INFORMATION
Volume 29, Number 1, pp.131-137
ISSN 1343-4500 eISSN 1344-8994


# Task-Conditioned Logit Fusion for Fine-Grained Few-Shot Classification

Derui Zheng*, Jiacheng Li**, Masato Noto**

* *Field of Electrical, Electronics and Information Engineering, Graduate School of Engineering, Kanagawa University, Yokohama, Japan*
E-mail: r202570166bj@jindai.jp

** *Department of Applied Systems and Mathematics, Kanagawa University, Yokohama, Japan*
E-mail: {lijiacheng, noto}@kanagawa-u.ac.jp

**Abstract**

Few-shot image classification remains challenging due to the scarcity of training samples and the fine-grained similarity between classes. Conventional metric-based methods, such as Prototypical Networks, rely on a single distance measure, which limits their ability to capture both scale- and angle-based discriminative cues. To address this, we propose Task-Conditioned Logit Fusion (TCLF), an extension to ProtoNet that jointly integrates Euclidean and cosine metrics through task-adaptive weighting. This approach enables interpretable, task-dependent reliability estimation for each metric branch. Experiments on *mini*-ImageNet 5-way 1-shot classification show that TCLF improves accuracy by approximately 3% over the ProtoNet baseline while maintaining stability across seeds.
**Keywords:** Few-shot learning, Metric fusion, Prototype networks, Task conditioning, Fine-grained recognition

## 1 Introduction

With the rapid development of the Internet of Things (IoT) and intelligent sensing systems, large numbers of smart devices have been deployed in critical domains such as industrial inspection and ecological observation, continuously generating real-time data. Although the overall data volume is massive, real-world data distributions are often highly imbalanced: most samples correspond to normal or frequent categories, while target classes, anomalies, or rare events are extremely scarce and expensive to annotate. For example, rare bird species (e.g., *Ciconia boyciana*) exhibit fine-grained characteristics with only a few available images, despite high interspecies similarity. This coexistence of "extremely few samples + fine-grained differentiation" makes few-shot learning (FSL) a crucial research paradigm [1].

To enable effective learning under limited supervision, FSL typically relies on meta-learning [2], acquiring fast adaptation ability through cross-task episodic training. Existing FSL methods can generally be grouped into three categories [1]: (1) gradient-based methods, (2) metric-based methods [3], and

(3) generative and augmentation-based methods. Among them, metric-based approaches have become widely adopted in fine-grained and distribution-shift scenarios due to their simplicity, stability, and strong generalization ability. Within the metric-learning paradigm, existing advances can be characterized by three mechanisms: (i) metric space enhancement [3,4,5,6]; (ii) prototype aggregation and structural modeling [4,5]; and (iii) feature-space adaptation and task-aware mechanisms [7,8,9,10].

Many existing methods operate within a single metric space (e.g., Euclidean or cosine), limiting their ability to simultaneously capture scale-based and angle-based discriminative cues. Although multi-metric designs have been explored, for example in BSNet [4], BMPN [5], and ASM [8], their fusion strategies are typically manually preset or globally learned, and thus fail to adapt to the geometric structure of each episode. This limitation leads to degraded generalization performance in fine-grained recognition and under task distribution shifts [7,11]. For instance, TADAM remains restricted to Euclidean space [7], and existing multi-metric models still struggle under complex task variation [9]. Recent analyses further confirm that episodic structure and task diversity significantly influence FSL stability [12,13].

To overcome these limitations, we propose the Task-Conditioned Logit Fusion (TCLF) method. TCLF extends ProtoNet by jointly leveraging Euclidean distance and cosine similarity, using a task-encoding network that automatically generates fusion weights from support-set statistics. This allows the model to dynamically select appropriate geometric cues for each episode.

The contributions of this work are threefold. First, we introduce a task-aware fusion framework that adaptively integrates Euclidean distance and cosine similarity in the logit space, extending prior metric-fusion approaches. Second, we design a lightweight task-encoding module that generates stable and interpretable per-episode fusion weights. Third, experiments on *mini*-ImageNet demonstrate consistent accuracy improvements in the 1-shot setting.

# 2 Methods

## 2.1 Prototypical Networks

Prototypical Networks [3] classify a query by comparing it with class prototypes in the embedding space. For class $k$, the prototype is defined as the mean of its support embeddings:

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\theta(x_i). \tag{1}$$

For a query sample $x$, its distance to each prototype is calculated using squared Euclidean distance, and the negative distance serves as the logit value:

$$d_k^{(E)} = \|f_\theta(x) - c_k\|_2^2, \quad l_k^{(E)} = -d_k^{(E)}. \tag{2}$$

The predicted probability of class membership is obtained through a softmax function:

$$p(y = k \mid x) = \frac{\exp(l_k^{(E)})}{\sum_j \exp(l_j^{(E)})}. \tag{3}$$

## 2.2 Multi-Metric Fusion

Prototypical Networks leverage Euclidean distance to measure similarity, but relying solely on a single distance metric may limit discriminative ability, especially in fine-grained recognition tasks where subtle angular differences between features can be informative. To enhance representational flexibility, we incorporate cosine similarity as an additional metric and perform fusion in the logit space, enabling complementary geometric cues to jointly influence classification.

Given the cosine similarity between a query sample $x$ and prototype $c_k$, the cosine-based score is computed as

$$l_k^{(C)} = s_k^{(C)} = \frac{f_\theta(x)^\top c_k}{\|f_\theta(x)\| \, \|c_k\|}. \tag{4}$$

The final classification logit for class $k$ is obtained by linearly combining Euclidean- and cosine-based logits:

$$l_k = \alpha \, l_k^{(E)} + \beta \, l_k^{(C)}, \tag{5}$$

where $\alpha$ and $\beta$ control the relative influence of each metric. When $\alpha = 1, \beta = 0$, the framework reduces to standard ProtoNet; when $\alpha = 0, \beta = 1$, classification relies solely on angular similarity.

## 2.3 Task-Conditioned Logit Fusion

In FSL, intra-class compactness, inter-class separability, and class-wise sample distribution can vary substantially across episodes. Using fixed metric weights for all tasks risks overfitting on simple episodes and underfitting on difficult ones. To address this issue, TCLF dynamically adjusts the relative importance of Euclidean and cosine metrics according to the geometric structure of each episode.

To characterize task difficulty, four statistics are computed from the support features and the resulting class prototypes: intra-class dispersion, inter-class separation, sample imbalance, and prototype-level mean cosine similarity. These descriptors jointly summarize the geometric configuration and distribution properties of the episode, enabling the model to select more appropriate metric weights.

Let $S_k$ denote the support set of class $k$, and $W$ denote the number of classes in the episode. The four statistics are defined as

$$\text{intra} = \frac{1}{W} \sum_{k=1}^{W} \frac{1}{|S_k|} \sum_{x_i \in S_k} \|f_\theta(x_i) - c_k\|_2^2, \tag{6}$$

$$\text{inter} = \frac{2}{W(W-1)} \sum_{i<j} \|c_i - c_j\|_2^2, \tag{7}$$

$$\text{imb} = \frac{\text{std}(|S_k|)}{\text{mean}(|S_k|)}, \tag{8}$$

$$C_{\text{mean}} = \frac{1}{W(W-1)} \sum_{i \neq j} \frac{c_i^\top c_j}{\|c_i\| \, \|c_j\|}. \tag{9}$$

Here, intra and inter capture compactness and separability, imb measures class imbalance, and $C_{\text{mean}}$ reflects the global directional consistency of prototypes.

The combined vector $t = [\text{intra}, \text{inter}, \text{imb}, C_{\text{mean}}]$ serves as the task descriptor input to a small task encoder $g_\phi$, implemented as a two-layer multilayer perceptron (MLP), which outputs task-dependent metric weights as

$$\alpha, \beta = g_\phi(t). \tag{10}$$

In the feasibility version, the fusion weights $\alpha$ and $\beta$ are generated via a softplus activation, ensuring non-negativity ($\alpha, \beta \geq 0$). While this unconstrained formulation allows metric branches to be flexibly scaled, it may lead to unstable magnitudes across episodes. To maintain consistency in the logit space, we further normalize the weights:

$$\tilde{\alpha} = \frac{\alpha}{\alpha + \beta}, \quad \tilde{\beta} = \frac{\beta}{\alpha + \beta}, \tag{11}$$

enforcing $\tilde{\alpha} + \tilde{\beta} = 1$. The final fused logit can then be written as

$$l_k = \tilde{\alpha}\, l_k^{(E)} + \tilde{\beta}\, l_k^{(C)}. \tag{12}$$

## 2.4 Product-of-Experts Interpretation

The fusion mechanism of TCLF can be theoretically interpreted through the Product-of-Experts (PoE) framework. In this view, the Euclidean and cosine branches act as two independent discriminative experts, each modeling similarity from a distinct geometric perspective.

Let $p_E(y = k \mid x)$ and $p_C(y = k \mid x)$ denote their respective posterior probabilities. Assuming conditional independence, the joint posterior is

$$p(y = k \mid x) \propto \left[p_E(y = k \mid x)\right]^{\alpha} \left[p_C(y = k \mid x)\right]^{\beta}. \tag{13}$$

Taking logarithms yields:

$$\log p(y = k \mid x) = \alpha \log p_E(y = k \mid x) + \beta \log p_C(y = k \mid x) + \text{const}. \tag{14}$$

Here, const denotes the class-independent softmax normalization term $\log \sum_j \exp(l_j)$, which does not affect energy comparisons.

Since the softmax probability satisfies $-\log p(y = k \mid x) = -l_k + \text{const}$, the negative logit is equivalent to an energy function up to a constant shift. This allows us to directly treat metric logits as energies in the PoE formulation. From an energy-based perspective, each metric defines an energy $E_m(y \mid x) = -l_k^{(m)}$, and the joint posterior can be written as

$$p(y = k \mid x) \propto \exp\left(-\alpha E_E - \beta E_C\right) \leftrightarrow p(y = k \mid x) \propto \exp\left(\alpha l_k^{(E)} + \beta l_k^{(C)}\right). \tag{15}$$

Under this interpretation, $\alpha$ and $\beta$ act as precision-like coefficients that encode the task-dependent reliability of each metric. This heteroscedastic formulation extends ProtoNet's isotropic assumption and provides a probabilistic explanation for TCLF's adaptive fusion of Euclidean (scale-based) and cosine (direction-based) cues. Furthermore, the PoE view clarifies that TCLF is not a heuristic weighted sum, but a principled probabilistic fusion of multiple metric experts.

# 3 Results and Discussion

## 3.1 Experimental Setup

Experiments are conducted on the standard *mini*-ImageNet benchmark, which contains 100 classes with 600 images each, split into 64, 16, and 20 categories for meta-training, meta-validation, and meta-testing, respectively. All images are resized to $84 \times 84$. For a fair comparison with Prototypical Networks, we adopt the commonly used Conv4 backbone, consisting of four convolutional blocks (64-filter Conv–BN–ReLU–MaxPool) without dropout. The network is trained from scratch using episodic meta-learning on the meta-train split, following standard FSL practice.

Training follows the standard episodic protocol: 5-way classification with 1-shot or 5-shot support sets, 15 query samples per class, and 30-way and 20-way training episodes for the 1-shot and 5-shot settings, respectively. For each model, 600 test episodes are sampled, and results are reported as mean accuracy with 95% confidence intervals across multiple seeds.

Optimization uses Adam with an initial learning rate of $1 \times 10^{-3}$, episodic learning rate scheduling, and validation-based early stopping. This configuration matches the ProtoNet implementation, ensuring reproducibility and a fair comparison between methods.

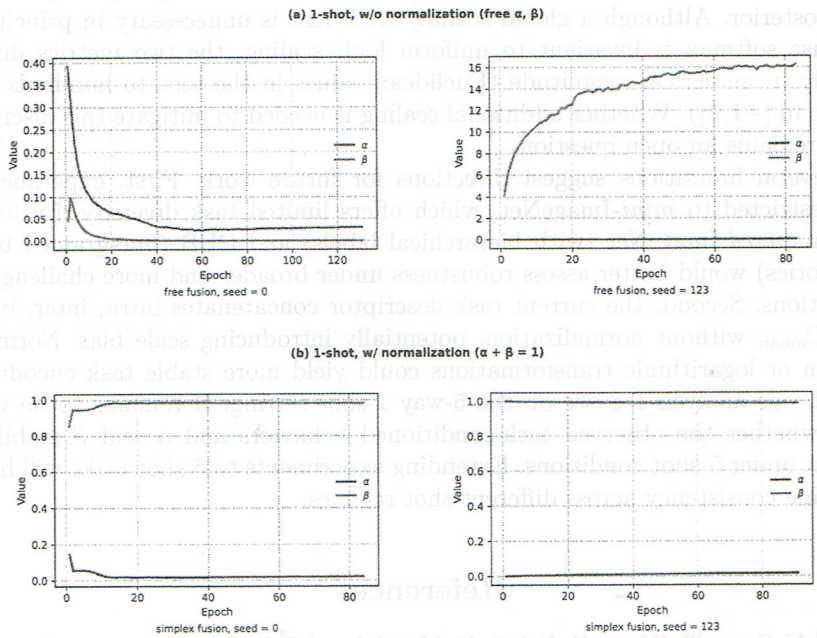


Figure 1: Evolution of the task-adaptive metric weights $\alpha$ and $\beta$ during training under different settings: (a) Unconstrained fusion (free $\alpha, \beta$) with representative seeds; (b) normalized fusion with $\alpha + \beta = 1$.

### 3.2 Quantitative Results

To evaluate task-conditioned metric fusion, we compare three variants on *mini*-ImageNet: ProtoNet (Clean), TCLF (non-norm., $\alpha, \beta \geq 0$), and TCLF (norm., $\alpha + \beta = 1$). The non-normalized variant achieves an average accuracy of 46.10% (about 3% above ProtoNet) but exhibits strong seed-dependent behavior. Across different seeds, $\alpha$ and $\beta$ fall into three regimes—cosine-dominant, Euclidean-dominant, or both small—indicating that the weights behave like unconstrained temperature factors rather than task-level confidences (Fig. 1 (a)). This instability leads to large performance variance.

In contrast, the normalized TCLF enforces a simplex constraint that stabilizes training. Across all seeds, $\alpha$ and $\beta$ converge to consistent values ($\beta \approx 0.98$, $\alpha \approx 0.02$), yielding interpretable and stable metric preferences without accuracy loss. TCLF (norm.) achieves 46.16% (+3% over ProtoNet) and avoids the unstable weight behavior observed in the unconstrained version (Fig. 1 (b)). These results show that task-conditioned metric fusion improves few-shot generalization, and that normalization is essential for obtaining stable and interpretable metric weights.

### 3.3 Discussion

From a theoretical perspective, the Euclidean and cosine experts in the PoE framework are treated as independent components whose logits jointly define the posterior. Although a global scaling coefficient is unnecessary in principle, because softmax is invariant to uniform logit scaling, the two metrics differ greatly in numerical magnitude (Euclidean values in the tens to hundreds vs. cosine in $[-1, 1]$). Whether additional scaling is needed to mitigate this discrepancy remains an open question.

Several limitations suggest directions for future work. First, experiments are restricted to *mini*-ImageNet, which offers limited task diversity. Evaluating on *tiered*-ImageNet (with hierarchical labels) and CUB (fine-grained bird categories) would better assess robustness under broader and more challenging conditions. Second, the current task descriptor concatenates intra, inter, imb, and $C_{\text{mean}}$ without normalization, potentially introducing scale bias. Normalization or logarithmic transformations could yield more stable task encoding. Third, our analysis focuses on the 5-way 1-shot setting. It remains to be verified whether the observed task-conditioned behaviors and $\alpha$ and $\beta$ stability persist under 5-shot conditions. Extending experiments to 5-shot tasks will help examine consistency across different shot regimes.

## References

[1] Y. Song, T. Wang, P. Cai, S. K. Mondal and J. P. Sahoo, A comprehensive survey of few-shot learning: evolution, applications, challenges, and opportunities, *ACM Computing Surveys*, 55(13s), 2023, 1–40.

[2] Y. Chen, Z. Liu, H. Xu, T. Darrell and X. Wang, Meta-baseline: exploring simple meta-learning for few-shot learning, *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, 9062–9071.

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

4077–4087.

[4] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao and J.-H. Xue, BSNet: bi-similarity network for few-shot fine-grained image classification, *IEEE Transactions on Image Processing*, 30(2020), 1318–1331.

[5] W. Fu, L. Zhou and J. Chen, Bidirectional matching prototypical network for few-shot image classification, *IEEE Signal Processing Letters*, 29(2022), 982–986.

[6] J. Lai, S. Yang, G. Jiang, X. Wang, Y. Li, Z. Jia, X. Chen, J. Liu, B.-B. Gao, W. Zhang, Y. Xie and C. Wang, Rethinking the metric in few-shot learning: from an adaptive multi-distance perspective, *Proc. ACM International Conference on Multimedia (ACM MM)*, 2022, 4021–4030.

[7] B. Oreshkin, P. Rodríguez López and A. Lacoste, TADAM: task dependent adaptive metric for improved few-shot learning, *Advances in Neural Information Processing Systems*, 31(2018), 721–731.

[8] H. Li, L. Li, Y. Huang, N. Li and Y. Zhang, An adaptive plug-and-play network for few-shot learning, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5.

[9] Y. Guo, R. Du, Y. Dong, T. Hospedales, Y. Z. Song and Z. Ma, Task-aware adaptive learning for cross-domain few-shot learning, *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, 1590–1599.

[10] X. He, F. Li and L. Liu, Task-adaptive relation dependent network for few-shot learning, *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2021, 1–8.

[11] S. Laenen and L. Bertinetto, On episodes, prototypical networks, and few-shot learning, *Advances in Neural Information Processing Systems*, 34(2021), 24581–24592.

[12] S. Sun and H. Gao, Meta-AdaM: a meta-learned adaptive optimizer with momentum for few-shot learning, *Advances in Neural Information Processing Systems*, 36(2023), 65441–65455.

[13] K. Topollai and A. Choromanska, Task-level contrastiveness for cross-domain few-shot learning, *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, 6489–6499.