International Journal of Innovative Information Systems & Technology Research 14(3):401-412, July-Sept, 2026

© SEAHI PUBLICATIONS, 2026 www.seahipublications.org **ISSN: 2467-8562**

DOI:10.5281/zenodo.22934307

A Lightweight Task-Adaptive Prototypical Network with MMD Feature Alignment for Cross-Domain Few-Shot Image Classification

[1]E. B. Owabie *[2]O. E. Bennett *[3]D. J. S. Sako

[1,2&3]Department of Computer Science, Rivers State University, Port-Harcourt, Nigeria
Email:* [1]*emmanuella.owabie@ust.edu.ng*, [2]*bennett.okoni@ust.edu.ng

ABSTRACT
Few-shot image classification aims to recognise novel classes from very limited labelled examples, but standard prototypical networks rely on fixed Euclidean distance and often generalise poorly when source and target visual domains differ. This study developed a lightweight enhanced prototypical-network framework for cross-domain few-shot image classification by integrating a ResNet-12 embedding backbone, prototype-based classification, task-adaptive channel-weighted distance, and training-time Maximum Mean Discrepancy (MMD) feature alignment. The metric module generated episode-specific channel weights from prototype statistics, allowing the classifier to adjust feature importance without replacing the efficient nearest-prototype inference path. The MMD component aligned source and target feature distributions during training and was excluded from inference. Python implementation was carried out with PyTorch and evaluated under a five-way one-shot episodic protocol with 15 query images per class. miniImageNet was used as the source and within-domain evaluation dataset, while CUB-200-2011 and Stanford Cars were used as fine-grained target-domain datasets with class-disjoint adaptation and evaluation partitions. Four configurations were compared: Baseline ProtoNet, +Learned Metric, +Domain Adaptation, and Full Framework. Each configuration completed 1000 training episodes and 150 evaluation episodes per dataset. +Learned Metric achieved the highest miniImageNet_test mean accuracy of 0.2831, while Full Framework achieved the highest target-domain means of 0.2932 on CUB-200-2011 and 0.2667 on Stanford Cars. The confidence intervals overlapped across configurations, so the observed differences are interpreted as modest rather than decisive. Full Framework recorded 8.21% inference overhead relative to Baseline ProtoNet, remaining below the 25% efficiency target. The study contributes a reproducible CPU-based implementation and cautious empirical evaluation of task-adaptive metric weighting with training-time feature alignment for cross-domain few-shot image classification.
Keywords: Few-shot learning; prototypical networks; cross-domain image classification; metric learning; task-adaptive distance.

I. INTRODUCTION

Deep learning image classifiers usually require large labelled datasets to achieve strong performance. This requirement limits their usefulness in domains where annotation is expensive, rare, or time-sensitive, such as medical imaging, industrial inspection, rare species identification, remote sensing, and specialised product recognition. [1] describe few-shot learning as a response to this practical data-scarcity problem,

covering its evolution, application areas, and persistent challenges. [2] place the same problem within the broader meta-learning literature, where models are trained to acquire transferable learning strategies rather than memorising a fixed set of classes. In image classification, the aim is therefore not only to fit known categories, but to recognise novel categories from one or a few labelled examples.

Metric-based few-shot learning has become one of the most influential approaches to this problem. Prototypical networks classify query images by embedding support and query examples into a shared feature space, computing one prototype per class from the support examples, and assigning each query to the nearest prototype [3]. This design is simple, sample-efficient, and computationally attractive because it avoids expensive task-specific optimisation during inference. However, standard prototypical networks commonly use fixed Euclidean distance, which treats every feature dimension as equally important across all classification episodes.

The fixed-distance assumption is limiting because few-shot tasks differ in their discriminative cues. In one episode, colour or texture features may separate classes; in another, shape or fine-grained part structure may matter more. [4] addressed this limitation by learning a relation function for comparing support and query features, showing that similarity itself can be treated as a learnable component. [5] later showed, through TADAM, that task-conditioned metric adaptation can improve few-shot classification by allowing the model to respond to the structure of the current episode. These studies establish the academic basis for adaptive similarity computation. Their practical implication is that a useful classifier should adapt its distance behaviour while avoiding excessive support-query pairwise computation.

Domain shift creates a second major challenge. A model trained on a source visual domain may perform weakly when evaluated on target domains with different visual statistics. [6] showed that conclusions drawn from within-domain few-shot benchmarks can weaken under more careful transfer evaluation. [7] broadened this concern by studying cross-domain few-shot learning across several target domains, demonstrating that domain gap is not a minor implementation issue but a central evaluation problem. Fine-grained domains such as birds and vehicles are particularly difficult because class differences may be subtle and visually localised. [8] further argued that prototypical networks may exploit source-domain shortcuts that fail under visual shift, while [9] identified cross-domain generalisation as an open challenge across modern few-shot visual recognition.

Against this background, the present work develops and evaluates a lightweight enhanced prototypical network for cross-domain few-shot image classification. The model combines task-adaptive channel-weighted prototype distance with training-time Maximum Mean Discrepancy feature alignment. The task-adaptive metric generates feature-channel weights from prototype statistics in each episode, while MMD encourages source and target embeddings to occupy more aligned feature distributions during training. The method preserves the standard prototype-based inference path, and the MMD branch is not used during prediction.

The paper is therefore positioned as an implementation and evaluation study rather than as a leaderboard comparison. Its contribution is a modular enhanced ProtoNet model, validated component behaviour, an ablation comparison across baseline, metric-only, adaptation-only, and full configurations, and an efficiency analysis showing that the full model remains within a modest inference-overhead target.

II. Related Works

Few-shot learning methods are commonly grouped into optimisation-based, metric-based, memory-based, and hybrid approaches. [2] provide this broad taxonomy and show that meta-learning methods differ mainly in what they transfer across tasks: parameters, similarity functions, memory structures, or optimisation behaviour. Optimisation-based methods such as Model-Agnostic Meta-Learning learn initial parameters that can adapt rapidly to new tasks using a small number of gradient updates [10]. Metric-based methods instead learn embedding spaces where examples from the same class are close and examples from different classes are separated. [12] demonstrated this principle early through Siamese

networks for one-shot image recognition. [11] extended metric-based learning through matching networks, where attention over labelled examples supports one-shot prediction. [4] introduced relation networks, replacing fixed distance with a learned comparison module. [3] simplified the metric-learning pipeline through prototypical networks, where each class is represented by the mean of its support embeddings.

Prototypical networks remain a strong and widely used metric-learning baseline because of their conceptual simplicity and efficient inference. In each episode, class prototypes are computed as the mean embeddings of support examples, and query examples are classified by distance to these prototypes [3]. The approach works well when the embedding function maps semantically similar images close together. However, its standard Euclidean distance function assumes equal feature importance and does not adapt to the current task.

Learned distance metrics address this limitation by allowing the similarity function to adapt. Relation networks learn a neural comparison function between support and query features, replacing fixed distance with a trainable relation score [4]. Task-adaptive methods such as TADAM condition the metric or embedding space on the support set, allowing the model to emphasise task-relevant feature dimensions [5]. These studies support the academic basis for replacing fixed metric behaviour with adaptive similarity computation. The remaining practical challenge is to achieve adaptation without large pairwise-comparison overhead.

Cross-domain few-shot learning extends the problem by evaluating models on target domains that differ from the source domain. [6] showed that domain shift can cause substantial degradation in few-shot classification, even when methods perform well on standard within-domain benchmarks. [7] broadened cross-domain few-shot evaluation and showed that no single method dominates under all domain shifts. [8] further argued that prototypical networks can suffer from simplicity bias, where features useful in the source domain fail to transfer to target domains.

Domain adaptation provides one route for addressing this issue. Feature-level alignment methods encourage source and target data to share more similar internal representations. Maximum Mean Discrepancy is a statistical distribution-discrepancy measure that compares source and target embeddings in a reproducing kernel Hilbert space [13]. MMD-based losses have been used in deep domain adaptation because they can be added to a training objective without requiring adversarial optimisation.

Recent cross-domain few-shot studies have moved beyond fixed-distance prototype classification in several directions. [8] revisited prototypical networks under domain shift and emphasised the role of feature bias in cross-domain failure. [9] surveyed cross-domain few-shot visual recognition and identified task adaptation, feature alignment, and domain-aware evaluation as recurring themes. [14] proposed adaptive parametric prototype learning, shifting attention from fixed prototype averages to learned prototype construction. [15] approached the problem through cross-level distillation and feature denoising, aiming to improve target-domain representation quality. [16] used adaptive transformer networks to learn transferable features between source and target domains. These methods show that cross-domain few-shot learning is an active research direction, but they also illustrate a trade-off: stronger adaptation often increases model complexity, requires extra target information, or changes the inference procedure substantially. The gap addressed in this paper is therefore specific: combining task-adaptive distance weighting and training-time MMD alignment within the efficient prototypical-network pipeline.

III. Methodology

Model Architecture

The architecture builds on the standard prototypical-network pipeline. Input images are passed through a ResNet-12 embedding backbone to produce 512-dimensional feature vectors. Support embeddings are averaged by class to form five class prototypes. Query embeddings are then compared with these

prototypes using either fixed Euclidean distance or the proposed task-adaptive weighted distance, depending on the configuration under evaluation.

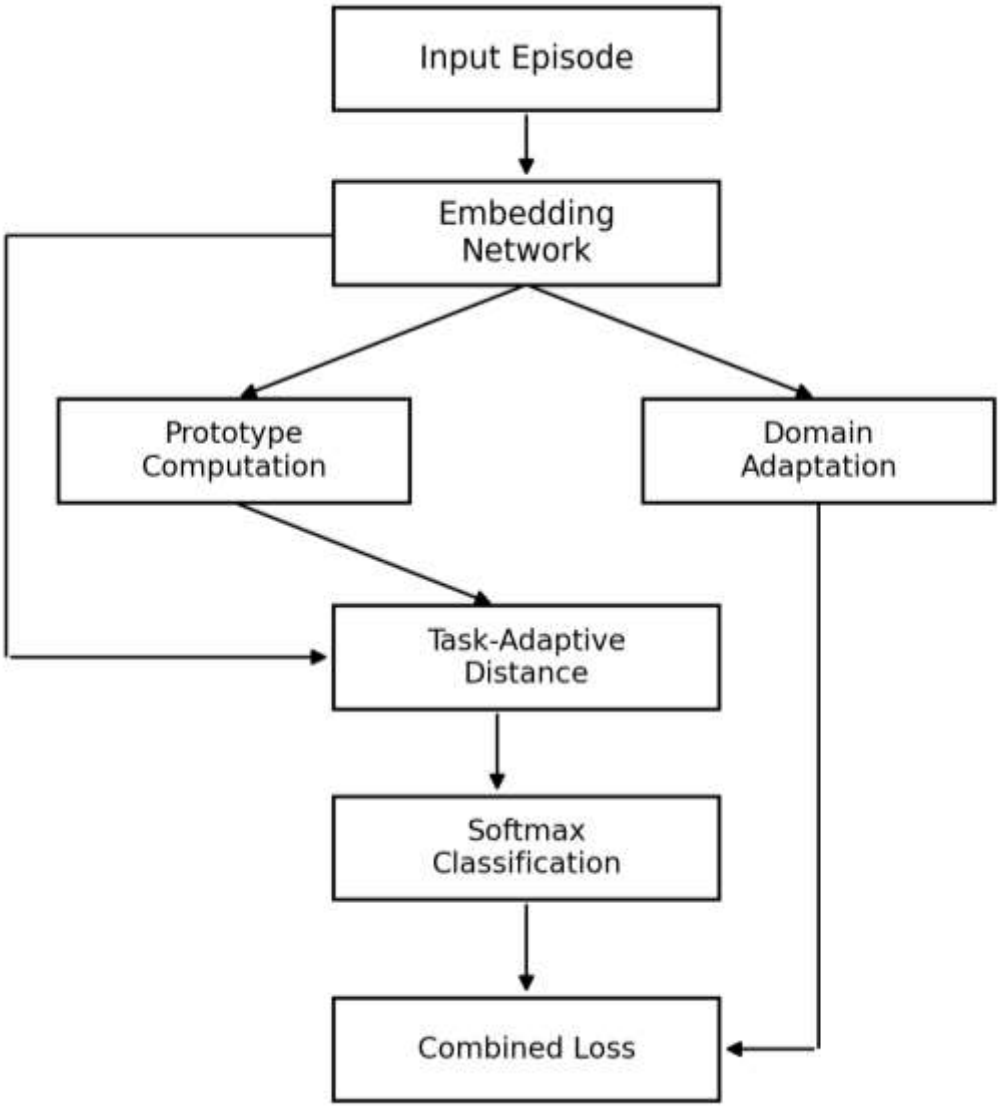

Figure 5: Architecture of the enhanced prototypical network

The Full Framework contains two enhancements. The first is the task-adaptive distance metric module. It computes prototype statistics for the current episode, generates a task descriptor by combining the mean and standard deviation of the prototypes, and passes this descriptor through a lightweight bottleneck network to produce 512 channel weights. These weights scale the squared distance between each query embedding and each prototype, allowing distance computation to change from one episode to another.

Model Formulation

The mathematical formulation defines how the developed model computes prototypes, derives task-level metric weights, classifies query images, and applies feature alignment during training. Let $f_\phi(x)$ denote the E

the proto

$$p_c = \frac{1}{|S_c|} \sum_{x_i \in S_c} f_\phi(x_i) \tag{1}$$

For the task-adaptive metric, the prototype mean and standard deviation are concatenated into a task descriptor. This task-conditioning step follows the principle used in task-adaptive metric learning, where support-set information is used to adjust metric behaviour for the current episode [5]:

$$t = [mean(p_1, \ldots, p_N) \parallel std(p_1, \ldots, p_N)] \tag{2}$$

A bottleneck neural network maps this task descriptor to channel weights. In this work, the task-conditioning idea is implemented as a lightweight channel-weighting module:

$$\alpha = g_\psi(t) \tag{3}$$

The weighted distance between a query embedding q and class prototype p_c extends the nearest-prototype distance principle of [3] by applying task-conditioned channel weights:

$$d(q, p_c) = \sum_i \alpha_i (q_i - p_{c,i})^2 \tag{4}$$

Classification is performed through softmax over negative distances. For configurations using domain adaptation, the training objective adds an MMD alignment term. MMD was introduced by [13] as a kernel two-sample test and is used here as a feature-distribution alignment loss:

$$L_{total} = L_{cls} + \lambda L_{MMD} \tag{5}$$

where L_{cls} is cross-entropy loss and L_{MMD} measures source-target feature-distribution discrepancy.

Datasets and Preprocessing

Three image datasets were used. miniImageNet served as the source training and within-domain evaluation dataset, following its common use in few-shot image classification studies. CUB-200-2011 served as a fine-grained bird-recognition target domain and provides 200 bird categories with visually subtle inter-class differences [17]. Stanford Cars served as a fine-grained vehicle-recognition target domain and is widely used for evaluating models on visually similar object categories [18]. The target datasets were divided into class-disjoint adaptation and evaluation partitions so that target-domain evaluation did not use the same classes available during adaptation.

Figure 6: Visual samples from miniImageNet, CUB-200-2011, and Stanford Cars

Table 1: Dataset Catalogue

Dataset	Domain	Role	Images	Classes
miniImageNet	General objects	Source and within-domain evaluation	50,400	84
CUB-200-2011	Fine-grained birds	Cross-domain target	11,796	200
Stanford Cars	Fine-grained vehicles	Cross-domain target	16,185	196

All images were resized to 84 x 84 pixels. Training augmentation included random crop, horizontal flip, and colour jitter. The evaluation protocol used five-way one-shot episodes with 75 query images per episode.

Experimental Configuration

Four configurations were evaluated to isolate the effect of each component.

Table 2: Experimental Configurations

Configuration	Description	Purpose
Baseline ProtoNet	ResNet-12, prototypes, fixed Euclidean distance	Standard baseline
+Learned Metric	Baseline with task-adaptive channel-weighted distance	Tests adaptive metric effect
+Domain Adaptation	Baseline with MMD training alignment	Tests alignment effect
Full Framework	Learned metric and MMD alignment combined	Tests integrated framework

Each configuration completed 1000 training episodes. Evaluation used 150 episodes per dataset and configuration. Accuracy was reported as mean episodic classification accuracy with 95% confidence interval half-width. Inference efficiency was measured as time per episode and overhead relative to the baseline. Only publicly available academic image datasets were used, no human subjects were recruited, and no personally identifiable information was collected.

IV. RESULTS & DISCUSSION

The implementation successfully indexed miniImageNet, CUB-200-2011, and Stanford Cars. The target-domain datasets were split into adaptation and evaluation partitions with disjoint classes. CUB-200-2011 used 100 adaptation classes and 100 evaluation classes, while Stanford Cars used 98 adaptation classes and 98 evaluation classes.

Automated validation checks confirmed that the prototype tensor had shape 5 x 512, the task descriptor had length 1024, the metric-weight vector had length 512, and the distance matrix had shape 75 x 5. MMD returned near-zero values for identical embeddings and positive values for shifted embeddings. The lambda schedule reached the configured maximum value of 0.1, and the inference path did not call the domain adapter. These checks are important because the numerical comparison is meaningful only if the metric branch, adaptation branch, and inference path behave as designed.

Classification Accuracy

Table 3 presents the five-way one-shot evaluation results across the four configurations and three evaluation datasets.

Table 3: Five-Way One-Shot Evaluation Results

Configuration	Evaluation Dataset	Episodes	Mean Accuracy	95% CI Half-Width
Baseline ProtoNet	miniImageNet_test	150	0.2731	0.0106
Baseline ProtoNet	CUB-200-2011	150	0.2898	0.0125
Baseline ProtoNet	Stanford Cars	150	0.2656	0.0132
+Learned Metric	miniImageNet_test	150	0.2831	0.0117
+Learned Metric	CUB-200-2011	150	0.2858	0.0126
+Learned Metric	Stanford Cars	150	0.2652	0.0131
+Domain Adaptation	miniImageNet_test	150	0.2816	0.0115
+Domain Adaptation	CUB-200-2011	150	0.2922	0.0112
+Domain Adaptation	Stanford Cars	150	0.2660	0.0122
Full Framework	miniImageNet_test	150	0.2812	0.0116
Full Framework	CUB-200-2011	150	0.2932	0.0120
Full Framework	Stanford Cars	150	0.2667	0.0129

All configurations performed above the 20% random-choice level expected for five-way classification. The +Learned Metric configuration produced the highest miniImageNet_test mean accuracy of 0.2831. The Full Framework produced the highest CUB-200-2011 mean accuracy of 0.2932 and the highest Stanford Cars mean accuracy of 0.2667.

The accuracy differences should be interpreted cautiously. The confidence intervals overlap within each dataset, so the experiment does not establish statistically clear superiority of any enhanced configuration over the baseline. The evidence supports a measured interpretation: the enhanced components were implemented, validated, and associated with modest observed improvements in selected settings under a consistent evaluation protocol.

Fig. 6. Few-shot classification accuracy across source and target domains

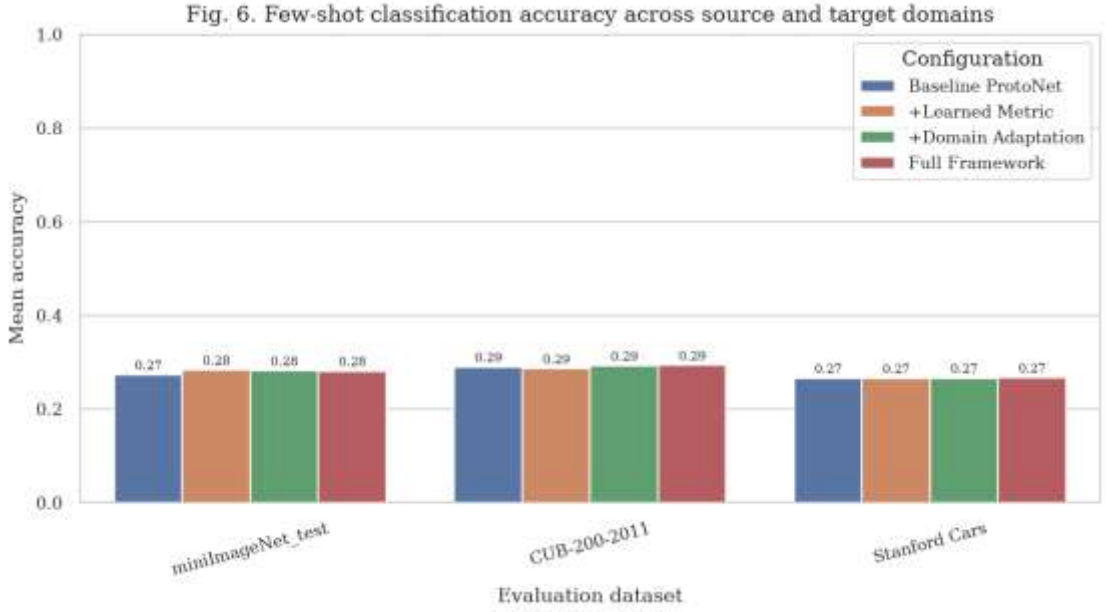

Figure 7: Classification accuracy across evaluation domains

Cross-Domain Retention

Cross-domain retention was computed by dividing target-domain accuracy by the corresponding miniImageNet_test accuracy for each configuration.

Table 4: Cross-Domain Retention Results

Configuration	Target Domain	Source Accuracy	Target Accuracy	Retention Ratio
Baseline ProtoNet	CUB-200-2011	0.2731	0.2898	1.0612
Baseline ProtoNet	Stanford Cars	0.2731	0.2656	0.9727
+Learned Metric	CUB-200-2011	0.2831	0.2858	1.0094
+Learned Metric	Stanford Cars	0.2831	0.2652	0.9369
+Domain Adaptation	CUB-200-2011	0.2816	0.2922	1.0376
+Domain Adaptation	Stanford Cars	0.2816	0.2660	0.9448
Full Framework	CUB-200-2011	0.2812	0.2932	1.0430
Full Framework	Stanford Cars	0.2812	0.2667	0.9485

CUB-200-2011 retention exceeded 1.0 for all configurations because CUB mean accuracy was higher than each configuration's own miniImageNet_test mean. Stanford Cars retention remained below 1.0 for all configurations. These ratios are useful for describing source-target balance, but they should not be treated alone as evidence of model superiority. The retention results must be read together with the absolute accuracies and confidence intervals.

Fig. 7. Cross-domain retention from miniImageNet to target domains

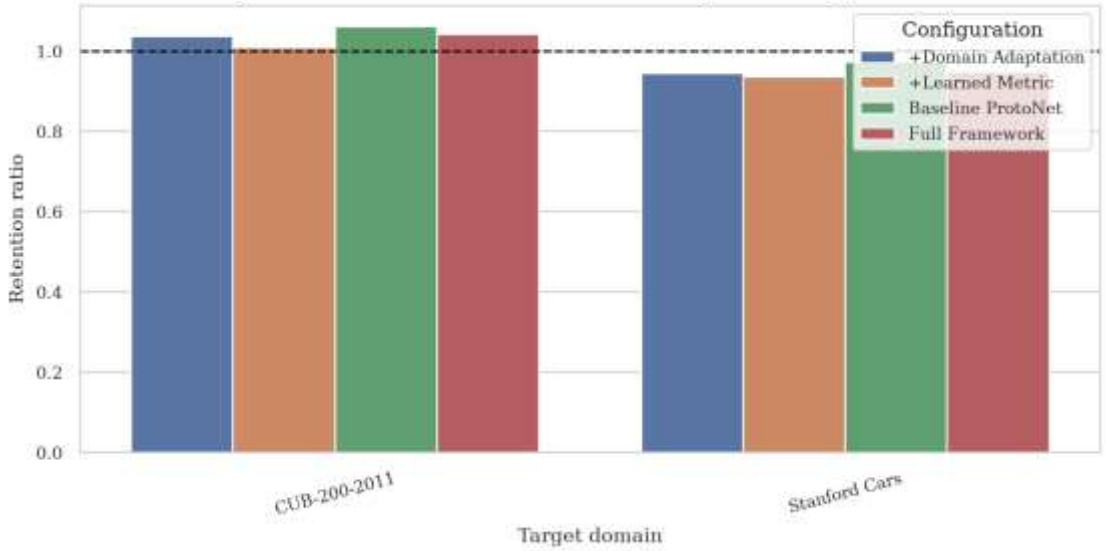

Figure 8: Cross-domain retention from miniImageNet to target domains

Inference Efficiency

The model was designed to preserve efficient prototype-based inference. Table 5 reports measured inference time and overhead.

Table 5: Inference Time and Computational Overhead

Configuration	Time per Episode (ms)	Time per Episode (s)	Overhead Ratio	Overhead (%)
Baseline ProtoNet	2409.50	2.41	1.0000	0.00
+Learned Metric	2364.43	2.36	0.9813	-1.87
+Domain Adaptation	2570.84	2.57	1.0670	6.70
Full Framework	2607.38	2.61	1.0821	8.21

The Full Framework recorded 8.21% inference overhead relative to the Baseline ProtoNet, which remained below the predefined 25% target. Since MMD was not used during inference, the additional prediction cost came mainly from the trained inference path and task-adaptive metric computation. This supports the efficiency objective of the framework.

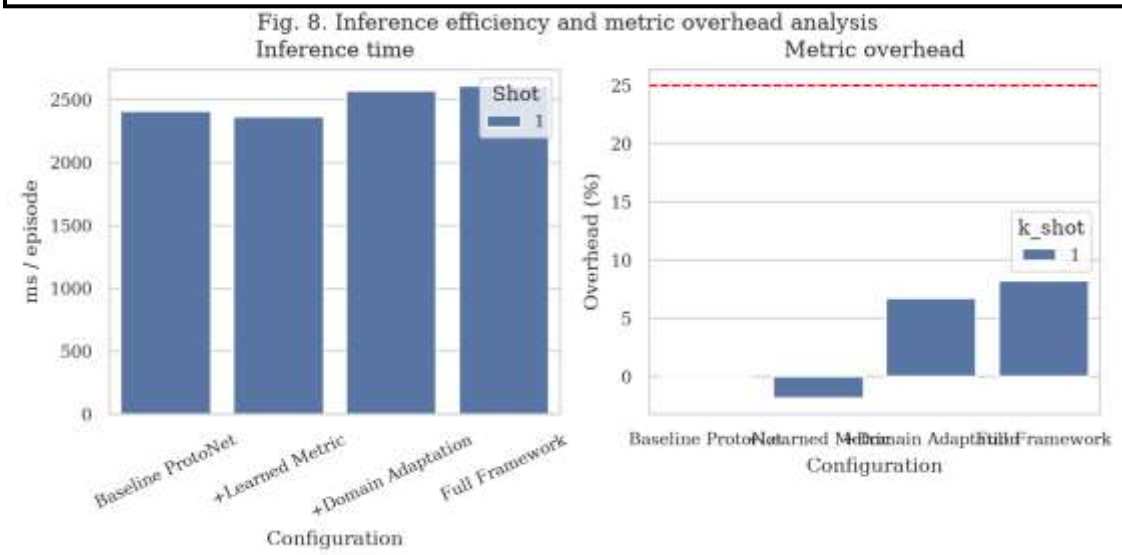

Figure 9: Inference time and baseline-relative overhead

Task-Adaptive Metric Behaviour

The metric module generated 512 channel weights from episode prototype statistics. The saved channel-weight distribution had mean 1.0000, standard deviation 0.6085, minimum 0.0154, median 1.3123, and maximum 1.5704. This confirms that the model did not apply one uniform weighting pattern across all feature dimensions. Instead, it produced non-uniform task-conditioned channel weights, supporting the intended adaptive-distance design.

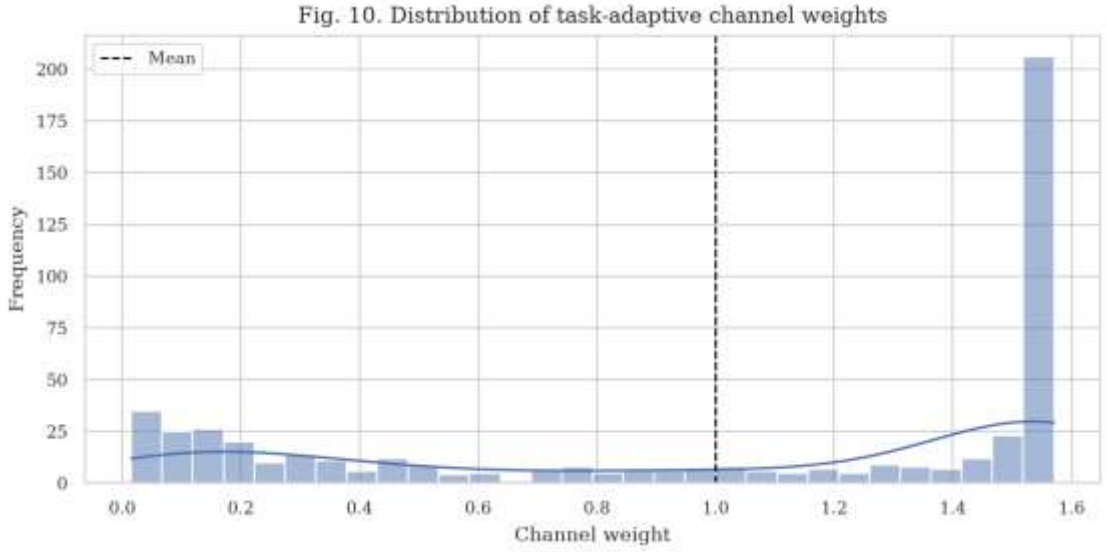

Figure 10: Distribution of task-adaptive channel weights

Taken together, the reported result sets support an implementation-level interpretation. The completed experiment provides evidence for component validity, controlled ablation comparison, cross-domain evaluation, and low inference overhead. The accuracy evidence remains limited by the overlap in confidence intervals, so the model is best discussed as a reproducible enhanced ProtoNet evaluation rather than as a benchmark-leading classifier.

V. CONCLUSION

This paper presented a lightweight enhanced prototypical network for cross-domain few-shot image classification. The model combined ResNet-12 embeddings, prototype computation, task-adaptive channel-weighted distance, and training-time MMD feature alignment while retaining the efficient nearest-prototype inference path.

Under the five-way one-shot protocol, all configurations performed above the random-choice level. +Learned Metric achieved the highest miniImageNet_test mean accuracy of 0.2831, while Full Framework produced the highest target-domain means of 0.2932 on CUB-200-2011 and 0.2667 on Stanford Cars. The confidence intervals overlapped across configurations, so these results are best interpreted as modest observed gains rather than statistically clear superiority over the baseline.

The main value of the model is therefore its balance between adaptation and efficiency. Full Framework added only 8.21% inference overhead relative to Baseline ProtoNet because MMD was restricted to training and the metric module used channel weighting rather than support-query pairwise relation computation. This makes the approach suitable as a reproducible ablation baseline for studying efficient cross-domain enhancements to prototypical networks.

The findings remain limited by the CPU-based training budget, one-shot-only evaluation, and overlapping confidence intervals. Further evaluation with longer GPU training, multiple seeds, paired episode testing, five-shot experiments, stronger pretrained backbones, and additional target domains would be needed before making broader claims about cross-domain few-shot performance.

REFERENCES

[1] Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2023). A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys, 56*(2), 40.

[2] Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5149-5169.

[3] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems, 30*, 4077-4087.

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

[5] Oreshkin, B. N., Rodriguez, P., & Lacoste, A. (2018). TADAM: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems, 31*, 721-731.

[6] Chen, W. Y., Liu, Y. C., Kira, Z., Wang, Y. C., & Huang, J. B. (2019). A closer look at few-shot classification. In *Proceedings of the International Conference on Learning Representations*.

[7] Guo, Y., Codella, N. C., Karlinsky, L., Codella, J. V., Smith, J. R., Saenko, K., Rosing, T., & Feris, R. (2020). A broader study of cross-domain few-shot learning. In *Proceedings of the European Conference on Computer Vision*, 124-141.

[8] Zhou, F., Wang, P., Zhang, L., Wei, W., & Zhang, Y. (2023). Revisiting prototypical network for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20061-20070.

[9] Zhang, C., Cai, Y., Lin, G., & Shen, C. (2024). Deep learning for cross-domain few-shot visual recognition: A survey. *ACM Computing Surveys, 57*(2), 45.

[10] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, 70*, 1126-1135.

[11] Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*.

[12] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one-shot learning. In *Advances in Neural Information Processing Systems, 29*, 3630-3638.

[13] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research, 13*, 723-773.

[14] Heidari, M., Alchihabi, A., En, Q., & Guo, Y. (2023). Adaptive parametric prototype learning for cross-domain few-shot classification. *arXiv*, 2309.01342.

[15] Zheng, H., Wang, R., Liu, J., & Kanezaki, A. (2023). Cross-level distillation and feature denoising for cross-domain few-shot classification. *arXiv*, 2311.02392.

[16] Paeedeh, M., Pratama, M., Ma'sum, M. A., Mayer, W., Cao, Z., & Kowalczyk, R. (2024). Cross-domain few-shot learning via adaptive transformer networks. *Knowledge-Based Systems, 288*, 111458.

[17] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 dataset* (Technical Report CNS-TR-2011-001). California Institute of Technology.

[18] Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554-561.