Multi-level Feature Learning For Multimedia Pattern Recognition In English Education

Xinjie Zhu

School of Foreign Languages, Zhengzhou University of Science and Technology, Zhengzhou, 450064, China

Corresponding author. E-mail: xjiezhu2024@163.com

Received: Mar. 21, 2024; Accepted: Sep. 01, 2024

In multimedia pattern recognition, particularly in English education, the sharing of local features between classes and their varying classification reliability are often overlooked in existing methods, which diminishes feature discrimination and complicates the handling of small inter-class variations. In this paper, a multi-level feature learning method based on enhanced local descriptors is proposed for mining multimedia patterns in English educations (MFL). Specifically, multi-scale global information is extracted using the pyramid aggregation network and fused with local features to enhance inter-class uniqueness. During classification, local descriptors that better distinguish between classes are emphasized, resulting in improved inter-class discrimination. MFL achieves a significant accuracy improvement over baseline methods across three datasets, offering a new perspective for analyzing multimedia content in English education.

Keywords: English education; multimedia image recognition; multi-level feature learning

http://dx.doi.org/10.6180/jase.202507_28(7).0007

1. Introduction

Multimedia image recognition in English education is a vital technology for processing large datasets and extracting valuable insights [1, 2]. Traditional approaches to multimedia content analysis struggle with high dimensionality, complexity, and the lack of well-annotated samples, especially as data scales increase. Few-shot multimedia image recognition has emerged as a promising solution, offering strong generalization capabilities despite limited annotated data. This approach not only addresses the limitations of traditional methods but also enhances deep understanding and accurate classification in multimedia content analysis [3–5]. It presents significant opportunities and challenges in information processing, marking a promising direction for future research and development in multimedia analysis.

In few-shot multimedia image recognition, each class is represented by only a few annotated samples, which are not uniformly sampled from the class distribution. This leads to a significant deviation between the labeled sample distribution and the true underlying distribution, resulting in small intra-class differences due to uneven data acquisition. Moreover, in few-shot image classification, the classes to be distinguished are not just coarse-grained categories with significant semantic differences but also fine-grained categories with similar semantics, differing mainly in local features. This is common across various datasets. For example, in the StanfordDogs dataset [1], Japanese Spitz and Maltese dogs share traits like round faces, long ears, and small body sizes. Here, samples from different classes often have similar global features, making subtle differences in local features, such as color, crucial for differentiation. Relying solely on global features for classification in such scenarios can lead to inadequate attention to distinguishing details, thus limiting classification accuracy.

To improve local feature learning in few-shot multimedia image recognition, some researchers model the corre-

lation between support and query sample representations, weighting feature matrices based on these correlations to enhance global feature learning [6, 7]. Others focus on local descriptor classification within the sample-to-class metric framework [8, 9]. For instance, DeepEMD introduces the Wasserstein distance instead of the usual cosine or Euclidean distances to improve similarity measurement among sets of local descriptors [8], while DN4 uses the sum of cosine distances between sample-local and class-local descriptors to quantify similarity between samples and classes [9].

Although previous works achieve encouraging performance, they often overlook the phenomenon of inter-class sharing that exists within local features, as well as the fact that different local features hold varying degrees of importance for correctly classifying samples, which weakens the discrimination of the representations. Specifically, the facial local patterning information can distinguish the maltese dog from the japanese spitz and afghan hound, but it cannot serve as discriminative information between the japanese spitz and afghan hound due to their shared characteristics like local patterning and long fur. If the local descriptors only encapsulate local features, it becomes challenging to differentiate such categories that possess numerous shared local attributes.

To address these challenges, a novel multi-level feature learning method (MFL) is proposed for few-shot multimedia image recognition, consisting of four components: the feature extractor module, the global feature learning module, the local learning module, and the semantic enhancement module. Specifically, MFL models the similarity relationships between the local descriptors of samples and classes to assign each sample accordingly, via using the vector at each position of the local feature matrix as a local descriptor to represent corresponding samples and aggregating local descriptors from samples of the same class to describe the class. To enhance the discriminative power of local descriptors, MFL captures multi-scale global information from samples using the pyramid aggregation network with a residual structure, effectively mitigating the adverse effects of inter-class similarity. Meanwhile, MFL learns intra-class similarity and inter-class distinctiveness for each local descriptor in the support set, assigning greater weights to descriptors with high intra-class similarity and inter-class distinctiveness. MFL ensures that more discriminative local descriptors play a significant role in classification. Comprehensive experiments on three real-world multimedia image recognition datasets demonstrate that MFL achieves state-of-the-art performance in few-shot multimedia image recognition.

MFL makes several notable contributions, including:

- MFL proposes a novel local descriptor-based feature learning by fusing multi-scale global information through the pyramid aggregation network and local features, to improve inter-class distinctiveness for enhancing the discriminative power of local descriptors.
- MFL assigns varying weights to enhance the impact of more discriminative descriptors in classification by incorporating both intra-class similarity and inter-class distinctiveness of each local descriptor in the support set, which mitigates the influence of irrelevant information on classification.
- The experimental results on three real world datasets showcase that MFL sets a new benchmark in the field of few-shot multimedia image recognition, which underscores the efficacy of MFL in advancing the performance frontier on demanding few-shot multimedia image recognition tasks.

The subsequent sections of MFL are conducted as follows: Section II shows the description of few-shot multimedia image recognition. Section III elaborates an in-depth exposition of the MFL. Section IV offers a comprehensive depiction of experiment evaluations. Lastly, Section V concludes MFL.

2. Related works

Currently, few-shot multimedia image recognition methods in English education based on deep models have emerged a lot, encompassing metric-based strategies, optimization-based strategies, data augmentation-based methods.

2.1. Metric-based strategies

Metric-based methods focus on learning features with good category representativeness, inter-class discriminability, and effective metric methods for feature comparison, which methods have become the most widely applied in the field of the few-shot multimedia image recognition due to their effectiveness and e of scalability. For instance, RelationNet employs a deep convolutional network as a metric, departing from the conventional Euclidean distance, to capture and model correlations between pairs of images for the purpose of classification [10]. AL-FS provides a valuable insight into algorithm design for few-shot classification and streamlining the design process via disentangling the training and adaptation processes [11]. DeepEMD utilizes the differentiable earth mover's distance as a measurement function, providing a means to capture patch-level similarities between image pairs of support and query,

allowing for the modeling of detailed patch-level relationships and enhancing the performance [8]. BSSD employs the Sinkhorn distance to establish a local-optimal matching between support-query image pairs, diverging from global-position matching. This strategic choice effectively addresses data scarcity problems by modeling local feature maps, fostering more consistent correlations between support and query images [12]. MFS employs set-based representations, incorporating lightweight self-attention matrices into established autoencoder networks, to amplify the transferability of representation vectors [13].

2.2. Optimization-based strategies

Optimization-based strategies want to capture a set of initial parameters to help models to converge quickly on unseen classes using only a few samples through several optimization steps. For instance, MAML (Model-Agnostic Meta-Learning) formulates a model-agnostic optimization objective designed to unveil optimal initialization parameters [14]. This approach ensures rapid adaptation to new tasks, requiring only one or a few annotated images. MAML's methodology focuses on achieving a robust and adaptable model initialization that facilitates efficient learning from limited data for various tasks. Reptile utilizes first-order approximation techniques in the optimization process, deviating from the previous MAML's approach of computing the costly second-order derivatives of the Hessian matrix [15]. This modification significantly enhances learning efficiency by circumventing the need for computationally expensive second-order derivatives, thereby making the training process more efficient and scalable. MetaOptNet updates the MAML framework by employing the optimization objective of the nonlinear classifier, diverging from the conventional linear classifier [16]. MELR, on the other hand, employs cross-episode attention and cross-episode consistency during meta-training to enhance robustness against poorly-sampled shots in meta-test [17]. LLAC takes a different approach by transforming task-agnostic classifiers into task-specific ones. It dynamically predicts classifier weights for each unique task, guided by the center-uniqueness loss [18].

2.3. Data augmentation-based strategies

Data augmentation-based strategies represent the most intuitive approach, involving the generation of additional training samples to address the small-sample problem. However, contemporary approaches often integrate data augmentation methods with metric-base or optimization-based strategies to collectively enhance model performance, rather than relying solely on data augmentation to tackle the challenges of few-shot multimedia image recognition. For instance, MFGN enables the generative model to generate features for the remaining samples belonging to each class in the support set based on the feature of one sample from that class in the training process [19]. Such a manner can capture intra-class variations and allow the model to generate additional features for the same class hinging on a restricted pool of labeled examples when addressing few-shot tasks. Label-Halluc first trains a classifier on few-shot tasks and then utilizes this classifier to generate pseudo-labels for samples in an auxiliary set [20]. The information from the auxiliary set, along with pseudo-labels and knowledge distillation, is employed to train a few-shot classifier. DAGAN first extracts the category features for generating new data where Gaussian noise is incorporated into these category features [21]. The noised category features are then input into the decoder to obtain the generated new samples. This method controls the output sample's category by inputting the category of the sample, enabling the generation of additional training samples for any new class. FFT can leverage continuous attributes in scene images, such as rainy/clear and day/night, to synthesize scene samples with varying degrees of other attributes [22]. For example, it can generate scenes with different lighting conditions or levels of rainfall intensity.

Differences between Current Methods and MFL. Current methods often focus on cross-entropy loss in sample classification, neglecting the high-level abstract semantic information inherent in the samples. This oversight results in insufficient feature learning and inadequate class representation. To address small inter-class differences, it is crucial to emphasize discriminative local features that amplify distinctions between classes. However, existing approaches often overlook the inter-class sharing of original local features and the presence of category-independent information, which limits the classification effectiveness of these local features. In contrast, MFL employs fast Fourier convolution to extract multi-scale global information from samples and integrates it with the original local features to generate local descriptors with stronger inter-class uniqueness. MFL then calculates the similarities of each local descriptor within the same class and across different classes, amplifying the impact of descriptors that show high intra-class similarity and low inter-class similarity during classification. This approach reduces the influence of irrelevant information, enhances the discriminative power of local descriptors, and ultimately enables the model to achieve more accurate classification results.

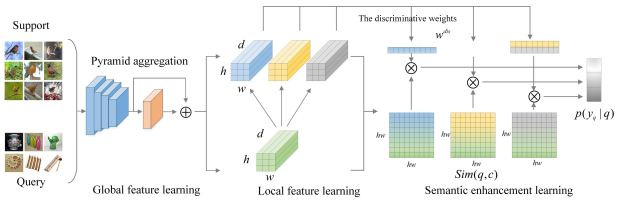

Fig. 1. The illustration of MFL.

3. Problem formulation

The few-shot multimedia image recognition is to develop a versatile model by leveraging a meticulously annotated and comprehensive base dataset. Subsequently, the model is finetuned by utilizing a restricted set of training images from a new dataset, aimed at proficiently classifying entirely new categories. Specifically, given a base dataset D_{base} and a new dataset D_{new} whose classes are disjoint, i.e., $Y^{\text{base}} \varnothing Y^{\text{new}}$, an episodic training strategy with support sets and query sets is utilized to train a classifier of the new dataset based on the base dataset. Formally, the support set is formed by the inclusion of n distinct classes, wherein each individual class is represented by k images, which is defined as $S = \{(x_{si}, y_{si})\}_{i=1}^{n\times k}$, where x_{si} denotes the i-th image in D_{base} (or D_{new}), and y_{si} denotes corresponding label. The query set is formed by the inclusion of n distinct classes, wherein each individual class is represented by m images, which is defined as $Q = \left\{\left(x_{qi}, y_{qi}\right)\right\}_{i=1}^{n\times m}$. The training process of the few-shot multimedia image recognition model involves utilizing S and Q sampled from D_{base} . Subsequently, the model undergoes evaluation using S and Q sampled from the D_{new} .

The few-shot multimedia image recognition model is trained on the support set S and the query set Q sampled from D_{base} , and tested on the support set S and the query set Q sampled from D_{new}.

4. Multi-level feature learning for multimedia pattern recognition in english education

As illustrated in Fig. 1, a novel multi-level feature learning method (MFL) is proposed for few-shot multimedia image recognition, comprising four key modules: the feature extractor, the global feature learning module, the local learning module, and the semantic enhancement module. MFL begins by using the feature extractor to generate local features for both support and query samples. These local features are then fed into the global feature learning module, where multi-scale global insights are derived using the pyramid aggregation network with a residual structure, resulting in enriched local descriptors. Next, the local learning module assigns weights based on the similarity between these enhanced local descriptors and the class descriptors. Finally, the weight matrix and enhanced local descriptors from the support and query samples are processed by the semantic enhancement module, which produces classification outcomes that emphasize critical details. The following section provides a detailed explanation of each module.

4.1. The feature extractor module

In the episodic task $E = \{S, Q\}$ of the few-shot multimedia image recognition, MFL utilizes the feature extractor, implemented by an any convolutional neural network without the fully connected layer, to derive the initial array of local descriptors from each sample in the dataset.

Specifically, given a sample $x_i \in [D, H, W]$ of task E, where D, H, and W denote channel number, the height, the width, respectively, MFL learns the latent representation z_i of x_i via the feature extractor f_θ :

$$z_i = f_\theta\left(x_i\right) \in [d, h, w] \tag{1}$$

Then, considering each d-dimensional vector corresponding to each position in the feature matrix as a local descriptor, we can obtain a set of M local descriptors:

$$F\left(x_i\right) = \left\{z_{i,1}, z_{i,2}, \ldots, z_{i,M}\right\} \in [d, M] \tag{2}$$

where $F\left(x_i\right)$ denotes the set of local descriptors of x_i , and M=h * w .

4.2. The global feature learning module

Local descriptors may suffer from insufficient inter-class discrimination due to their focus on local features. To address this, MFL introduces a pyramid multi-scale feature learning network designed to extract and integrate global and semi-global information from representations, enhancing local descriptors with multi-scale information.

Given a representation $z_i \in [d, h, w]$, the pyramid multi-scale feature learning network enhances local descriptors through the following steps: (1) The network first partitions the representation into local features $z_i^l \in [(1-\alpha)d, h, w]$ and global features $z_i^g \in [\alpha d, h, w]$ using the hyper-parameter α. Eq. (2) The global features z_i^g are processed through a pyramid structure to capture multi-scale information: A series of down-sampling operations, such as max-pooling, followed by convolutional layers, are applied to generate a pyramid of feature maps at different scales. Each level of the pyramid represents information at a different level of abstraction, from fine details to coarse features. Eq. (3) The network then integrates features from different scales: Features from each pyramid level are

concatenated or summed to create a comprehensive multi-scale representation. This approach combines detailed local features with global context, enriching the feature representation. These multi-scale features are fused with the local features z_i^l using additional convolutional layers to capture the interactions between local and global information. Eq. (4) the network concatenates the enhanced local features and multi-scale global features along the channel dimension to produce the final multi-scale feature representation $z_i^{pMS} \in [d, h, w]$. This output is then combined with the original features z_i through residual connections to generate the enhanced local descriptors:

$$\bar{F}(x_i) = \left\{\bar{z}_{i,1}, \bar{z}_{i,2}, \ldots, \bar{z}_{i,M}\right\} \in [d, M] \tag{3}$$

4.3. The local learning module

For each local descriptor of support images, two representative scores are defined: the intra-class representative score w^{intra} that reflects intra-class similarity and the cross-class representative score w^{inter} that reflects inter-class differentiation. Prior to providing explicit formulae for the two scores, it is imperative to establish the local descriptor enhancement:

$$F(S) = \{F(x_1), F(x_2), \ldots F(x_{nk})\} \in [nk, d, M] \tag{4}$$

At the same time, using prototypes as representatives for each class, the calculation for the i-th class prototype is:

$$\bar{F}(i) = \frac{1}{k}\sum_{j=1}^{k} \bar{F}\left(x_{(i-1)j+j}\right) \tag{5}$$

For each class, a prototype of the local descriptor enhancement $\bar{z}_i^p$ is calculated to represent the significant local features of the i-th class, i.e., $z_i^p = \frac{1}{M}\sum_{m=1}^{M} z_{i,j}^p$, where $z_{i,j}^p$ is the j-th enhanced local description of the i-th class prototype. Following that, the representativeness score for intra-class similarities is computed for the m-th enhanced local descriptor of the k-th sample belonging to the i-th class:

$$w_{i,k,m}^{\text{intra}} = \frac{\left(\bar{z}_{i,k,m}\right)^T \bar{z}_i^p}{\left\|\bar{z}_{i,k,m}\right\|_2 \left\|\bar{z}_i^p\right\|_2} \tag{6}$$

This score represents the matching degree between the highly activated region of the local descriptor enhancement $\bar{z}_{i,k,m}$ and the classification saliency region represented by the prototype $\bar{z}_i^p$, that is, to what extent the information contained in the local descriptor enhancement can represent class features.

The cross-class representativeness score is as follows:

$$w_{i,k,m}^{\text{inter}} = \frac{1}{N-1}\sum_{n=1;n\neq i}^{N} \frac{\left(\bar{z}_{i,k,m}\right)^T \bar{z}_n^p}{\left\|\bar{z}_{i,k,m}\right\|_2 \left\|\bar{z}_n^p\right\|_2} \tag{7}$$

The score represents the degree to which the highly activated region of $\bar{z}_{i,k,m}$ matches the significant regions of other categories, that is, to what extent the information contained in the enhanced local descriptor can confuse the current category with other categories. Intuitively, when the intra-class representativeness score of an enhanced local descriptor is high and the cross-class representativeness score is low, that is, it can effectively represent the features of this class without being confused with other class features, the enhanced local descriptor should dominate the classification process. Therefore, this section uses these two scores to define the discriminative weights for local descriptor enhancement:

$$w_{i,k,m}^{\text{dis}} = w_{i,k,m}^{\text{intra}} - w_{i,k,m}^{\text{inter}} \tag{8}$$

In summary, the discriminative weights defined on the support set in this subsection can be assigned corresponding weights based on whether the enhanced local descriptor contains discriminative information that is beneficial for classification. Because the support set samples have ground-truth labels, it is possible to determine which enhanced local descriptors have more class representativeness and are more reliable in classification according to Eq. (8); However, the query set samples do not have label information, and the value obtained from Eq. (8) can only reflect the similarity of enhanced local descriptors to different classes. It is not possible to infer which enhanced local descriptors contain more real class information, so only the discriminative weights of enhanced local descriptors in the support sets are calculated. Finally, the discriminative weights are concatenated to obtain a discriminative attention matrix with enhanced local descriptors of support sets, i.e., $w_{i,k,m}^{\text{dis}} \in [n, kM]$, which participates in the calculation of classification results in the sample-class metric.

4.4. The semantic enhancement module

According to the definition of the local descriptor enhancement in support samples, the local descriptor enhancement in query samples is obtained via:

$$\bar{F}(q) = \left\{\bar{z}_{q,1}, \bar{z}_{q,2}, \ldots, \bar{z}_{q,M}\right\} \in [d, M] \tag{9}$$

Firstly, we calculate the similarity between the local descriptor enhancement of the query sample $\bar{F}(q)$ and the class c descriptor $\bar{F}(c)$, and obtains the following similarity matrix:

$$\text{sim}(q,c) = \cos(\bar{F}(q), \bar{F}(c)) \in [M, kM] \tag{10}$$

where $\cos(\cdot)$ denotes the cosine similarity. we utilize the discriminative attention matrix of class c to weight the similarity matrix and incorporate significance information into the similarity matrix:

$$\text{sim}'(q,c) = \text{sim}(q,c) \circ w^{\text{dis}}[c,:] + \text{sim}(q,c) \tag{11}$$

Next, we select j of the most significant kM similarities obtained from each query sample descriptor to obtain a similarity matrix, i.e., top $\left(\text{sim}'(q,c)\right) \in [M,k]$, and then sum the selected similarities to obtain the similarity, denoted as:

$$\phi(q,c) = \sum_{i=1}^{M}\sum_{j=1}^{k} \text{topk}\left(\text{sim}'(q,c)\right)[i,j] \tag{12}$$

We convert the similarity into probability through softmax normalization:

$$P(c \mid q) = \frac{\exp(\phi(q,c))}{\sum_{c' \in N} \exp\left(\phi\left(q,c'\right)\right)} \tag{13}$$

The similarity between the query sample q and other classes can be obtained in the same manner. Finally, we train MFL via the cross entropy loss:

$$L = -\frac{1}{nk}\sum_{q=1}^{k}\sum_{c=1}^{n} y \log P(c \mid q) \tag{14}$$

After obtaining the optimal model parameters through training, fix all parameters and complete the few-shot multimedia image recognition in the test set according to Eq. (14). Calculate the average accuracy of the model on all classification tasks to obtain its classification performance on the new class.

5. Experiments

5.1. Setup

Dataset and metric: The classification performance of the MFL is evaluated using three few-shot multimedia image recognition datasets, i.e., CUB-200-2011 [8] , StanfordDogs [8] , and StanfordCars [8] . The scale and partition details of each dataset are presented in Table 1. Consistent with prior researches, classification performance is evaluated via the accuracy (ACC), where higher values signify superior performance.

$$\text{ACC} = \frac{N_{\text{correct}}}{N_{\text{total}}} \tag{15}$$

where N_{correct} and N_{total} denote the number of correctly classified samples and the total samples.

Implementation Details: In MFL, a four-layer convolutional network is used as the feature extractor, denoted as θ_f. Each module in the feature extractor consists of a convolutional layer with 643×3 kernels, a LeakyReLU activation function, and a batch normalization layer. Depending on the input sample size of 84×84 or 32×32, the output feature matrix dimensions are $64 \times 10 \times 10$ or $64 \times 8 \times 8$, respectively. This results in each sample having either 100 or 64 local descriptors, each with a dimension of 64 . MFL is trained on each dataset for 200 epochs, with 200 few-shot tasks per epoch. The Adam optimizer is used for parameter updates, starting with an initial learning rate of 0.001 , which is halved every 20 epochs. Given the strong representational power of fast Fourier convolution, the model may overfit the training set, so different weight decay parameters are applied based on the specific dataset's training conditions.

5.2. Comparison with baselines

Comparison baselines: Twelve baseline methods are compared in the experiments, including DN4 (CVPR, 2019) [9], FEAT (CVPR, 2020) [7], BSNet (TIP, 2020) [23], BSNet+ (TIP, 2020)[23] , ATL-net (IJCAI, 2021) [24], IFLT (ECCV, 2022) [25], TOAN (TCSVT, 2022) [26], and DeepEMD (TPAMI, 2022) [8].

Comparison results: Table 2 presents a comparison of average accuracy across three real-world multimedia datasets for the 5 -way 1 -shot and 5 -way 5 -shot settings. The results indicate that MFL either outperforms other methods or matches the top-performing methods in most cases, demonstrating its effectiveness and superiority. Meanwhile, there exist three observations: (1) Performance Comparison: MFL, ATL-Net, and BSNet generally outperform IFTL, DN4, and DeepEMD. This is likely because the former methods impose additional constraints on local descriptors, while the latter methods use each local descriptor equally. This suggests that some local descriptors contain more useful information for classification, while others may provide less relevant information. Selectively using different local descriptors can lead to more accurate results. (2) Relative Improvement Over ATL-Net: Compared to ATL-Net, MFL achieves comparable or better accuracy in all cases. Notably, MFL demonstrates a relative improvement of 11.66% in the 5 -way 1 -shot setting on CUB-200-2011. Both methods aim to mitigate the impact of irrelevant information, but ATL-Net's local descriptors contain only local features, while MFL's descriptors integrate multi-scale global semantic information. This results in MFL showing relatively better accuracy in multiple settings, as simple local information often lacks sufficient inter-class discrimination. (3) Robustness Across Datasets: Some methods perform well only under certain settings. For example, ATL-Net per-

Table 1. Three real world datasets.

dataset	class	sample	train/val/test	type
CUB-200-2011	200	11788	100/50/50	84*84
StanfordDogs	120	20580	70/20/30	84*84
StanfordCars	193	16185	130/17/49	84*84

forms well on StanfordDogs and StanfordCars but shows significantly lower performance on CUB-200-2011. In contrast, MFL achieves cutting-edge accuracy across all three datasets, indicating that the issue addressed by MFL is a common concern inherent in the data. This highlights the robust universality of MFL.

5.3. Ablation Study

This section conducts five ablation experiments about the feature extractor module, the global feature learning module, the local feature learning module, and the semantic enhancement module, to further analyze component effectiveness on the CUB-200-2011 dataset. Specifically, (1) MFL utilizes the feature extractor module as backbone network to obtain classification results (backbone). (2) MFL combines the feature extractor module and the global feature learning module to obtain classification results (backone+G). (3) MFL combines the feature extractor module and the local feature learning module to obtain classification results (backone+L). (4) MFL combines the feature extractor module and the semantic enhancement module to obtain classification results (backone+S). (5) MFL combines the feature extractor module, the global feature learning module, the local feature learning module to obtain classification results (backone+G+L). As shown in Table 3, every component is vital for harnessing the full potential of the model, and omitting any part would compromise its effectiveness.

6. Conclusion

A novel multi-level feature learning approach (MFL) is proposed for few-shot multimedia image recognition, effectively addressing the challenge of small inter-class differences. MFL enhances the discriminative power of local descriptors by incorporating global information and modeling both intra-class and inter-class relationships, leading to better classification accuracy and a deeper understanding of class similarities and differences. Experimental results on three datasets demonstrate that MFL outperforms existing methods, highlighting its potential to enhance classification accuracy in scenarios with limited annotated data. However, the method's reliance on the strong representational capacity of fast Fourier convolution may make it prone to overfitting, and its dependence on handcrafted feature extraction could limit adaptability across diverse datasets. Future research could explore the integration of automated feature selection methods to mitigate overfitting, expand MFL's application to other multimedia content types such as audio or video, and optimize the dynamic balance between global and local features during training for further performance improvements.

References

[1] J. Y. Lim, K. M. Lim, C. P. Lee, and Y. X. Tan, (2023) *"SCL: Self-supervised Contrastive Learning for Few-shot Image Classification"* **Neural Networks**: 19–30. DOI: 10.1016/j.neunet.2023.05.037.

[2] J. Gao, P. Li, A. A. Laghari, G. Srivastava, T. R. Gadekallu, S. Abbas, and J. Zhang, (2024) *"Incomplete multiview clustering via semidiscrete optimal transport for multimedia data mining in IoT"* **ACM Transactions on Multimedia Computing, Communications and Applications *20***(6): 1–20. DOI: 10.1145/3625548.

[3] Y. Dong, H. Zhang, C. Wang, and Y. Wang. "Fine-grained Ship Classification Based on Deep Residual Learning for High-resolution SAR Images". In: *International Conference on Learning Representations, Remote Sens.* 2019, 1095–1104. DOI: 10.1080/2150704X.2019.1650982.

[4] X. Wei et al., (2022) *"Fine-grained Image Analysis with Deep Learning: A Survey"* **IEEE Trans. Pattern Anal. Mach. Intell.** 8927–8948. DOI: 10.1109/TPAMI.2021.3126648.

[5] P. Li, A. A. Laghari, M. Rashid, J. Gao, T. R. Gadekallu, A. R. Javed, and S. Yin, (2022) *"A deep multimodal adversarial cycle-consistent network for smart enterprise system"* **IEEE Transactions on Industrial Informatics *19***(1): 693–702. DOI: 10.1109/TII.2022.3197201.

[6] P. Chikontwe, S. Kim, and S. H. Park. "CAD: Co-adapting Discriminative Features for Improved Few-shot Classification". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 14554–14563.

Table 2. Comparison of MFL to prior works on three few-shot datasets in terms of the accuracy. - denotes the results are not available in the official articles.

	CUB-200-2011		StanfordDogs		StanfordCars	
Method	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
DN4	66.46	83.26	51.14	70.55	61.88	88.71
FEAT	64.82	80.73	50.98	70.15	62.86	81.99
BSNet	55.81	76.34	43.16	62.61	44.56	63.72
BSNet+	62.84	85.39	43.42	71.90	40.89	86.88
ATL-net	60.91	77.05	50.49	73.20	67.95	89.16
IFLT	64.25	76.41	41.35	56.57	36.70	64.25
TOAN	66.10	82.27	49.77	69.29	70.28	87.45
DeepEMD	64.08	80.55	46.73	65.74	61.63	72.95
MFL	**71.12**	**85.85**	**52.12**	73.11	**771.19**	**88.92**

Table 3. Ablation experiments of each component in MFL (CUB-200-2011).

	backbone	backone + G	backone + L	backone + S	backone + L + G	MFL
5-way 1-shot	55.78	62.55	62.55	60.24	65.81	71.12
5-way 5-shot	65.11	71.26	73.44	74.21	80.12	85.85

[7] H. J. Ye, H. Hu, D. C. Zhan, et al. "Few-shot Learning via Embedding Adaptation with Set-to-set Functions". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 8808–8817.

[8] C. Zhang, Y. Cai, G. Lin, and C. Shen. "DeepEMD: Few-shot Image Classification with Differentiable Earth Mover's Distance and Structured Classifiers". In: *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. 2020, 12203–12213.

[9] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo. "Revisiting Local Descriptor Based Image-to-Class Measure for Few-shot Learning". In: *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. 2019, 7260–7268.

[10] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. "Learning to Compare: Relation Network for Few-shot Learning". In: *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. 2018, 1199–1208.

[11] X. Luo, H. Wu, J. Zhang, L. Gao, J. Xu, and J. Song. "A closer look at few-shot classification again". In: *International Conference on Machine Learning*. 2023, 23103–23123.

[12] Y. B. Liu, L. C. Liu, X. H. Wang, M. Yamada, and Y. Yang, (2023) *"Bilaterally Normalized Scale-consistent Sinkhorn Distance for Few-shot Image Classification"* **IEEE Trans. Neural Netw. Learn. Syst.** 12203–12213.

[13] A. Afrasiyabi, H. Larochelle, J.-F. Lalonde, and C. Gagné. "Matching Feature Sets for Few-shot Image Classification". In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 2022, 9014–9024.

[14] C. Finn, P. Abbeel, and S. Levine. "Model Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. 2017, 1126–1135.

[15] K. Lee, S. Maji, A. Ravichandran, and S. Soatto. "Meta-Learning with Differentiable Convex Optimization". In: *CVPR*. 2019, 10657–10665.

[16] A. Nichol, J. Achiam, and J. Schulman, (2018) *"On First-Order Meta-Learning Algorithms"* **CoRR *abs/1803.02999***:

[17] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola. "Rethinking Few-shot Image Classification: A Good Embedding is All You Need?" In: *Computer Vision – ECCV*. 2020.

[18] N. Lai, M. Kan, C. Han, X. Song, and S. Shan, (2021) *"Learning to Learn Adaptive Classifier-Predictor for Few-Shot Learning"* **IEEE Trans Neural Netw Learn Syst**: 3458–3470. DOI: 10.1109/TNNLS.2020.3011526.

[19] Y. Yu, D. Zhang, and Z. Ji. "Masked Feature Generation Network for Few-Shot Learning". In: *IJCAI*. 2022, 7005–7014.

[20] Y. Jian and L. Torresani. "Label Hallucination for Few-Shot Classification". In: *AAAI*. 2022, 3695–3701. DOI: 10.1609/aaai.v36i6.20659.

[21] A. Antoniou, A. Storkey, and H. Edwards. "Data Augmentation Generative Adversarial Networks". In: *AAAI*. 2018, 1050.

[22] R. Kwitt, S. Hegenbart, and M. Niethammer. "One-shot Learning of Scene Locations via Feature Trajectory Transfer". In: *CVPR*. 2016, 78–86.

[23] L. X, W. J, S. Z, et al., (2020) *"BSNet: Bi-similarity Network for Few-shot Fine-grained Image Classification"* **IEEE Transactions on Image Processing**: 1318–1331. DOI: 10.1109/TIP.2020.3043128.

[24] C. Dong, W. Li, J. Huo, et al. "Learning Task-aware Local Representations for Few-shot Learning". In: *IJCAI*. 2021, 716–722.

[25] B. Q, R. I, and A. R. "Improving Few-Shot Learning Through Multi-task Representation Learning Theory". In: *Computer Vision–ECCV*. 2022, 435–452.

[26] H. Huang, J. Zhang, L. Yu, J. Zhang, Q. Wu, and C. Xu, (2021) *"TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples"* **IEEE Transactions on Circuits and Systems for Video Technology** *32*(2): 853–866.