Efficient Few-Shot Image Classification with Unbalanced Sinkhorn Distance for Robust Feature Alignment and Large-Scale Applications

Yun Pang and Hayati A. Rahman *

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia
Email: yunpang.guet.edu@gmail.com (Y.P.); hayatiar@tmsk.uitm.edu.my (H.A.R.)
*Corresponding author

***Abstract*—The core challenge of few-shot image classification lies in efficiently learning and accurately classifying with limited labeled data. Current metric-based meta-learning methods primarily rely on computing the structural distance between the query and support sets for matching. However, traditional metric learning typically employs Euclidean distance or standard Sinkhorn Distance (SD) for feature matching, assuming that the total mass of the source and target distributions is equal. This assumption overlooks the imbalance in data distribution in real-world tasks. To overcome these challenges, this study systematically introduces Unbalanced Sinkhorn Distance (USD) into Few-Shot Learning (FSL) to enhance the model's ability to match features map of query set and support set under imbalanced distributions. USD allows dynamic adjustment of matching distributions during metric computation. Moreover, this method effectively reduces the interference of background noise when matching features between the support and query sets while maintaining low computational costs. Experimental results demonstrate that our method achieves state-of-the-art performance on four FSL benchmark datasets, significantly outperforming existing Optimal Transport (OT)-based methods.**

***Keywords*—few-shot learning, Unbalanced Sinkhorn Distance, optimal transmission theory, meta-learning**

I. Introduction

In recent years, deep learning has achieved remarkable success in computer-vision tasks such as image classification, detection, and segmentation. Nevertheless, these achievements rely heavily on large-scale labeled datasets, which are costly and time-consuming to create. In practical scenarios, only a handful of annotated examples are often available per class, leading to the Few-Shot Learning (FSL) problem.

To address this data-scarcity challenge, researchers have proposed a variety of FSL methods, notably meta-learning and metric-learning approaches. Metric-based models aim to learn a transferable embedding space in which similarities between support and query samples can be measured reliably. Although techniques such as Prototypical Networks [1] perform well under balanced and clean conditions, they depend on simple Euclidean or cosine distances that struggle to capture complex spatial relations and are vulnerable to feature misalignment.

Optimal Transport (OT) theory has recently been introduced to FSL to achieve finer structural alignment. For example, Deep Earth Mover's Distance (DeepEMD) employs the Earth Mover's Distance (EMD) to compute soft correspondences between feature maps, delivering significant gains over Euclidean-based metrics [2]. However, DeepEMD and similar balanced-OT methods assume that all transport mass must be preserved—a restriction that hampers robustness when class imbalance, background clutter, or missing data are present.

To overcome this limitation, we propose a new FSL framework built on Unbalanced Sinkhorn Distance (USD). By incorporating Kullback-Leible (KL) -divergence penalties into the Sinkhorn iterations, USD relaxes the mass-conservation constraint and allows partial matching between support and query distributions, following the unbalanced-OT formulation of Chizat *et al.* [3]. This design enables dynamic mass adjustment and yields more reliable alignment under noisy or imbalanced conditions.

To validate our algorithm, we conducted extensive experiments across multiple datasets to demonstrate its effectiveness. The results showcase the robustness and accuracy of our approach. Our main contributions are summarized as follows:

(1) Proposing a metric learning framework based on USD. This work introduces USD with KL-divergence regularization, allowing dynamic adjustment of matching distributions. This improves feature robustness, reduces background noise interference, and ensures efficient classification even in imbalanced data scenarios.
(2) Optimizing feature matching and improving computational efficiency. Our method relaxes the strict mass conservation constraint and integrates Sinkhorn entropy regularization with KL-divergence optimization, reducing computational cost while maintaining classification accuracy.

Manuscript received March 11, 2025; revised April 25, 2025; accepted June 17, 2025; published October 17, 2025.

(3) Achieving state-of-the-art performance on multiple benchmark datasets. Our method outperforms existing state-of-the-art approaches in 1-shot and 5-shot tasks on Mini-ImageNet, Tiered-ImageNet, FC100, and CUB datasets.

II. Related Works

Early studies in FSL cast the problem as meta-learning: a model is trained on many small "episodes" so that it can adapt rapidly to novel classes. The seminal Siamese Network [4] and the cognitive one-shot learner of Lake *et al.* [5] showed that pair-wise similarity and episodic supervision permit generalisation from very few labelled samples. Subsequent work sharpened the optimization perspective. MetaOptNet formulates the inner loop as a differentiable convex programme [6], while Meta-Baseline demonstrates that a simple cosine classifier atop a fixed backbone can rival more sophisticated meta-learners [7]. Despite their efficiency, these approaches depend on global Euclidean or cosine distances and therefore remain vulnerable when support and query distributions are misaligned or imbalanced.

To compensate for prototype bias and sparse supervision, a second strand focuses on feature refinement. Category-Traversal Networks search for task-relevant channels across classes [8], whereas Manifold Mixup interpolates hidden representations to smooth decision boundaries [9]. TADAM [10] introduces task-dependent metric scaling, and Prototype Rectification (PR) employs label propagation to correct biased prototypes in transductive settings [11]. Transfer Based FSL (TB-FSL) further leverages feature-distribution statistics to calibrate decision boundaries under covariate shift [12]. Nonetheless, these methods still rely on heuristic global pooling and cannot explicitly align fine-grained structures when parts of the query image lie outside the support manifold.

OT metrics provide a principled way to minimise the cost of aligning two feature distributions. DeepEMD [2] incorporates Earth-Mover's Distance to establish pixel-level correspondences, whereas the Bilaterally Normalised Sinkhorn Distance (BSSD) accelerates entropic OT via scale-consistent matrix balancing [13]. Brownian Distance Covariance OT captures higher-order dependence beyond pairwise similarity [14, 15]. A common shortcoming of these balanced OT methods is the mass-conservation constraint, which forces all query mass to be matched to the support set; background noise, occluded regions, and severe class imbalance therefore lead to over-alignment and degraded robustness.

Unbalanced OT theory relaxes this rigidity by penalizing, rather than forbidding, mass variation. Chizat *et al.* [3] derive a scalable KL-regularized formulation, and Chapel *et al.* [16] extend partial transport to positive–unlabeled learning. These advances suggest that allowing unmatched mass to be discarded is key to robust alignment. Veilleux *et al.* [17] further argue that standard, class-balanced benchmarks over-estimate FSL performance and introduce a Dirichlet query-imbalance protocol to emulate realistic deployments. In safety-critical domains, compact Convolution Neural Networks (CNNs) have been applied to multi-class skin-lesion triage with limited annotations [18], and transfer-learning pipelines detect knee-joint synovial fluid from Magnetic Resonance Imaging (MRI) scans under severe data scarcity [19]; both studies underline the demand for interpretable, distribution-aware matching.

Motivated by the brittleness of balanced OT in DeepEMD [2] and BSSD [13], we adopt the unbalanced-OT principle [3, 16] and introduce an USD for few-shot classification. USD preserves Sinkhorn's efficiency yet permits partial-mass transport, enabling the model to ignore unmatched or noisy regions. As our experiments show, this optimization flexibility translates into superior accuracy and more coherent transport plans under class imbalance, occlusion, and background clutter. As illustrated in Fig. 1, the iterative process of USD alignment in feature space is visualized between the support and query sets. As the number of iterations increases (from $it = 0$ to $it = 200$), the initially disordered matching evolves into a structurally coherent alignment, demonstrating the effectiveness of USD in modeling fine-grained correspondences between distributions with unequal mass.

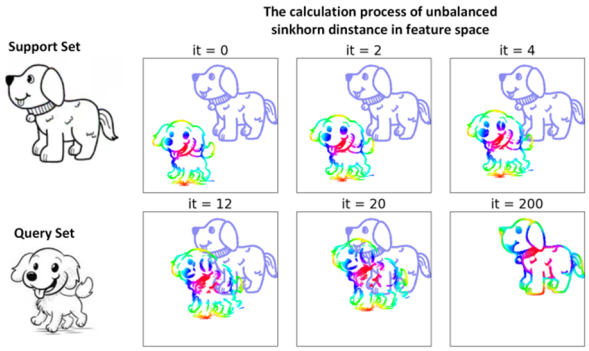

Fig. 1. This image illustrates the process of calculating the USD in feature space and demonstrates how the USD iteratively adjusts feature point matching to optimize the structural distance between the query set and the support set.

III. Proposed Method

In this section, we first derive the formula transition from Earth Mover's Distance (EMD) to USD. Starting with the EMD, we incorporate an entropic regularization term to facilitate computational efficiency. This leads to the Sinkhorn Distance, an efficient approximation of EMD. To handle distributions with unequal total mass, we introduce a penalty term for mass differences, resulting in the USD. This unbalanced variant maintains the benefits of entropic regularization while addressing real-world data imbalances. Subsequently, we integrate the optimal transport theory based on USD into a few-shot learning model. This integration allows for improved feature matching and classification performance.

A. Earth Mover Distance Theory

This section provides a comprehensive derivation of EMD, Sinkhorn Distance (SD), and USD to clarify the mathematical evolution behind each step.

EMD, also known as Wasserstein Distance, measures the optimal transport cost between two probability distributions. Given two discrete distributions:

Source distribution: $a = (a_1, a_2, \ldots, a_m)$ over the set $X = \{x_1, x_2, \ldots, x_m\}$, satisfying.

$$\sum_{i=1}^{m} a_i = 1 \tag{1}$$

Target distribution: $b = (b_1, b_2, \ldots, b_n)$ over the set $Y = \{y_1, y_2, \ldots, y_n\}$, satisfying.

$$\sum_{j=1}^{n} b_j = 1 \tag{2}$$

Given the transport matrix T, where $T_{i,j}$ represents the amount of mass transported from x_i to y_j, the EMD optimization problem is formulated as:

$$\min_{T \in R^{m \times n}} \sum_{i=1}^{m} \sum_{j=1}^{n} C_{i,j} T_{i,j} \tag{3}$$

where: $C_{i,j}$ is the cost matrix, representing the cost of moving mass from x_i to y_j. T satisfies the marginal constraints: $T1 = a$, $T^T = b$, $T \geq 0$.

EMD solves the OT problem, but its major drawback is its high computational complexity of $O(n^3)$, making it impractical for large-scale problems.

B. Sinkhorn Distance (SD)

To accelerate EMD computation, SD introduces an entropy regularization term, making the optimization problem solvable using the efficient Sinkhorn-Knopp algorithm. The regularized optimal transport problem is given by:

$$\min_{T \in R^{m \times n}} \sum_{i=1}^{m} \sum_{j=1}^{n} C_{i,j} T_{i,j} + \lambda \sum_{i=1}^{m} \sum_{j=1}^{n} T_{i,j} \log T_{i,j} \tag{4}$$

where the first term represents the optimal transport cost. The second term is the entropy regularization term, which smooths the transport matrix T, improving computational efficiency.

The regularization parameter λ controls the smoothness. When $\lambda \to 0$, SD reduces to EMD.

Solution method: Using Lagrange dual optimization, we introduce dual variables u and v, allowing us to express T as:

$$T_{i,j} = \exp(-\frac{C_{i,j}}{\lambda}) u_i v_j \tag{5}$$

Defining the kernel matrix:

$$K_{i,j} = \exp(-\frac{C_{i,j}}{\lambda}) \tag{6}$$

We can rewrite T as:

$$T = \mathrm{diag}(u) K \mathrm{diag}(v) \tag{7}$$

Iterative updates:

$$u^{(t+1)} = \frac{a}{(Kv^t)}, v^{(t+1)} = \frac{b}{K^T u^{(t+1)}} \tag{8}$$

Computational complexity is reduced to $O(n^2)$ making SD significantly faster than EMD.

C. Unbalanced Sinkhorn Distance (USD)

SD assumes that the total mass of source and target distributions is equal:

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j = 1 \tag{9}$$

However, in many real-world applications different distributions may have unequal total mass. To address this, USD introduces Kullback-Leibler (KL)-divergence regularization, allowing mass adjustment during the matching process.

USD Optimization Problem:

$$\min_{T \in R^{m \times n}} \sum_{i=1}^{m} \sum_{j=1}^{n} C_{i,j} T_{i,j} + \lambda \sum_{i=1}^{m} \sum_{j=1}^{n} T_{i,j} \log T_{i,j} + \tau KL(T1 \| a) + \tau KL(T^T 1 \| b) \tag{10}$$

The additional KL-divergence terms:

$$KL(T1 \| a) = \sum_i T_i \log(\frac{T_i}{a_i}) - (T_i - a_i) \tag{11}$$

Allow mass adjustments rather than enforcing strict equality constraints. The regularization parameter τ controls the smoothness. When $\tau \to \infty$, USD reduces to SD.

Solution method: USD uses Sinkhorn-Knopp-like iterative updates, but now KL-divergence must be considered:

$$u^{(t+1)} = \frac{a}{(Kv^{(t)})^{\tau/(\lambda+\tau)}} \tag{12}$$

$$v^{(t+1)} = \frac{a}{(K^T u^{(t+1)})^{\tau/(\lambda+\tau)}} \tag{13}$$

D. Unbalanced Sinkhorn Distance for Few-Shot Learning

In a FSL scenario based on meta-learning, the training samples are divided into a Support Set $\mathcal{S} = \{\mathcal{S}_i\}_{i=1}^{k}$ and a Query Set $\mathcal{Q} = \{\mathcal{Q}_j\}_{j=1}^{m}$. Each sample $\mathcal{S}_i \in \mathcal{S}$ and $\mathcal{Q}_j \in \mathcal{Q}$ is passed through the feature extractor network, ResNet12, resulting in feature maps $\mathcal{F}_{\mathcal{S}_i}$ and $\mathcal{F}_{\mathcal{Q}_j}$ for the Support Set and Query Set, respectively. Specifically, $\mathcal{F}_{\mathcal{S}_i} \in \mathbb{R}^{5\times640\times25}$ and $\mathcal{F}_{\mathcal{Q}_j} \in \mathbb{R}^{75\times640\times25}$. The feature

maps are used to calculate the cost matrix M_{ij} using the cosine similarity, given by $M_{ij} = 1 - \frac{\mathcal{F}_{S_i} \cdot \mathcal{F}_{Q_j}}{\|\mathcal{F}_{S_i}\| \|\mathcal{F}_{Q_j}\|}$. This cost matrix is then used to calculate the USD.

IV. Experiments

To assess the effectiveness of our proposed algorithm in few-shot classification, we begin by introducing the benchmark datasets and outlining the implementation details. Next, we conduct an ablation study to analyze the contribution of each component and present qualitative results for deeper insights. Finally, we compare our approach against state-of-the-art methods on widely used benchmark datasets and evaluate its performance in cross-domain experiments.

A. Dataset Description

The experiments are conducted on four widely used few-shot learning datasets: Mini-ImageNet, Tiered-ImageNet, FC100, and CUB. Mini-ImageNet contains 100 classes with 600 images per class, split into 64 for training, 16 for validation, and 20 for testing. Tiered-ImageNet includes 608 classes, structured to enhance domain differences, with 20 super-classes for training, 6 for validation, and 8 for testing. FC100, derived from CIFAR100, consists of 36 super-classes divided into 60 training, 20 validation, and 20 testing classes. CUB, originally for fine-grained bird classification, has 200 classes and 11,788 images, split into 100 for training, 50 for validation, and 50 for testing. The dataset splits ensure a structured evaluation of few-shot learning models, as shown in Fig. 2.

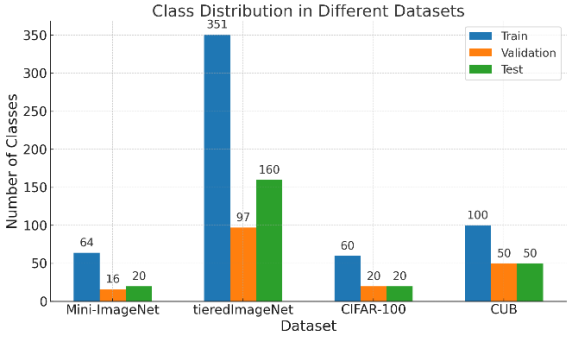

Fig. 2. This bar chart illustrates the class distribution across different datasets used for training, validation, and testing in few-shot learning tasks. The datasets include Mini-ImageNet, Tiered-ImageNet, CIFAR-100, and CUB.

B. Implementation Details

Following standard practice in metric-based few-shot learning [1, 2], we apply uniform training strategies across all evaluated methods. Data augmentation includes random cropping, horizontal flipping, and color jittering. We use cosine learning rate scheduling and train all models under an episodic meta-learning framework with identical pretraining settings to ensure fair comparison. All experiments were conducted on an Ubuntu 22.04 operating system with an NVIDIA RTX 4090 GPU, implemented using the PyTorch framework. All methods were evaluated under the same hardware configuration without any specialized acceleration or optimization. The model maintains high computational efficiency during both training and inference.

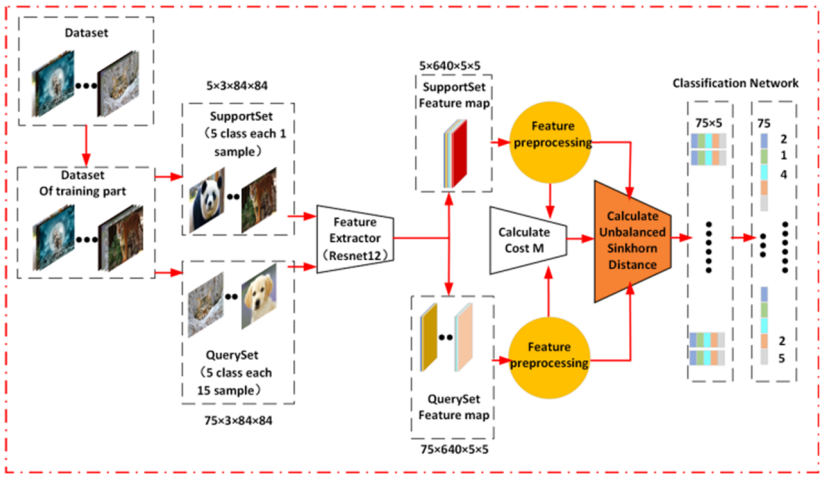

Fig. 3. Few-shot learning framework based on Unbalanced Sinkhorn Distance.

Fig. 3 illustrates the training process for FSL using USD. Initially, the dataset is divided into a support set and a query set. The support set consists of 5 classes with 1 sample each class, and the query set consists of 5 classes with 15 samples each, both having the input size of 5×3×84×84 for support and 75×3×84×84 for query. Features are extracted from both sets using the ResNet12 feature extractor, producing feature maps with sizes 5×640×25 for the Support Set and 75×640×25 for the Query Set. After feature preprocessing, a cost matrix (Cost M) is computed between the support and query feature maps. USD is used to calculate the similarly between each query sample and the support samples, resulting in a 75×5 distance matrix. This process demonstrates how USD is effectively utilized for feature matching and classification in few-shot learning, enhancing training efficiency and accuracy.

C. Analysis

Fig. 4 illustrates the classification accuracy trends of

different methods on the Mini-ImageNet dataset as the number of shots varies.

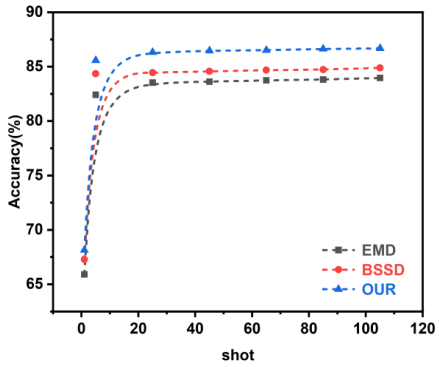

Fig. 4. This graph compares the accuracy of different few-shot learning methods across varying numbers of shots (from 1 to 105).

In this experiment, all other conditions were kept constant, and only the shot number was changed to observe the classification performance of different methods under varying data availability. Overall, the accuracy of all methods increases with the number of shots and stabilizes after a certain point, indicating that having more training samples improves classification performance, but the benefit diminishes at higher shot numbers. Compared to other methods, OUR method consistently achieves the highest accuracy across all shot settings, with a particularly significant advantage in low-shot scenarios. This demonstrates its superior feature alignment capability in few-shot learning tasks, allowing it to handle imbalanced data more effectively. The BSSD method improves upon EMD in low-shot cases, but as the shot number increases, the performance gap between them narrows, suggesting that BSSD mainly enhances feature matching in low-data scenarios. In contrast, EMD stabilizes in high-shot tasks but maintains lower overall accuracy. In summary, OUR outperforms other methods across different data scales, proving its effectiveness and robustness in few-shot learning tasks.

1）Analysis of Unbalanced Sinkhorn Distance (USD) parameters under distributional imbalance

The blur (λ) and scaling (τ) in USD not only affect numerical stability and convergence but also help address distributional mismatch between the support and query sets in FSL.

In our context, distributional imbalance refers to situations where query samples fall outside the support feature distribution due to intra-class variation, occlusion, or background differences. Traditional OT methods like EMD and SD require full mass transport, which can lead to incorrect matches under such mismatch.

USD alleviates this problem by allowing partial matching. The scaling (τ) controls how much unmatched mass is tolerated—larger values enable more flexibility in ignoring poorly aligned query features. The blur (λ) controls the smoothness of transport updates, helping stabilize the matching when the distributions are misaligned.

Together, these parameters make USD more robust in few-shot scenarios where full alignment between support and query is not always possible. Fig. 5 shows the performance of USD under different blur and scaling parameters, confirming that proper tuning enhances robustness in the presence of support-query distribution mismatch—exactly the imbalance our method aims to handle.

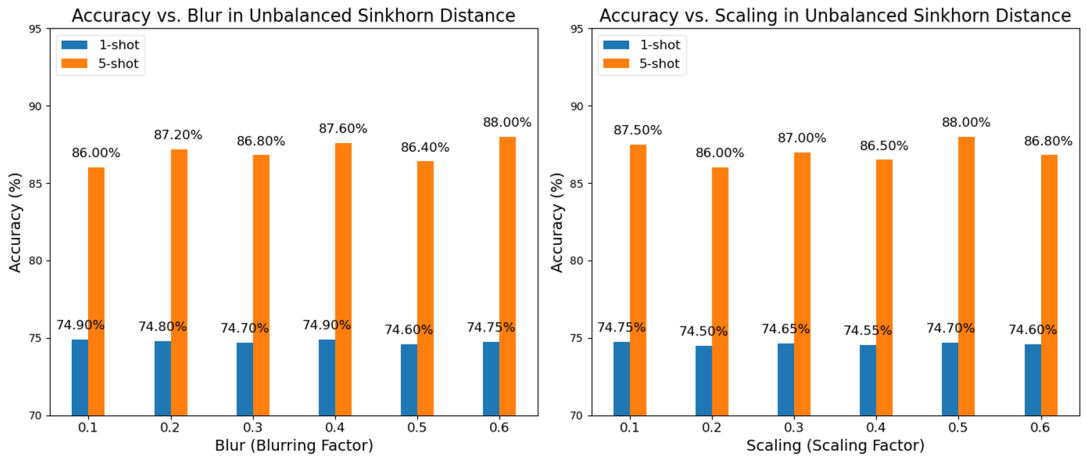

Fig. 5. These graphs illustrate the impact of varying blurring and scaling factors on the accuracy of USD for both 1-shot and 5-shot learning tasks.

2）Comparison performance with different backbone networks

ResNet is a type of artificial neural network introduced by Kaiming He *et al*. [20] in 2015 to address the vanishing gradient problem. This architecture significantly improves performance and has become a fundamental building block in many state-of-the-art deep learning models. To study the impact of different ResNet architectures on model performance in the context of few-shot learning, ResNet12, ResNet18, and ResNet34s were evaluated using the Tiered-ImageNet dataset.

By keeping all other testing conditions constant only changing the backbone network, the study aimed to isolate the effects of network depth and complexity on few-shot learning performance. ResNet12 is expected to offer faster training and lower computational cost but might sacrifice some accuracy due to its limited capacity. ResNet18 represents a balanced approach, potentially offering a good

trade-off between computational efficiency and accuracy. ResNet34s, the deepest network among the three, is anticipated to capture more complex features and provide higher accuracy, especially beneficial for 5-shot learning where more information is available. The results indicate that despite the increasing number of parameters from ResNet12 to ResNet34s, the accuracy improvement is minimal. The experimental results are shown in Table I.

TABLE I. THIS TABLE COMPARES THE PERFORMANCE OF THREE DIFFERENT RESNET ARCHITECTURES (RESNET12, RESNET18, AND RESNET34S) ON 1-SHOT AND 5-SHOT LEARNING TASKS IN TIERED-IMAGENET

Backbone	Parameter	1-shot (%)	5-shot (%)
ResNet12	1,249,200	75.58±0.27	88.21±0.36
ResNet18	23,728,320	75.73±0.16	88.37±0.25
ResNet34s	33,412,800	75.86±0.42	88.41±0.51

3）Various unbalanced optimization transmission algorithms and transmission matrix visualization

Fig. 6 compares four Unbalanced Optimal-Transport (UOT) strategies—Partial-OT [16], L2-UOT [21], KL-UOT [16], and Entropic-KL-UOT [3]—applied to the same support–query feature maps under Euclidean and cosine costs. Although all methods share the identical cost matrix, they differ in the way they regularize the row and column sums, and this choice dictates how much mass can be discarded or smoothed during transport. Partial-OT imposes a hard upper bound on the transported mass (m = 0.7), yielding an extremely sparse plan that effectively ignores background clutter but occasionally drops useful query mass, which explains the lower recall observed in Table II. Replacing the hard cap with an L2 penalty produces denser couplings that retain more query information; however, the quadratic term penalizes large local deviations too strongly, so mismatched regions are still partially over-aligned. KL-UOT switches to a logarithmic divergence, encouraging relative mass preservation; visually its transport matrix becomes more structured, and class accuracy improves correspondingly. Our USD adds an entropic term on top of the KL-divergence, combining smooth assignment with the freedom to discard mismatched mass. As shown in Fig. 6(a), using Euclidean distance, the Entropic KL-UOT method produces clearer, semantically aligned transport paths and a well-structured transport matrix. Fig. 6(b) demonstrates similar results with cosine distance, where Entropic KL-UOT still maintains coherent matching and compact transport patterns. These results highlight the robustness and effectiveness of our USD approach under different distance settings.

TABLE II. THIS TABLE COMPARES THE PERFORMANCE OF DIFFERENT REGULARIZED UOT METHODS USING COSINE AND EUCLIDEAN DISTANCES AS COST FUNCTIONS

Different UOT methods	Cost(cosine) (%)	Cost (Euclidean) (%)
Partial UOT	73.56±0.52	67.24±0.34
L2-UOT	73.27±0.17	67.38±0.53
KL-UOT	73.15±0.35	67.28±0.17
Entropic KL-UOT	75.58±0.38	67.49±0.26

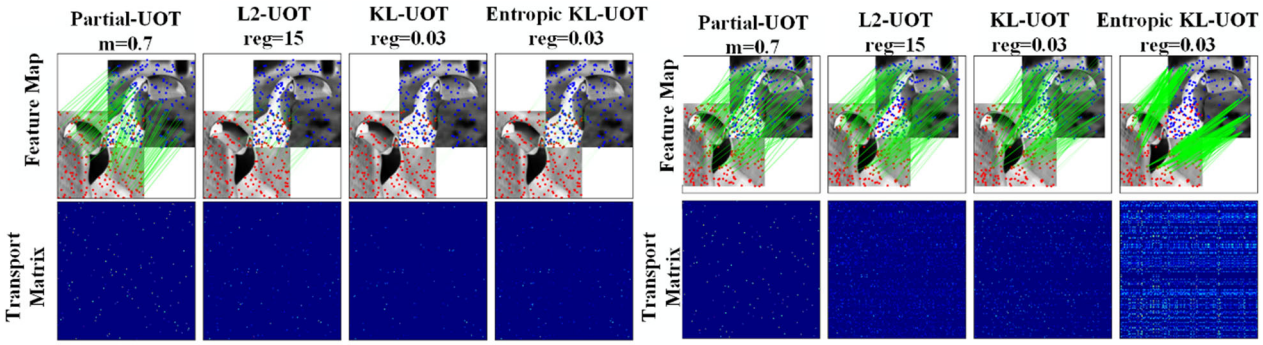

Fig. 6. Visualization of transport plans using different UOT methods under Euclidean (a) and cosine (b) distance costs. Each method yields distinct feature alignments and transport matrix patterns, reflecting their sensitivity to cost definitions.

4) Feature visualization and semantic consistency analysis

Fig. 7 contrasts the t-SNE embeddings produced by the SD baseline (a) with those obtained using our USD (b). In the SD plot, points from the five test classes intermingle and several outliers spread across the space—a direct consequence of the *mass-conservation* constraint, which forces every query feature to be matched to some support mass even when they are semantically unrelated. By contrast, USD introduces a KL-based mass-penalty that allows unmatched or noisy features to be partially discarded. This relaxation yields two visible effects: tighter intra-class clusters—the variance within each color patch is markedly reduced—and larger inter-class margins—clusters are more clearly separated with fewer overlapping samples. The visual evidence confirms the theoretical advantage of USD: by optimizing a flexible transport plan, it mitigates over-alignment and suppresses background artefacts, thereby providing semantically coherent feature alignment that underpins the quantitative gains.

D. Comparison with State-of-the-Art Methods

Tables III and IV present the performance comparison of various few-shot learning methods across multiple datasets. Some of the data presented in these tables have been sourced from relevant studies such as [2, 14]. In Table III, USD achieved the best performance on Mini-ImageNet and Tiered-ImageNet, with 1-shot accuracy reaching 68.14% and 75.58%, and 5-shot accuracy reaching 85.57% and 88.21%, respectively. Table IV demonstrates the robustness of our methods on the CIFAR100 and CUB datasets. USD leads with a 1-shot accuracy of 47.56% on CIFAR100 and 76.68% on CUB, and a 5-shot accuracy of 64.62% on CIFAR100 and 88.89% on CUB.

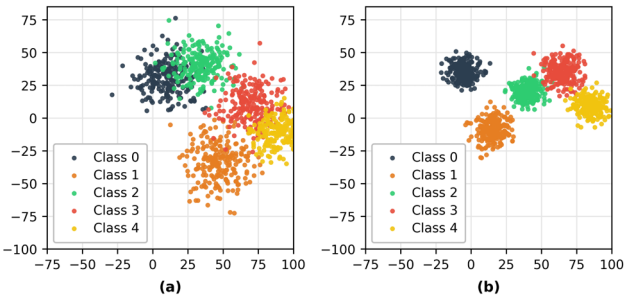

Fig. 7. Feature visualization and semantic consistency analysis.

TABLE III. RESULTS ON MINI-IMAGENET AND TIERED-IMAGENET DATASETS

Method	Backbone	Mini-ImageNet 1-shot (%)	Mini-ImageNet 5-shot (%)	Tiered-ImageNet 1-shot (%)	Tiered-ImageNet 5-shot (%)
CTM [8]	ResNet18	64.12±0.82	80.51±0.13	68.41±0.39	84.28±1.73
S2M2 [9]	ResNet18	64.06±0.18	80.58±0.12	-	-
TADAM [10]	ResNet12	58.50±0.30	76.70±0.38	-	-
MetaOptNet [6]	ResNet12	62.64±0.44	78.63±0.46	65.99±0.72	81.56±0.63
DN4 [20]	ResNet12	64.73±0.44	79.85±0.31	-	-
Baseline++ [22]	ResNet12	60.56±0.45	77.40±0.34	65.10±0.92	80.39±0.69
Good-Embed [23]	ResNet12	64.82±0.60	82.14±0.43	71.52±0.69	86.03±0.58
FEAT [14]	ResNet12	66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16
Meta-Baseline [7]	ResNet12	63.17±0.23	79.26± 0.17	68.62±0.27	83.29±0.18
MELR [24]	ResNet12	67.40±0.43	83.40±0.28	72.14±0.51	87.01±0.35
FRN [25]	ResNet12	66.45±0.19	82.83±0.13	71.16±0.22	86.01±0.15
IEPT [26]	ResNet12	67.05±0.44	82.90±0.30	72.24±0.50	86.73±0.34
BML [27]	ResNet12	67.04±0.63	83.63±0.29	68.99±0.50	85.49±0.34
ProtoNet [1]	ResNet12	62.11±0.44	80.77±0.30	68.31±0.51	83.85±0.36
ADM [28]	ResNet12	65.87±0.43	82.05±0.29	70.78±0.52	85.70±0.43
Convert [25]	ResNet12	64.59±0.45	82.02±0.29	69.75±0.52	84.21±0.26
DeepEMD [2]	ResNet12	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58
BSSD [13]	ResNet12	67.28±0.20	83.48±0.14	71.55±0.23	86.13±0.16
USD (ours)	ResNet12	**68.14±0.26**	**85.57±0.16**	**75.58±0.38**	**88.21±0.36**

TABLE IV. RESULTS ON CIFAR100 CUB DATASETS

Method	Backbone	CIFAR100 1-shot (%)	CIFAR100 5-shot (%)	CUB 1-shot (%)	CUB 5-shot (%)
cosine classifier [7]	ResNet12	38.47±0.70	57.67±0.77	67.30±0.86	84.75±0.60
TADAM [10]	ResNet12	40.10±0.40	56.10±0.40	66.09±0.92	82.50±0.58
MetaOptNet [6]	ResNet12	41.10±0.60	55.35±0.60	65.36±0.28	81.28±0.41
ProtoNet [1]	ResNet12	41.54±0.76	57.08±0.76	64.25±0.34	82.23±0.36
Match Net [29]	ResNet12	43.88±0.75	57.05±0.71	71.87±0.85	85.08±0.57
MTL [3]	ResNet12	45.10±0.38	57.62±0.59	70.87±0.46	84.65±0.32
Relation Net [30]	ResNet34s	44.62±0.18	56.47±0.39	66.20±0.99	82.30±0.58
DeepEMD [2]	ResNet12	46.47±0.78	63.22±0.71	75.65±0.83	88.69±0.50
BSSD [13]	ResNet12	45.36±0.37	63.57±0.29	72.46±0.34	85.28±0.19
USD (ours)	ResNet12	**47.56±0.83**	**64.62±0.52**	**76.68±0.83**	**88.89±0.63**

E. Cross-domain classification (Mini-ImageNet → CUB)

Table V shows the cross-domain FSL experiments conducted under the 5-way 1-shot and 5-way 5-shot scenarios, our methods, USD, demonstrated significant superiority over existing approaches, particularly when training on Mini-ImageNet and testing on the CUB dataset. As presented in Table Ⅴ, USD achieved the highest accuracy in the 5-shot scenario, reaching 74.61%, which surpasses all other compared methods. In the 1-shot scenario, USD also performed exceptionally well, leading with accuracies of 54.48%. These results highlight the effectiveness of our proposed methods in addressing the challenges of cross-domain few-shot learning, particularly when applied to diverse datasets such as Mini-ImageNet and CUB. The superior performance of USD emphasizes their capability in capturing and transferring relevant features across domains, making them robust solutions for few-shot learning tasks in varying contexts.

Although the results from DeepBDC is competitive, the differences in image resolution (84×84×3 in our case versus 224×224×3 in theirs) make direct comparison difficult [31]. Therefore, we did not include a comparison with DeepBDC. Our experiments underscore the effectiveness of our approach, particularly in the challenging 1-shot setting.

TABLE V. THE TABLE PRESENTS A COMPARISON OF CROSS-DOMAIN FEW-SHOT LEARNING RESULTS UNDER 1-SHOT AND 5-SHOT CONDITIONS, USING MINI-IMAGENET FOR TRAINING AND CUB FOR TESTING

Method	Backbone	1-shot (%)	5-shot (%)
Baseline [22]	ResNet-18	45.31±0.59	65.57±0.70
Baseline++ [22]	ResNet-18	46.52±0.73	62.04±0.76
GNN+FT [32]	ResNet-12	48.26±0.56	66.98±0.68
BML [27]	ResNet-12	51.47±0.63	72.42±0.54
ProtoNet [5]	ResNet-12	48.24±0.68	67.19±0.38
Good-Embed [23]	ResNet-12	48.39±0.73	67.43±0.44
ADM [28]	ResNet-12	50.61±0.48	70.55±0.43
DeepEMD [2]	ResNet-12	52.36±0.52	72.36±0.58
BSSD	ResNet-12	52.47±0.38	71.46±0.34
USD (ours)	ResNet-12	**54.48+0.47**	**74.61+0.46**

V. CONCLUSION

In this paper, we propose a simple yet effective method for few-shot image classification. Our approach addresses the challenge of mismatched source and target distributions by effectively aligning the features between the query set and the support set using USD. By reshaping the positions of query samples within the feature maps, our method enhances the distinguishability and diversity of features relative to the support set. We have carefully designed the embedding layer to enable end-to-end training. Our method was benchmarked against existing few-shot learning approaches based on OT theory, demonstrating significant improvements in both computational efficiency and accuracy. Extensive experiments show that our proposed method outperforms state-of-the-art approaches, establishing a new standard in FSL. Our work highlights the largely overlooked potential of USD and encourages its future application in deep learning.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare that are relevant to the content of this article.

AUTHOR CONTRIBUTIONS

Yun Pang was responsible for the conceptualization of this study, methodology, software development, and writing the original draft. Hayati Abd Rahman was responsible for data analysis, reviewing, and editing. All authors had approved the final version.

REFERENCES

[1] J Snell, K Swersky, and R Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, 2017. doi: 10.48550/arXiv.1703.05175

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

[3] L. Chizat, G. Peyré, B. Schmitzer, and F. X. Vialard, "Scaling algorithms for unbalanced optimal transport problems," *Mathematics of Computation*, vol. 87, no. 314, pp. 2563–2609, 2018. doi: 10.1090/mcom/3303

[4] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *ICML Deep Learning Workshop*, vol. 2, no. 1, pp. 1–30, 2015.

[5] B. Lake, R. Salakhutdinov, and J. Gross, "One shot learning of simple visual concepts," in *Proc. of the Annual Meeting of the Cognitive Science Society*, vol. 33, no. 33, 2011.

[6] K. Lee, S. Maji, and A. Ravichandran, "Meta-learning with differentiable convex optimization," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp.10657–10665. doi:10.1109/CVPR.2019.0109 1

[7] Y. Chen, Z. Liu, H. Xu, *et al.*, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2021. pp. 9062–9071. doi:10.1109/ICCV48922.2021.00893

[8] H. Li, D. Eigen, S. Dodge, *et al*., "Finding task-relevant features for few-shot learning by category traversal," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. https://doi.org/10.48550/arXiv.1905.11116

[9] P. Mangla, N. Kumari, A. Sinha, *et al.*, "Charting the right manifold: Manifold mix up for few-shot learning," in *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2218–2227.

[10] B. Oreshkin, P. Rodríguez López, A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," *Advances in Neural Information Processing Systems*, p. 31, 2018. doi: 10.48550/arXiv.1805.10123

[11] J. Liu, L. Song, and Y. Qin, "Prototype rectification for few-shot learning," in *Proc. Computer Vision–ECCV 2020*, 2020, pp. 741–756.

[12] Y. Hu, V. Gripon, and S. Pateux, "Leveraging the feature distribution in transfer-based few-shot learning," in *Proc. International Conference on Artificial Neural Networks*. *Cham: Springer International Publishing*, 2021, pp. 487–499.

[13] Y. Liu, L. Zhu, X. Wang, *et al.*, "Bilaterally normalized scale-consistent Sinkhorn distance for few-shot image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11475–11485, 2023.

[14] J. Xie, F. Long, J. Lv, *et al.*, "Joint distribution matters: Deep brownian distance covariance for few-shot classification," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7972–7981. doi: 10.48550/arXiv.2204.04567

[15] B. Schmitzer, "Stabilized sparse scaling algorithms for entropy regularized transport problems," *SIAM Journal on Scientific Computing*, vol. 41, no. 3, pp. A1443–A1481, 2019. doi: 10.1137/16M1106018

[16] L. Chapel, M. Z. Alaya, and G. Gasso, "Partial optimal transport with applications on positive-unlabeled learning," *Advances in Neural Information Processing Systems 33*, pp. 2903–2913, 2020.

[17] O. Veilleux, M. Boudiaf, P. Piantanida, *et al*. "Realistic evaluation of transductive few-shot learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9290–9302, 2021.

[18] I. Iqbal, M. Younus, K. Walayat, *et al*., "Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images," *Computerized Medical Imaging and Graphics,* vol. 88, 101843, 2021.

[19] I. Iqbal *et al*., "Deep learning-based automated detection of human knee joint's synovial fluid from magnetic resonance images with transfer learning," *IET Image Processing,* vol. 14, no. 10, pp. 1990–1998, 2020.

[20] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[21] J. A. Black, A. Paez, and P. A. Suthanaya, "Sustainable urban transportation: Performance indicators and some analytical approaches," *Journal of Urban Planning and Development*, vol. 128, no. 4, pp. 184–209, 2002.

[22] M. Hou and I. Sato. "A closer look at prototype classifier for few-shot image classification," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25767–25778, 2022.

[23] Y. Tian *et al*., "Rethinking few-shot image classification: A good embedding is all you need?" in *Proc. Computer Vision–ECCV 2020*, 2020, pp. 266–282.

[24] N. Fei *et al*. "MELR: Meta-learning via modeling episode-level relationships for few-shot learning," in *Proc. International Conference on Learning Representations*, 2021.

[25] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proc. of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[26] M. Zhang *et al*., "IEPT: Instance-level and episode-level pretext tasks for few-shot learning," in *Proc*. *International Conference on Learning Representations*, 2021.
[27] Z. Zhou *et al*., "Binocular mutual learning for improving few-shot classification," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8382–8391.
[28] W. Li, L. Wang, J. Huo, *et al*., "Asymmetric distribution measure for few-shot learning," in *Proc. of the Twenty-Ninth International Joint Conference on Artificial Intelligenc*e, vol. 407, 2020, pp. 2957–2963.
[29] L. Chapel, R. Flamary, H. Wu, *et al*., "Unbalanced optimal transport through non-negative penalized linear regression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23270–23282, 2021.
[30] M. Cuturi, "Sinkhorn distances: lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems*, vol. 2, pp. 2292–2300, 2013.
[31] Y. Liu, L. Zhu, X. Wang, *et al*. "Bilaterally-normalized scale-consistent Sinkhorn distance for few-shot image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11475–11485, 2024.
[32] H. Y. Tseng, H. Y. Lee, J. B. Huang, *et al*., "Cross-domain few-shot classification via learned feature-wise transformation," arXiv preprint, arXiv:2001.08735, 2020.