Neurocomputing 669 (2026) 132432

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Optimizing few-shot distribution estimation with negative calibration mitigation

Juan Zhao [a], Lili Kong [a], Chenwei Tang [a,b], Wei Ju [a], Deng Xiong [c], Jiancheng Lv [a,b,*]

[a] College of Computer Science, Sichuan University, Chengdu, 610065, China
[b] Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, Chengdu, 610065, China
[c] School of Mechanical Engineering, Stevens Institute of Technology, Hoboken, NJ, 07030, USA

ARTICLE INFO

Communicated by J. Yu

Keywords:
Few-shot learning
Distribution fusion
Data augmentation
Negative calibration
Distribution calibration

ABSTRACT

Few-shot learning is a challenging task that has attracted increasing research attention and achieved significant advances in recent years. Data augmentation–based methods have been proposed to calibrate biased feature distributions in few-shot settings by leveraging base-class statistics based on class similarity, thereby better approximating the true novel-class distribution. However, these methods overlook whether calibration benefits all samples, potentially degrading learning performance, which is known as negative calibration. This study systematically investigates this issue and introduces a distributional fusion framework to mitigate its adverse effects. The framework enhances effectiveness by pre-screening samples prone to negative calibration based on their forgetting frequency. Additionally, existing distribution calibration methods typically measure similarity between novel-class and base-class distributions using feature means, capturing positional differences while ignoring the geometric structure encoded in covariance. The proposed framework dynamically refines novel-class distribution estimation using an optimal transport strategy, leveraging the Bures–Wasserstein distance to accurately capture distributional differences via covariance comparison. On a series of Few-Shot Learning benchmarks, our method achieves a performance improvement of 2.77 %−4.69 % compared to existing methods, and effectively mitigates negative calibration.

1. Introduction

Deep learning algorithms have demonstrated remarkable achievements across various domains, including speech recognition [36,79], computer vision [35,83], video captioning [9,82], and natural language processing [32]. However, these methods commonly rely on large amounts of labeled samples to achieve high performance, which is challenging and prohibitively expensive in practical scenarios. To tackle this problem, Few-Shot Learning (FSL) was proposed, aiming to learn from the limited samples with supervised information and then effectively generalize during the testing phase [16,39,66].

Few-Shot Learning is constrained by the scarcity of examples per novel class, making it challenging to faithfully approximate the true underlying distribution. Although existing data augmentation methods can increase the number of training samples, they often fail to accurately capture the distribution of novel classes, leading to distortion in the estimated novel-class distribution. To alleviate this issue, researchers have proposed various data-augmentation-based distribution calibration methods, aiming to reduce distributional bias and more accurately estimate the feature distributions of novel classes. Inspired by the observation in [77] that semantically similar classes tend to share similar mean and variance statistics in the feature representation space, several distribution calibration approaches [21,33,34,44,53,77] leverage the distributional statistics of base classes to calibrate the biased distributions of novel classes according to class similarity. By sampling from the calibrated distributions, these methods generate synthetic samples that more closely approximate the true distributions.

However, existing methods overlook the fact that calibration may not benefit all samples, potentially degrading model performance. Preliminary experiments were performed to examine the impact of calibration on individual samples. As illustrated in Fig. 1, the results indicate that calibration is not universally effective and may even induce negative calibration for certain samples.

* Corresponding author at: College of Computer Science, Sichuan University, Chengdu, 610065, China.
Email addresses: zhaojuan1@stu.scu.edu.cn (J. Zhao), konglili@stu.scu.edu.cn (L. Kong), tangchenwei@scu.edu.cn (C. Tang), juwei@scu.edu.cn (W. Ju), dxiong@stevens.edu (D. Xiong), lvjiancheng@scu.edu.cn (J. Lv).

https://doi.org/10.1016/j.neucom.2025.132432
Received 7 August 2025; Received in revised form 28 October 2025; Accepted 13 December 2025
Available online 17 December 2025

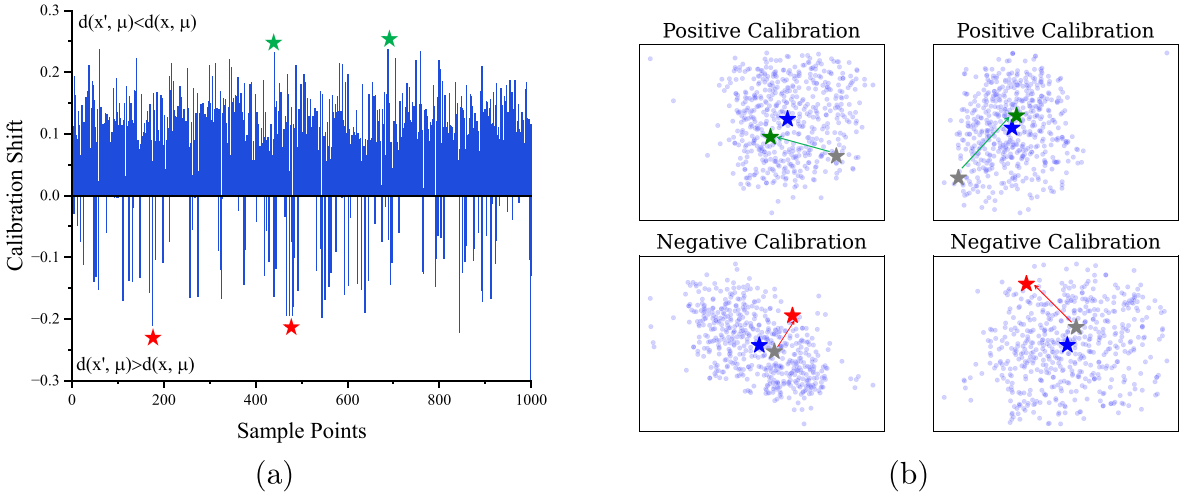


Fig. 1. (a) The calibration shift of 200 5-way 1-shot tasks on *mini*Imagenet with 1000 samples. It can be observed that there are 10 % of samples with negative calibration. (b) The t-SNE visualization of feature distributions of positive and negative calibration examples: the ★ represents the support sample in 5-way 1-shot tasks, while ★ represents the prototype of the ground-truth distribution of the support set. The ★ represents negative calibration sample, deviating from the prototype after distribution transform. Conversely, the ★ represents positive calibration sample, which is close to the prototype of the ground-truth distribution after distribution transform. (Best viewed in color).

To quantify this effect, the concept of calibration shift is defined as the Euclidean distance between the original sample x and the prototype of the ground-truth distribution μ, denoted as $d(x, \mu)$, relative to the distance after calibration, denoted as $d(x', \mu)$, where x' denotes the calibrated sample. Fig. 1(a) illustrates the calibration shift of 200 5-way 1-shot tasks on *mini*Imagenet, each containing 1000 samples. Approximately 10 % of samples exhibit negative calibration. Examples of positive and negative calibration for 5-way 1-shot tasks on *mini*Imagenet are shown in Fig. 1(b). Negative calibration occurs when samples deviate from the prototype after a distributional transformation, whereas positive calibration indicates that samples closely align with the prototype of the ground-truth distribution following the distributional transformation.

Preliminary experiments confirm the existence of negative calibration and its adverse impact on model performance. Consequently, identifying negative calibration samples is essential, and a mitigation strategy is employed that pre-filters such samples before distribution calibration to reduce their interference with novel-class distribution estimation. Negative calibration samples typically have centered, large target objects with simple, semantically clear backgrounds. These samples can be recognized by the model with high confidence even without calibration, and including them in calibration may impair distribution estimation accuracy. While these characteristics provide intuition, they are not easily quantifiable or generalizable, necessitating an objective metric to systematically identify negative calibration samples.

Inspired by Toneva et al. [67], forgetting frequency during training is used as the selection criterion. This simple metric requires no additional networks or parameters and has low correlation with class labels [61], providing an objective measure of sample quality. Distribution calibration is then applied only to positive calibration samples, enhancing the accuracy and robustness of novel-class distribution estimation.

Moreover, existing distribution calibration methods typically measure similarity between novel class and base class distributions using feature means. This captures positional differences but ignores the geometric structure represented by covariance. To address this limitation, theoretical studies provide guidance. Olkin et al. [51] demonstrated that incorporating covariance matrices [11] extends traditional point-to-point metrics into a holistic measure of distributional differences, enabling more accurate evaluation of distribution similarity. Han et al. [22] further pointed out that the Bures–Wasserstein distance combines Euclidean and Mahalanobis distances, capturing differences in both position and geometric structure of Gaussian distributions. This provides a theoretical foundation for fine-grained distribution alignment.

Inspired by these findings, the proposed method introduces a Bures–Wasserstein based approach [1] that considers both mean and covariance to quantify the similarity between novel-class and base-class distributions. Furthermore, sample richness is crucial for building robust classifiers in few-shot scenarios. Gaussian sampling from the calibrated distributions is performed to augment training data, enhancing training accuracy and reducing overfitting. The main contributions of this paper are as follows:

- To the best of current knowledge, the adverse effects of negative calibration occurring during the distribution calibration process are systematically identified and analyzed for the first time. A distributional fusion framework is introduced to mitigate these effects by identifying and filtering negatively calibrated samples, which is accomplished through pre-screening support samples based on their forgetting frequency.
- A novel positive distribution transformation model is introduced to dynamically adjust the estimation of novel-class distributions using optimal transport techniques, specifically the Bures-Wasserstein distance, allowing accurate capture of both the position and geometric structure of pairwise Gaussian distributions.
- Extensive experiments are conducted on three widely used benchmark datasets for Few-Shot Learning, demonstrating robust performance across all benchmarks. The proposed method achieves a performance improvement of 2.77 %–4.69 % over existing methods while effectively mitigating negative calibration.

The subsequent sections of this paper are organized as follows. The related works or areas that are closely connected are discussed in Section 2. We then elaborate on the formulation and corresponding process of the proposed methods in Section 3. Section 4 presents extensive experiments and ablation studies aimed at evaluating the performance of the proposed methods. Finally, the analysis and limitations are presented in Section 5, while the conclusions are discussed in Section 6.

2. Related work

In this section, we delineate the existing studies or domains most pertinent to our approach. This section encompasses two parts, specifically, Few-Shot Learning and sample hardness assessment.

2.1. Few-shot learning

Few-Shot Learning (FSL) addresses the challenge of recognizing unseen classes with only a few labeled samples. Existing FSL methods can be broadly categorized into four groups: algorithm-optimized, model-based, metric-based, and optimization-based methods.

Traditional data augmentation techniques, such as rotation [37], scaling [40], and mixup [45], expand the dataset through simple geometric transformations. In contrast, generative approaches such as GPRN [46], WeditGAN [13], Ensemble FSC [31], and Variational Auto-Encoders (VAEs) [74] generate synthetic samples from learned latent representations. Model-based approaches enhance learning capacity by leveraging external memory [81] or metric-based models such as MIFN [17], TST-MFL [64], Prototypical Network [60], and Relation Network [56]. These methods are selected for comparison with our approach. Moreover, algorithm-based approaches, such as MAML [14] and MetaEDL [43], facilitate rapid adaptation across diverse tasks.

Additionally, methods designed to enhance data augmentation through distribution calibration have been widely adopted. The Distribution Calibration (DC) method [77] modifies the support set distribution by transferring statistics from base classes. However, the DC method neglects the adaptive transfer of transferable knowledge during the calibration process. Adaptive Distribution Calibration (ADC) [44] introduces a method to automatically determine cost functions between base and novel classes by incorporating optimal transport, but it overlooks covariance information within the distributions. The Hierarchical Optimal Transport (H-OT) method [21] further introduces a dual-layered design, consisting of high-level OT for learning cost functions and low-level OT for assessing similarities between base and novel samples. Nevertheless, this method incurs significant computational costs as it considers all base and novel samples, yielding only marginal performance gains.

Inspired by DC [77] and transductive learning, TDO [6] and PDC [33] attempt to reduce estimation bias in feature distributions. However, their similarity matching relies on the Euclidean distance, which fails to capture feature correlations. RTDC [34] and DDC [3] capture underlying distribution information by estimating the covariance matrix of each novel class. However, covariance estimation becomes unreliable in 5-way 1-shot scenarios due to the lack of sample diversity. Additionally, DDWM [72] can produce an infinite number of samples for few-shot classes. AdaBC [75] adopts a secondary selection strategy based on the blank center to identify approximate base classes, while ConCM [71] proposes memory-aware prototype calibration to address prototype bias and structure fixity. However, all of these methods overlook the issue of negative calibration, where over-adjustment of feature distributions leads to biased representations.

2.2. Sample hardness assessment

Recent studies assess example difficulty based on the training loss, particularly near convergence. These approaches include curriculum learning [5,15,25] and self-paced learning [80]. Curriculum learning advocates sequential training with progressively harder examples, among which the widely adopted Transfer Teacher method [50] plays a key role in guiding task difficulty.

Self-paced learning (SPL) [20,28,85] employs a difficulty measurer that shares parameters with the classification model, thereby reducing computational cost and improving robustness. Meta-task learning (MTL) [63] identifies low-accuracy classes as difficult ones and constructs additional training tasks centered on them. Additionally, Online Hard Example Mining (OHEM) [2,42,59] selects the most challenging samples within each mini-batch to improve neural network performance, effectively reweighting the data distribution toward harder instances. Meanwhile, MaxUp [19] minimizes the average risk under worst-case data augmentations by generating multiple random perturbations or transformations. This study is inspired by Toneva et al. [67], who empirically demonstrated that unforgettable examples can be removed without affecting model generalization.

2.3. Discussion

The methodologies discussed in this section have profoundly influenced our research in two primary aspects: (1) Our approach is fundamentally categorized as an augmentation-based method, enhancing the diversity of training samples. (2) The distribution calibration process, which is informed by the prior selection of negative calibration samples, draws substantial inspiration from the foundational work of [67]. Furthermore, the findings in [61] reveal relatively low class correlation based on the analysis of forgetting events, indicating a diminished sensitivity to the inherent quality of the class samples.

3. Methodology

This section first introduces the preliminaries of few-shot learning in Section 3.1. Next, the proposed approach is detailed, which employs a negative calibration sample selection strategy to identify negative calibration samples based on the frequency of forgetting events, as discussed in Section 3.2. Following that, the transformation of positive support samples from the base classes using an optimal transport strategy is described in Section 3.3. Finally, the sample augmentation and training procedure adopted in the proposed framework are explained in Section 3.4. An illustration of the proposed method is shown in Fig. 2, and the key symbols with their corresponding definitions are summarized in Table 1.

3.1. Preliminary

A typical Few-Shot Learning (FSL) setting involves base dataset and novel dataset. The base dataset denoted as $\mathcal{D}_{\text{base}} = \left\{\left(X_i, Y_b\right), Y_b \in \mathcal{C}_{\text{b}}\right\}_{i=1}^{N_{\text{base}}}$, which consists of samples labeled with classes from $\mathcal{C}_{\text{b}}$, where $b = 1, 2, \ldots, B$. This dataset is used to train the model. The novel dataset, denoted as $\mathcal{D}_{\text{novel}} = \left\{\left(X_j, Y_n\right), Y_n \in \mathcal{C}_{\text{n}}\right\}_{j=1}^{N_{\text{novel}}}$, consists of samples labeled with classes from $\mathcal{C}_{\text{n}}$, where $n = 1, 2, \ldots, N$. The labels Y_b and Y_n represent the class labels for instances X_i and X_j, respectively. The two class label sets $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{novel}}$ are disjoint: $\mathcal{D}_{\text{base}} \cap \mathcal{D}_{\text{novel}} = \emptyset$. N_{base} and N_{novel} denote the total numbers of observations in $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{novel}}$. During testing, a task $\mathcal{T}$ contains N classes sampled from $\mathcal{D}_{\text{novel}}$, with each class having K support samples and q query samples. The support set denoted as $\mathcal{S} = \left\{\left(X_j, Y_j\right)\right\}_{j=1}^{N\times K}$ and the query set denoted as $\mathcal{Q} = \left\{\left(X_j, Y_j\right)\right\}_{j=N\times K+1}^{N\times K+N\times q}$. Through training the model on the base classes, FSL employs the labeled support set to adjust to the task, subsequently assessing its performance on the unannotated query set, marked as an "N-way K-shot" task.

We pre-train our feature extractor on the base dataset D_{base} using self-supervision. The objective function of the embedding module combines the self-supervision loss with the classification loss. Subsequently, the feature extractor parameters are configured to extract features from novel classes D_{novel}.

3.2. Negative calibration sample selection

This section aims to identify negative calibration samples that induce calibration shift and consequently degrade model performance during distribution transformation. Fig. 3 presents examples of negative and positive calibration samples from the CUB dataset [70], highlighting their distinguishing features. Negative calibration samples are typically centrally positioned and occupy a large portion of the image, exhibiting clear and easily recognizable features such as a bird centered against a clear sky, which can be correctly classified without calibration. Conversely, positive calibration samples often display ambiguous characteristics that necessitate calibration, such as a bird partially occluded or surrounded by cluttered backgrounds, which complicate classification in few-shot settings.

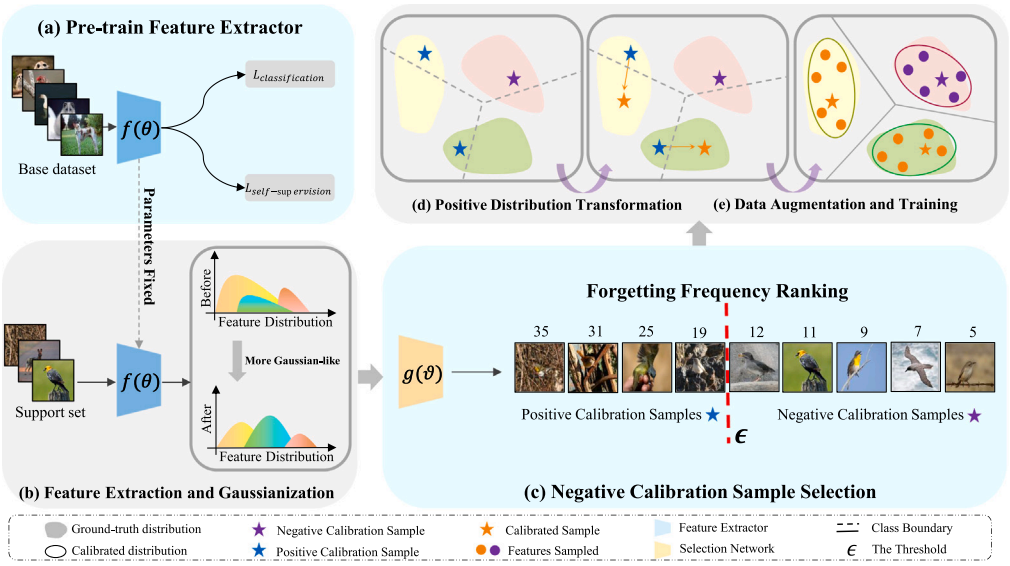


Fig. 2. The illustration of our method comprises four key steps: (a) **Pre-train feature extractor** (Section 3.1): Pre-train the backbone network with self-supervision on base dataset to extract features. (b) **Feature extraction and gaussianization** (Section 3.3): Extract novel class features using a pre-trained feature extractor and apply Tukey's Ladder of Powers transformation to Gaussianize the distribution. (c) **Negative calibration sample selection** (Section 3.2): aims to mitigate adverse effects on model calibration by pre-screening samples based on the frequency of forgetting events. (d) **Positive distribution transformation** (Section 3.3): We calibrate the support set distribution using optimal transport for positive samples. (e) **Data augmentation and training** (Section 3.4): During data augmentation, feature vectors are sampled from the calibrated or original distributions for training.

Table 1
Main symbols and instructions.

Symbol	Instruction
N	The number of categories in a few-shot task.
K	The number of samples per class in the support set.
B	The number of categories in base dataset.
b, n	The base and novel class.
S, Q	The support and query set.
$\mathbb{T}$	The set of negative calibration samples.
ϵ	The negative calibration sample selection threshold.
$\tilde{S}, \tilde{Q}$	The transformed support and query set.
$\boldsymbol{\mu}_b$, $\boldsymbol{\mu}_n$	The mean feature vectors of the base and novel classes.
$\boldsymbol{\Sigma}_b$, $\boldsymbol{\Sigma}_n$	The variance feature vectors of the base and novel classes.
λ	The power in Tukey's transformation.
M	The number of generated features per class.
$\mathrm{W}_2^2(b, n)$	The Bures-Wasserstein distance.
$\mathcal{X}_B$	The discrete uniform distribution over B base classes.
$\mathcal{X}_N$	The discrete uniform distribution over N novel classes.
$\mathbf{T} \in \mathbb{R}^{B \times N}$	The transfer weight from $\mathcal{X}_B$ to $\mathcal{X}_N$.
$\mathbf{C} \in \mathbb{R}_{\geq 0}^{B \times N}$	The transfer cost from $\mathcal{X}_B$ to $\mathcal{X}_N$.
γ	The entropy regularization parameter.
$\boldsymbol{\mu}', \boldsymbol{\Sigma}'$	The mean and variance of the calibrated distribution.
$\mathbb{S}_j'$	The augmentation set $\mathbb{S}_j'$ based on negative calibration sample X_j'.
$\mathbb{S}_j$	The augmentation set $\mathbb{S}_j$ based on positive calibration sample X_j.
$\mathbb{D}$	All augmentation samples for support set.

Based on this analysis, the selection of negative calibration samples is guided by their classification difficulty. Samples that are easily learned and consistently classified correctly are more likely to induce negative calibration, whereas difficult or outlier samples typically require calibration to better represent the class distribution. While these observations provide intuition, they lack an objective criterion for systematic identification. To address this, the forgetting frequency of each example is analyzed within a standard classification framework, following [67]. This metric quantifies how consistently a sample is remembered during training, facilitating reliable detection of negative calibration samples.

The support set is defined as $S = (X_j, Y_j)_{j=1}^{N \times K}$, as described in the previous sections. This approach utilizes label pairs to model the conditional probability distribution $p(y \mid x; \vartheta)$ with a deep neural network equipped with parameters ϑ. The primary training objective for the network is to minimize the empirical risk, which is defined as follows:

$$R = \frac{1}{|S|} \sum_j \ell \left(p \left(y \mid \mathbf{X}_j; \vartheta \right), Y \right), \tag{1}$$

here, ℓ represents the cross-entropy loss function and the minimization of this loss is executed via the stochastic gradient descent (SGD). Regarding the dynamics of learning events, these occurrences are denoted as follows:

$$\hat{y}_j^t = \arg\max_k p \left(y_{jk} \mid \mathbf{X}_j; \vartheta^t \right). \tag{2}$$

The predicted label for sample $\mathbf{X}_j$ is obtained after t steps. Additionally, we define $\mathrm{acc}_j^t = \mathbb{1}_{\hat{y}_j^t = y_j}$ as a binary variable that indicates whether the example $\mathbf{X}_j$ is correctly classified at time step t. A forgetting event for example j is detected when acc_j^t shows a decrease between two consecutive updates:

$$\mathrm{acc}_j^t > \mathrm{acc}_j^{t+1}. \tag{3}$$

According to this definition, examples that are forgotten at least once during training are termed forgettable. During training, the forgetting frequency of each sample X_j in the support set, denoted by T_j, is tracked over all epochs τ:

$$T_j = \sum_{t=1}^{\tau} (\mathrm{acc}_j^t > \mathrm{acc}_j^{t+1}), j = 1, \ldots, N \times K. \tag{4}$$

The threshold ϵ for negative calibration sample selection is determined according to the frequency of forgetting events observed during training:

$$\mathbb{T} = \sum_{j=1}^{N \times K} (T_j \leq \epsilon), \tag{5}$$

here, $\mathbb{T}$ represents the set of negative calibration samples identified based on their forgetting frequency during training.

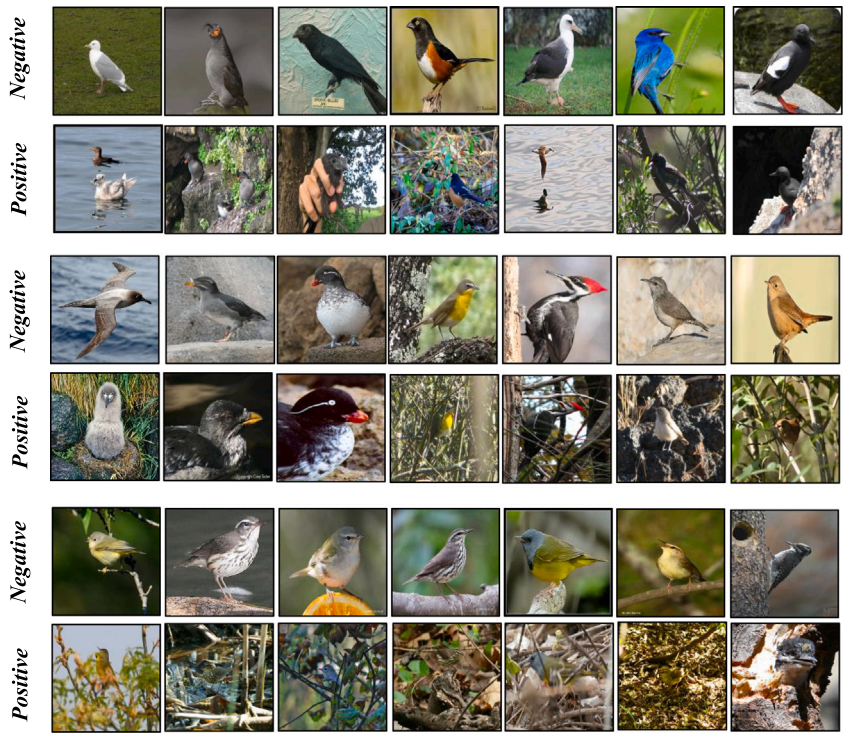


Fig. 3. Examples of negative and positive calibration samples in CUB dataset.

Analyzing the frequency of forgetting events allows us to easily identify negative calibration samples. The less frequently a sample is forgotten during training, the higher the likelihood that it is a negative calibration sample. By recognizing these samples prior to distribution transformation, negative calibration can be effectively prevented. Conversely, samples that are frequently forgotten may inherently pose greater challenges in accurate classification or may deviate from accurately representing the general class, thus necessitating calibration. This renders them less probable candidates for negative calibration examples.

3.3. Positive distribution transformation

As discussed earlier, base classes are characterized by abundant data, whereas support sets contain only a few labeled samples. Consequently, transferring statistics from well-sampled base classes allows for a more accurate estimation of the support set's distribution. Building on this idea, a positive distribution calibration strategy is proposed for the positive calibration samples introduced in Section 3.2.

For simplicity, it is assumed that the feature vectors follow a Gaussian distribution. Thus, the mean and covariance matrix statistics can capture the distribution information accurately. To make the feature distributions more Gaussian-like, Tukey's Ladder of Powers transformation [68] is applied to the features of both base and novel classes. This transformation, part of a family of power transformations, effectively reduces skewness and enhances the Gaussian-like properties of the distributions. Tukey's Ladder of Powers transformation is formulated as:

$$\tilde{\mathbf{x}} = \begin{cases} \boldsymbol{x}^{\lambda} & \text{if } \lambda \neq 0, \\ \log(x) & \text{if } \lambda = 0, \end{cases} \tag{6}$$

here, λ is a hyper-parameter used to adjust the skewness of distributions. Decreasing λ makes the distributions less positively skewed, and conversely, increasing λ enhances positive skewness. Note that setting $\lambda = 1$ results in almost no alteration to the original features. After the transformation, the base dataset samples are updated to $\tilde{X}_i$. Additionally, the support set and the query set are updated to $\tilde{S}$ and $\tilde{Q}$, respectively.

After that, the mean and variance of the feature vectors from a base class b are calculated as follows:

$$\begin{aligned} \boldsymbol{\mu}_b &= \frac{\sum_{i=1}^{|C_b|} \left(\tilde{X}_i\right)}{|C_b|}, \\ \boldsymbol{\Sigma}_b &= \frac{1}{|C_b|-1} \sum_{i=1}^{|C_b|} \left(\tilde{X}_i - \boldsymbol{\mu}_b\right)\left(\tilde{X}_i - \boldsymbol{\mu}_b\right)^T. \end{aligned} \tag{7}$$

Similar to base classes, the mean $\boldsymbol{\mu}_n$ and covariance $\boldsymbol{\Sigma}_n$ of novel classes are calculated. Once the distributional statistics are obtained, the more reliably estimated base class statistics are used to calibrate the biased distributions of novel classes. The calibration is guided by the similarity between base and novel classes, quantified using the Bures–Wasserstein distance [1], which accounts for differences in both location (means) and geometric structure (covariances) of the distributions:

$$\mathrm{W}_2^2(b,n) = \left\|\boldsymbol{\mu}_b - \boldsymbol{\mu}_n\right\|_2^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_n - 2\left(\Sigma_b^{\frac{1}{2}} \boldsymbol{\Sigma}_n \boldsymbol{\Sigma}_b^{\frac{1}{2}}\right)^{\frac{1}{2}}\right), \tag{8}$$

where $\Sigma^{\frac{1}{2}}$ is the matrix square root, while $tr(\cdot)$ represents the trace operation of matrix.

However, a challenging issue exists: determining how to properly distribute this information. To address this, the similarity between base and novel classes is quantified by minimizing an optimal transport distance, with the cost function defined using the Bures–Wasserstein distance. For this purpose, $\mathcal{X}_B$ and $\mathcal{X}_N$ are modeled as discrete uniform distributions over Gaussian representations of B base classes and N novel classes, respectively:

$$\begin{aligned} \mathcal{X}_B &\sim \sum_{b=1}^{B} \frac{1}{B} \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b), \\ \mathcal{X}_N &\sim \sum_{n=1}^{N} \frac{1}{N} \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n). \end{aligned} \tag{9}$$

Let $\mathbf{T}$ denote the transfer weight from the base class distribution $\mathcal{X}_B$ to the novel class distribution $\mathcal{X}_N$. After that, we use sinkhorn algorithm [10] to compute transfer weight matrix with a minimal transportation cost:

$$\mathbf{T}^* = \arg\min_{\mathbf{T}\in\Pi} \langle \mathbf{T}, \mathbf{C} \rangle - \gamma \mathbf{H}(\mathbf{T}), \tag{10}$$

where $\mathbf{H}(\mathbf{T})$ denotes the entropy of $\mathbf{T}$ regularized by γ, a smaller γ would force the entropy to become smaller. $\mathbf{C} \in \mathbb{R}_{\geq 0}^{B\times N}$ represents the transport cost matrix by leveraging $\mathrm{W}_2^2(b, n)$. Π is a set contains all possible transfer weight matrices:

$$\Pi(\mathcal{X}_B, \mathcal{X}_N) := \left\{ \mathbf{T} \in \mathbb{R}_+^{B\times N} \middle| \mathbf{T}\mathbf{1}_N = \frac{1}{B}, \mathbf{T}^\top \mathbf{1}_B = \frac{1}{N} \right\}. \tag{11}$$

The uniform marginal constraints in Eq. (11) inherently normalize the transport plan, and the Sinkhorn iterations are implemented in the log-domain to ensure numerical stability. The Bures-Wasserstein cost is smooth over positive semi-definite matrices, allowing stable optimization even when $\gamma = 0$. Once we obtain the transport weight $\mathbf{T}^*$ by minimizing the optimal transport problem in Eq. (10), the novel class distribution can be calibrated through the following process:

$$\boldsymbol{\mu}' = \frac{\Sigma_{b=1}^{B} \mathbf{T}_{bn}^* \boldsymbol{\mu}_b + \tilde{X}_j}{2}, \boldsymbol{\Sigma}' = \frac{\Sigma_{b=1}^{B} \mathbf{T}_{bn}^* \boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_n}{2}. \tag{12}$$

In theory, when there is only one support sample for a class, it is not feasible to compute the class covariance $\boldsymbol{\Sigma}_n$. Consequently, we define $\boldsymbol{\Sigma}_n$ as a zero matrix of the same dimensionality as the feature space. Additionally, note that $\tilde{X}_j \notin \mathbb{T}$, indicating that the distribution transformation is applied only to positive calibration samples, while negative calibration samples remain unchanged. For negative calibration samples, the feature vector of the sample is used as the mean, while the variance is employed to define the variance for data augmentation. This approach not only reduces computational costs but also avoids negative transformation calibration.

3.4. Data augmentation and training

In Few-Shot Learning scenarios involving more than one shot, the distribution calibration process should be repeated multiple times, each time utilizing a different feature vector from the support set. This approach mitigates bias associated with relying on a single sample and aims to achieve a more diverse and precise distribution estimation.

For an N-way-K-shot task, the augmentation set $\mathbb{S}_j'$ based on negative calibration sample $\tilde{X}_j'$ is generated without calibration as follows:

$$\mathbb{S}_j' \sim \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right), \ \tilde{X}_j' \in \mathbb{T}, \tag{13}$$

similarly, the augmentation set $\mathbb{S}_j$ based on the calibrated positive sample $\tilde{X}_j$ is generated as follows:

$$\mathbb{S}_j \sim \mathcal{N}\left(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j'\right), \ \tilde{X}_j \notin \mathbb{T}, \tag{14}$$

thus, all augmentation samples are denoted as:

$$\mathbb{D} = \left\{ (\mathbf{x}, y) \mid \mathbf{x} \in \mathbb{S}_j \cup \mathbb{S}_j', y \in \mathcal{Y}^T \right\}, \tag{15}$$

where $\mathcal{Y}^{\mathcal{T}}$ represents the set of classes for task $\mathcal{T}$. Moreover, the number of generated features per class, denoted as M, is configured as a hyperparameter. The classifier is trained by minimizing the cross-entropy loss, incorporating features from both the support set $\tilde{S}$ and the generated features $\mathbb{D}$, and is parameterized by Θ:

$$\ell = \sum_{(\mathbf{x},y)\sim\tilde{S}\cup\mathbb{D}} -\log \Pr(y \mid \mathbf{x}; \Theta). \tag{16}$$

Algorithm 1 Optimizing Few-Shot Distribution Estimation with Negative Calibration Mitigation.

Input: Base dataset $\mathcal{D}_{\text{base}}$ features, Support set $\mathcal{S}$.
Output: The optimal parameter Θ of the classifier.
Train a WideResNet to classify the base classes, along with a self-supervised task;
Extract support set features and query set features by the pre-trained WideResNet;
Select the negative calibration samples with negative calibration sample selection threshold ϵ by counting forgotten events during the training process with Eq. (5);
Transform samples in base dataset and novel dataset with Tukey's Ladder of Powers transformation;
Calculate base classes's statistics μ_b, Σ_b according to the transformed features;
for negative calibration samples $\tilde{X}_j' \in \mathbb{T}$ **do**
Form an augmentation set $\mathbb{S}_j'$ based on negative calibration samples $\tilde{X}_j'$ according to Eq. (13);
end for
for positive calibration samples $\tilde{X}_j \notin \mathbb{T}$ **do**
Calibrate the mean $\boldsymbol{\mu}_j$ and the covariance $\boldsymbol{\Sigma}_j$ for positive calibration sample $\tilde{X}_j$ with Eq. (12);
Form an augmentation set $\mathbb{S}_j$ based on positive calibration samples $\tilde{X}_j$ with Eq. (14);
end for
Train a classifier using both the support set $\tilde{S}$ and sampled features $\mathbb{D}$ with Eq. (16).

3.5. Workflow

To provide a comprehensive overview, the workflow is detailed in Algorithm 1.

4. Experiments and discussions

In this section, we introduce our experimental setups, including datasets, evaluation metrics, implementation details, network architecture and parameter settings. We then evaluate the proposed method by comparing it to typical and state-of-the-art few-shot methods. Additionally, we discuss the effectiveness of different modules, the visualization and the influence of hyper-parameters.

4.1. Datasets and evaluation metric

In this paper, we conduct experiments on three standardized few-shot image classification benchmarks, including the *mini*ImageNet, the *tiered*ImageNet and the CUB dataset.

(1) *mini*ImageNet [69], a refined version of ImageNet, comprises 100 classes, each with 600 84×84 images. Generally 64 classes are used for training, 16 for validation, and 20 for testing.
(2) *tiered*ImageNet [54], introduced by Mengye et al. in 2018, is also derived from ImageNet. *tiered*ImageNet differs from *mini*ImageNet in that it comprises 608 categories, with each class having about $1,300$ images of size 84×84 and divided into 351, 97 and 160 classes for training, validation and testing, respectively.
(3) CUB-200–2011[70] (Caltech-UCSD Birds) is an image dataset of birds, which contains 200 species of birds and a total of $11,788$ images. Generally 100 classes are used for training, 50 for validation and 50 for testing.

As mentioned earlier, recent works widely adopt the standard Few-Shot setting of 5-way 1-shot and 5-way 5-shot. Accordingly, we employ the same settings and use top-1 accuracy as the evaluation metric, along with the corresponding 95 % confidence interval across all three datasets. The training subset is referred to as the base dataset, while

the testing subset is used as the novel dataset. We randomly generate $10,000$ tasks from the novel dataset, each comprising 5 classes ($N = 5$) with either 1 or 5 support samples per class and 15 query samples.

4.2. Implementation details

In this section, we discuss the parameters of different modules as follows.

Feature extraction module: We adopt WideResNet-28–10, a wide residual network consisting of 28 layers with a widening factor of 10, as the feature extraction backbone. The model is trained following the S2M2 framework [49], which has demonstrated strong performance in few-shot learning tasks. It produces 640-dimensional latent feature representations. The model is trained with a batch size of 16, an initial learning rate of 0.001, and a dropout rate of 0.5. All experiments are implemented in PyTorch and optimized using the Adam optimizer with standard weight decay.

Negative calibration sample selection: We employ a lightweight network (two convolutional layers and one fully connected layer) trained with SGD and dropout following [67]. The training runs for $\tau = 100$ epochs across all datasets. Moreover, samples that are never correctly classified are treated as though they have been forgotten infinitely many times, for sorting purposes. The selection threshold ϵ is set to 20 for CUB, and 15 for both *mini*ImageNet and *tiered*ImageNet.

Feature distribution transform and augmentation module: In this module, we set the Tukey transformation parameter λ to 0.5 across all benchmarks. The Sinkhorn algorithm uses a regularization coefficient $\gamma = 0$ and a maximum of 1000 iterations. We generate 600, 700, and 750 samples per class for CUB, miniImageNet, and tieredImageNet, respectively, which yield optimal results. For the cross-entropy-based classifier, we use Logistic Regression with default settings from scikit-learn.

4.3. Comparison with state-of-the-art methods

In this section, we compare our method with four groups of representative few-shot learning methods, including *metric-based, model-based, optimization-based*, and *augmentation-based* approaches. Among them, DC [77] is most closely related to our approach and can be regarded as our baseline. Table 2 shows the performance on the *mini*ImageNet and CUB datasets, while Table 3 presents the results on the *tiered*ImageNet dataset. The results of metric-based methods are cited from [29], which adopt the same backbones and training schemes for all compared methods. The results of optimization-based, model-based, and augmentation-based methods are cited from their original papers. The second-best performance is achieved by H-OT [21], AMMD [73], and DDC [3].

Our method consistently achieves higher accuracy rates across all evaluated datasets compared to other approaches, indicating that it effectively mitigates calibration shift. Notably, augmentation–based methods, such as DC [77], ADC [44], and H-OT [21], outperform other categories of few-shot learning methods. Moreover, our approach further surpasses these methods, showing a significant improvement over the state of the art, with a $1.56\,\%-3.44\,\%$ increase in the 1-shot scenario and a $2.72\,\%-3.74\,\%$ increase in the 5-shot scenario compared with the second-best results. Furthermore, our model performs particularly well in the 5-way 5-shot settings. This is mainly because previous studies have often overlooked the importance of capturing covariance information for distribution calibration, potentially missing opportunities for performance enhancement. Additionally, the performance gain on the fine-grained CUB dataset is more pronounced compared with other benchmarks, indicating that accurate distribution calibration is pivotal for few-shot tasks. This is achieved by precisely capturing both the position and geometric shape of pairwise Gaussian distributions using the Bures–Wasserstein distance.

4.4. Ablation study

Role of each module. Table 4 presents an ablation study comparing the effects of different modules in our method on classification accuracy using the *mini*ImageNet dataset. Specifically, the results in the first, second, and third rows demonstrate that the Tukey transformation

Table 2
5-way 1-shot and 5-way 5-shot classification accuracy (%) on *mini*ImageNet and CUB with 95 % confidence intervals. All results in this table are reported as provided in the original papers. The number in **bold** indicates the best performance, and the second-best is underlined.

	Method	*mini*ImageNet		CUB	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Algorithm-optimized	MAML [14]	48.70 ± 1.84	63.10 ± 0.92	67.45 ± 0.97	83.47 ± 0.59
	Meta-SGD [38]	50.47 ± 1.87	64.03 ± 0.94	53.34 ± 0.97	67.59 ± 0.82
	LEO [55]	61.76 ± 0.08	77.59 ± 0.12	–	–
	RISE [76]	53.22 ± 0.81	69.41 ± 0.67	67.42 ± 0.96	82.86 ± 0.59
	MetaTEDL [43]	60.40 ± 0.80	78.00 ± 0.60	–	–
Metric-based	RFPN [27]	67.43 ± 0.51	83.69 ± 0.43	77.26 ± 0.45	91.09 ± 0.33
	ProtoNet [60]	49.42 ± 0.78	68.20 ± 0.66	72.99 ± 0.88	86.64 ± 0.51
	RelationNet [65]	50.44 ± 0.82	65.32 ± 0.70	66.65 ± 0.91	82.12 ± 0.63
	CAN [26]	63.85 ± 0.48	79.44 ± 0.34	–	–
	AAP2S [47]	64.82 ± 0.12	81.31 ± 0.22	77.64 ± 0.19	90.43 ± 0.18
	MIFN [17]	65.42 ± 0.65	80.85 ± 0.41	48.21 ± 0.60	65.33 ± 0.54
	TST-MFL [64]	60.42 ± 0.82	80.03 ± 0.58	80.79 ± 0.84	91.15 ± 0.40
Model-based	Baseline++ [4]	53.97 ± 0.79	75.90 ± 0.61	69.55 ± 0.89	85.17 ± 0.50
	MGGN [84]	65.73 ± 0.52	83.29 ± 0.37	–	–
	Negative-Cosine [41]	63.20 ± 0.80	80.94 ± 0.59	72.66 ± 0.85	89.40 ± 0.43
	S2M2 [49]	64.93 ± 0.18	83.18 ± 0.11	–	–
Augmentation-based	Delta-Encoder [57]	59.90	69.70	69.80	82.60
	Meta Variance Transfer [52]	–	67.67 ± 0.70	69.61 ± 0.46	84.1 ± 0.35
	DC [77]	68.57 ± 0.55	82.88 ± 0.42	79.56 ± 0.87	90.67 ± 0.35
	ADC [44]	68.67 ± 0.21	84.37 ± 0.11	80.20 ± 0.09	91.42 ± 0.08
	H-OT [21]	69.04 ± 0.29	84.36 ± 0.41	80.26 ± 0.35	$\underline{91.45 \pm 0.38}$
	FA-adapter [62]	67.79 ± 0.42	83.24 ± 0.28	80.49	90.33
	AMMD [73]	$\underline{70.31 \pm 0.45}$	85.22 ± 0.29	–	–
	DDC [3]	69.17 ± 1.25	$\underline{85.23 \pm 0.97}$	$\underline{81.47 \pm 1.77}$	90.92 ± 0.98
	Ours	$\mathbf{73.34 \pm 0.33}$	$\mathbf{87.95 \pm 0.25}$	$\mathbf{83.03 \pm 0.46}$	$\mathbf{94.71 \pm 0.11}$

Table 3
5-way 1-shot and 5-way 5-shot classification accuracy (%) on *tiered*ImageNet with 95 % confidence intervals. All results in this table are reported as provided in the original papers. The number in **bold** indicates the best performance, and the second-best is underlined.

Method	5-way 1-shot	5-way 5-shot
Algorithm-optimized		
MAML [14]	51.67 ± 1.81	70.30 ± 1.75
Meta-SGD [38]	–	–
LEO [55]	66.76 ± 0.08	81.44 ± 0.12
RISE [76]	51.73 ± 0.90	69.65 ± 0.74
Boost-MT [30]	69.73 ± 0.71	84.91 ± 0.49
Metric-based		
ProtoNet [60]	53.42 ± 0.78	72.20 ± 0.66
RelationNet [65]	54.44 ± 0.82	71.32 ± 0.70
CAN [26]	69.89 ± 0.51	84.23 ± 0.37
AAP2S [47]	70.83 ± 0.15	84.15 ± 0.29
MIFN [17]	69.54 ± 0.67	84.04 ± 0.49
RFPN [27]	73.02 ± 0.50	87.27 ± 0.42
CPN [29]	74.20 ± 0.20	85.63 ± 0.30
Model-based		
Baseline++ [4]	–	–
Negative-Cosine [41]	62.33 ± 0.82	–
S2M2 [49]	66.80 ± 0.93	68.03 ± 0.40
MGGN [84]	70.12 ± 0.75	86.53 ± 0.95
Augmentation-based		
Delta-Encoder [57]	67.83	75.30
Meta Variance Transfer [52]	72.60 ± 0.30	81.54 ± 0.25
DC [77]	74.19 ± 0.25	87.90 ± 0.14
ADC [44]	74.47 ± 0.10	88.36 ± 0.15
H-OT [21]	75.91 ± 0.35	89.33 ± 0.48
FA-adapter [62]	72.86 ± 0.50	86.60 ± 0.32
AMMD [73]	74.22 ± 0.50	87.55 ± 0.34
Ensemble FSC [31]	75.49 ± 0.52	88.38 ± 0.32
Ours	**79.35 ± 0.12**	**93.07 ± 0.33**

significantly enhances model performance. It is noteworthy that applying the Tukey transformation solely to either base classes or novel classes results in inferior performance compared to not applying it to both classes, due to a larger discrepancy between the Gaussian-like distribution and the skewed distribution. The results in the fourth row demonstrate that the negative calibration sample selection module significantly improves model performance by 1.61 % and 3.27 %. Additionally, the results in the fifth and sixth rows demonstrate that both distribution calibration and training with generated features significantly reduce the mismatch of distributions and align the feature distributions more closely with the Gaussian assumption.

Ablation studies with different backbones. Table 5 displays the performance of our model with various pre-trained backbones in 5-way 1-shot classification accuracy on the *mini*ImageNet dataset, demonstrating that our method is robust across different backbones. We select four popular backbones: four convolutional layers (conv4) [18], ResNet18 [23], WRN28 [78], and WRN28 trained with rotation loss [49]. To ensure a fair comparison, we standardize the output dimensions of all backbones. In conclusion, superior representation learning methods can significantly improve performance in subsequent classification tasks. Deeper backbone networks often yield more powerful feature representations. By incorporating self-supervision techniques into these backbones, we enhance the robustness of extracted features, leading to improved classification performance. Additionally, our method demonstrates improved accuracy when enhancing the cross-entropy-based classifier with Logistic Regression (LR) [24] compared to Support Vector Machine (SVM) [8].

Table 5
5-way 1-shot and 5-way 5-shot classification accuracy (%) on *mini*ImageNet with different backbones and classifiers. The number in **bold** indicates the best performance.

Backbones	Classifiers	*mini*ImageNet	
		5-way 1-shot	5-way 5-shot
Conv4	SVM	59.75 ± 0.35	74.37 ± 0.30
Conv4	LR	60.42 ± 0.16	75.56 ± 0.18
ResNet18	SVM	65.33 ± 0.23	80.46 ± 0.36
ResNet18	LR	66.58 ± 0.12	81.57 ± 0.40
WRN28	SVM	69.32 ± 0.21	84.61 ± 0.22
WRN28	LR	71.27 ± 0.10	85.35 ± 0.31
WRN28 + Rotation Loss	SVM	72.82 ± 0.23	86.66 ± 0.22
WRN28 + Rotation Loss	LR	**73.34 ± 0.33**	**87.95 ± 0.25**

Ablation studies for negative calibration sample selection. Fig. 4 presents ablation results on the threshold setting and temporal stability of forgetting events. As shown in the left subfigure, increasing the threshold for selecting negative calibration samples initially improves 5-way 1-shot classification accuracy, which peaks at 20 for CUB and 15 for *mini*ImageNet, before declining. To assess the temporal consistency of forgetting events, we compute the Spearman rank correlation between the forgetting events observed at epoch 150 and those at earlier epochs. The middle subfigure shows that the correlation stabilizes after 100 epochs, indicating convergence in forgetting frequency ordering.

Ablation study on the number of generated features per class. The right side of Fig. 4 examines the effect of increasing the number of generated features per class on classification accuracy in the CUB (red) and *mini*ImageNet (blue) datasets for 5-way 1-shot tasks. We observed that when the number of generated features per class exceeds 600 and 700 for the two datasets, respectively, the performance gains in classification accuracy begin to plateau.

Ablation studies for hyper-parameters. The left side of Fig. 5 shows the 5-way 1-shot classification accuracy under varying power parameters λ in Tukey's transformation. As λ increases, the test accuracy first improves and then declines, with $\lambda = 0.5$ yielding the best performance. Notably, when $\lambda = 1$, the transformation preserves the original feature representations. The right side of Fig. 5 shows the 5-way 1-shot classification accuracy with different values of the entropy regularization coefficient γ in the optimal transport strategy.

Table 4
5-way 1-shot and 5-way 5-shot classification accuracy (%) on *mini*ImageNet with different modules. Note that sample selection, and generated features are abbreviated as S.S. and G.F. in this table. The number in **bold** indicates the best performance.

Tukey Transformation		Negative Calibration S.S.	Distribution Calibration	Training with G.F.	*mini*ImageNet	
base class	novel class				5-way 1-shot	5-way 5-shot
		✓	✓	✓	72.52 ± 0.15	85.81 ± 0.31
	✓	✓	✓	✓	72.34 ± 0.22	85.67 ± 0.14
✓		✓	✓	✓	72.03 ± 0.52	85.23 ± 0.26
✓	✓		✓	✓	71.73 ± 0.33	84.68 ± 0.44
✓	✓	✓		✓	71.21 ± 0.11	83.04 ± 0.23
✓	✓	✓	✓		70.37 ± 0.42	83.69 ± 0.30
✓	✓	✓	✓	✓	**73.34 ± 0.33**	**87.95 ± 0.25**

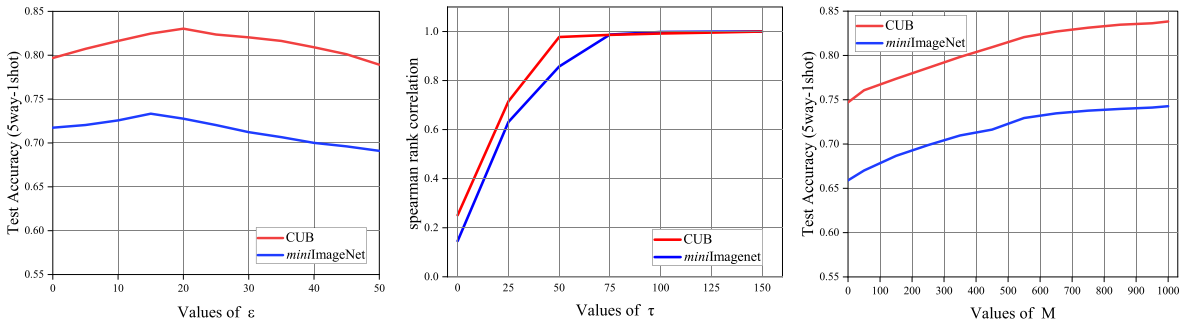


Fig. 4. Left: Classification accuracy with varying negative calibration sample selection threshold for selecting negative calibration samples. Middle: Spearman rank correlation of forgetting events across training epochs. Right: Classification accuracy with varying numbers of generated features per class.

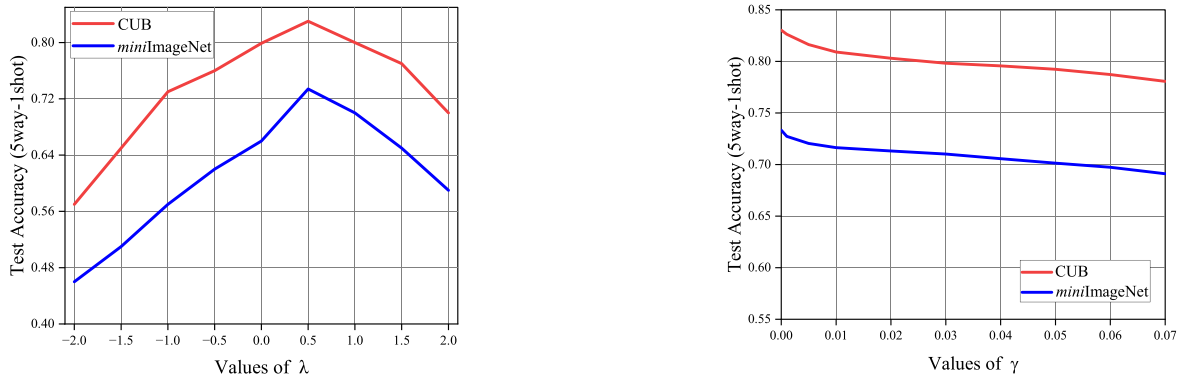


Fig. 5. Left: Classification accuracy with varying power parameters in Tukey's transformation. Right: Classification accuracy with varying entropy regularization coefficient in optimal transport.

Table 6
5-way 1-shot and 5-way 5-shot classification accuracy (%) on *mini*ImageNet with different matching distributions. The number in **bold** indicates the best performance.

Distribution	5-way 1-shot	5-way 5-shot
None-distribution	59.37 ± 0.68	81.03 ± 0.51
Laplacian distribution	68.55 ± 0.32	84.21 ± 0.41
Multimodal distribution	73.05 ± 0.33	87.81 ± 0.53
Gaussian distribution	**73.34 ± 0.33**	**87.95 ± 0.25**

As γ increases, test accuracy consistently decreases. Interestingly, $\gamma = 0$ yields the best performance without exhibiting numerical instability, confirming that explicit entropy regularization is unnecessary due to the inherent normalization and stability of the Sinkhorn updates.

Ablation studies for different matching distributions. Table 6 shows an ablation study to compare the effect of different matching distributions: including non-distribution, Laplacian distribution, multimodal distribution, and Gaussian distribution. We leverage two feature augmentation methods including rotation [58] and scaling [12] as the different modes of multimodal. Additionally, the mean and variance of the multimodal distribution are calculated as the average values of Gaussian distributions across the different modes. The model achieves the best performance with Gaussian distribution for 5-way 1-shot and 5-way 5-shot tasks, respectively. This demonstrates that the assumption of feature Gaussian distribution is rational.

Ablation studies for different distance measures. Table 7 shows an ablation study to compare the effect of different distance measures in our method: including Wasserstein distance, Kullback–Leibler divergence (KL divergence), and Euclidean distance. We observed that the accuracy is highest when utilizing the Bures-Wasserstein distance compared to the other distances due to its ability to accurately capture the position and geometric shape of pairwise Gaussian distributions of base and novel classes. Compared with the baseline method proposed by Yang et al. [77], which uses Euclidean distance, our method significantly improves model performance by 0.79 % and 1.79 % on the CUB dataset.

Robustness across lightweight network architectures. To study how network architecture affects example forgetting, we compare a simple CNN and a deeper ResNet18. We track how often each training example is forgotten (i.e., transitions from correct to incorrect classification) and rank examples by this count, designating the least-forgotten as "unforgettable." We then assess if the simple CNN identifies the same unforgettable examples as ResNet18. Using the CNN-based ranking, we compute precision and recall and consider the top-17 k least-forgotten examples according to the CNN and measure: Precision: the fraction of these examples that are also unforgettable in ResNet18; Recall: the fraction of ResNet18's unforgettable examples that appear in the top-17 k of the CNN's list. The results are visualized in Fig. 6 as a precision-recall curve. We observe a high degree of overlap between the two curves, indicating strong agreement in the sets of unforgettable examples identified by both models. This suggests that the intrinsic difficulty of learning a given example is largely independent of model architecture. Even a shallow, non-residual network captures similar learning dynamics as a deeper ResNet, particularly in identifying the simplest and most consistently learnable examples.

Sensitivity to label noise. Furthermore, we evaluate the impact of label noise on the final few-shot classification performance. As shown in Table 8, when injecting 5 % and 10 % noisy labels into the support set, the accuracy degrades by 1.37 %–2.05 %. The relatively small performance drop indicates that the overall learning framework exhibits a degree of robustness to moderate label noise.

Table 7
5-way 1-shot and 5-way 5-shot classification accuracy (%) on *mini*ImageNet with different distance measures. The number in **bold** indicates the best performance.

Dataset	Bures-Wasserstein distance		Euclidean distance		KL divergence	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
*mini*Imagenet	**73.34 ± 0.33**	**87.95 ± 0.25**	71.76 ± 0.63	86.31 ± 0.13	71.88 ± 0.53	86.62 ± 0.20
*tiered*Imagenet	**79.35 ± 0.12**	**93.07 ± 0.33**	78.75 ± 0.66	92.33 ± 0.15	78.64 ± 0.53	92.64 ± 0.30
CUB	**83.03 ± 0.46**	**94.71 ± 0.11**	82.24 ± 0.25	92.92 ± 0.55	82.43 ± 0.46	91.27 ± 0.32

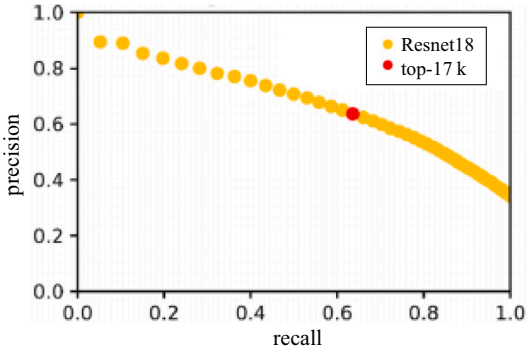


Fig. 6. Precision-Recall curve for retrieving ResNet18's unforgettable examples using a simpler CNN's ordering.

Table 8
Effect of random label noise on classification accuracy (%) in *mini*ImageNet.

Noise Level	*mini*Imagenet	
	5-way 1-shot	5-way 5-shot
0 % (Clean)	**73.34 ± 0.33**	**87.95 ± 0.25**
5 %	72.70 ± 0.18	86.94 ± 0.29
10 %	71.97 ± 0.29	85.90 ± 0.36

4.5. Computational complexity

Chizat et al. [7] compute the time complexity bound of the general optimal transport (OT) distance between two discrete distributions of size m as $m^2 \log(m)/\varepsilon^2$ to reach ε-accuracy with Sinkhorn's algorithm. In this paper, we approximate the OT distance between the distribution P consisting of B samples for base classes, and distribution Q consisting of $N \times K$ samples for the N-way-K-shot task, with the condition that $B > N \times K$. Thus, the time complexity for using Sinkhorn's algorithm to approximate the OT distance is $O(B^2 \log(B)/\varepsilon^2)$ to achieve ε-accuracy. In Table 9, we compare the computational cost of the DC [77], H-OT [21], and our algorithm on a single NVIDIA TITAN P8 GPU. We observed that our model's computational cost lies between that of the DC [77] method and H-OT [21] across three datasets. While our model incurs a higher computational cost than the DC [77] due to the introduction of OT, it is lower than that of H-OT [21] thanks to pre-screening negative calibration samples. Moreover, our model outperforms both methods while maintaining acceptable costs. Therefore, our model delivers superior performance with acceptable overhead on large-scale datasets with more samples.

4.6. Visualization

Visualization plays a crucial role in understanding our method. In this section, we show the visualization of the adaptive cost function and the associated transport plan, visualization of generated samples, distributions of forgetting events across training examples, and negative/positive calibration samples as follows.

4.6.1. Visualization of adaptive cost and transport plan

To further examine whether our method can capture the similarity between the base classes and novel classes, Fig. 7(a) depicts the heatmap of transport probability matrix on a randomly selected task in *mini*Imagenet, while Fig. 7(b) shows the heatmap of the transport cost for the same task. We note that the transport probability matrix between a base class and a novel class is usually larger when their transport cost is smaller. As classes become more similar, the weight value in the transport probability matrix increases, which can reflect the different contributions of the base classes. This approach facilitates the transfer of statistical information from base classes to novel classes in an adaptive manner.

4.6.2. Visualization of generated samples

In this section, we randomly select a 5-way 1-shot task sampled from the *mini*ImageNet dataset to visualize the t-SNE [48] representations of the features generated by our method alongside the ground-truth distributions. In Fig. 8(a), it becomes evident that the support set consists of only five samples for the 5-way-1-shot scenario, highlighting the limited data available for training. Fig. 8(b) shows the feature distributions generated without negative calibration sample selection, illustrating that adopting an equal distribution transform could result in negative calibration, which significantly deviates from the ground-truth distribution, compared to Fig. 8(d). In Fig. 8(c), we can observe that colors blue and green represent the negative calibration samples selected by our approach, which are set to undergo no transformation. The other three samples adopt distribution transformation. As a result, the distribution of the generated samples closely resembles the ground-truth distribution. Notably, Fig. 8 illustrates that our method not only accurately estimates the ground-truth distribution but also effectively prevents the negative calibrations by negative calibration sample selection in advance, significantly enhancing the model's generalization capabilities.

Table 9
Comparison of computational cost when testing 5-way 1-shot and 5-way 5-shot classification tasks, where s denotes seconds.

Dataset	DC		**H-OT**		Our method	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
*mini*Imagenet	2.34*s*	8.41*s*	2.91*s*	10.06*s*	2.73*s*	8.03*s*
*tiered*Imagenet	5.76*s*	24.61*s*	9.31*s*	41.12*s*	8.12*s*	32.55*s*
CUB	2.11*s*	8.24*s*	2.60*s*	11.12*s*	2.21*s*	10.18*s*

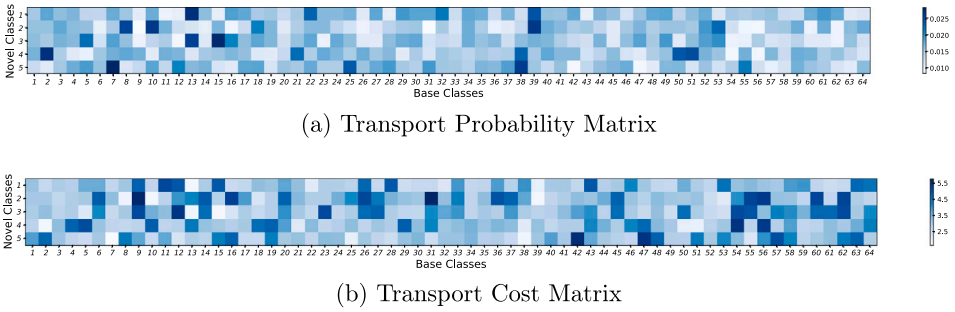


Fig. 7. Heatmap of transport probability matrix (a) and transport cost matrix (b) on a randomly selected task on *mini*Imagenet. The lighter the color is, the smaller the value is.

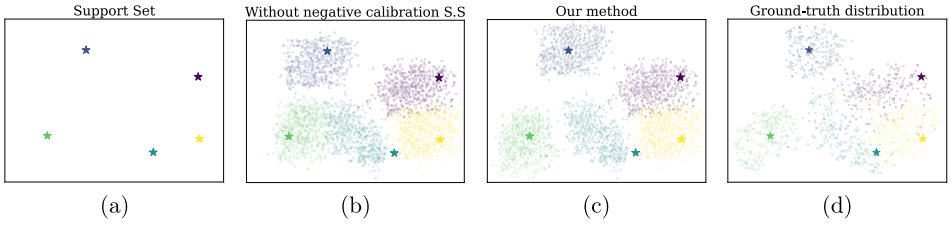


Fig. 8. The t-SNE visualization of feature distributions learned by our method and ground-truth distributions on a random 5-way-1-shot task sampled from *mini*ImageNet. ★ represents support set features, • in figure (b) represents the generated features without negative calibration sample selection. Note that sample selection are abbreviated as S.S. ▲ in figure (c) represents the generated features by our method, × in figure (d) represents the features of ground-truth distributions.

5. Analysis and limitations

In this section, we will discuss the applicability and the limitations of our method.

Applicability of our method. We demonstrate that our model, which assumes that samples from the same class follow a Gaussian distribution, achieves strong performance in image classification tasks. However, while we have not explored the validity of the Gaussian distribution assumption for other datasets, our learning framework is flexible and can be applied to any distribution, such as Laplacian or multi-modal distributions. Although our current work is built on a Gaussian assumption, our framework is adaptable and can be extended to other distribution types. This assumption allows our method to be applicable to other image-based computer vision tasks, including object detection, action recognition, and anomaly detection.

Challenges and limitations of the proposed approach. Although our model reduces computational overhead through the negative calibration sample selection strategy, the inclusion of optimal transport techniques still results in higher computational costs compared to standard methods. This presents challenges, especially in large-scale datasets or resource-constrained environments, such as edge devices with limited computational power. Moreover, the proposed method assumes that the distribution of each class follows a Gaussian distribution. We apply Tukey's Ladder of Powers transformation to reduce skewness and enhance Gaussian-like properties. While the method has shown effectiveness in computer vision tasks, its applicability to natural language processing (NLP) remains uncertain due to the distinct characteristics of each modality. Further research is required to evaluate its performance in the NLP domain.

Flexible thresholding for negative calibration samples. The threshold ε, used to identify negative calibration samples, is empirically determined on a per-dataset basis according to validation performance, providing a practical balance between stability and interpretability. Nonetheless, a more adaptive strategy, such as a task-specific or quantile-based thresholding scheme, could further improve generalizability across diverse datasets. Further research is required to develop and evaluate adaptive thresholding mechanisms for negative calibration sample selection.

6. Conclusion

Building upon the aforementioned observations on calibration shift and distribution similarity, we propose a distributional fusion framework. First, we systematically delve into the negative calibration issues and present an approach to mitigate adverse effects on distribution calibration, facilitating positive transformation gains in the process by pre-screening negative calibration samples based on the frequency of forgetting events. Additionally, our method dynamically adjusts the estimation of novel class distributions using an optimal transport strategy. Specifically, we utilize the Bures-Wasserstein distance to accurately capture the position and shape of pairwise Gaussian distributions of base and novel classes. The experimental results demonstrate that our method surpasses the existing state-of-the-art by margins of 2.77 %–4.69 % on the *mini*ImageNet, *tiered*ImageNet, and CUB datasets.

CRediT authorship contribution statement

Juan Zhao: Writing – review & editing, Writing – original draft. **Lili Kong:** Data curation. **Chenwei Tang:** Data curation. **Wei Ju:** Funding acquisition. **Deng Xiong:** Methodology. **Jiancheng Lv:** Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jiancheng Lv reports financial support was provided by National Major Scientific Instruments and Equipment Development Project of National Natural Science Foundation of China. Chenwei Tang reports financial support was provided by Natural Science Foundation of Sichuan. Jiancheng Lv reports financial support was provided by Fundamental Research Funds for the Central Universities. Chenwei Tang reports financial support was provided by Tianfu Yongxing Laboratory Organized

Research Project Funding. Jiancheng Lv reports financial support was provided by Science Fund for Creative Research Groups of Sichuan Province Natural Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the Fundamental Research Funds for the Central Universitiesunder Grant 1082204112364, National Major Scientific Instruments and Equipment Development Project of National Natural Science Foundation of China under Grant 62427820, Science Fund for Creative Research Groups of Sichuan Province Natural Science Foundation 2024NSFTD0035, and Natural Science Foundation of Sichuan 24NSFSC3404.

Data availability

Data will be made available on request.

References

[1] R. Bhatia, T. Jain, Y. Lim, On the bures–wasserstein distance between positive definite matrices, Expo. Math. 37 (2019) 165–191.
[2] P. Chakraborty, M. Alfadel, M. Nagappan, Blaze: cross-language and cross-project bug localization via dynamic chunking and hard example learning, IEEE Trans. Softw. Eng. (2025).
[3] L. Chen, Y. Gu, Y. Guo, F. Dong, D. Jiang, Y. Chen, Ddc: dynamic distribution calibration for few-shot learning under multi-scale representation, Knowl.-Based Syst. 311 (2025) 113030.
[4] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C.F. Wang, J.-B. Huang, A closer look at few-shot classification, arXiv preprint arXiv:1904.04232, 2019.
[5] N. Cheng, J. Xu, C. Guan, J. Gao, W. Wang, Y. Li, F. Meng, J. Zhou, B. Fang, W. Han, Touch100k: a large-scale touch-language-vision dataset for touch-centric multimodal representation, Inf. Fusion (2025) 103305.
[6] Z. Chi, Z. Wang, M. Yang, D. Li, W. Du, Learning to capture the query distribution for few-shot learning, IEEE Trans. Circuits Syst. Video Technol. 32 (2021) 4163–4173.
[7] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, G. Peyré, Faster Wasserstein distance estimation with the sinkhorn divergence, Adv. Neural Inf. Process. Syst. 33 (2020) 2257–2269.
[8] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[9] F. Cui, Y. Zhang, X. Wang, X. Tian, J. Yu, Enhancing target-unspecific tasks through a features matrix, arXiv preprint arXiv:2505.03414, 2025.
[10] M. Cuturi, Sinkhorn distances: lightspeed computation of optimal transport, Adv. Neural Inf. Process. Syst. 26 (2013).
[11] D.C. Dowson, B. Landau, The fréchet distance between multivariate normal distributions, J. Multivar. Anal. 12 (1982) 450–455.
[12] L.L. Duan, J.E. Johndrow, D.B. Dunson, Scaling up data augmentation MCMC via calibration, J. Mach. Learn. Res. 19 (2018) 2575–2608.
[13] Y. Duan, L. Niu, Y. Hong, L. Zhang, Weditgan: few-shot image generation via latent space relocation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 1653–1661.
[14] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1126–1135.
[15] Y.-G. Fu, X. Chen, S. Xu, J. Li, X. Yao, Z. Huang, Y.-M. Wang, Gsscl: a framework for graph self-supervised curriculum learning based on clustering label smoothing, Neural Networks 181 (2025) 106787.
[16] P. Ganesan, S.K. Jagatheesaperumal, M.M. Hassan, F. Pupo, G. Fortino, Few-shot image classification using graph neural network with fine-grained feature descriptors, Neurocomputing 610 (2024) 128448.
[17] R. Gao, H. Su, S. Prasad, P. Tang, Few-shot classification with multisemantic information fusion network, Image Vis. Comput. 141 (2024) 104869.
[18] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2414–2423.
[19] C. Gong, T. Ren, M. Ye, Q. Liu, Maxup: lightweight adversarial training with data augmentation improves neural network training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2474–2483.
[20] D. Guan, Y. Xing, J. Huang, A. Xiao, A. El Saddik, S. Lu, S2match: self-paced sampling for data-limited semi-supervised learning, Pattern Recognit. 159 (2025) 111121.
[21] D. Guo, L. Tian, H. Zhao, M. Zhou, H. Zha, Adaptive distribution calibration for few-shot learning with hierarchical optimal transport, Adv. Neural Inf. Process. Syst. 35 (2022) 6996–7010.
[22] A. Han, B. Mishra, P. Jawanpuria, J. Gao, Generalized bures-wasserstein geometry for positive definite matrices, arXiv preprint arXiv:2110.10464, 2021.
[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[24] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, John Wiley & Sons, 2013.
[25] M. Hou, K. Hindriks, A.E. Eiben, K. Baraka, Active robot curriculum learning from online human demonstrations, in: 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2025, pp. 810–818.
[26] R. Hou, H. Chang, B. Ma, S. Shan, X. Chen, Cross attention network for few-shot classification, Adv. Neural Inf. Process. Syst. 32 (2019).
[27] Y. Huang, H. Hao, W. Ge, Y. Cao, M. Wu, C. Zhang, J. Guo, Relation fusion propagation network for transductive few-shot learning, Pattern Recognit. 151 (2024) 110367.
[28] X. Ji, X. Cheng, P. Zhou, Self-paced learning for anchor-based multi-view clustering: a progressive approach, Neurocomputing 635 (2025) 129921.
[29] M. Jiang, J. Fan, J. He, W. Du, Y. Wang, F. Li, Contrastive prototype network with prototype augmentation for few-shot classification, Information Sciences 686 (2025a) 121372.
[30] W. Jiang, G. Liu, D. He, K. He, Boosting meta-training with base class information for robust few-shot learning, Eng. Appl. Artif. Intell. 152 (2025b) 110780.
[31] Z. Jiang, N. Tang, J. Sun, Y. Zhan, Combining various training and adaptation algorithms for ensemble few-shot classification, Neural Networks (2025c) 107211.
[32] H. Lane, M. Dyshel, Natural Language Processing in Action, Simon and Schuster, 2025.
[33] B. Li, C. Liu, M. Shi, X. Chen, X. Ji, Q. Ye, Proposal distribution calibration for few-shot object detection, IEEE Trans. Neural Netw. Learn. Syst. (2023).
[34] J. Li, C. Ye, F. Wang, J. Pan, A robust transductive distribution calibration method for few-shot learning, Pattern Recognit. 163 (2025a) 111488.
[35] M. Li, J. Jiang, H. Yao, Drop inherent biases: multi-level attention calibration for robust cross-domain few-shot classification, Neurocomputing 636 (2025b) 130056.
[36] M. Li, H. Yao, Fmvp: fine-grained meta visual prompt enabled domain-specific few-shot classification, Neurocomputing 633 (2025) 129688.
[37] Y. Li, L. Chen, W. Li, N. Wang, Few-shot fine-grained classification with rotation-invariant feature map complementary reconstruction network, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–12.
[38] Z. Li, F. Zhou, F. Chen, H. Li, Meta-sgd: Learning to learn quickly for few-shot learning, arXiv preprint arXiv:1707.09835, 2017.
[39] J.Y. Lim, K.M. Lim, C.P. Lee, Y.X. Tan, A review of few-shot image classification: approaches, datasets and research trends, Neurocomputing (2025) 130774.
[40] Z. Lin, W. Yang, H. Wang, H. Chi, L. Lan, J. Wang, Scaling few-shot learning for the open world, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 13846–13854.
[41] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, H. Hu, Negative margin matters: understanding margin in few-shot classification, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer, 2020, pp. 438–455.
[42] L. Liu, Y. Liang, X. Yan, L. Huangfu, S. Samtani, Z. Yu, Y. Zhang, D.D. Zeng, Hard sample mining: a new paradigm of efficient and robust model training, IEEE Trans. Neural Netw. Learn. Syst. (2025a).
[43] T. Liu, C. Wen, Q. Xiong, J. Li, Meta transfer evidence deep learning for trustworthy few-shot classification, Expert Syst. Appl. 259 (2025b) 125371.
[44] X. Liu, K. Zhou, P. Yang, L. Jing, J. Yu, Adaptive distribution calibration for few-shot learning via optimal transport, Inf. Sci. 611 (2022) 1–17.
[45] Y. Liu, M. Li, F. Giunchiglia, L. Huang, X. Li, X. Feng, R. Guan, Dual-level mixup for graph few-shot learning with fewer tasks, in: Proceedings of the ACM on Web Conference 2025, 2025c, pp. 2646–2656.
[46] L. Luo, Z. Yang, Z. Chen, R. Cao, Y. Liu, Gprn: gan-based prototype refinement network for few-shot learning, Neural Comput. Appl. (2025) 1–24.
[47] R. Ma, P. Fang, T. Drummond, M. Harandi, Adaptive poincaré point to set distance for few-shot classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 1926–1934.
[48] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (2008).
[49] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, V.N. Balasubramanian, Charting the right manifold: manifold mixup for few-shot learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2218–2227.
[50] A. Marini, R. Genereux, The challenge of teaching for transfer, in: Teaching for Transfer, Routledge, 2013, pp. 1–19.
[51] I. Olkin, F. Pukelsheim, The distance between two random vectors with given dispersion matrices, Linear Algebra Appl. 48 (1982) 257–263.
[52] S.-J. Park, S. Han, J.-W. Baek, I. Kim, J. Song, H.B. Lee, J.-J. Han, S.J. Hwang, Meta variance transfer: learning to augment from the others, in: International Conference on Machine Learning, PMLR, 2020, pp. 7510–7520.
[53] H. Ran, C. Jia, X. Li, Z. Zhang, Few-shot learning with distribution calibration for event-level rumor detection, Neurocomputing 618 (2025) 129034.
[54] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J.B. Tenenbaum, H. Larochelle, R.S. Zemel, Meta-learning for semi-supervised few-shot classification, arXiv preprint arXiv:1803.00676, 2018.
[55] A.A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, R. Hadsell, Meta-learning with latent embedding optimization, arXiv preprint arXiv:1807.05960, 2018.
[56] A. Santoro, D. Raposo, D.G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, Adv. Neural Inf. Process. Syst. 30 (2017).
[57] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, A. Kumar, R. Feris, R. Giryes, A. Bronstein, Delta-encoder: an effective sample synthesis method for few-shot object recognition, Adv. Neural Inf. Process. Syst. 31 (2018).

[58] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) 1–48.
[59] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 761–769.
[60] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Adv. Neural Inf. Process. Syst. 30 (2017).
[61] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, A. Morcos, Beyond neural scaling laws: beating power law scaling via data pruning, Adv. Neural Inf. Process. Syst. 35 (2022) 19523–19536.
[62] J. Sun, J. Li, Few-shot classification with fork attention adapter, Pattern Recognition 156 (2024) 110805.
[63] Q. Sun, Y. Liu, Z. Chen, T.-S. Chua, B. Schiele, Meta-transfer learning through hard tasks, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2020) 1443–1456.
[64] Z. Sun, W. Zheng, P. Guo, M. Wang, Tst_mfl: two-stage training based metric fusion learning for few-shot image classification, Inf. Fusion. 113 (2025) 102611.
[65] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H.S. Torr, T.M. Hospedales, Learning to compare: relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.
[66] W. Tang, H. Pei, X. Wang, Z. He, L. Yu, X. Yang, Reducing hubness to improve inductive few-shot learning, Neurocomputing (2025) 130879.
[67] M. Toneva, A. Sordoni, R.T.D. Combes, A. Trischler, Y. Bengio, G.J. Gordon, An empirical study of example forgetting during deep neural network learning, arXiv preprint arXiv:1812.05159, 2018.
[68] J.W. Tukey, et al., Exploratory Data Analysis, vol. 2, Reading, MA, 1977.
[69] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, Adv. Neural Inf. Process. Syst. 29 (2016).
[70] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-Ucsd birds-200-2011 dataset (2011).
[71] Q. Wang, Z. Chen, K. Huang, X. Su, C. Yang, C. Xu, Concm: Consistency-driven calibration and matching for few-shot class-incremental learning, arXiv preprint arXiv:2506.19558, 2025.
[72] X. Wei, W. Du, H. Wan, W. Min, Feature distribution fitting with direction-driven weighting for few-shot images classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 10315–10323.
[73] J. Wu, S. Wang, J. Sun, Ammd: attentive maximum mean discrepancy for few-shot image classification, Pattern Recognition 155 (2024) 110680.
[74] X. Xiao, Y. Zhang, Y. Li, X. Li, T. Wang, J. Hamm, X. Wang, M. Xu, Visual variational autoencoder prompt tuning, arXiv preprint arXiv:2503.17650, 2025.
[75] J. Xu, B. Wang, H. Wan, P. Su, X. Wei, Adaptive blank compensation for few-shot image classification, Neurocomputing 637 (2025) 130127.
[76] Z. Yan, Y. An, H. Xue, Reinforced self-supervised training for few-shot learning, IEEE Signal Process. Lett. (2024).
[77] S. Yang, L. Liu, M. Xu, Free lunch for few-shot learning: Distribution calibration, arXiv preprint arXiv:2101.06395, 2021.
[78] S. Zagoruyko, N. Komodakis, Wide residual networks, arXiv preprint arXiv:1605.07146, 2016.
[79] Y. Zahid, C. Zarges, B. Tiddeman, J. Han, Adversarial diffusion for few-shot scene adaptive video anomaly detection, Neurocomputing 614 (2025) 128796.
[80] C. Zhang, W. Fan, B. Wang, C. Chen, H. Li, Self-paced semi-supervised feature selection with application to multi-modal alzheime's disease classification, Inf. Fusion. 107 (2024a) 102345.
[81] L. Zhang, L. Zuo, Y. Du, X. Zhen, Learning to adapt with memory for probabilistic few-shot learning, IEEE Trans. Circuits Syst. Video Technol. 31 (2021) 4283–4292.
[82] Y. Zhang, B. Liu, J. Bao, Q. Huang, M. Zhang, J. Yu, Learnability matters: active learning for video captioning, Adv. Neural Inf. Process. Syst. 37 (2024b) 37928–37954.
[83] J. Zhao, L. Kong, J. Lv, An overview of deep neural networks for few-shot learning, Big Data Min. Anal. 8 (2024) 145–188.
[84] P. Zheng, X. Guo, E. Chen, L. Qi, L. Guan, Edge-labeling based modified gated graph network for few-shot learning, Pattern Recognition 150 (2024) 110264.
[85] P. Zhou, X. Wang, L. Du, Bi-level ensemble method for unsupervised feature selection, Inf. Fusion. 100 (2023) 101910.

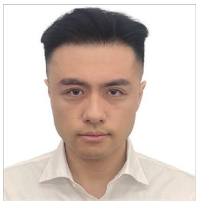
Author biography

Juan Zhao received the MEng degree in Software Engineering from ChongQing University, ChongQing, China in 2019. She is currently a PhD candidate at the College of Computer Science, Sichuan University, Chengdu, China. Her research focuses on the Few-Shot Learning in neural networks.

Lili Kong received the M.S. degree in College of Computer Science, Sichuan University, Chengdu, China, in 2018. She is currently pursuing the Ph.D. degree with the Data Intelligence and Computing Art Laboratory, College of Computer Science, Sichuan University, Chengdu. Her research focuses on the small data learning in neural networks.

Chenwei Tang received the PhD degree in computer science and technology from Sichuan University, Chengdu, China in 2020. She is currently a postdoctoral researcher at the College of Computer Science, Sichuan University, Chengdu, China. Her research focuses on the neural network methods for zero-shot learning, computer vision, and industrial intelligence.

Wei Ju is currently an associate professor with the College of Computer Science, Sichuan University, Chengdu, China. Prior to that, he worked as a postdoc research fellow and received his Ph.D. degree in the School of Computer Science from Peking University, Beijing, China, in 2022. He received the B.S. degree in Mathematics from Sichuan University, Sichuan, China, in 2017. His current research interests lie primarily in the area of machine learning on graphs including graph representation learning and graph neural networks, and interdisciplinary applications such as bioinformatics, drug discovery, recommender systems and knowledge graphs. He has published more than 90 papers in top-tier venues and has won the best paper finalist in IEEE ICDM 2022.



Deng Xiong received his M.S. in Computer Science from Stevens Institute of Technology, NJ, USA, in 2022, and his M.Eng. in Mechanical Engineering from Stevens Institute of Technology, NJ, USA, in 2017. He served as both a Teaching Assistant and Research Assistant in the Department of Mechanical Engineering. Currently, he is working as an independent researcher. His research interests encompass artificial intelligence, big data, cloud computing, database, robotics, soft materials, and additive manufacturing.

Jiancheng Lv received the PhD degree in computer science and engineering from the University of Electronic Science and Technology of China, Chengdu, China in 2006. He was a research fellow at the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He is currently a professor at the Data Intelligence and Computing Art Laboratory, College of Computer Science, Sichuan University, Chengdu, China. His research interests include neural networks, machine learning, and big data.