicted in "***Fig. 4:***" employs depth-separable convolutional modules [25] as a feature extractor. This module comprises convolutional modules across three branches, each operating at different scales.

Branch 1 features undergo a 1×1 convolution. Branch 2 processes the feature extractor through a 3×3 convolution followed by a 1×1 convolution, resulting in its output. Branch 3 utilizes two 3×3 convolutions and a 1×1 convolution. Using two 3×3 convolutions achieves the same receptive field as a single 5×5 kernel:

Following the multi-granularity generation module, each sample can acquire three feature vectors at distinct scales. The number of branches (representing granularity), de-noted as r, can be adjusted based on the image's scale. In this context, we set r=3.

2.3 Multi-granularity self-attention module

The self-attention mechanism inherent in the transformer architecture constitutes a feature representation technique that captures global dependencies within data representations [26]. Efforts to capture long-xplore the integration of self-attention into convolutional neural network (CNN) architectures, thereby facilitating the fusion of self-attention and convolutional operations[27-28]. The Swin Transformer[29] has introduced a hierarchical feature representation scheme to address the need to amalgamate multi-granularity features within image data. Additionally, the Pyramid Vision Transformer (PVT2)[30] integrates the Vision Trans-former (ViT) paradigm into a CNN-inspired pyramid structure, thereby enabling PVT2 to excel particularly in dense prediction tasks.

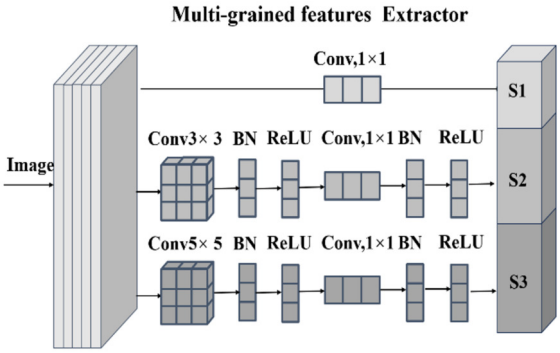

Fig.4: *Structure of multi-grained feature generation module.*

CrossViT[31] introduces the Transformer attention mechanism across layers to capture multi-granularity feature information at different levels. While Transformers have achieved remarkable results on datasets with limited samples, it is worth noting that their training process is relatively intricate, demanding substantial computational resources and time.

We develop a multi-granularity self-attention method to learn and capture global information between input sequences. With the help of our suggested self-attention aggregation module, the local features in various scales are weighted and fused into the global information, improving the image's local feature representation and the connection between sample details, which makes it possible to acquire contextual information and larger sensory fields.

According to "***Fig. 5:***", the image x goes through the multi-granularity feature module to generate n-scale local features $f(x_i \in)R^{1\times d}, i \in [1, r]$, then enters the self-attention module for feature fusion.

The query set and support set are represented by the set $T = \{(x_i, y_i, n), n = 1, \ldots, r, i = 1, \ldots, l_k + l_q\}$, where r stands for the number of pieces in the feature map at different scales and l_q: indicates the number of samples in the support set. d is the feature dimension. Create the feature tensor $F \in R^{(l_k+l_q)\times r\times d}$. The input feature triad (F, F, F) is first linearly transformed by the Transformer module into Q_i, K_i, and V_i, where Q_i, K_i, and V_i Stand for the query matrix,

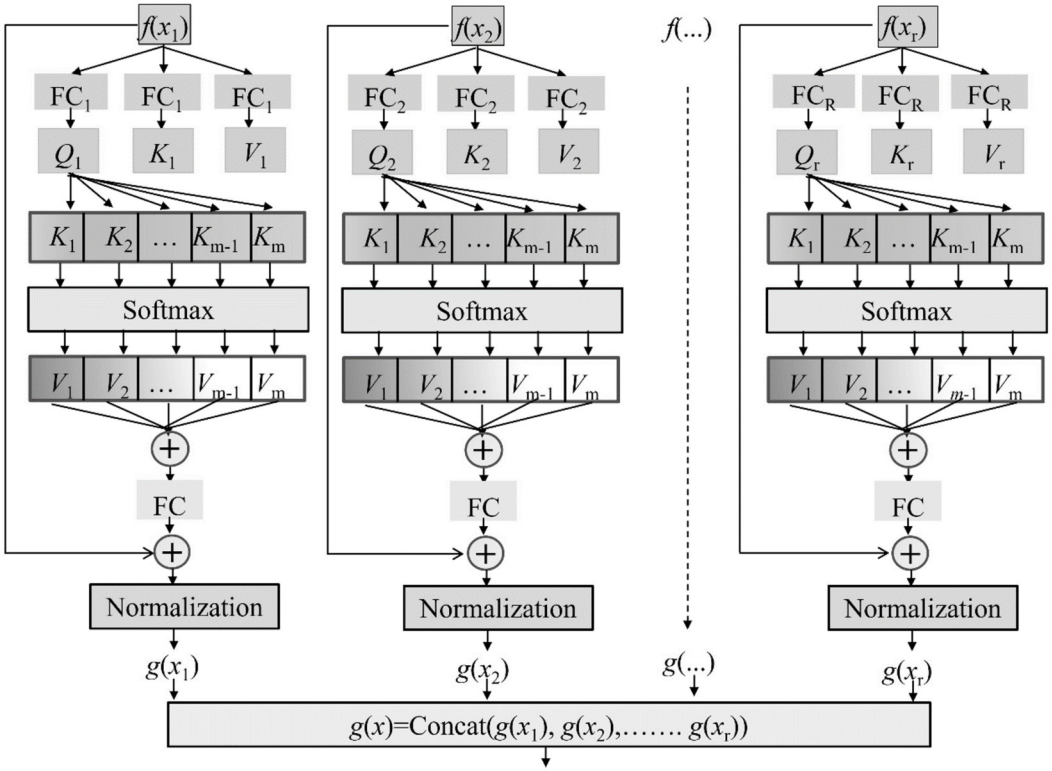

Fig.5: *Self-Attention Fusion Module for Multi-Granularity Features.*

key-value matrix, and value matrix, respectively, and $i \in [1, r]$. The attention module is defined as follows based on the matrix operation of self-attention:

$$\left(f_Q^{(i)}, f_K^{(i)}, f_V^{(i)}\right) = \left(f^{(i)}W_Q, f^{(i)}W_k, f^{(i)}W_V\right) \quad (1)$$

$$g_{Att}^{(i)} = Norm\left(f^{(i)} + Soft\left(\frac{f_Q^{(i)}(f_K^{(i)})^T}{\sqrt{d_K}}\right) f_V^{(i)}\right) \quad (2)$$

Where $d_K = d.W_Q, W_K$, and W_V Represent the parameters of the whole connection layer. The Transformer module fuses the multi-granularity feature maps and uses this as the output feature information for the samples:

$$g(x) = \sum_{i=1}^{r} g_{Att}^{(i)} \in R^{(l_k+l_q)\times r\times d} \quad (3)$$

2.4 Weighted metrics module

The metric module calculates a relationship score by measuring the interclass similarity as a function in the multi-granularity feature space created from the sample set categories C. The implementation process of this module includes feature extraction, feature fusion, calculation of relationship scores, and output of category probabilities.

The set C is divided into a support set. x_i and a query set x_j, which together define an Episode. In each episode, the model learns the features of a small number of labeled samples in the support set to infer the class of samples in the query set.

To calculate the scores of the classes in this research, we utilize metric learning in the relational network [32] module.

$$\boldsymbol{S_{i,j}} = R(\varphi(f_\tau(x_i), f_\tau(x_j))) \quad (4)$$

Where f_τ is the embedding module (x_i), the support set labeled sample feature vector, and (x_j) is the query set sample feature vector. $\varphi(f_\tau(x_i), f_\tau(x_j)$ is the descriptor connection of intensely localized multi-granularity features, R is the metric learning network. The weights of each module are derived through cross-validation and weighted fusion of image affiliation probabilities and relationship scores.

We thus derive the final prediction of the model. Suppose that an image-to-image metric module has a prediction for the category k, which has a categorical probability of subordination of $p_k(a)$, and two image-to-class metric modules for category k. The relationship score is $p_k(b)$,$p_k(c)$. It will be the case that $p_k(a)$, $p_k(b)$,and $p_k(c)$ are viewed as three probability vectors of length when considering the number of categories. The weighted sum of the three vectors is calculated, and the cumulative maximum is taken as the final prediction as in Equation (5):

$$p_k = \text{argmax}(\alpha p_k(a) + \beta p_k(b) + \gamma p_k(c) \quad (5)$$

Where α, β, γ is the weight of each module, $\alpha+\beta+\gamma = 1$.

2.5 Multi-Granularity Feature Fusion Architecture

The Few-shot Classification Model for Multi-Granularity Feature Enhancement and Fusion (MGFEA) comprises three modules with "***Fig. 6:***". It includes a multi-granularity feature extraction module, a Transformer self-attentive feature fusion module, and a metric module. The following describes the implementation methods and steps of the multi-granularity feature fusion model MGFEA.

1. Basic feature extraction. First, we extract the initial feature representation of all input samples S.

2. Multi-granularity feature generation module. Provide the base feature S to the multi-granularity creation tool as input to create the multi-granularity feature F for the given samples.

3. Self-attentive feature aggregation module. Fuse the multiscale features in F with the weighted global features extracted by the transformer as the enhanced feature expression E of the input sample.

4. Metric Module for Weighted Distance. Calculate the weighted difference in categories between the samples in the query set and the support set.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Experimental Dataset

We used two datasets often used in few-shot learning: CIFAR-FS [33] and miniImageNet[34]. The CIFAR-FS dataset comes from the CIFAR 100 dataset, where the 100 categories are divided into three groups: a test set (20 categories), a validation set (16 categories), and a training set (64 categories). A subset of ImageNet called MiniImageNet has 100 categories total, 64 of which are utilized for training, 16 for validation, and 20 for testing.

3.2 Training Strategy for Few-shot Learning

Our training technique is episode-style[35], which we use to learn migration knowledge. Coordinating the training process with the testing settings improves the model's generalization ability. To help with the classification model's training, we model a few-shot target classification issue and produce an episode for each training cycle. An auxiliary set A contains an arbitrary selection of query sets Q and support set S for each episode. K-labeled samples are included in each N class, making up the support set S.

Tens of thousands of episodes drawn from the auxiliary set A are created throughout the training phase to fine-tune the classification model. Each episode may be thought of as an independent work. When the created query set's correct labels are applied, these episodes are linked to supervised learning. The model trained on the support set S may be utilized directly to sort every unlabeled sample in the question set Q during testing. The 5-way 2-shot scenario illustrated in "***Fig. 7:***" is worth considering.

3.3 Experimental Methods and Results

For the trials in our few-shot technique, we employ a 5-way, 1-shot strategy (i.e., five categories, one sample per category) and a 5-way, 5-shot strategy (i.e., five categories, five pieces each category). This section compares the image classification job experimentally on CIFAR-FS and Minimagenet datasets. *Table 1* displays the experimental findings.

Table 1 demonstrates that on the CIFAR-FS and Mini-imagine datasets, 5-shot Accuracy is superior to 1-shot Accuracy. It indicates that more features are discovered when training data increases and the classification impact improves. Our model performs better than other models in the 1-shot and 5-shot work scenarios on the CIFAR-FS dataset by over 1.75% and 2.63%, respectively; our classification performance is 1.62% and 1.27%, respectively, on the Mini-ImageNet dataset.

Table 1: *Comparison of different few-shot learning models.*

Models Backbones	1-shot	5-shot
Resnet12	55.43	68.16
Resnet18	59.65	69.65
Resnet34	61.15	71.43
Resnet50	61.87	73.83
MGFEA - Resnet12	65.45	86.72
MGFEA - Resnet18	69.88	86.95
MGFEA - Resnet34	70.05	87.12
MGFEA - Resnet50	71.92	87.86

"***Fig. 8:***" illustrates model training on the CIFAR-FS dataset, where classification accuracy quickly exceeds 80% at epoch=20, and the loss function quickly converges at epoch=40 with "***Fig. 9:***". The studies demonstrate the efficiency and speed with which the multi-granularity feature augmentation model can extract multiscale features from tiny sample datasets.

3.4 Ablation Experiments

In this section, we conduct ablation comparison tests on the mini-magnet few-shot dataset. Initially, we select the residual network with various parameters as the feature extraction module. When the multiscale branching $r = 3$, we perform a comparison experiment using the original network (with the feature enhancement module removed) and the proposed new network.

The results are displayed in Table 2. The classification performance of few-shot images improves to

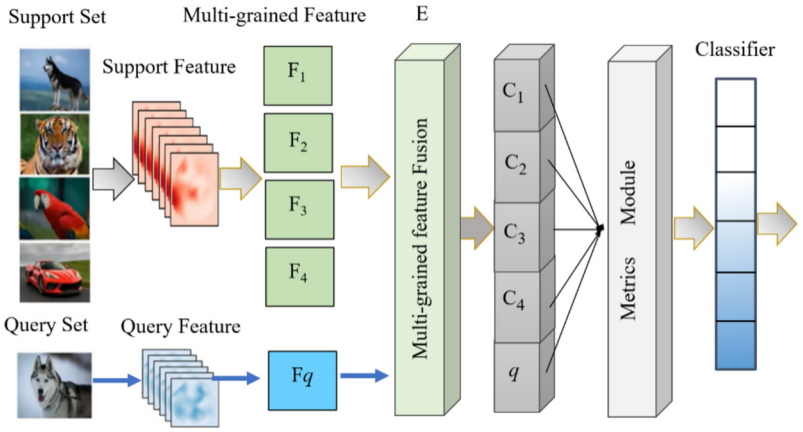

Fig.6: *Structure of Multi-Granularity Enhanced Network.*

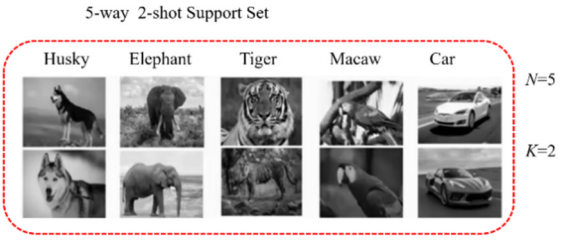

Fig.7: *Example of an N-way K-shot.*

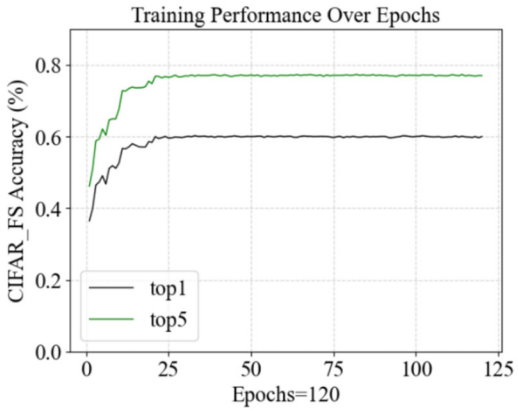

Fig.8: *Accuracy for 1-shot and 5-shot.*

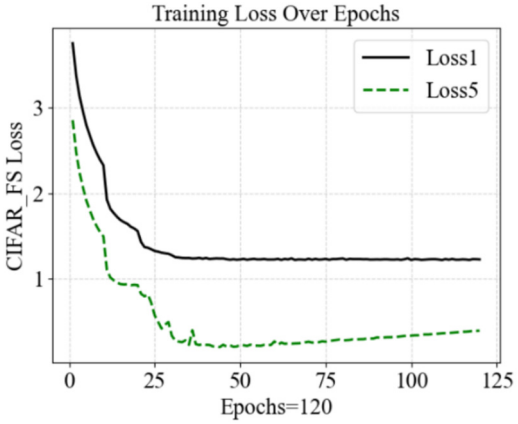

Fig.9: *Loss for 1-shot and 5-shot.*

Table 2: *Comparison of different backbones on dataset CIFAR-FS.*

Model	CIFAR-FS		Miniimagenet	
	1-shot	5-shot	1-shot	5-shot
ProTyNet (2017)[36]	55.20±0.70	83.50±0.50	49.42±0.78	68.20±0.57
TADAM (2019) [37]	70.30±0.40	84.10±0.10	58.50±0.30	76.70±0.30
MetaSVM (2019) [38]	72.80±0.70	85.00±0.50	62.64±0.61	78.63±0.46
Deepend (2020) [39]	75.65±0.83	86.79±0.50	65.91±0.82	82.41±0.56
Metabase (2021) [40]	74.28±0.50	85.90±0.50	63.17±0.23	79.26±0.17
QSFormer (2022) [41]	75.40±0.50	86.36±0.30	65.24±0.28	79.96±0.20
DeepBDC (2022) [42]	76.75±0.30	87.20±0.20	67.83±0.43	85.45±0.29
MGFEA (Ours)	78.50±0.64	89.83±0.25	69.45±0.20	86.72±0.12

some extent as the backbone network's layer count rises but gradually tends to a stable value. Experiments show that the features of the original network affect the model, while its performance is significantly improved after we add a multi-granularity feature fusion module with "***Fig. 10:***".

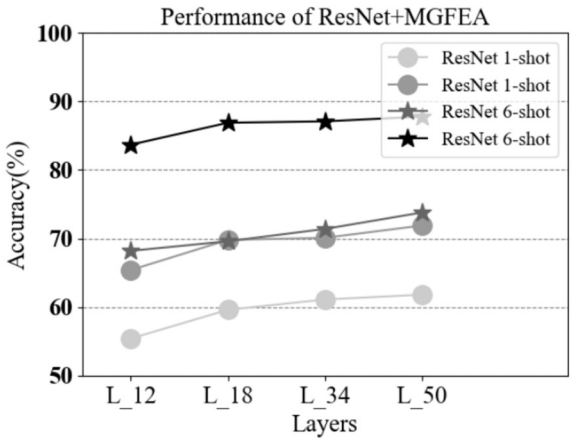

Fig.10: *Performance comparison of feature enhancement with different parameters.*

Table 3: *Comparison of different branching parameters r.*

Models Backbones	Para: r	1-shot	5-shot	s/episode
MGFEA – Resnet12	1	58.77	68.26	5.7
MGFEA – Resnet18	2	65.60	75.40	7.2
MGFEA – Resnet34	3	69.45	86.72	8.5
MGFEA – Resnet50	4	69.52	86.78	13.7

According to Table 3, by selecting various branching parameters ($r = 1, 2, 3, 4$), the studies assess the impact of various granularity features on the network performance. The time (s/episode) spent on a single model iteration is the basis for the temporal complexity analysis. Our research findings utilize Resnet12 as a model network.

Table 3 shows that the single-scale self-attention module performs least effectively with features extracted directly from the backbone at $r = 1$. As multiscale branching is increased, the performance of the fused multi-grained features improves to a stable value, but the computational time cost increases due to the increased parameterization

4. CONCLUSION

In this research, we provide a multi-granular feature improvement strategy for small sample picture classification. The approach utilizes the multiscale local features produced by the prototype network. It extracts the feature improvement information comprising global information through a fusing module based on Transformer architecture, which enhances the model's feature extraction and generalization capabilities. On the CIFARF10 and MiniImageNet few-shot datasets, the suggested technique performs better in classification than previous models, and the ablation experiments further support its efficacy.

The increased representation of visual characteristics in our suggested Transformer self-attentive feature fusion module can be included in various few-shot classification techniques. The computing cost of the transformer grows for multi-granular features of an image as the number of scale branches increases. Therefore, balancing classification effectiveness against computational complexity is critical when determining the ideal number of multi-branches for a given job. To further enhance the model's classification performance, future studies will investigate how to develop a more effective multi-granularity feature fusion module.

AUTHOR CONTRIBUTIONS

Conceptualization, Wang Jing; methodology, Wang Jing; software, Wang Jing; validation, Wang Jing and Chia S. Lim; formal analysis, Wang Jing; investigation, Wang Jing; data curation, Wang Jing; writing—original draft preparation, Wang Jing; writing—review and editing, Wang Jing and Chia S. Lim; visualization, Wang Jing; supervision, Chia S. Lim; funding acquisition, Wang Jing. All authors have read and agreed to the published version of the manuscript.

References

[1] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang and T. Tan, "Notice of Removal: Semantic Prompt for Few-Shot Image Recognition," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 23581-23591, 2023.

[2] S. Li, G. Yang, X. Liu, K. Huang and Y. Liu, "Few-shot object detection based on global context and implicit knowledge decoupled head," *IET Image Processing*, vol. 18, no. 6, pp. 1460-1474, 2024.

[3] Z. Cheng, S. Wang, T. Xin, T. Zhou, H. Zhang and L. Shao, "Few-Shot Medical Image Segmentation via Generating Multiple Representative Descriptors," in *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 2202-2214, June 2024.

[4] Y. L. Chang, T. H. Tan, W. H. Lee, L. Chang, Y. N. Chen, K. C. Fan, and M. Alkhaleefah, "Consolidated convolutional neural network for hyperspectral image classification," *Remote Sensing*, vol. 14, no. 7, pp. 1571, 2022.

[5] X. Zhang, D. Meng, H. Gouk, and T.M.Hospedales, "Shallow Bayesian meta-learning for real-world few-shot recognition," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 651-660, 2021.

[6] Y. Gu, X. Han, Z. Liu, and M. Huang, "Ppt: Pre-trained prompt tuning for few-shot learning," *arXiv preprint*, arXiv:2109.04332, 2022.

[7] X. Chao and L. Zhang, "Few-shot imbalanced classification based on data augmentation," *Multimedia Systems*, vol. 29, pp. 2843–2851, 2024.

[8] Q. Zhang, X. Yi, J. Guo, Y. Tang, T. Feng, and R. Liu, "A few-shot rare wildlife image classification method based on style migration data augmentation," *Ecological Informatics*, vol. 77, 2023.

[9] R. Das, Y.-X. Wang, and J. M. Moura, "On the importance of distractors for few-shot classification," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9030-9040, 2021.

[10] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91-99, 2022.

[11] X. Li, X. Yang, Z. Ma, and J. Xue, "Deep metric learning for few-shot image classification: a selective review," arXiv e-prints, arXiv-2105, 2021.

[12] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analy-*

sis and Machine Intelligence, vol. 44, no. 9, pp. 5149-5169, 2021.

[13] J. Zhao, Y. Yang, X. Lin, J. Yang, and L. He, "Looking wider for better adaptive representation in few-shot learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10981-10989, 2021.

[14] D. Wang, Y. Cheng, M. Yu, X. Guo, and T. Zhang, "A hybrid approach with optimization-based and metric-based meta-learner for few-shot learning," *Neurocomputing*, vol. 349, pp. 202-211, 2019.

[15] W. Zheng, B. Zhang, J. Lu, and J. Zhou, "Deep relational metric learning," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12065-12074, 2021.

[16] X. Wu, T. Thitipong, and J. Wang, "Image classification based on multi-granularity convolutional Neural network model," *the 19th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 1-4, 2022.

[17] Y. Wu, B. Wu, Y. Zhang, and S. Wan, "A novel method of data and feature enhancement for few-shot image classification," *Soft Computing*, vol. 27, no. 8, pp. 5109-5117, 2023.

[18] S. H. Sun, M. M. Cheng, K. Zhao, X.-Y. Zhang, M. Harandi, and P. H. S. Torr, "Res2Net: A New Multiscale Backbone Architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652-662, 2021.

[19] X. G. Wu and T. Tanprasert, "A Multi-Grained Attention Residual Network for Image Classification," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 17, no. 2, pp. 215-224, 2023.

[20] B. Dong, R. Wang, J. Wang, and L. Xue, "Multi-scale features self-enhancement network for few-shot learning," *Multimedia Tools and Applications*, vol. 80, no. 25, pp. 33865-33883, 2021.

[21] H. Chen, H. Li, Y. Li, and C. Chen, "Multi-scale adaptive task attention network for few-shot learning," *26th International Conference on Pattern Recognition (ICPR)*, pp. 4765-4771, 2022.

[22] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *Computing Surveys*, vol. 55, no. 13s, pp. 1-40, 2023.

[23] W. Wang, J. Zhang, Y. Cao, Y. Shen, and D. Tao, "Towards Data-Efficient Detection Transformers," *arXiv preprint*, arXiv:2203.09507, 2022.

[24] Y. Tian, X. Zhao, and W. Huang, "Meta-learning approaches for learning-to-learn in deep learning: a survey," *Neurocomputing*, vol. 494, pp. 203-223, 2022.

[25] Z. Y. Khan and Z. Niu, "CNN with depthwise separable convolutions and combined kernels for rating prediction," *Expert Systems with Applications*, vol. 170, p. 114528, 2021.

[26] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. ul Khan, and M. Shah, "Transformers in vision: a survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1-41, 2022.

[27] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815-825, 2022.

[28] J. Fang, H. Lin, X. Chen, and K. Zeng, "A hybrid network of CNN and transformer for lightweight image super-resolution," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1103-1112, 2022.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012-10022, 2021.

[30] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415-424, 2022.

[31] C.-F. R. Cheng-Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multiscale vision transformer for image classification," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 357-366, 2021.

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

[33] L. Bertinetto, J. F. Henriques, P. H. S. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," *arXiv preprint*, arXiv:1805.08136, 2018.

[34] D. Chen, Y. Wang, Y. Li, F. Mao, Y. He, and H. Xue, "Self-Supervised Learning for Few-Shot Image Classification," *ICASSP 2021, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, pp. 1745-1749, 2021.

[35] P. Zhu, Z. Zhu, Y. Wang, J. Zhang, and S. Zhao, "Multi-granularity episodic contrastive learning for few-shot learning," *Pattern Recognition*, vol. 131, p. 108820, 2022.

[36] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[37] S. W. Yoon, J. Y. Seo, and J. K. Moon, "Tapnet: Neural network augmented with task-adaptive projection for few-shot learning," *International Conference on Machine Learning*, pp. 7115-7123, May 2019.

[38] K. Sohn, H. Lee, and X. Yan, "Meta-learning with differentiable convex optimization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10657-10665, 2019.

[39] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable Earth mover's distance and structured classifiers," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12200-12210, 2020.

[40] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9042-9051, 2021.

[41] X. Wang, X. Wang, B. Jiang, and B. Luo, "Few-Shot Learning Meets Transformer: Unified Query-Support Transformers for Few-Shot Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7789-7802, 2023.

[42] J. Xie, F. Long, J. Lv, Q. Wang, and P. Li, "Joint distribution matters: Deep Brownian distance covariance for few-shot classification," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7962-7971, 2022.

Wang Jing received the M.S.degree in computer science from the Lincoln University College, Malaysia with a Master's degree in Computer Science in 2023. She is currently pursuing the PhD. Degree with the Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia. Her main research areas are artificial intelligence applications and computer vision.

Chia S. Lim received the B.S. degree (Hons.) in mathematics and the M.S. degree in mathematics from Oklahoma State University, Stillwater,OK, USA, in 1994 and 1997, respectively, and the Ph.D. degree in mathematics from Michigan State University, East Lansing, MI, USA, in 2002. He is currently heading the Graduate School of Technology, Asia Pacific University of Technology and Innovation. His research interests include linear algebra and its application to machine learning and data science.