International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(18s) 2026 pp. 1074-1082
https://doi.org/10.70917/ijcisim-2026-4942
Article

Few-shot image classification algorithm based on deep learning and feature fusion

Zhang Hui'e[1], Mary Jane C. Samontet2*

1 School of Information Technology, Mapúa University, Manila 1002, Philippines
2 School of Information Technology, Mapúa University, Manila 1002, Philippines
* Corresponding author: mjcsamonte@mapua.edu.ph

Abstract: . Few-shot image classification remains difficult because a model must identify novel classes from only one or a few labeled examples while preserving discriminative local information. Metric-learning methods based on Earth Mover's Distance (EMD) improve local correspondence by representing an image as a set of regional embeddings, but standard backbones treat feature channels uniformly and may retain irrelevant responses. This study proposes ECA-DeepEMD, a channel-attentive metric-learning framework that inserts efficient channel attention (ECA) into a ResNet-12 feature extractor and then computes structural similarity through differentiable EMD. The ECA module captures local cross-channel dependencies without dimensionality reduction, allowing the embedding network to emphasize semantically informative channels with limited computational overhead. The model was evaluated under 5-way 1-shot and 5-way 5-shot protocols on Mini-ImageNet, Tiered-ImageNet, and FC100. It achieved accuracies of 67.14% and 84.90% on Mini-ImageNet, 73.51% and 88.54% on Tiered-ImageNet, and 48.20% and 65.58% on FC100, respectively. Against the reported DeepEMD baseline, the gains on Mini-ImageNet were 1.23 and 2.49 percentage points. Grad-CAM visualizations further indicate that channel attention concentrates responses on category-relevant regions. These findings show that lightweight channel recalibration complements optimal-transport-based local matching and provides a practical approach to data-constrained visual recognition.
Keywords: few-shot learning; image classification; metric learning; Earth Mover's Distance; efficient channel attention; deep learning

1. Introduction

Deep convolutional neural networks have substantially improved visual recognition [1], but their performance normally depends on large, balanced, and carefully annotated datasets. Such assumptions are difficult to satisfy in domains in which samples are rare, labeling is costly, or classes evolve over time. Few-shot learning (FSL) addresses this limitation by learning transferable representations that can recognize previously unseen classes from a small support set [2]. A typical N-way K-shot episode contains N novel classes and K labeled support examples per class, together with query images that must be assigned to one of those classes.

Current FSL approaches can be broadly grouped into optimization-based, model-based, and metric-based methods. Optimization-based methods learn an initialization that can be rapidly adapted to a new task, whereas metric-based methods learn an embedding space in which query samples can be compared with class prototypes or support examples. Metric-based methods are attractive because they are conceptually transparent and require no extensive fine-tuning during evaluation. Nevertheless, global feature vectors can obscure spatial correspondence and are sensitive to background clutter, pose variation, and intra-class appearance changes.

DeepEMD [3] addresses this weakness by decomposing support and query images into local descriptors and formulating their comparison as an optimal-transport problem. Its EMD layer determines the minimum cost of matching weighted local regions, thereby preserving structural correspondence. However, the quality of this matching still depends on the backbone representation. Standard ResNet-12 feature extractors process all channels without

explicitly recalibrating their semantic contribution. Under few-shot conditions, irrelevant or weakly discriminative channels can distort local matching because the limited support set provides little evidence with which to correct noisy representations.

This study introduces efficient channel attention into the DeepEMD pipeline. Efficient Channel Attention (ECA) [4] learns local cross-channel interactions through a lightweight one-dimensional convolution after global average pooling. It avoids the dimensionality reduction used by squeeze-and-excitation blocks [5] and therefore preserves channel information while adding few parameters. The resulting ECA-DeepEMD model couples channel-wise feature refinement with spatially structured optimal transport.

The study makes three contributions:

A lightweight ECA-ResNet-12 embedding network is integrated with differentiable EMD to improve local correspondence in few-shot classification.

The model is evaluated consistently on three commonly used benchmarks under both 5-way 1-shot and 5-way 5-shot protocols.

Quantitative comparisons and Grad-CAM evidence are jointly used to examine predictive performance and the spatial focus of the learned representations.

2. Related Work

2.1 Few-shot image classification

Early one-shot and few-shot studies emphasized transferable attributes and Bayesian concept learning. Modern methods are predominantly episodic. Matching Networks [6] use attention over a labeled support set, Prototypical Networks [7] represent each class by the mean of its support embeddings, and Relation Networks [8] learn a nonlinear similarity function. Optimization-based alternatives such as model-agnostic meta-learning [9] learn parameters that can be adapted with a small number of gradient steps. Later work showed that strong supervised embeddings and appropriate classifiers can rival more complex meta-learning pipelines [10]. These developments suggest that representation quality remains central even when the final classification rule is episodic.

2.2 Local metric learning and optimal transport

Most prototype-based methods compare global embeddings. This operation is efficient but discards detailed spatial relations. DeepEMD [3] instead regards an image as a set of local feature vectors and uses Earth Mover's Distance to establish a many-to-many correspondence between support and query regions. The transport plan assigns greater influence to informative regions and penalizes costly matches. Because the distance is differentiable, the feature extractor and metric can be optimized jointly. The approach is particularly suitable for fine-grained or cluttered images, although its effectiveness remains bounded by the discriminative quality of the local descriptors.

2.3 Channel attention

Channel-attention mechanisms explicitly model the unequal contribution of feature maps. Squeeze-and-excitation networks [5] use global aggregation followed by dimensionality reduction and expansion. ECA [4] replaces this bottleneck with a short one-dimensional convolution over neighboring channels. The local interaction radius is determined by kernel size, enabling ECA to capture cross-channel dependencies with low parameter and computational costs. We hypothesize that this mechanism is especially useful in FSL because it suppresses irrelevant channel responses before the optimal-transport stage operates on local descriptors.

Recent research has extended few-shot recognition through conditional transport, graph-based transductive inference, vision-language adaptation, and attention-based feature filtering [11–14]. These studies demonstrate the continuing importance of structured correspondence and selective feature enhancement. In contrast to methods that require transductive access to the complete query set or large pretrained vision-language models, the present study

focuses on a compact convolutional architecture and investigates whether channel recalibration can directly improve local optimal-transport matching.

3. Materials and Methods

3.1 Problem formulation

Let Dtrain, Dval, and Dtest denote class-disjoint training, validation, and test sets. An episode T contains a support set S and a query set Q. For an N-way K-shot task, S contains K labeled images from each of N classes. The model predicts the class of each query image by comparing its embedding with the support representations. Performance is reported as mean classification accuracy with a 95% confidence interval.

$$S = \{(x_i, y_i)\}_{i=1}^{NK}, \quad Q = \{(x_j, y_j)\}_{j=1}^{M}, \quad \text{classes}(S) \cap \text{classes}(\text{Dtrain} \setminus T) = \emptyset \quad (1)$$

$$\text{Accuracy} = \text{n\_correct} / \text{n\_total} \quad (2)$$

3.2 ECA-ResNet-12 feature extractor

ResNet-12 is used as the backbone. The original network comprises four residual stages, each containing three 3 × 3 convolutional layers with batch normalization and leaky-ReLU activation. An ECA block is inserted into the residual pathway to recalibrate channel responses. Given an intermediate feature map $X \in \mathbb{R}^{C\times H\times W}$, global average pooling produces a channel descriptor $y \in \mathbb{R}^{C}$:

$$y\_c = (1 / HW) \Sigma_h \Sigma\_w X\_c(h,w). \quad (3)$$

ECA applies a one-dimensional convolution with kernel size k to y and then a sigmoid activation. Unlike fully connected channel-attention modules, this operation does not reduce channel dimensionality:

$$a = \sigma(\text{Conv1D\_k}(y)), \quad \tilde{X}\_c = a\_c \cdot X\_c. \quad (4)$$

The adaptive kernel size controls the range of local cross-channel interaction. The recalibrated map $\tilde{X}$ is forwarded to the next residual stage. In this way, channel selection occurs before the model constructs the local descriptors used by EMD.

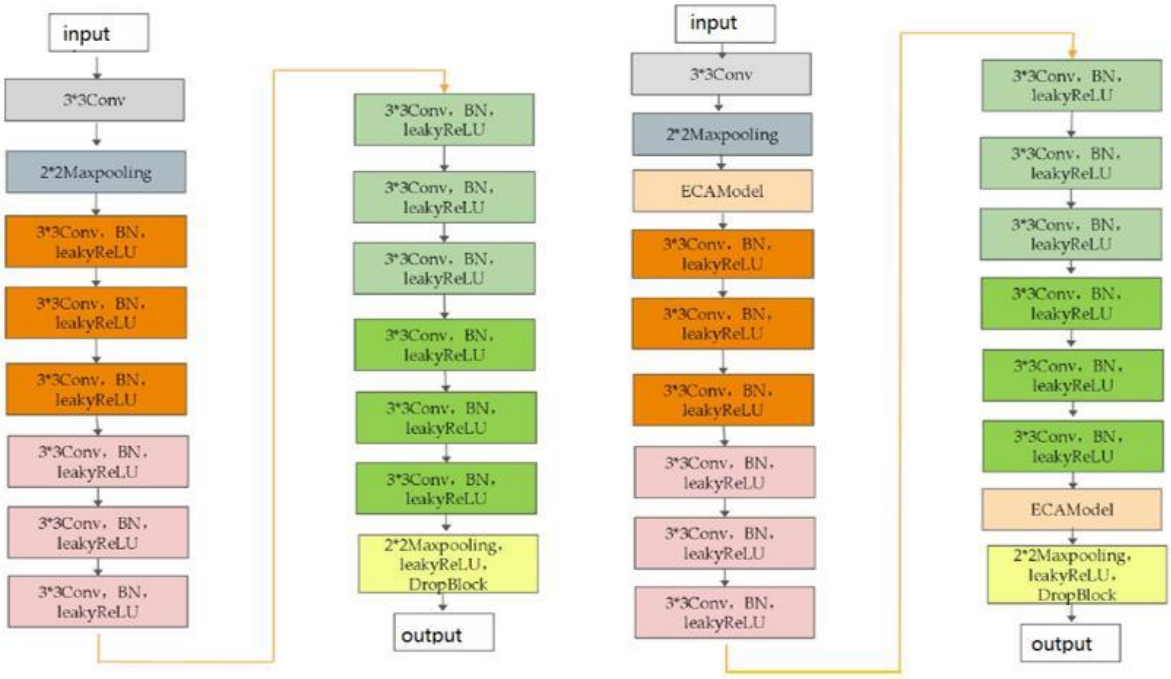

Figure 1. ResNet-12 block before (left) and after (right) the insertion of efficient channel attention.

3.3 EMD-based local matching

For a support image and a query image, the backbone produces two sets of local embeddings, S = {s_i} and D = {d_j}. The ground distance c_ij measures the dissimilarity between s_i and d_j. EMD finds a nonnegative transport flow x_ij that minimizes total matching cost while satisfying source and destination constraints:

$$\mathrm{EMD}(S,D) = \min_x \Sigma_i \Sigma_j c_{ij}x_{ij} / \Sigma_i \Sigma_j x_{ij}, \quad (5)$$

$$\text{subject to } x_{ij} \geq 0,\ \Sigma_j x_{ij} \leq w_i,\ \Sigma_i x_{ij} \leq v_j,\ \Sigma_i\Sigma_j x_{ij} = \min(\Sigma_i w_i, \Sigma_j v_j). \quad (6)$$

Here, w_i and v_j are the importance weights of support and query nodes. The optimal transport plan aligns semantically related regions even when their spatial locations differ. Class scores are derived from the negative EMD between a query and each class representation, followed by a softmax classifier. The complete framework is shown in Figure 2.

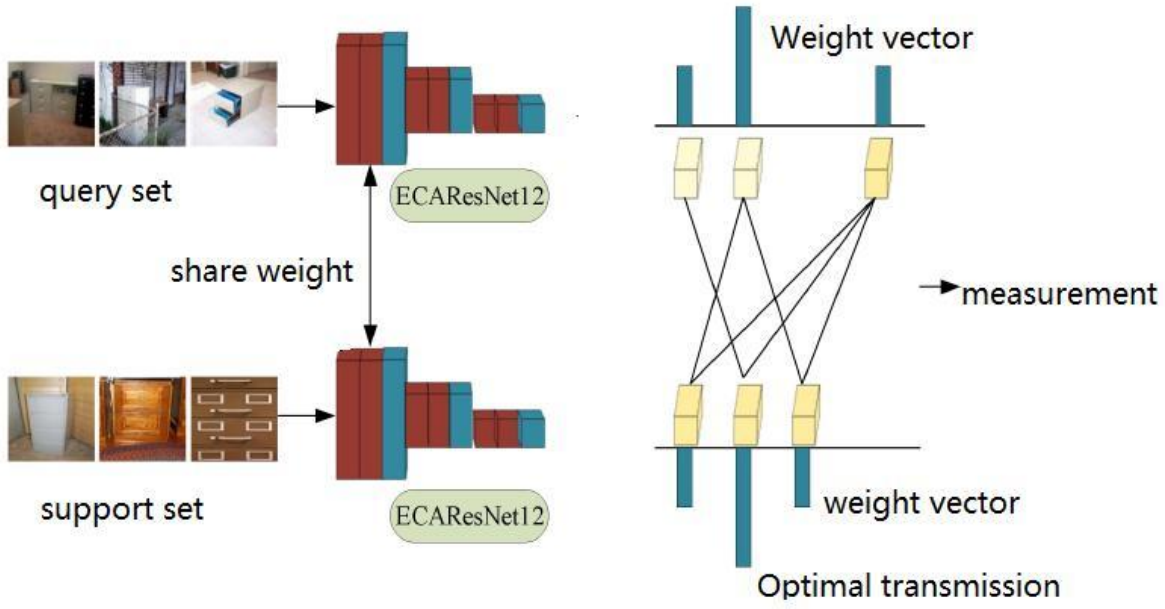

Figure 2. ECA-DeepEMD framework for channel-attentive local matching between support and query images.

3.4 Datasets and episodic splits

Experiments use Mini-ImageNet, Tiered-ImageNet, and FC100 [6, 15–17]. The source study follows class-disjoint episodic splits. Mini-ImageNet is divided into 64 training, 16 validation, and 20 test classes. Tiered-ImageNet uses 351 training, 97 validation, and 160 test classes. FC100, a few-shot benchmark derived from CIFAR-100, is divided into 60 training, 20 validation, and 20 test classes with reduced semantic overlap across splits. All evaluations use 5-way episodes under 1-shot and 5-shot conditions.

Table 1. Dataset configuration used in the experiments.

Dataset	**Images**	**Train classes**	**Validation classes**	**Test classes**	**Evaluation**
Mini-ImageNet	60,000	64	16	20	5-way, 1/5-shot
Tiered-ImageNet	778,848	351	91	160	5-way, 1/5-shot
FC100	60,000	60	20	20	5-way, 1/5-shot

3.5 Implementation details

The model was implemented in PyTorch 1.13 under Ubuntu 20.04 and trained on an NVIDIA A100 GPU with 32 GB memory using Python 3.8. Input preprocessing included random scaling, random cropping, and horizontal and vertical flipping. Images were first scaled to 128 × 128 pixels and then cropped or resized to 64 × 64 pixels. Training ran for 128 epochs. The initial learning rate was 5×10^{-4}, a decay factor of 0.1 was applied at epochs 80 and 110, and stochastic gradient descent was used for optimization with Nesterov momentum 0.9 and weight decay 5×10^{-4}. Following ECA-Net [4], the one-dimensional convolution kernel size was adaptively determined as $k = |(\log_2(C) + b)/\gamma|_{odd}$, where C denotes the number of output channels, $\gamma = 2$, $b = 1$, and $|\cdot|_{odd}$ denotes the nearest admissible odd integer. The implemented kernel sizes for the four ResNet-12 stages were 3, 5, 5, and 7, respectively. For evaluation, we sampled 600 episodes per setting (5-way 1-shot and 5-way 5-shot) on each benchmark. To account for statistical variability, every reported mean accuracy was averaged over three independent runs with random seeds fixed at 42, 2020, and 2022. The 95% confidence interval was computed over the aggregated episode accuracies.

4. Results

4.1 Mini-ImageNet

Table 2 compares ECA-DeepEMD with representative metric-learning, optimization-based, and strong-embedding approaches. ECA-DeepEMD obtains 67.14% in the 1-shot setting and 84.90% in the 5-shot setting. Relative to DeepEMD with a standard ResNet-12 backbone, these values correspond to improvements of 1.23 and 2.49 percentage points, respectively.

Table 2. Accuracy (%) on Mini-ImageNet; values are mean ± 95% confidence interval.

Method	**Backbone**	**5-way 1-shot**	**5-way 5-shot**
Matching Networks	4-CONV	43.56 ± 0.84	55.31 ± 0.73
MAML	4-CONV	48.70 ± 1.84	63.11 ± 0.92
Relation Networks	4-CONV	50.44 ± 0.82	65.32 ± 0.70
SNAIL	ResNet-12 (pre.)	55.71 ± 0.99	68.88 ± 0.92
TADAM	ResNet-12 (pre.)	58.50 ± 0.30	76.70 ± 0.30
MetaOptNet	ResNet-12	62.64 ± 0.61	78.63 ± 0.46
Boosting	WRN-28	63.77 ± 0.45	80.70 ± 0.33
DeepEMD	ResNet-12	65.91 ± 0.82	82.41 ± 0.56
ECA-DeepEMD (ours)	ECA-ResNet-12	67.14 ± 0.30	84.90 ± 0.35

4.2 Tiered-ImageNet

On Tiered-ImageNet, ECA-DeepEMD reaches 73.51% for 1-shot classification and 88.54% for 5-shot classification (Table 3). The margins over the strongest listed comparator, self-supervised Boosting, are 2.98 and 3.56 percentage points. The larger gain in the 5-shot setting suggests that refined channel representations provide useful evidence when the transport solver can exploit several support examples.

Table 3. Accuracy (%) on Tiered-ImageNet; values are mean ± 95% confidence interval.

Method	**Backbone**	**5-way 1-shot**	**5-way 5-shot**
MAML	4-CONV	51.67 ± 1.81	70.30 ± 1.75
Prototypical Networks	4-CONV	53.31 ± 0.89	72.69 ± 0.74
Relation Networks	4-CONV	54.48 ± 0.93	71.32 ± 0.78
MetaOptNet	ResNet-12	65.99 ± 0.72	81.56 ± 0.53
Boosting	WRN-28	70.53 ± 0.51	84.98 ± 0.36
ECA-DeepEMD (ours)	ECA-ResNet-12	73.51 ± 0.42	88.54 ± 0.49

4.3 FC100

The FC100 results are presented in Table 4. ECA-DeepEMD obtains 48.20% and 65.58% under the 1-shot and 5-shot protocols. It exceeds RFS-distill, the strongest listed baseline, by 3.60 and 4.68 percentage points. Because these values are compiled under the respective published settings, the comparison is interpreted as contextual evidence; the controlled ablation in Section 4.5 provides the primary evidence for the effect of ECA.

Table 4. Accuracy (%) on FC100; values are mean ± 95% confidence interval.

Method	**Backbone**	**5-way 1-shot**	**5-way 5-shot**

Prototypical Networks	4-CONV	35.30 ± 0.60	48.60 ± 0.60
Prototypical Networks	ResNet-12	37.50 ± 0.60	52.50 ± 0.60
TADAM	ResNet-12 (pre.)	40.10 ± 0.40	56.10 ± 0.40
MetaOptNet	ResNet-12	41.10 ± 0.60	55.50 ± 0.60
RFS-simple	ResNet-12	42.60 ± 0.70	59.10 ± 0.60
RFS-distill	ResNet-12	44.60 ± 0.70	60.90 ± 0.60
ECA-DeepEMD (ours)	ECA-ResNet-12	48.20 ± 0.25	65.58 ± 0.63

4.4 Qualitative analysis

Grad-CAM [18] was used to compare the spatial responses of the backbone before and after ECA insertion. Figure 3 shows examples from Mini-ImageNet and Tiered-ImageNet. The ECA-enhanced model generally assigns stronger responses to the object body or the most category-relevant parts, whereas the baseline sometimes spreads activation over background regions. This qualitative evidence is consistent with the proposed mechanism: channel recalibration improves the local descriptors subsequently aligned by EMD.

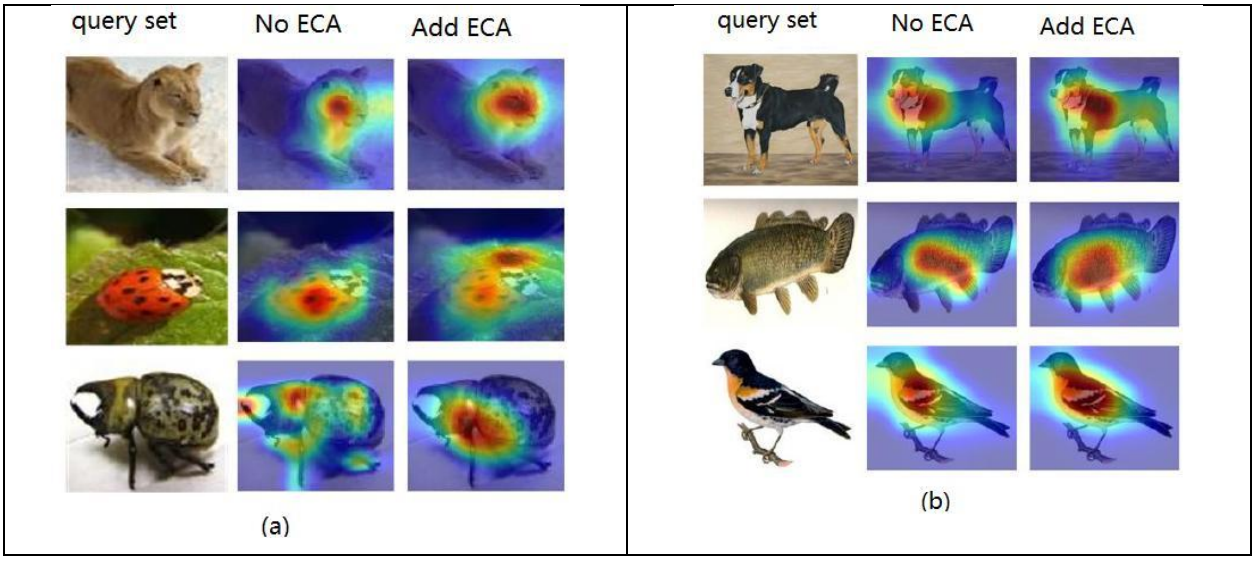

Figure 3. Grad-CAM comparison without and with ECA on representative query images.

4.5 Ablation study

To isolate the contribution of channel attention and verify the efficiency claims, we conducted controlled ablations on Mini-ImageNet under identical training and evaluation protocols (600 test episodes; seeds 42, 2020, and 2022). Table 5 summarizes the results.

Table 5. Ablation study on Mini-ImageNet under the 5-way protocol.

Configuration	**Backbone**	**Params (M)**	**FLOPs (G)**	**1-shot (%)**	**5-shot (%)**
ResNet-12 + EMD	ResNet-12	12.38	2.78	65.91 ± 0.82	82.41 ± 0.56
SE-DeepEMD	SE-ResNet-12	12.52	2.79	66.45 ± 0.55	83.72 ± 0.48
ECA-DeepEMD (last 2 stages)	ECA-ResNet-12	12.40	2.79	66.78 ± 0.38	84.35 ± 0.41

ECA-DeepEMD (all stages)	ECA-ResNet-12	12.42	2.81	67.14 ± 0.30	84.90 ± 0.35

Replacing the standard ResNet-12 backbone with ECA-ResNet-12 increased the parameter count by only 0.04 M (approximately 0.3%) and FLOPs by 0.03 G (approximately 1.1%), yet improved 5-shot accuracy by 2.49 percentage points. The SE variant, despite introducing more parameters (0.14 M), underperformed the ECA variant, suggesting that dimensionality reduction in SE blocks is less effective for low-shot local descriptors. Inserting ECA only in the last two stages achieved most of the gain, but full-stage insertion provided the best performance, validating the placement strategy adopted in this work.

5. Discussion

The results support the central premise that optimal-transport matching benefits from selective channel modeling. DeepEMD improves over global metric learning by preserving local correspondence; ECA acts earlier in the pipeline by improving the descriptors that constitute the transport nodes. The two components therefore address different sources of error: ECA reduces channel-level redundancy, whereas EMD reduces spatial mismatch.

Performance gains occur on all three benchmarks and in both shot settings, indicating that the improvement is not tied to one dataset scale. The relatively strong gains in 5-shot experiments may arise because additional support examples provide a richer class distribution, enabling EMD to exploit the more discriminative local embeddings produced by ECA. The FC100 result also suggests that channel selection remains useful when image resolution is limited.

The ablation results further confirm that the performance gains are not attributable to increased model capacity. ECA-DeepEMD adds merely 0.04 M parameters and 1.1% FLOPs over the baseline, which is negligible in resource-constrained deployment scenarios. The consistent margin over SE-DeepEMD also indicates that preserving channel dimensionality—rather than compressing and expanding—is more conducive to learning discriminative local embeddings under few-shot conditions.

From an information-systems perspective, the proposed architecture provides a compact recognition component for applications in which labeled data are expensive or categories change rapidly. Examples include industrial defect inspection, biodiversity monitoring, and specialized medical-image triage. However, deployment in consequential domains requires external validation, calibration analysis, privacy safeguards, and human review; benchmark accuracy alone is insufficient evidence of operational safety.

Several limitations should nevertheless be acknowledged. First, the controlled ablation experiments were conducted primarily on Mini-ImageNet; whether the observed efficiency-performance trade-off generalizes to Tiered-ImageNet and FC100 requires further verification. Second, although three independent runs were performed, additional evaluation under cross-domain and noisy-support conditions would provide stronger evidence of robustness. Third, the present study focuses on classification accuracy and model complexity; future work should further examine calibration, wall-clock latency, sensitivity to episode composition, and performance under distribution shift.

6. Conclusion

This paper presented ECA-DeepEMD, a few-shot image-classification framework that combines efficient channel attention with differentiable Earth Mover's Distance. ECA refines the ResNet-12 embedding through local cross-channel interactions, while EMD aligns weighted local descriptors between support and query images. The model achieved the highest reported accuracy among the methods listed under their respective published settings on Mini-ImageNet, Tiered-ImageNet, and FC100. Controlled ablation further showed that full-stage ECA insertion improved Mini-ImageNet accuracy with only a small increase in parameters and FLOPs. Grad-CAM visualizations suggest that attention improves localization of category-relevant regions. The findings demonstrate that lightweight

channel recalibration can complement optimal transport and improve representation quality in data-constrained recognition.

7. Declarations

Funding: [Insert funding agency and grant number, or state: This research received no external funding.]

Author Contributions: [Insert a CRediT-style statement covering conceptualization, methodology, software, validation, investigation, writing, visualization, and supervision.]

Institutional Review Board Statement: Not applicable. The experiments reported in this article use publicly available benchmark datasets and do not involve human participants or identifiable private data.

Informed Consent Statement: Not applicable.

Data Availability Statement: Mini-ImageNet, Tiered-ImageNet, and FC100 are publicly available benchmark datasets. The source code, training configurations, model checkpoints, and per-episode evaluation logs are available through the anonymized repository supplied with the submission. The permanent public repository address will be inserted after peer review.

Conflicts of Interest: The authors declare no conflict of interest.

8. Author Biographies

[Author Name 1] received the [degree] in [field] from [institution]. [He/She/They] is currently [position] at [institution]. Research interests include few-shot learning, computer vision, and intelligent information systems.

[Author Name 2] received the [degree] in [field] from [institution]. [He/She/They] is currently [position] at [institution]. Research interests include deep learning and industrial applications of artificial intelligence.

REFERENCES

1. K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016. doi:10.1109/CVPR.2016.90.
2. Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. “Generalizing from a Few Examples: A Survey on Few-Shot Learning”, ACM Computing Surveys, 53(3), Article 63, 2020. doi:10.1145/3386252.
3. C. Zhang, Y. Cai, G. Lin, and C. Shen. “DeepEMD: Few-Shot Image Classification with Differentiable Earth Mover’s Distance”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12203–12213, 2020.
4. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. “ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542, 2020.
5. J. Hu, L. Shen, and G. Sun. “Squeeze-and-Excitation Networks”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, 2018.
6. O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. “Matching Networks for One Shot Learning”. In Advances in Neural Information Processing Systems, 29, pp. 3630–3638, 2016.
7. J. Snell, K. Swersky, and R. S. Zemel. “Prototypical Networks for Few-shot Learning”. In Advances in Neural Information Processing Systems, 30, pp. 4077–4087, 2017.
8. 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, pp. 1199–1208, 2018.
9. C. Finn, P. Abbeel, and S. Levine. “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”. In Proceedings of the 34th International Conference on Machine Learning, pp. 1126–1135, 2017.
10. 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, pp. 266–282, 2020. doi:10.1007/978-3-030-58568-6_16.
11. L. Tian, P. Zhou, Z. Wang, and R. Wang. “Prototypes-Oriented Transductive Few-Shot Learning with Conditional Transport”. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16317–16326, 2023.
12. H. Zhu and P. Koniusz. “Transductive Few-Shot Learning with Prototype-Based Label Propagation by Iterative Graph Refinement”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23996–24006, 2023.

13. S. Martin, Y. Fu, A. K. S. D. S. Kumar, and L. Van Gool. “Transductive Zero-Shot and Few-Shot CLIP”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28816–28826, 2024.
14. J. R. Cumplido, J. M. Buenaposada, and L. Baumela. “Slot Attention-Based Feature Filtering for Few-Shot Learning”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2025.
15. M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel. “Meta-Learning for Semi-Supervised Few-Shot Classification”. In International Conference on Learning Representations, 2018.
16. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “ImageNet: A Large-Scale Hierarchical Image Database”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009. doi:10.1109/CVPR.2009.5206848.
17. A. Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”, Technical Report, University of Toronto, Toronto, 2009.
18. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626, 2017.
19. B. N. Oreshkin, P. Rodríguez, and A. Lacoste. “TADAM: Task Dependent Adaptive Metric for Improved Few-Shot Learning”. In Advances in Neural Information Processing Systems, 31, pp. 721–731, 2018.
20. K. Lee, S. Maji, A. Ravichandran, and S. Soatto. “Meta-Learning with Differentiable Convex Optimization”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10657–10665, 2019. doi:10.1109/CVPR.2019.01091.
21. S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord. “Boosting Few-Shot Visual Learning with Self-Supervision”. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8059–8068, 2019. doi:10.1109/ICCV.2019.00815.
22. D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In International Conference on Learning Representations, 2015.
23. N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. “A Simple Neural Attentive Meta-Learner”. In International Conference on Learning Representations, 2018.
24. Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele. “Meta-Transfer Learning for Few-Shot Learning”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 403–412, 2019.
25. H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha. “Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8808–8817, 2020.
26. M. A. Jamal and G.-J. Qi. “Task-Agnostic Meta-Learning for Few-Shot Learning”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11719–11727, 2019.
27. A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”, Communications of the ACM, 60(6), pp. 84–90, 2017. doi:10.1145/3065386.
28. S. Belongie, J. Malik, and J. Puzicha. “Shape Matching and Object Recognition Using Shape Contexts”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(4), pp. 509–522, 2002. doi:10.1109/34.993558.