IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 March 2024, accepted 6 April 2024, date of publication 9 April 2024, date of current version 18 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3386808

RESEARCH ARTICLE

NegCosIC: Negative Cosine Similarity-Invariance-Covariance Regularization for Few-Shot Learning

WEI HAN LIU, KIAN MING LIM, (Senior Member, IEEE), THIAN SONG ONG, (Senior Member, IEEE), AND CHIN POO LEE, (Senior Member, IEEE)
Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia

Corresponding author: Kian Ming Lim (kmlim@mmu.edu.my)

This work was supported by the Malaysian Ministry of Higher Education through the Fundamental Research Grant Scheme under Grant FRGS/1/2021/ICT02/MMU/02/4.

ABSTRACT Few-shot learning continues to pose a challenge as it is inherently difficult for visual recognition models to generalize with limited labeled examples. When the training data is limited, the process of training and fine-tuning the model will be unstable and inefficient due to overfitting. In this paper, we introduce NegCosIC: Negative Cosine Similarity-Invariance-Covariance Regularization, a method that aims to improve the mean accuracy from the perspective of stabilizing the fine-tuning process and regularizing variance. NegCosIC incorporates a negative simple cosine similarity loss to stabilize the parameters of the feature extractor during fine-tuning. In addition, NegCosIC integrates invariance loss and covariance loss to regularize the embeddings in order to reduce overfitting. Experimental results demonstrate that NegCosIC is able to bring substantial improvements over the current state-of-the-art methods. An in-depth worse case analysis is also conducted and shows that NegCosIC is able to outperform state-of-the-art methods on worst case accuracy. The proposed NegCosIC achieved 2.15% and 2.13% higher accuracy on miniImageNet 1-shot and 5-shot tasks, 3.22% and 2.67% higher accuracy on CUB 1-shot and 5-shot tasks, and 2.13% and 7.74% higher accuracy on CIFAR-FS 1-shot and 5-shot tasks in terms of worst-case accuracies.

INDEX TERMS Few-shot learning, negative cosine similarity, invariance, covariance, regularization.

I. INTRODUCTION

Over the years, the field of deep learning has seen significant advancements in standard computer vision tasks, such as object recognition. However, there remain many challenges in maintaining high accuracy when training data is limited. For example, in the field of medical imaging, labeled training data is difficult to obtain. This leads to the issue of overfitting that affects the robustness and generalization of the model [1], [2], [3]. Although there are different augmentation methods based on generative models and image transformations such as rotation and scaling, they are not efficient because their effectiveness depends on the specific sub-domains. This prompts researchers to explore the prospect of few-shot learning. The goal of few-shot learning is to identify new objects with just a small number of training examples per class, which closely resembles real-world situations where obtaining labeled data could be challenging and costly. Similar to human intelligence that is capable of learning based on a few examples, developing machine learning models that are able to learn with a few training samples across different classes is also an important step to advance artificial intelligence in general.

Common methods to mitigate the challenge of few-shot learning are based on the "learning to learn" mechanism or meta-learning. In meta-learning, the model is trained based on a series of different few-shot classification tasks and subsequently evaluated on test data in order to learn

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

parameters that easily generalize to new tasks [4], [5], [6]. In recent years, another branch of few-shot learning methods called metric-based methods have garnered more interests from researchers due to their better performance over other few-shot learning methods. In general, many metric-based methods use a feature extractor pre-trained on the base classes for feature extraction, and then train a classifier based on a selected metric to compute the differences between feature embeddings of test data for classification. Some of the popular metric-based methods are: matching networks [7] that compared the query features with support features based on cosine distance and memory mechanism; prototypical networks [8] that compared the query features with the embedding prototype of support features from each class with Euclidean distance; relation networks [9] that examined the query features with the embedding prototype of support features from each class using a relation module where its parameters were fine-tuned.

Despite making some progress in few-shot learning, researchers still find ways to rethink the efficiency and the way evaluations are made in the field. When the training data is limited, the process of training and fine-tuning the model will be unstable and inefficient due to overfitting. In view of this, we introduce a new few-shot learning framework that incorporates a negative cosine similarity loss together with invariance loss and covariance loss during the training of few-shot classifiers. Negative cosine similarity loss improves the model by preventing the fine-tuned representations from deviating significantly from the original learned representations during the fine-tuning phase of feature extractor. Meanwhile, invariance loss and covariance loss improve the model by regularizing the variance of the feature representations. This in turn will boost the performance of the model.

In addition, [10] recently argued that the usual evaluation metrics, namely mean accuracy and 95% confidence interval are not practical because they are the average accuracy over a series of different episodes of few-shot tasks. The authors showed that the accuracy of individual episode varies a lot and is unstable, with the worst being 37.33% accuracy and the best being close to 100%, while the average accuracy was 68.96% and the 95% confidence interval was [68.07, 69.85]. In most real world applications, it is not practical to run many experiments on few-shot tasks and pick the best or average episode among them, which is the way few-shot learning models are currently evaluated in the field. Therefore, it is more important to focus on one episode instead of the average of many episodes. Thus, we also adopted worst case accuracy additionally in evaluating our model during the experiments.

The main contributions of this paper are summarized as follows:

1) To stabilize the parameters of the feature extractor during fine-tuning, a negative cosine similarity loss is introduced in order to encourage the original learned representations to remain similar to the fine-tuned representations.
2) In addition, to reduce overfitting, invariance loss and covariance loss are incorporated to regularize the variance of the learned feature representations. With features that are more robust, we then use a cosine classifier to make predictions for the few-shot tasks.
3) Through extensive experiments, we show that the proposed NegCosIC is able to achieve higher average accuracy on few-shot recognition datasets.
4) We also conducted an intensive worse case analysis and show the proposed NegCosIC is able to achieve significantly higher worst case accuracy on few-shot recognition datasets.

II. RELATED WORK

Conventionally, few-shot learning is commonly done in an inductive manner. The model is first trained on a set of train data, and subsequently evaluated and used on other test data without the need of leveraging other unlabeled data for fine-tuning. Generally, few-shot learning is categorized into 3 groups, which are gradient-based methods, hallucination-based methods and metric-based methods.

A. GRADIENT-BASED FEW-SHOT CLASSIFICATION

Gradient-based methods aim to fine-tune the model based on a small number of data samples to address the few-shot learning problems [4], [5], [6], [11], [12], [13], [14], [15], [16], [17]. These methods can be categorized into two types: initialization-based methods [5], [6], [11], [13], [14] and optimization-based methods [4], [12], [15], [16], [17]. Initialization-based methods learn a good initialization for the model's parameters across multiple tasks. By doing so, the model is able to perform well in new tasks with few data samples and parameter updates. For example, Model-Agnostic Meta-Learning (MAML) [5] aimed to find optimal parameters for the model based on the loss from a set of tasks, making the fine-tuning process for new tasks more efficient.

In contrast, optimization-based methods aim to learn a good optimizer that allows the model to adapt rapidly to new tasks with few data samples and parameter updates. These methods often replace the standard optimizer with a new kind of optimizer, such as a Long Short-Term Memory-based meta-learner [4] or a mechanism with external memory for updating the parameters [12]. Recently, an end-to-end framework called GCLR-SVM [15] was proposed to embed representations into a latent space and enhances the representations by reconstructing the latent codes using variational information. In addition, the authors in [16] introduced a Multi-level Second-order (MISo) network that included a Second-order Pooling (SoP) and Feature Matching (FM) module to reweight their respective branches for relation learning. They also added a self-supervised discriminator to improve the distinctiveness of the representation by predicting the index of abstraction and scale. Moreover, [17] proposed a new learning paradigm A-MET that adaptively removes unwanted and incomplete features learned during the pre-training stage and tackles the objective discrepancy

between the transfer learning and meta-learning. They also introduced a GSCM metric that represent samples by jointly re-embedding the sample features to get prediction results that are more consistent. Reference [18] proposed Adaptive Learning Knowledge Networks (ALKN) that has an adaptive learning knowledge module to store the memory of learned knowledge and a decoder that ingests the query representation and the data from the adaptive learning knowledge module for classification.

B. HALLUCINATION-BASED FEW-SHOT CLASSIFICATION

In few-shot learning, data deficiency is a common issue. To address this, hallucination-based techniques have been proposed in recent years. These methods, as described in [19], [20], [21], and [22], aim to augment the available data to produce more training samples. The techniques can be divided into two categories: the first type transfers appearance variations from the initial data categories, as seen in [19] and [21], while the second type uses generative adversarial networks (GANs) to transfer the style, as seen in [20]. These techniques are often used together with other few-shot learning methods, resulting in increased complexity. Reference [22] proposed to transform the base classes into Gaussian form with power transformation for Maximum A Posteriori (MAP) estimation. After that, the Gaussian mean of the novel classes are estimated under the Gaussian prior given few samples from it. In the end, every novel classes is represented by a unique Gaussian distribution where sufficient trainable features can be sampled and this in turn improves the prediction.

C. METRIC-BASED FEW-SHOT CLASSIFICATION

Metric-based methods have gained progressively more attention in recent literature as they are the most effective few-shot learning techniques. These methods learn to differentiate between objects using limited examples by leveraging the information about the similarity of the limited available data. Typically, a CNN-based feature extractor is first trained on a larger dataset. The feature extractor is then used to extract features from the limited data of novel classes. Subsequently, a metric-based classifier is trained to recognize the objects given the features. The metric can be any of the following: cosine similarity [7], Euclidean distance [8], custom convolutional neural network-based distance module [9], [23], [24], or graph neural network [25], [26], [27].

Matching Network [7] used an end-to-end weighted nearest neighbor classifier based on an attention mechanism that utilized cosine similarity of two feature embeddings. Next, Prototypical Network [8] calculated the mean of the extracted features of the support data and compared the Euclidean distance between the class mean and the query data for classification. In addition, Relation Network [9] concatenated the feature maps of the training set and passed them into a relation module that contains the score for each class and converted them into one-hot vectors later. The relation module was optimized through mean square error (MSE) by regressing the value of the score to the true label. Moreover, Task Dependent Adaptive Metric (TADAM) [23] introduced a dynamic task-conditioning module that extracts task representations, which were used to improve the feature extractor. They also applied metric scaling and auxiliary task co-training to improve the few-shot learning algorithm. Recently, DeepEMD [24] adopted Earth Mover's Distance (EMD) to generate the minimum matching cost between the feature vectors of the support and query images for few-shot classification. Reference [28] proposed a multi-scale decision network (MSDN) that utilizes feature fusion and feature weighting to enhance the fitting ability of Relation Network during feature concatenation.

On the other hand, the authors of [25] formulated a graph neural network (GNN) framework for few-shot learning, where the extracted features act were used as an input to a GNN that consists of different layers of nodes and graph convolutional layers. Reference [26] improved on [25] by introducing Edge Graph Neural Network (EGNN) to predict the edge-label on the graph based on the similarity within the cluster and dissimilarity between different clusters. Additionally, Distribution Propagation Graph Network (DPGN) [27] introduced a dual complete network made of a point graph and a distribution graph. The label information was then propagated from labeled data to unlabeled data for a number of updates.

D. TRANSDUCTIVE FEW-SHOT CLASSIFICATION

Transductive few-shot learning is a subfield of metric-based few-shot learning. It has been shown to achieve significant improvement over other few-shot learning methods such as inductive metric-based methods, gradient-based methods, and hallucination-based methods, as evidenced in recent studies [29], [30], [31], [32], [33], [34], [35], [36], [37]. In inductive few-shot learning settings, models are first trained on observed and labeled training data, and subsequently used to make predictions on unobserved and unlabeled test data. In contrast, transductive few-shot learning models are trained using both observed and labeled training data, as well as observed and unlabeled test data, and then used to classify the test data.

The Transductive Propagation Network (TPN) [31] was the first work to explicitly model transductive inference in few-shot learning settings. TPN proposed a framework for learning to propagate labels between data instances for unseen classes through episodic meta-learning. In [32], a simple method was proposed that minimizes the entropy of model predictions on unlabeled query samples, which surprisingly achieved competitive performance over complex meta-learning methods. Another study [33] proposed using pseudo-labeling and feature shifting in a prototypical network based on cosine similarity. PT-MAP [34] applied Power Transform (PT) to the data to better align it with typical distri-

bution assumptions, and used Maximum A Posteriori (MAP) to compute class centers for classification. Reference [36] obtained a regularized manifold by leveraging the unlabeled query data and using non-parametric embedding propagation to smooth decision boundaries by outputting a set of features interpolations based on a similarity graph. Similarly, [29] performed parameterless feature fusion between the query data and the support data to propagate information across features for better feature representations. Later, [35] proposed to minimize a quadratic binary-assignment function, which achieved competitive performance. The function contains a unary term assigning query samples to the nearest class prototype and a pairwise Laplacian term. In doing so, it encouraged nearby query samples to have consistent label assignments. Additionally, [30] proposed a method that maximizes mutual information between the query features and predictions of a few-shot task while subjecting to supervision constraints from the support set. In another study [37], a transductive clustering procedure based on a conditional neural-adaptive feature extractor was developed to produce better class means for few-shot classification.

E. PRESERVING INFORMATION CONTENT OF FEATURE EMBEDDINGS

When the pre-trained models are fine-tuned or trained again based on a limited set of data, there is an inherent bias-tradeoff going on in the few-shot scenario. Recent work by [38] proposed a novel loss term to prevent an issue known as informational collapse, in which variables in a neural network carry redundant information, leading to inefficient and ineffective learning. The proposed loss term aimed to produce decorrelated embedding vectors by minimizing the distance between the normalized cross-correlation matrix of the embedding vectors from the two branches and the identity matrix. By minimizing this distance, the network was encouraged to learn independent and discriminative features, which can enhance the network's ability to generalize and improve performance on downstream tasks. In addition, the authors in [39] preserved the information content of the embeddings in the process of training joint embedding architectures by using invariance, variance and covariance loss terms. This explicitly prevents a collapse due to a shrinkage of the embedding vectors towards zero. Moreover, it also prevents an informational collapse due to redundancy between the embedding variables. On the other hand, The authors in [10] tackled this issue by only allowing the parameters of the feature extractor's final layer to be updatable to reduce bias. They also utilize the base images and the novel images together to update the feature extractor. A simple negative cosine similarity loss function was employed to prevent the learned representations from deviating dramatically from the subsequent fine-tuning, which reduces variance and overfitting. These strategies help to improve not only the mean accuracy of few-shot models but also the worst-case accuracy.

III. METHODOLOGY

In this section, the common few-shot setting is first introduced. Then, the details of the proposed method NegCosIC are described. A summary of NegCosIC is depicted in Fig. 1.

A. FEW-SHOT SETTING

We consider few-shot learning in the context of a labeled training set $D_{base} = \{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^{N_{base}}$, where each sample is represented by its raw images $\mathbf{x}_j$ and its corresponding one-hot encoded label $\mathbf{y}_j$. The set of classes for this base dataset is denoted by Y_{base}. In few-shot settings, there is a distinct test dataset $X_{test} = \{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^{N_{test}}$ with a set of classes Y_{test} such that $Y_{base} \cap Y_{test} = \emptyset$. For the few-shot classification tasks, the labeled data samples are randomly sampled from the test dataset. Each task consists of N distinct classes, with K_s labeled samples from each class, resulting in an N-way K_s-shot task. The set of these labeled samples is referred to as the support set, denoted by s. The size of the support set is $|s| = K_s \cdot N$. Additionally, each task has an unlabeled query set q composed of K_q examples from each of the N classes, resulting in a query set size of $|Q| = K_q \cdot N$. These examples are typically unseen. After training the models on the base classes, few-shot learning methods utilize the labeled support sets to adapt to new tasks, with evaluations performed on the unlabeled query sets. Meanwhile, the raw images from the support set and query set are X_s and X_q respectively, with their actual labels Y_s and Y_q. The predicted labels of the support set are known as $\hat{Y}_s$, whereas the predicted labels of the query set are $\hat{Y}_q$.

B. REGULARIZATION WITH NEGATIVE COSINE SIMILARITY LOSS, INVARIANCE LOSS, AND COVARIANCE LOSS

There are two stages in the proposed NegCosIC, which are the pre-training stage and the fine-tuning stage. During the pre-training stage, a feature extractor is first pre-trained on the base dataset. After that, the classification layer is removed. A new learnable classification head W is added at the final layer, and the few-shot model is fine-tuned based on the novel set during the fine-tuning stage. The embeddings from both the pre-training stage feature extractor $f_\theta(\mathbf{x})$ and $\hat{f}_\theta(\mathbf{x})$ are defined as $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ and $\mathbf{z}' = [\mathbf{z}'_1, \ldots, \mathbf{z}'_n]$ respectively, with θ being the parameters of the feature extractor. During fine-tuning, θ and W are updated based on the loss and learning rate η. Note that the parameters of the feature extractor $f_\theta(\mathbf{x})$ will be fixed, while the parameters of $\hat{f}_\theta(\mathbf{x})$ will be updated. Then, the predicted labels $\hat{Y}$ are obtained through a cosine similarity classifier based on the feature embedding input $\widetilde{\mathbf{z}}$ and classification head input $\widetilde{W}$:

$$P(\widetilde{\mathbf{z}}, \widetilde{W}) = \tau \cdot \frac{\widetilde{\mathbf{z}} \cdot \widetilde{W}}{\|\widetilde{\mathbf{z}}\| \left\|\widetilde{W}\right\|} \tag{1}$$

where τ is a scaling factor.

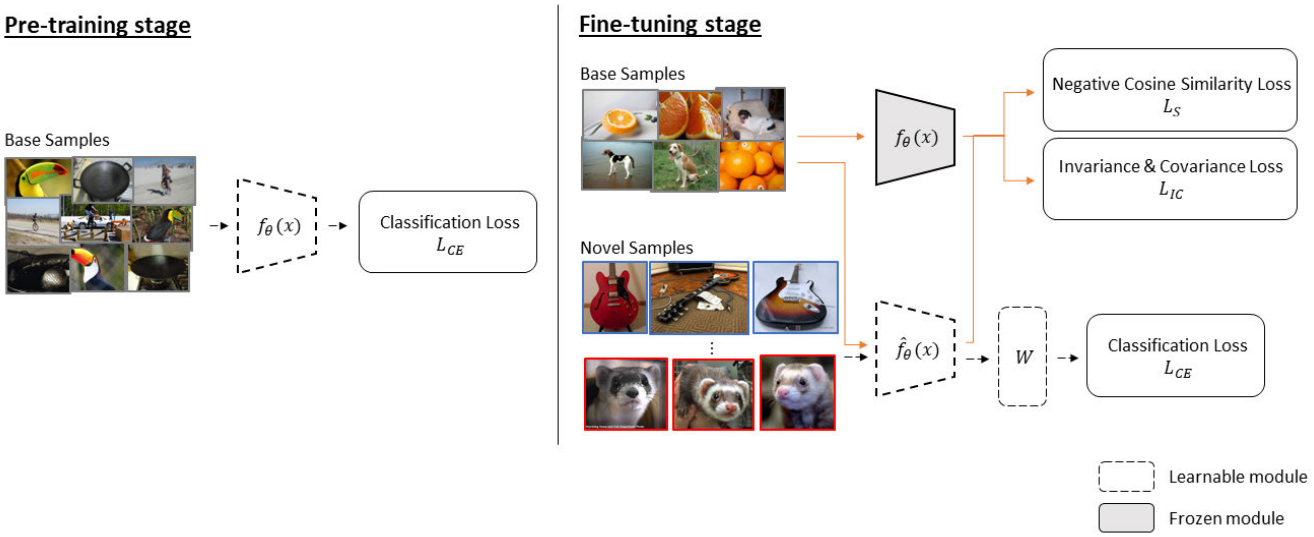

FIGURE 1. The proposed few-shot recognition method NegCosIC. First, a feature encoder network $f_\theta(\mathbf{x})$ is trained using the data from base set. After that, the final classification layer is removed. The backbone is subsequently fine-tuned into $\hat{f}_\theta(\mathbf{x})$ based on the samples from the novel set, with an additional new classification layer W. The stability of $\hat{f}_\theta(\mathbf{x})$ is achieved by the negative cosine similarity loss L_S. Through L_{IC}, each embedding's variance is maintained above a threshold, and the covariance between the embeddings are attracted to zero, de-correlating the variables from each other. In addition, the standard cross entropy loss L_{CE} minimizes error to increase the average accuracy.

In many metric-based few-shot methods, the parameters of the feature extractor are frozen and not updated during the fine-tuning phase based on the query data to reduce overfitting on the limited data. In the proposed NegCosIC, the final layer of the feature extractor is unfrozen for updating the parameters. In order to reduce severe overfitting, a negative cosine similarity loss function is proposed to stabilize the fine-tuning phase by enforcing the fine-tuned representations to remain similar to the original learned representations:

$$L_S(\mathbf{z}, \mathbf{z}') = -\frac{\mathbf{z} \cdot \mathbf{z}'}{\|\mathbf{z}\| \|\mathbf{z}'\|} \tag{2}$$

During the fine-tuning phase, instead of using the novel data for computing L_S, a set of 256 randomly sampled data from the base set is used. L_S is minimized when $\hat{f}_\theta(x)$ and $f_\theta(x)$ are similar to each other. By doing so, the parameter update becomes more stable and thereby reduce the overfitting. In addition to L_S, invariance loss and covariance loss terms that are introduced as parts of the final loss.

The invariance loss term L_I between $\mathbf{z}$ and $\mathbf{z}'$ as the mean-squared euclidean distance between each pair of vectors, without any normalization:

$$L_I(\mathbf{z}, \mathbf{z}') = \frac{1}{n} \sum_j \|\mathbf{z}_j - \mathbf{z}'_j\|_2^2 \tag{3}$$

After that, the covariance matrix of $\widetilde{\mathbf{z}}$ is formulated as:

$$C(\widetilde{z}) = \frac{1}{n-1} \sum_{j=1}^{n} (\widetilde{\mathbf{z}}_j - \bar{\widetilde{\mathbf{z}}})(\mathbf{z}_j - \bar{\widetilde{\mathbf{z}}})^T, \quad \bar{\widetilde{\mathbf{z}}} = \frac{1}{n} \sum_{j=1}^{n} \widetilde{\mathbf{z}}_j \tag{4}$$

where $\widetilde{\mathbf{z}}$ is a feature vector input. Then, the covariance regularization term c can be defined as the sum of the squared off-diagonal of $C(\widetilde{\mathbf{z}})$, accompanied by a factor $1/d$ with the purpose of scaling the criterion as a function of the dimension:

$$c(\widetilde{\mathbf{z}}) = \frac{1}{d} \sum_{j \neq k} [C(\widetilde{\mathbf{z}})]_{j,k}^2 \tag{5}$$

where d is the dimension of the feature embedding. This covariance regularization term decorrelates the different dimensions of the embeddings and prevents them from encoding similar information by encouraging the off-diagonal coefficients of $C(\widetilde{\mathbf{z}})$ to be close to 0.

The covariance loss is subsequently defined as:

$$L_C(\mathbf{z}, \mathbf{z}') = c(\mathbf{z}) + c(\mathbf{z}') \tag{6}$$

The overall invariance-covariance loss is a weighted average of invariance loss and covariance loss:

$$L_{IC}(\mathbf{z}, \mathbf{z}') = \lambda \cdot L_I(\mathbf{z}, \mathbf{z}') + \rho \cdot L_C(\mathbf{z}, \mathbf{z}') \tag{7}$$

where λ and ρ are hyper-parameters controlling the importance of each term in the loss function.

Finally, a standard cross entropy loss L_{CE} that aims to maximizing the average accuracy is used as well:

$$L_{CE}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{n} \sum_{j=1}^{n} \mathbf{y}_j \log(\hat{\mathbf{y}}_j) \tag{8}$$

where $\hat{\mathbf{y}}$ is the predicted label.

Overall, the final loss is defined as:

$$\mathcal{L} = \alpha \cdot L_S + \beta \cdot L_{IC} + \epsilon \cdot L_{CE} \tag{9}$$

with α, β, and ϵ set to 1, 0.0001, and 0.1 respectively in all experiments. The training procedure of the fine-tining stage is presented in Algorithm 1.

Algorithm 1 The Training Procedure During The Fine-Tuning Stage of the Proposed NegCosIC

input : $X_{base}, X_s, Y_s, f_\theta, \hat{f}_\theta, \eta$
receive θ from pre-training stage
randomly initialize W
$\mathbf{z}_{base} \leftarrow f_\theta(X_{base})$
$\mathbf{z}'_{base} \leftarrow \hat{f}_\theta(X_{base})$
$\mathbf{z}_s \leftarrow f_\theta(X_s)$
for *epoch* $\leftarrow 0$ **to** 100 **do**
  $\hat{Y}_s \leftarrow P(\mathbf{z}_s, W)$ (1)
  Compute loss $L_S(\mathbf{z}_{base}, \mathbf{z}'_{base})$ (2)
  Compute loss $L_{IC}(\mathbf{z}_{base}, \mathbf{z}'_{base})$ (3, 4, 5, 6, 7)
  Compute loss $L_{CE}(Y_s, \hat{Y}_s)$ (8)
  Compute final loss $\mathcal{L}$ (9)
  $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$
  $W \leftarrow W - \eta\nabla_W\mathcal{L}$
end

C. SUMMARY

In summary, the proposed NegCosIC incorporates negative cosine similarity loss that encourage the original learned representations to remain similar to the fine-tuned representations. This stabilizes the parameter updating of the pre-trained feature encoder. In addition, the proposed NegCosIC utilizes invariance loss and covariance loss that regularize the variance of the extracted features. As a result, the feature representations are more robust and this in turn improves the accuracy of the model.

IV. EXPERIMENTS

In this section, the datasets, evaluation protocols, implementation details, and results obtained by the proposed NegCosIC are described. Three common benchmarks are used to evaluate the performance, which are miniImageNet, CIFAR-FS and CUB.

A. DATASETS

miniImagenet: This is a widely used dataset [4], [7] in few-shot image classification. It is a subset of ILSVRC-12 [40] that contains 60,000 randomly selected images from 100 classes with the size of 84 × 84 pixels. The dataset is split into 64 training classes, 16 validation classes, and 20 test classes.

CIFAR-FS: This dataset is a randomly sampled subset of CIFAR-100 [41]. It is partitioned into 64 base, 16 validation and 20 novel classes. For each class, there are 600 random images with the size of 32 × 32 pixels.

CUB: This fine-grained classification dataset consists of a total of 6,033 images from 200 classes [42]. Following the protocol of [43], it is split into 100 training classes, 50 validation classes, and 50 test classes with the images resized to 84 × 84 pixels.

B. EVALUATION PROTOCOLS

The experiments are evaluated based on the standard few-shot classification settings, which are 5-way 1-shot and 5-way 5-shot tasks. The training data consist of 1 or 5 labeled data from each of the 5 classes, while the test data consist of 15 instances randomly selected from the same classes.

The experimental results are reported in average accuracy (ACC_m). The average accuracy is based on the reported results from respective work. In addition, we also adopted the worst case accuracy metric inspired from [10]. The worst case accuracy metrics are the worst 1 accuracy (ACC_1), average of worst 10 accuracy (ACC_{10}), and average of worst 100 accuracy (ACC_{100}). The worst accuracy of other compared methods are based on the respective official implementation of each work. Note that the experiments are conducted based on the same 500 episodes for each dataset in order for making the comparison fair.

C. TRAINING PROCEDURE AND HYPERPARAMETERS

During the pre-training phase of feature extractor, ResNet-18 [43], [44] and WRN-28-10 (Wide Residual Network with 28 layers and width factor of 10) [45] network architectures are used. At the last block of the network architecture, average pooling is utilized to obtain the feature vectors. The network is trained by using Adam optimizer [46] for 400 epochs with batch size of 256. The learning rate is set to 0.001, β_1 and β_2 are set to 0.9 and 0.999 respectively, and ϵ is set to 1e-8.

In the fine-tuning phase, all the hyperparameters are kept constant throughout all the experiments for simplicity. In summary, the coefficient λ of invariance loss L_I and the coefficient ρ of covariance loss L_C are set to 1 and 0.04 respectively. Meanwhile, the coefficient β of L_{IC} is set to 0.0001. Similarly, the objective function is optimized with Adam optimizer with the parameters recommended in [46], with the learning rate of 0.001, β_1 and β_2 of 0.9 and 0.999 respectively, and ϵ of 1e-8. A cosine classifier is employed as the classification head. The fine-tuning process is executed for 100 iterations in each episode.

D. COMPARISON WITH THE STATE-OF-THE-ART METHODS

Experiments of standard 5-way 1-shot and 5-way 5-shot classification tasks are carried out on three datasets: miniImageNet, CUB and CIFAR-FS. For a fair comparison, only the results based on the backbone ResNet-18 and WRN-28-10 are compared. As shown in Table 1, the proposed NegCosIC outperforms other state-of-the-art methods in terms of average accuracy (ACC_m) on miniImageNet and CUB datasets. With WRN-28-10 backbone, it achieves 69.41% on miniImageNet 5-way 1-shot task and 85.98% on miniImageNet 5-way 5-shot task. Meanwhile, the accuracy of the model on CUB 5-way 1-shot and 5-way 5-shot task is 85.68% and 94.66% respectively. Furthermore, it achieves

TABLE 1. Mean accuracy ACC_m (%) on mini-ImageNet, CUB, and CIFAR-FS 5-way 1-shot and 5-way 5-shot classification tasks. ACC_m is the average accuracy from respective published papers unless otherwise noted.

Dataset	Method	Network	1-shot ACC_m	5-shot ACC_m
miniImageNet	ProtoNet† [8]	ResNet-18	54.16±0.82	73.68±0.65
	MixtFSL [47]	ResNet-18	60.11±0.73	77.76±0.58
	Negative-Cosine [48]	ResNet-18	62.33±0.82	80.94±0.59
	$\mathrm{S2M2}_R$ [49]	ResNet-18	64.06±0.18	80.58±0.12
	NegCosIC (ours)	ResNet-18	**65.20±0.50**	**83.64±0.38**
	Negative-Cosine [48]	WRN-28-10	61.72±0.81	81.79±0.55
	MixtFSL [47]	WRN-28-10	64.31±0.79	81.66±0.60
	$\mathrm{S2M2}_R$ [49]	WRN-28-10	64.93±0.18	83.18±0.11
	PT+NCM [34]	WRN-28-10	65.35±0.20	83.87±0.13
	Transductive-FT [32]	WRN-28-10	65.73±0.68	78.40±0.52
	CGCS [50]	WRN-28-10	67.02±0.20	82.32±0.14
	LR-DC [51]	WRN-28-10	68.57±0.55	82.88±0.42
	AC+SR [10]	WRN-28-10	69.38	85.87
	NegCosIC (ours)	WRN-28-10	**69.41±0.52**	**85.98±0.40**
CUB	$\mathrm{S2M2}_R$ [49]	ResNet-18	71.43±0.28	85.55±0.52
	ProtoNet† [8]	ResNet-18	71.88±0.91	87.42±0.48
	Negative-Cosine [48]	ResNet-18	72.66±0.85	89.40±0.43
	MixtFSL [47]	ResNet-18	73.94±1.10	86.01±0.50
	NegCosIC (ours)	ResNet-18	**74.54±0.81**	**88.21±0.45**
	CGCS [50]	WRN-28-10	74.66±0.21	88.37±0.12
	LR-DC [51]	WRN-28-10	79.56±0.87	90.67±0.35
	PT+NCM [34]	WRN-28-10	80.57±0.20	91.15±0.10
	$\mathrm{S2M2}_R$ [49]	WRN-28-10	80.68±0.81	90.85±0.44
	AC+SR [10]	WRN-28-10	85.14	94.42
	NegCosIC (ours)	WRN-28-10	**85.68±0.78**	**94.66±0.41**
CIFAR-FS	$\mathrm{S2M2}_R$ [49]	ResNet-18	63.66±0.17	76.07±0.19
	MTUNet [52]	ResNet-18	66.31±0.50	80.16±0.39
	NegCosIC (ours)	ResNet-18	**68.98±0.66**	**84.62±0.50**
	ProtoNet◇ [8]	WRN-28-10	61.60	79.08
	Inductive-FT [32]	WRN-28-10	68.72±0.67	86.11±0.47
	Negative-Cosine◇ [48]	WRN-28-10	68.90	83.82
	MixtFSL◇ [47]	WRN-28-10	69.42	81.05
	LR-DC◇ [51]	WRN-28-10	72.52	83.92
	CGCS [50]	WRN-28-10	73.00±0.70	85.80±0.50
	PT+NCM [34]	WRN-28-10	74.64±0.21	**87.64±0.15**
	AC+SR [10]	WRN-28-10	74.00	86.65
	$\mathrm{S2M2}_R$ [49]	WRN-28-10	**74.81±0.19**	87.47±0.13
	NegCosIC (ours)	WRN-28-10	73.98±0.65	87.33±0.48

† ACC_m from [43] ◇ ACC_m from [10]

higher accuracy than most of the methods on CIFAR-FS 5-way 1-shot and 5-way 5-shot task at 73.98% and 87.33%.

Typically, the model performance of existing methods is evaluated by calculating the mean accuracy of a set of episodes (ACC_m) along with the 95% confidence interval. As shown in [10], this model performance evaluation is not stable as the accuracy of individual episode varies a lot, which does not translate well to the practicality in real-world applications. In contrast, metrics that focus on worst-case scenarios, such as the accuracy of the worst case (ACC_1), the average accuracy of the 10 worst cases (ACC_{10}), and the average accuracy of the 100 worst cases (ACC_{100}), are more representative of the actual challenges faced in real-world applications. As a result, these metrics can more useful for evaluating few-shot learning models. Thus, we have also evaluated our proposed NegCosIC based on these metrics. Noticeably, the proposed NegCosIC with WRN-28-10 backbone achieves much higher accuracy in worst case accuracy relatively to compared methods as shown in Table 2.

On miniImagenet, NegCosIC outperforms other methods with the highest accuracy of 42.67%, 45.20%, 55.00% in 1-shot tasks and 65.33%, 67.60%, 76.68% in 5-shot tasks respectively at ACC_1, ACC_{10}, ACC_{100}, with improvement as much as over 2%.

Likewise, in comparison with other methods on CUB, NegCosIC obtains higher accuracy at 56.00%, 60.00%, 72.11% in 1-shot tasks and 78.67% and 82.67%, 88.29% in 5-shot tasks respectively at ACC_1, ACC_{10}, ACC_{100}, with improvement as much as over 3%.

In addition, the proposed NegCosIC improves accuracy as much as close to 8% compared to other methods on CIFAR-FS dataset. A performance of 38.67%, 44.93%, 57.24% in 1-shot tasks and 64.00%, 67.20%, 76.08% in 5-shot tasks respectively, is achieved at ACC_1, ACC_{10}, ACC_{100}.

It can be seen that the proposed method has significant improvement over the worst case and the worst 10 cases. This demonstrates that the proposed NegCosIC is able to perform better over harder and more extreme recognition tasks compared to other methods. In addition, this also shows

TABLE 2. Worst case accuracy ACC_1, ACC_{10}, ACC_{100}(%) on mini-ImageNet, CUB, and CIFAR-FS 5-way 1-shot and 5-way 5-shot classification tasks.

Dataset	Method	Network	1-shot			5-shot		
			ACC_1	ACC_{10}	ACC_{100}	ACC_1	ACC_{10}	ACC_{100}
miniImageNet	ProtoNet [8]	ResNet-18	19.76	26.08	37.62	43.74	49.78	59.46
	Negative-Cosine [48]	ResNet-18	23.72	32.96	43.48	50.87	56.94	66.90
	MixtFSL [47]	ResNet-18	28.42	34.30	43.76	44.10	57.94	66.58
	S2M2_R [49]	ResNet-18	37.58	42.87	53.40	58.66	66.21	74.73
	NegCosIC (ours)	ResNet-18	**40.53**	**44.12**	**53.98**	**62.83**	**66.42**	**74.82**
	Negative-Cosine [48]	WRN-28-10	24.27	36.13	46.92	53.30	58.12	68.86
	MixtFSL [47]	WRN-28-10	30.67	35.07	46.68	46.67	60.13	71.23
	PT+NCM [34]	WRN-28-10	32.00	38.13	48.41	56.00	64.00	73.89
	Transductive-FT [32]	WRN-28-10	24.00	33.73	47.60	50.67	53.73	63.09
	CGCS [50]	WRN-28-10	38.70	44.00	53.50	49.30	56.30	67.30
	LR-DC [51]	WRN-28-10	37.33	42.72	53.54	60.52	64.98	74.24
	AC+SR [10]	WRN-28-10	40.52	44.51	54.97	63.20	66.51	76.28
	NegCosIC (ours)	WRN-28-10	**42.67**	**45.20**	**55.00**	**65.33**	**67.60**	**76.68**
CUB	ProtoNet [8]	ResNet-18	28.00	35.14	47.90	53.33	60.65	70.20
	Negative-Cosine [48]	ResNet-18	36.00	42.20	55.22	70.70	72.69	79.51
	MixtFSL [47]	ResNet-18	40.00	44.93	55.76	57.33	66.40	76.65
	S2M2_R [49]	ResNet-18	50.71	55.30	68.43	72.28	76.59	84.06
	NegCosIC (ours)	ResNet-18	**53.92**	**57.72**	**71.20**	**76.31**	**81.03**	**87.15**
	CGCS [50]	WRN-28-10	50.67	56.00	68.59	57.33	63.33	72.76
	LR-DC [51]	WRN-28-10	44.00	54.80	66.52	68.80	76.16	84.30
	PT+NCM [34]	WRN-28-10	40.00	52.67	65.83	69.33	76.13	84.67
	S2M2_R [49]	WRN-28-10	52.00	56.56	69.70	73.86	77.72	85.47
	AC+SR [10]	WRN-28-10	52.78	57.46	71.31	76.00	80.83	87.76
	NegCosIC (ours)	WRN-28-10	**56.00**	**60.00**	**72.11**	**78.67**	**82.67**	**88.29**
CIFAR-FS	MTUNet [52]	ResNet-18	27.43	34.61	48.36	53.74	59.06	68.97
	S2M2_R [49]	ResNet-18	35.74	42.06	56.49	56.26	63.38	72.17
	NegCosIC (ours)	ResNet-18	**36.81**	**43.39**	**56.58**	**61.40**	**65.33**	**74.92**
	ProtoNet [8]	WRN-28-10	25.34	30.62	43.08	51.20	57.40	66.62
	Inductive-FT [32]	WRN-28-10	33.33	38.40	50.32	56.00	61.47	71.73
	Negative-Cosine [48]	WRN-28-10	26.66	39.76	52.58	52.00	61.54	71.88
	MixtFSL [47]	WRN-28-10	29.33	38.53	51.51	57.33	62.40	69.19
	LR-DC [51]	WRN-28-10	32.26	42.08	54.72	57.60	63.30	72.48
	CGCS [50]	WRN-28-10	33.33	41.33	55.53	56.00	62.13	69.16
	PT+NCM [34]	WRN-28-10	34.67	42.53	56.60	54.67	62.13	71.57
	AC+SR [10]	WRN-28-10	36.54	42.73	56.83	56.26	65.97	75.23
	NegCosIC (ours)	WRN-28-10	**38.67**	**44.93**	**57.24**	**64.00**	**67.20**	**76.08**

TABLE 3. Results of ablation studies on 5-way 1-shot and 5-way 5-shot classification tasks on miniImageNet with WRN-28-10 backbone.

Loss			1-shot				5-shot			
L_S	L_I	L_C	ACC_m	ACC_1	ACC_{10}	ACC_{100}	ACC_m	ACC_1	ACC_{10}	ACC_{100}
✗	✗	✗	67.74	36.00	42.53	52.95	84.54	57.33	64.93	74.53
✗	✓	✗	67.86	37.33	43.07	53.13	84.62	59.70	65.73	74.75
✗	✗	✓	67.84	38.67	42.23	52.10	84.59	60.00	66.13	74.61
✗	✓	✓	67.96	40.00	44.13	53.21	84.68	60.33	66.27	74.88
✓	✗	✗	69.38	40.52	44.51	54.97	85.87	63.20	66.51	76.28
✓	✓	✗	69.14	42.00	44.00	54.39	85.97	64.00	67.20	76.59
✓	✗	✓	69.08	41.67	42.80	53.99	85.86	64.00	66.80	76.19
✓	✓	✓	**69.41**	**42.67**	**45.20**	**55.00**	**85.98**	**65.33**	**67.60**	**76.68**

that the stabilization of the feature extractor's parameters with L_S as well as the variance regularization with L_{IC} during the fine-tuning process are helpful to boost the performance of the few-shot model not only in terms of mean accuracy but also worst case accuracy.

E. ABLATION STUDIES AND DISCUSSIONS

To investigate the effects of the major components of the proposed NegCosIC, an ablation study is conducted on mini-ImageNet, CUB and CIFAR-FS with WRN-28-10 backbone to study the effects of negative cosine similarity loss L_S, invariance loss L_I and covariance loss L_C. Table 3, 4 and 5 show that L_S is crucial to improve the mean accuracy and worst-case accuracy of the model. In both scenarios where L_S is or is not used, optimal performance is obtained when L_I and L_C are used together. This suggests that incorporating both losses in the method is most beneficial when the variance is regularized. In most cases, including either invariance or covariance loss in the method improves the performance of the model in terms of mean accuracy and also the worst case metrics.

Additionally, an ablation study is conducted to study the effect of invariance loss and covariance loss L_{IC}. Table 6 shows the effect of different values of β on L_{IC}. When β is too big, it brings a slightly negative effect on the performance

TABLE 4. Results of ablation studies on 5-way 1-shot and 5-way 5-shot classification tasks on CUB with WRN-28-10 backbone.

Loss			1-shot				5-shot			
L_S	L_I	L_C	ACC_m	ACC_1	ACC_{10}	ACC_{100}	ACC_m	ACC_1	ACC_{10}	ACC_{100}
✗	✗	✗	83.58	48.00	55.47	69.15	92.82	70.67	77.07	85.20
✗	✓	✗	83.72	51.00	56.13	69.43	92.84	72.00	77.87	85.39
✗	✗	✓	83.67	51.67	56.27	69.59	92.84	72.43	77.53	85.46
✗	✓	✓	83.89	53.33	57.33	69.73	92.89	74.67	78.00	85.81
✓	✗	✗	85.14	52.78	57.46	71.31	94.42	76.00	81.12	87.99
✓	✓	✗	85.55	53.33	59.33	71.73	94.58	77.33	81.47	88.25
✓	✗	✓	85.66	52.00	58.27	71.60	94.58	**78.67**	82.27	88.27
✓	✓	✓	**85.68**	**56.00**	**60.00**	**72.11**	**94.66**	**78.67**	**82.67**	**88.29**

TABLE 5. Results of ablation studies on 5-way 1-shot and 5-way 5-shot classification tasks on CIFAR-FS with WRN-28-10 backbone.

Loss			1-shot				5-shot			
L_S	L_I	L_C	ACC_m	ACC_1	ACC_{10}	ACC_{100}	ACC_m	ACC_1	ACC_{10}	ACC_{100}
✗	✗	✗	72.87	32.00	41.87	56.05	84.97	56.17	65.87	74.19
✗	✓	✗	72.90	35.67	42.40	56.37	85.07	57.00	66.10	74.60
✗	✗	✓	72.97	36.00	42.53	56.07	85.05	57.33	66.13	74.25
✗	✓	✓	72.99	37.33	42.67	56.44	85.10	58.67	67.20	74.60
✓	✗	✗	**74.00**	36.54	42.73	56.83	86.65	56.26	65.97	75.23
✓	✓	✗	72.70	36.00	43.20	56.94	87.22	60.00	66.27	75.81
✓	✗	✓	72.70	36.00	44.00	56.89	87.21	58.67	66.27	75.67
✓	✓	✓	73.98	**38.67**	**44.93**	**57.24**	**87.33**	**64.00**	**67.20**	**76.08**

TABLE 6. Effects of different β values of L_{IC} on 5-way 5-shot classification tasks on CIFAR-FS.

β	ACC_m	ACC_1	ACC_{10}	ACC_{100}
0.05	87.01	57.33	66.83	75.57
0.01	87.21	57.60	66.93	76.01
0.005	87.25	61.33	66.00	75.91
0.001	87.18	62.67	66.80	75.85
0.0005	87.14	62.67	66.80	75.91
0.0001	**87.33**	**64.00**	**67.20**	**76.08**
0.00005	87.21	61.33	66.80	75.93
0.00001	87.15	61.33	66.53	75.87

TABLE 7. Effects of different ρ values on L_C on 5-way 5-shot classification tasks on CIFAR-FS with WRN-28-10 backbone.

ρ	ACC_m	ACC_1	ACC_{10}	ACC_{100}
0.04	**87.33**	**64.00**	**67.20**	**76.08**
0.4	87.19	62.67	66.47	75.68
4	87.08	60.00	66.53	75.69
40	87.14	60.00	66.50	75.65

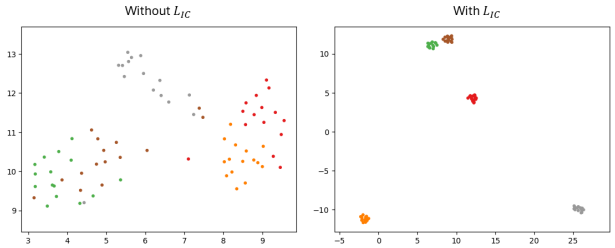

FIGURE 2. UMAP 2-dimensional visualization [53] of the features of 75 query images based on a randomly sampled 5-way 1-shot few-shot classification task from miniImageNet.

of the model as it restricts the model to be less adaptive from the base and novel samples due to variance regularization. Conversely, if the value of β is too low, there may not be a significant improvement in the model's performance whether or not L_{IC} is used. From the experimental results, we found that the optimal value is 0.0001, where the model is able to perform the best consistently throughout different datasets.

Moreover, we utilized 2-dimensional UMAP [53] for feature visualization to show the effect of employing L_{IC} in the model. The UMAP graph of the feature representations from novel images based on a randomly sampled 5-way 1-shot task from miniImageNet is shown in Figure 2. From the visualization, it can be seen that without L_{IC}, the data points of each class are more scattered and have longer intra-class distance and shorter inter-class distance. With L_{IC}, it can be seen that the data points of each class are more clustered together and have shorter intra-class distance and longer inter-class distance. This signifies that the proposed loss terms are useful to improve the ability of the model to have better class discrimination in low-data settings.

On top of that, another ablation study is done on the values of ρ on L_C. Table 7 demonstrates that when the value of ρ is higher, there is a marginal deterioration in the performance of the model. When the value of ρ is 0.04, the proposed NegCosIC performs the best in terms of ACC_m, ACC_1, ACC_{10} and ACC_{100}. This shows that NegCosIC is more effective when ρ is not too big in order to avoid overregularization.

Although several hybrid loss functions have been applied in different works in the literatures [54] and [55], they have been used in different architectures and with different data types. Ablation studies in this work indicated the efficiency of the proposed loss function. As an extension of this work, the proposed approach can be tested with medical image types such as dermoscopy images.

V. CONCLUSION

In this paper, NegCosIC is proposed for few-shot learning. The proposed NegCosIC utilizes a negative cosine similarity loss term to encourage the original learned representations to remain similar to the fine-tuned representations. This in turn stabilizes the parameters of the feature extractor during fine-tuning. In addition, to reduce overfitting during fine-tuning, we incorporate invariance loss and covariance loss to regularize the variance of the learned feature representations. This helps to reduce the possibilities of high variance when the pre-trained feature encoder is fine-tuned. By doing so, the learned representations become more robust, which allows the few-shot recognition model to achieve good performance not only in conventional mean accuracy over multiple few-shot episodes but also in the neglected worst-case accuracy. Through extensive experiments, the performance of the proposed NegCosIC is shown to be able to outperform many state-of-the-art methods in few-shot learning in both mean accuracy and worst-case's metrics. Based on the result, the proposed NegCosIC is able to improve the performance of the model across different network architectures, which are the commonly used ResNet-18 and WRN-28-10. Moreover, through UMAP visualization, it is shown that the proposed NegCosIC improves the class discrimination ability of the model because the data points of each class have shorter intra-class distance and longer inter-class distance. Thus, the proposed framework in this work is applicable to many practical problems.

REFERENCES

[1] E. Goceri, "Image augmentation for deep learning based lesion classification from skin images," in *Proc. IEEE 4th Int. Conf. Image Process., Appl. Syst. (IPAS)*, Dec. 2020, pp. 144–148.
[2] E. Goceri, "Medical image data augmentation: Techniques, comparisons and interpretations," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12561–12605, Nov. 2023.
[3] E. Goceri, "Comparison of the impacts of dermoscopy image augmentation methods on skin cancer classification and a new augmentation method with wavelet packets," *Int. J. Imag. Syst. Technol.*, vol. 33, no. 5, pp. 1727–1744, Sep. 2023.
[4] A. Srinivasan, A. Bharadwaj, M. Sathyan, and S. Natarajan, "Optimization of image embeddings for few shot learning," in *Proc. 10th Int. Conf. Pattern Recognit. Appl. Methods*, 2021, pp. 1–17.
[5] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1126–1135.
[6] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, *arXiv:1803.02999*.
[7] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 3630–3638.
[8] J. Wang and Y. Zhai, "Prototypical Siamese networks for few-shot learning," in *Proc. IEEE 10th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2020, pp. 178–181.
[9] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.
[10] M. Fu, Y.-H. Cao, and J. Wu, "Worst case matters for few-shot recognition," in *Comput. Vision–ECCV*. Cham, Switzerland: Springer, 2022, pp. 99–115.
[11] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–14.
[12] T. Munkhdalai and H. Yu, "Meta networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2554–2563.
[13] H. Xu, J. Wang, H. Li, D. Ouyang, and J. Shao, "Unsupervised meta-learning for few-shot learning," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107951.
[14] B. Zhang, K.-C. Leung, X. Li, and Y. Ye, "Learn to abstract via concept graph for weakly-supervised few-shot learning," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107946.
[15] X. Zhong, C. Gu, M. Ye, W. Huang, and C.-W. Lin, "Graph complemented latent representation for few-shot image classification," *IEEE Trans. Multimedia*, vol. 25, pp. 1979–1990, 2023, doi: 10.1109/TMM.2022.3141886.
[16] H. Zhang, H. Li, and P. Koniusz, "Multi-level second-order few-shot learning," *IEEE Trans. Multimedia*, vol. 25, pp. 2111–2126, 2023, doi: 10.1109/TMM.2022.3142955.
[17] Y. Zheng, X. Zhang, Z. Tian, W. Zeng, and S. Du, "Detach and unite: A simple meta-transfer for few-shot learning," *Knowl.-Based Syst.*, vol. 277, Oct. 2023, Art. no. 110798.
[18] M. Yan, "Adaptive learning knowledge networks for few-shot learning," *IEEE Access*, vol. 7, pp. 119041–119051, 2019.
[19] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3037–3046.
[20] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. Workshops*, 2018, pp. 1–11.
[21] Q. Luo, L. Wang, J. Lv, S. Xiang, and C. Pan, "Few-shot learning via feature hallucination with variational inference," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3962–3971.
[22] J. Wu, N. Dong, F. Liu, S. Yang, and J. Hu, "Feature hallucination via maximum a posteriori for few-shot learning," *Knowl.-Based Syst.*, vol. 225, Aug. 2021, Art. no. 107129.
[23] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–10.
[24] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12200–12210.
[25] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–17.
[26] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11–20.
[27] L. Yang, L. Li, Z. Zhang, X. Zhou, E. Zhou, and Y. Liu, "DPGN: Distribution propagation graph network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13387–13396.
[28] X. Wang, B. Ma, Z. Yu, F. Li, and Y. Cai, "Multi-scale decision network with feature fusion and weighting for few-shot learning," *IEEE Access*, vol. 8, pp. 92172–92181, 2020.
[29] W. Cui and Y. Guo, "Parameterless transductive feature re-representation for few-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2212–2221.
[30] M. Boudiaf, I. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. B. Ayed, "Information maximization for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–15.
[31] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang, "Learning to propagate labels: Transductive propagation network for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.
[32] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–14.
[33] J. Liu, L. Song, and Y. Qin, "Prototype rectification for few-shot learning," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2020, pp. 741–756.
[34] Y. Hu, V. Gripon, and S. Pateux, "Leveraging the feature distribution in transfer-based few-shot learning," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, 2021, pp. 487–499.
[35] I. Ziko, J. Dolz, E. Granger, and I. B. Ayed, "Laplacian regularized few-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11660–11670.
[36] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 121–138.

[37] C. Chen, X. Yang, C. Xu, X. Huang, and Z. Ma, "ECKPN: Explicit class knowledge propagation network for transductive few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6592–6601.
[38] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.
[39] A. Bardes, J. Ponce, and Y. Lecun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. ICLR 2022-10th Int. Conf. Learn. Represent.*, 2022, pp. 1–11.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[41] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.
[42] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, *Caltech-UCSD Birds 200*, document CNS-TR-2010-001, 2010.
[43] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–13.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[45] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. BMVC*, 2016, pp. 1–15.
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[47] A. Afrasiyabi, J.-F. Lalonde, and C. Gagné, "Mixture-based feature space learning for few-shot image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9021–9031.
[48] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. Hu, "Negative margin matters: Understanding margin in few-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 438–455.
[49] P. Mangla, M. Singh, A. Sinha, N. Kumari, V. N. Balasubramanian, and B. Krishnamurthy, "Charting the right manifold: Manifold mixup for few-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2207–2216.
[50] Z. Gao, Y. Wu, Y. Jia, and M. Harandi, "Curvature generation in curved spaces for few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8671–8680.
[51] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–12.
[52] B. Wang, L. Li, M. Verma, Y. Nakashima, R. Kawasaki, and H. Nagahara, "Match them up: Visually explainable few-shot image classification," *Int. J. Speech Technol.*, vol. 53, no. 9, pp. 10956–10977, May 2023.
[53] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.
[54] E. Goceri, "An application for automated diagnosis of facial dermatological diseases," *İzmir Kâtip Çelebi Üniversitesi Sağlik Bilim. Fakültesi Derg.*, vol. 6, no. 3, pp. 91–99, 2021.
[55] E. Goceri, "Polyp segmentation using a hybrid vision transformer and a hybrid loss function," *J. Imag. Informat. Med.*, pp. 1–13, Jan. 2024.

WEI HAN LIU received the bachelor's degree from Multimedia University, Malaysia, where he is currently pursuing the Ph.D. degree with the Faculty of Information Science and Technology. His research interests include deep learning, computer vision, and few-shot learning.

KIAN MING LIM (Senior Member, IEEE) received the B.I.T. degree (Hons.) in information systems engineering, the master's degree in engineering science, and the Ph.D. degree in IT from Multimedia University. He is currently an Associate Professor with the Faculty of Information Science and Technology, Multimedia University. His research interests include machine learning, computer vision, and pattern recognition.

THIAN SONG ONG (Senior Member, IEEE) is currently a Professor with the Faculty of Information Science and Technology, Multimedia University, Malaysia. His research interests include biometric security and machine learning. He has published more than 60 international refereed journals and conference papers in the related fields.

CHIN POO LEE (Senior Member, IEEE) received the Masters of Science and Ph.D. degrees in the area of abnormal behavior detection and gait recognition. She is currently an Associate Professor with the Faculty of Information Science and Technology, Multimedia University, Malaysia. Her research interests include action recognition, computer vision, gait recognition, and deep learning.