MDFM: Multi-Decision Fusing Model for Few-Shot Learning

Shuai Shao†, Lei Xing†, Rui Xu, Weifeng Liu, *Senior Member, IEEE,* Yan-Jiang Wang*, Bao-Di Liu*, *Member, IEEE*

***Abstract*—In recent years, researchers pay growing attention to the few-shot learning (FSL) task to address the data-scarce problem. A standard FSL framework is composed of two components: i) Pre-train. Employ the base data to generate a CNN-based feature extraction model (FEM). ii) Meta-test. Apply the trained FEM to the novel data (category is different from base data) to acquire the feature embeddings and recognize them. Although researchers have made remarkable breakthroughs in FSL, there still exists a fundamental problem. Since the trained FEM with base data usually cannot adapt to the novel class flawlessly, the novel data's feature may lead to the distribution shift problem. To address this challenge, we hypothesize that even if most of the decisions based on different FEMs are viewed as *weak decisions*, which are not available for all classes, they still perform decent in some specific categories. Inspired by this assumption, we propose a novel method Multi-Decision Fusing Model (MDFM), which comprehensively considers the decisions based on multiple FEMs to enhance the efficacy and robustness of the model. MDFM is a simple, flexible, non-parametric method that can directly apply to the existing FEMs. Besides, we extend the proposed MDFM to two FSL settings (e.g., supervised and semi-supervised settings). We evaluate the proposed method on five benchmark datasets and achieve significant improvements of 3.4%-7.3% compared with state-of-the-arts.**

***Index Terms*—Few-Shot Learning (FSL), distribution shift problem, Multi-Decision Fusing Model (MDFM)**

I. Introduction

In recent years, deep learning, as a powerful tool, has helped machines reached or even surpass human beings' level in various visual tasks, such as image classification [1]–[3], person re-identification [4]–[6], blind image quality assessment [7]–[9], vision-and-language navigation [10]–[12]. One indispensable factor is attributed to the large-scale labeled data. However, as the limitation of actual circumstances, it may be infeasible to collect large amounts of labeled data in the real world. Thus, few-shot learning (FSL), targets to address this problem with scarce labeled samples, has attracted growing attention. Generally, the current popular FSL model usually includes two components: i) Pre-train. Employ the base data $\mathcal{D}_{base}$ to generate a CNN-based feature extraction model (FEM). ii) Meta-test. First, extract the feature embeddings of novel data $\mathcal{D}_{novel} = \{\mathcal{S}, \mathcal{U}, \mathcal{Q}\}$, where $\mathcal{S}$, $\mathcal{U}$ and $\mathcal{Q}$ denote support set, unlabeled set and query set. Next, design a classifier to recognize the query samples. For more details, please refer to Section III.

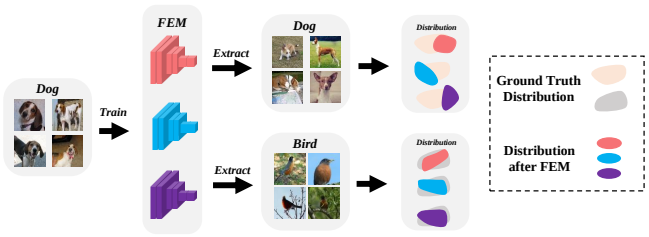


Fig. 1: An example to introduce distribution shift problem. The feature extraction model (FEMs) are trained with some dog images. We can achieve appropriate feature distributions when using these FEMs to extract other dog images but deviated feature distributions on extracting bird images.

To improve the performance of FSL, most works pay attention on designing a more robust and powerful FEM by introducing lots of strategies, such as self-supervised learning [13] [14], meta-learning [15] [16], graph structure [17] [18], knowledge distillation [19] [20]. Actually, these FEMs (trained from the base set) perform well when extracting base features, which makes the distribution based on any FEM close to the ground truth sample distribution. Just like the *dogs' distribution* in Figure 1. However, there is a fundamental problem in FSL: no matter how well the FEMs perform on the base class, they can not adapt to the novel class flawlessly due to the cross-domain limitation. Therefore, the novel samples' distributions based on the FEMs usually have a certain degree of deviation compared with the ground truth distributions, and they are very different from each other. An example of *birds' distribution* is illustrated in Figure 1. This is a typical distribution shift problem in transfer learning and domain adaption.

To address this issue, it sounds like we only need to fine-tune the network structure to accommodate the new class. However, as the scarce of labeled novel data (as an example, on typically 1-shot or 5-shot case, each category only has 1 or 5 labeled sample), this kind of method performs poorly on FSL, which has been proved in MAML [21]. Following, [19] proposed ensemble method to fuse multiple designed networks to tackle distribution shift problem. But this approach relies on the specialized FEM in the pre-train phase and specific classifier in the meta-test phase and only has a limited promotion for FSL. To this end, dedicated technology is

Shao Shuai, Rui Xu, Weifeng Liu, Yan-Jiang Wang, and Bao-Di Liu are with the College of Control Science and Engineering, China University of Petroleum (East China), 266580, China. Lei Xing is with the College of Oceanography and Space Informatics, China University of Petroleum (East China), 266580, China.

†Shao Shuai and Lei Xing are co-first authors.

*Bao-Di Liu (Email: thu.liubaodi@gmail.com) and Yan-Jiang Wang (Email: yjwang@upc.edu.cn) are corresponding authors.

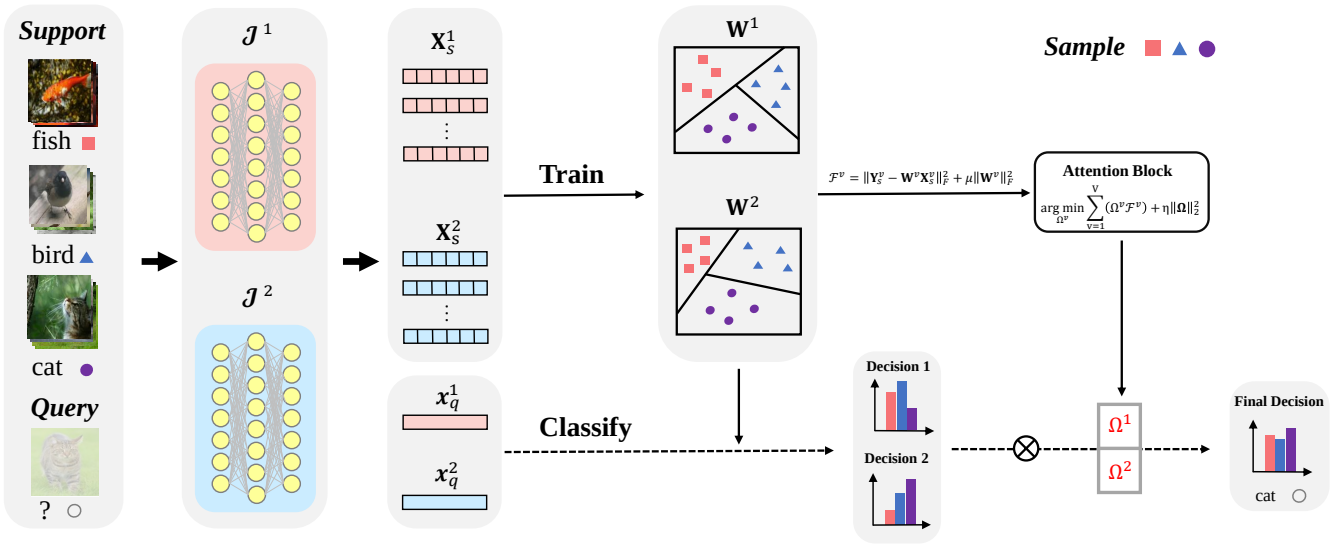


Fig. 2: The complete framework of our Multi-Decision Fusing Model (MDFM) on supervised setting. Assume we have two views of feature extraction models (FEMs), e.g., $\mathcal{J}^v$, where $v = [1, 2]$ denotes the v_{th} view. There are a total of 4 steps. **(1)** Input images to FEMs and obtain the support feature $\mathbf{X}_s^v$ and query feature $\mathbf{x}_q^v$. **(2)** Exploit support feature $\mathbf{X}_s^v$ to train classifiers $\mathbf{W}^v$. **(3)** Learn the combination weights $\mathbf{\Omega}$ for each view. **(4)** Use $\mathbf{W}^v$ to classify query data and achieve multi-decisions (use probability charts to represent). **(5)** Assert Ω^v to the corresponding decision and fuse them for final prediction.

necessary.

In this paper, we propose a Multi-Decision Fusing Model (MDFM) to solve this challenge from the perspective of fusing multi-decision. Specifically, assume we have multiple views. Each view corresponds to a strategy to design the FEM in the pre-train stage (for example, the first view's FEM follows MetaOptNet [16], and the second follows ICINet [22]). From Figure 1, we find that the bird samples' features distribute variants on different views, which means we can achieve multiple kinds of decisions through these features. Despite all the decisions are viewed as the *weak decisions* due to distribution shift problem, which may be not available for all classes, they still perform decent in some specific categories (we have evaluated this conclusion through confusion matrix in Figure 5). Therefore, we attempt to select the proper *weak decisions* for each class and then fuse all classes' decisions.

To achieve this purpose, we consider the training loss as the criterion and design a weighting mechanism (more details, please refer to Section IV). This way helps the *weak decisions* positively impact the final results. Different from [19], our MDFM is a simple, flexible, non-parametric model that can directly apply to all kinds of FEMs and classifiers. Besides, according to the data adopted in the design of the classifier, researchers categorize the FSL-based approaches as two sorts: i) Supervised Few-Shot Learning (SFSL), ii) Semi-Supervised Few-Shot Learning (SSFSL). The difference is that SSFSL uses unlabeled training data, while SFSL does not. In this paper, we extend our MDFM to the two settings. For convenience, we list some crucial abbreviations and notations in Table I, and illustrate the overall flowchart of MDFM in Figure 2.

In summary, the main contributions focus on:

- We propose a novel method for FSL, dubbed as Multi-Decision Fusing Model (MDFM). It comprehensively considers the *weak decisions* on multiple views to increase the efficacy and robustness of the FSL framework and solve the distribution shift problem.
- Compared with the existed ensemble method [19], MDFM is a simple, flexible, non-parametric method that is not restricted by the FEM and classifier. This highlight gives us more options to update the components on the FSL framework to improve classification accuracy, which may help apply the FSL framework in reality.
- We evaluate the proposed method on four benchmark datasets (mini-ImageNet, tiered-ImageNet, CIFAR-FS, FC100) and achieve significant improvements of 1.9%-7.1% compared with other state-of-the-art methods. Besides, to evaluate the robustness of the proposed method, we design the cross-domain experiments on the CUB dataset and achieve far better performance than state-of-the-art methods of at least 7.9%.

II. Related Work

A. Few-Shot Learning

In the past decade, FSL based works have attracted lots of attention. Researchers have proposed various classical frameworks to solve this problem. We list the two most popular types, including i) Meta-learning based methods, such as MAML [21], Reptile [23], LEO [24], which purpose to obtain a universal model to rapidly adapt to new tasks. ii) Metric learning based methods, focusing on looking for ideal distance metrics to strengthen model's robustness, including ProtoNet [15], MetaOpt [16], TADAM [25], MSML [26] *et.al.*.

TABLE I: Some important abbreviations and notations.

Abbreviation and Notation	Definition
FSL	few-shot learning
FEM	feature extraction model
MDFM	multi-decision fusing model
Std-Dec	standard decision
Meta-Dec	meta decision
SS-R-Dec	self-supervised rotation decision
SS-M-Dec	self-supervised mirror decision
$\mathcal{D}_{base}$, $\mathcal{D}_{novel}$	base data, novel data
$\mathcal{S}$, $\mathcal{Q}$, $\mathcal{U}$	support set, query set, unlabeled set
$\mathcal{J}^v(\cdot)$	CNN-based FEM on the v_{th} view
$\mathbf{X}^v$, $\mathbf{x}^v_{ts}$	features of training and testing data on the v_{th} view
$\mathbf{X}^v_s$, $\mathbf{X}^v_u$, $\mathbf{X}^v_q$	features of support, unlabeled, query data on the v_{th} view
$\mathbf{x}^v_{select}$, $\mathbf{y}^v_{select}$	most confident sample feature and corresponding label on the v_{th} view in self-training process
$\mathbf{Y}$	label matrices of training data
$\mathbf{Y}^v_s$, $\mathbf{Y}^v_u$	label matrices of support, unlabeled data
$\mathbf{W}^v$	classifier on the v_{th} view
$\mathbf{\Omega} = [\Omega^1, \Omega^2, \ldots, \Omega^V]^T$	combination weights for different decisions

In addition, all these methods can be split into another taxonomy, e.g. supervised few-shot learning (SFSL), and semi-supervised few-shot learning (SSFSL). For example, MAML [21], LEO [24], S2M2 [14], DPGN [18], TEAM [27], SIB [28], IPBT [29] *et.al.* are based on supervised setting, only use the labeled support data to train the classifier; and LST [30], EPNet [13], ICI [22], MHFC [31] *et.al.* are based on semi-supervised setting, employ both support and unlabeled data to train the classifier. Our method is applicable to both the two settings and achieves outstanding performance.

B. Distribution Shift Problem in FSL

Distribution shift problem is a typical problem in many fields, such as transfer learning, domain adaption, domain generalization, which also exist in the FSL. In FSL, researchers usually address this problem from two perspectives. i) On the one hand, researchers design more robust FEM, make it adapt to the novel, unseen classes, such as introducing self-supervised learning [14] and meta-learning [13] strategies . ii) On the other hand, researchers fix the FEM and pay attention on processing the extracted feature embeddings, make them more discriminative, such as the distribution calibration [32] and instance credibility inference [22].

C. Multi-View Learning

Just as every coin has two sides, it would be incomplete to define objects from a single perspective. Therefore, multi-view learning has received wide attention in recent years. There exist lots of classical methods and corresponding applications. For example, Liu *et.al.* proposed a sparse coding based multi-view method MHDSC [33] for image annotation task; Liu *et.al.* proposed SPM-CRC [34], which improve the collaborative representation model from multi-view learning to classify remote sensing images; Jan *et.al.* proposed MVCCA [35] and employed it in natural language processing. Liu *et.al.* proposed MHL [36] to solve Alzheimer's Disease Predicting problem; Zhang *et.al.* proposed IMHL [37], which is an inductive hypergraph learning from multi-view and applied it for $3D$ object recognition. All these methods may help FSL, and some multi-view based works have been proposed.

DenseCls [38] splits the feature map into different blocks and predicts the corresponding label. DivCoop [39] employs various datasets to train the FEMs and fuse them to a multi-domain representation. DWC [19] design an ensemble model with a cooperation strategy to fuse multiple information. URT [40] is the improvement of DivCoop [39], which introduces a transformer layer to better employing the different datasets. Just like our MDFM, all of these methods are based on multi-view learning. But they are restricted by the fixed FEMs and classifiers, which lose scalability. This paper, inspired by the traditional multi-view method and ensemble learning strategy, mainly focuses on designing a novel, flexible ensemble-based multi-view framework to address distribution shift problem and extending it to two FSL settings.

III. Problem Formulation

In this section, we focus on introducing the few-shot learning model. Two components, such as pre-train and meta-test, are involved in the FSL procedure. In pre-train phase, we assume that $\mathcal{D}_{base} = \{(x_i, y_i) | y_i \in \mathcal{C}_{base}\}_{i=1}^{N_{base}}$ represents the base dataset, where $\mathcal{C}_{base}$ denotes the base category set, x and y indicate the sample and corresponding label, respectively. N_{base} denotes the total number of base data. We train the CNN-based FEM $\mathcal{J}(\cdot)$ on $\mathcal{D}_{base}$. In this paper, we design several kinds of FEMs from different views, and define the FEM on the v_{th} view as $\mathcal{J}^v(\cdot)$, where $v = 1, 2, \cdots, V$. More details please refer to Section IV-D.

Next to the meta-test phase, utilising the $\mathcal{J}^v(\cdot)$ to extract features for novel dataset $\mathcal{D}_{novel} = \{(x_j, y_j) | y_j \in \mathcal{C}_{novel}\}_{j=1}^{N_{novel}}$, where $\mathcal{C}_{novel}$ denotes the novel category set, $C_{base} \cap C_{novel} = \emptyset$. N_{novel} denotes the number of novel data. Besides, the novel dataset composed of three components, e.g., $\mathcal{D}_{novel} = \{\mathcal{S}, \mathcal{U}, \mathcal{Q}\}$, where $\mathcal{S}$, $\mathcal{U}$ and $\mathcal{Q}$ denote support set, unlabeled set and query set, $\mathcal{S} \cap \mathcal{U} = \emptyset$, $\mathcal{S} \cap \mathcal{Q} = \emptyset$, $\mathcal{Q} \cap \mathcal{U} = \emptyset$. Finally, we design a classifier to classify $\mathcal{Q}$. The to-be-designed classifier includes two settings, e.g., supervised setting, semi-supervised setting. Section IV-B, IV-C show

more details. We follow standard C-way-M-$shot$ per episode as [22] for classification, where C-way denotes C classes, and M-$shot$ indicates M samples per class. We average the accuracies of all the episodes with 95% confidence intervals as the final result.

IV. Methodology

In this section, first, we propose a Multi-Decision Fusing Model (MDFM) for few-shot learning and show the details in Section IV-A. Then, we extend MDFM to varied settings (e.g., supervised and semi-supervised settings) in Section IV-B, IV-C. Next, we discuss the details about multiple decisions and define the corresponding FEMs in Section IV-D. Finally, we analyse the complex in Section IV-E.

A. Multi-Decision Fusing Model

It is worth noting that the proposed model can integrate all types of conventional classification strategies (such as support vector machine, logistic regression, linear regression). In this paper, we merely consider a simple linear regression model as an example. The objective function of the linear regression classifier is as follows:

$$\arg\min_{\mathbf{W}} \mathcal{F} = \|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_F^2 + \mu \|\mathbf{W}\|_F^2 \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{dim \times N}$, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N] \in \mathbb{R}^{C \times N}$, dim and N indicate the dimension and number of labeled samples, respectively. C denotes the number of categories. $\mathbf{x}_i, \mathbf{y}_i$ $(i = 1, 2, \dots)$ denote the feature embedding vector and one-hot label vector of the i_{th} sample. $\mathbf{W} \in \mathbb{R}^{C \times dim}$ represents the to-be-learned classifier. We directly optimize the objective function and obtain the $\mathbf{W}$ as:

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T + \mu \mathbf{I}\right)^{-1} \tag{2}$$

Following, given a testing sample feature $\mathbf{x}_{ts} \in \mathbb{R}^{dim \times 1}$, we predict the $\mathbf{x}_{ts}$'s category by:

$$\mathcal{C}(\mathbf{x}_{ts}) = onehot\left\{id_{max}\left\{\mathbf{W}\mathbf{x}_{ts}\right\}\right\} \tag{3}$$

where id_{max} denotes an operator to obtain the index of the max value in the vector. $onehot$ indicates the operator to generate a one-hot label.

To sufficiently extract more information of few-shot data in real applications, we introduce multiple feature representations for samples from different views. Assume that we have V views in total, each view has the corresponding feature embedding and classifier, e.g., $\left[(\mathbf{X}^1, \mathbf{W}^1), (\mathbf{X}^2, \mathbf{W}^2), \dots, (\mathbf{X}^V, \mathbf{W}^V)\right]$, where $(\cdot)^v, (v = 1, 2, \dots, V)$ denotes the variable on the v_{th} view. And each view obtains a decision by using Equation (3). We try to find the combination weights $\mathbf{\Omega} = [\Omega^1, \Omega^2, \dots, \Omega^V]^T$ to make the weak classifiers have a positive impact on the final decision, the objective function is formulated as:

$$\begin{aligned} \arg\min_{\mathbf{\Omega}} \mathcal{G} &= \sum_{v=1}^{V} (\Omega^v \mathcal{F}^v) + \eta \|\mathbf{\Omega}\|_2^2 \\ \text{s.t. } & \sum_{v=1}^{V} \Omega^v = 1, \ \Omega^v \geq 0 \end{aligned} \tag{4}$$

where Ω^v indicates the weight of v_{th} view. $\mathcal{F}^v$ indicates the (Equation (1))'s loss of v_{th} view. We introduce the Lagrangian to solve the problem, the Equation (4) is rewritten as:

$$\begin{aligned} \arg\min_{\mathbf{\Omega}, \zeta, \mathbf{\Lambda}} \mathcal{G} = &\sum_{v=1}^{V} (\Omega^v \mathcal{F}^v) + \eta \|\mathbf{\Omega}\|_2^2 \\ &-\zeta \left(\sum_{v=1}^{V} \Omega^v - 1\right) - \mathbf{\Lambda}^T \mathbf{\Omega} \end{aligned} \tag{5}$$

where ζ is a constant, $\mathbf{\Lambda} = [\Lambda^1, \Lambda^2, \dots, \Lambda^V]^T$ is a vector. Assume $\hat{\mathbf{\Omega}}$, $\hat{\zeta}$, $\hat{\mathbf{\Lambda}}$ are the optimal solutions, we solved this problem as:

$$\hat{\Omega}^v = \frac{1}{2\eta} max\left\{\frac{\sum_{v=1}^{V} \mathcal{F}^v}{V} + \frac{2\eta}{V} - \mathcal{F}^v - \hat{\Lambda}_{avg}, 0\right\} \tag{6}$$

where $\hat{\Lambda}_{avg}$ is a constant, denotes the average of $\hat{\mathbf{\Lambda}}$. For the detailed optimization process, please refer to Supplementary Material. Then, we employ the proposed MDFM to predict the testing sample $\mathbf{x}_{ts}$, the Equation (3) is rewritten as:

$$\mathcal{C}(\mathbf{x}_{ts}) = onehot\left\{id_{max}\left\{\sum_{v=1}^{V} \Omega^v \mathbf{W}^v \mathbf{x}_{ts}^v\right\}\right\} \tag{7}$$

where $\mathbf{W}^v = \mathbf{Y}\mathbf{X}^{vT}\left(\mathbf{X}^v\mathbf{X}^{vT} + \mu\mathbf{I}\right)^{-1}$, $\mathbf{x}_{ts}^v$ is the feature embedding of $\mathbf{x}_{ts}$ on the v_{th} view.

B. Supervised MDFM

Define the feature embedding of $\mathcal{D}_{novel}$ on the v_{th} view as $\mathbf{X}_{novel}^v = [\mathbf{X}_s^v, \mathbf{X}_u^v, \mathbf{X}_q^v]$, where $\mathbf{X}_s^v = \mathcal{J}^v(\mathcal{S})$, $\mathbf{X}_u^v = \mathcal{J}^v(\mathcal{U})$, and $\mathbf{X}_q^v = \mathcal{J}^v(\mathcal{Q})$ denote the feature embeddings of support, unlabeled, and query data on the v_{th} view. Researchers employ different data to design the classifier, and these methods can be split into two settings, e.g., supervised setting and semi-supervised setting.

Supervised setting in few-shot learning adopt the support set $\mathcal{S}$ to train the classifier and directly predict the query set $\mathcal{Q}$'s category. We directly utilize the MDFM to achieve this purpose by:

$$\begin{cases} \mathcal{F}^v = \|\mathbf{Y}_s^v - \mathbf{W}^v\mathbf{X}_s^v\|_F^2 + \mu\|\mathbf{W}^v\|_F^2 \\ \mathbf{W}^v = \mathbf{Y}_s^v\mathbf{X}_s^{vT}\left(\mathbf{X}_s^v\mathbf{X}_s^{vT} + \mu\mathbf{I}\right)^{-1} \\ \hat{\Omega}^v = \frac{1}{2\eta} max\left\{\frac{\sum_{v=1}^{V} \mathcal{F}^v}{V} + \frac{2\eta}{V} - \mathcal{F}^v - \hat{\Lambda}_{avg}, 0\right\} \\ \mathcal{C}(\mathbf{X}_q^v) = onehot\left\{id_{max}\left\{\sum_{v=1}^{V} \Omega^v \mathbf{W}^v \mathbf{X}_q^v\right\}\right\} \end{cases} \tag{8}$$

where $\mathbf{Y}_s^v$ denotes the one-hot label matrix of support data on the v_{th} view. Note that, on the supervised setting, the $\mathbf{Y}_s$ of different views are the same.

C. Semi-Supervised MDFM

Unlike supervised few shot learning, on semi-supervised setting, besides the support data's feature embeddings and label information, researchers also apply the feature embeddings of unlabeled data to construct the classifier and then predict the query label. This paper extends MDFM to semi-supervised setting by introducing a simple self-training

Algorithm 1: Semi-Supervised MDFM

Input: Base set $\mathcal{D}_{base}$, Novel set $\mathcal{D}_{novel} = \{\mathcal{S}, \mathcal{U}, \mathcal{Q}\}$
Output: Query label

Design the multi-view feature extraction model $\mathcal{J}^v(\cdot)$, and obtain feature embeddings by $\mathbf{X}_s^v = \mathcal{J}^v(\mathcal{S})$, $\mathbf{X}_u^v = \mathcal{J}^v(\mathcal{U})$, $\mathbf{X}_q^v = \mathcal{J}^v(\mathcal{Q})$.
repeat
  Train a basic classifier $\mathbf{W}^v$ through $\mathbf{X}_s^v$, and use it to predict the unlabeled data by Equation (9)
  Select the most confident sample and expand it to the support set by Equation (10).
until *the performance of to-be-learned classifiers are stable.*
Utilize the optimal classifier to predict the query label by Equation (8).

strategy to strengthen the classifier. We show the detailed steps as:

i) Train a basic classifier by employing the support data $\mathcal{S}$, and then utilize the trained classifier to predict the unlabeled data $\mathcal{U}$ by:

$$\begin{cases} \mathbf{W}^v = \mathbf{Y}_s^v \mathbf{X}_s^{vT} \left(\mathbf{X}_s^v \mathbf{X}_s^{vT} + \mu \mathbf{I} \right)^{-1} \\ \mathbf{Y}_u^v = \mathbf{W}^v \mathbf{X}_u^v \end{cases} \tag{9}$$

where $\mathbf{Y}_u^v$ denotes the predicted soft label matrix of unlabeled data on the v_{th} view.

ii) Follow traditional self-training strategy [41], select one most confident sample $\mathbf{x}_{select}^v$ through the $\mathbf{Y}_u^v$ without putting back, the corresponding one-hot pseudo label is denoted as $\mathbf{y}_{select}^v$. Then, expand it to the support data by:

$$\begin{cases} \mathbf{X}_s^v = [\mathbf{X}_s^v, \mathbf{x}_{select}^v] \\ \mathbf{Y}_s^v = [\mathbf{Y}_s^v, \mathbf{y}_{select}^v] \end{cases} \tag{10}$$

iii) Repeat **i)** and **ii)** until the performance of classifiers are stable. Finally, employ the optimal classifiers on different views to predict the query label by Equation (8). Note that, when we start updating the basic classifier, the label matrices $\mathbf{Y}_s^v, (v = 1, 2, \dots)$ would be different on different views. We summarize the steps in Algorithm 1.

D. Discussion about Multiple Decisions

The to-be-fused decisions are determined by the corresponding feature extraction models (FEMs), which have a large number of choices. As examples: **(1)** Standard decision (Std-Dec), the FEM utilizes a standard CNN-based classification structure, such as [22]. **(2)** Meta decision (Meta-Dec), the FEM introduces the meta-learning strategy to the network, just like [16]. **(3)** Self-supervised decision (SS-Dec), the FEM adds auxiliary losses to the standard CNN-based classification structure from a self-supervised perspective to strengthen the robustness of the network, similar as [14]. We show the results of all kinds of fusing ways in Table V.

In this paper, most experimental results are merely based on two kinds of SS-Decs. For the first category, we design the FEM by introducing standard classification loss $\mathcal{L}_c$ to predict the sample labels, and auxiliary rotation loss $\mathcal{L}_r$ (e.g., rotate the dataset to r degree and $r \in \mathcal{C}_{\mathcal{R}} = \{0°, 90°, 180°, 270°\}$) to predict image rotations. The decision based on this kind of FEM is dubbed as SS-R-Decision (SS-R-Dec). The first loss function is defined as $\mathcal{L}_c + \mathcal{L}_r$, we define $\mathcal{L}_c$ and $\mathcal{L}_r$ as:

$$\mathcal{L}_c = -\sum_c y_{(c,x)} log(p_{(c,x)}) \tag{11}$$

where $c \in \mathcal{C}_{base}$ denotes the c_{th} class. $y_{(c,x)}$, $p_{(c,x)}$ indicate the probabilities that the truth label and predicted label of x_{th} sample belongs to c_{th} class.

$$\mathcal{L}_r = -\sum_r y_{(r,x)} log(p_{(r,x)}) \tag{12}$$

where $y_{(r,x)}$, $p_{(r,x)}$ indicate the probabilities that the truth label and predicted label of x_{th} sample belongs to r_{th} class.

The second decision is SS-M-Decision (SS-M-Dec), which adopts another strategy to train the FEM. Specifically, we design the SS-M-Dec by adding the loss $\mathcal{L}_c$ and auxiliary mirror loss $\mathcal{L}_m$ (e.g., mirror the samples with m ways and $m \in \mathcal{C}_{\mathcal{M}} = \{vertically, horizontally, diagonally\}$) to the network to predict image mirrors. We summarize the second loss function as $\mathcal{L}_c + \mathcal{L}_m$, and define the $\mathcal{L}_m$ as:

$$\mathcal{L}_m = -\sum_m y_{(m,x)} log(p_{(m,x)}) \tag{13}$$

where $y_{(m,x)}$, $p_{(m,x)}$ indicate the probabilities that the truth label and predicted label of x_{th} sample belongs to m_{th} class.

E. Complexity Analyse

In our method, as that the pre-trained network in each view is based on the ResNet-12 backbone, their floating point operations (FLOPs) are fixed. Next, we merely talk about the complexity in the meta-test stage. It can be predicted that the complexity of our multi-view fusion method is related to the number of views. Assume we have 4 views. Our model takes about 4 times as long to process a single image as a single-view model. But fortunately, on the Tesla-V100 GPU, the single-view model just spends 9 milliseconds to process an image, and our method needs 36 milliseconds. This kind of consumption is acceptable. Besides the improvement of classification accuracy, our method maybe a feasible way in reality.

V. Experiments

In this section, we first briefly review the benchmark datasets and show the implementation details. Then, we list the supervised experimental results in Table II, III, and semi-supervised results in Table IV, then analyse them. Next, we perform ablation studies and discuss the multiple decisions to study what influences the performance. In the following, we conduct a cross-domain experiment to further evaluate the ability and robustness of the proposed method. We conduct all the experiments on a Tesla-$V100$ GPU with $32G$ memory.

TABLE II: The 5-way supervised few-shot classification accuracies on mini-ImageNet and tiered-ImageNet with 95% confidence intervals over 600 episodes. 4CONV, ResNet12, ResNet18 and WRN are the exploited FEM's architectures. "Dec" is the abbreviation of "Decision".

Method	Venue	Backbone	mini-ImageNet		tiered-ImageNet	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
ProtoNet [15]	NeurIPS,2017	4CONV	49.42	68.20	-	-
MAML [21]	ICML,2018	4CONV	48.70	63.11	-	-
RelationNet [42]	CVPR,2018	ResNet18	52.48	69.83	-	-
Baseline [43]	ICLR,2019	ResNet18	51.75	74.27	-	-
Baseline++ [43]	ICLR,2019	ResNet18	51.87	75.68	-	-
LEO [24]	ICLR,2019	WRN	61.76	77.59	66.33	81.44
TPN [17]	ICLR,2019	4CONV	52.78	66.42	55.74	71.01
AM3 [44]	NeurIPS,2019	ResNet12	65.30	78.10	69.08	82.58
TapNet [45]	ICML,2019	ResNet12	61.65	76.36	63.08	80.26
CTM [46]	CVPR,2019	ResNet18	64.12	80.51	-	-
DenseCls [38]	CVPR,2019	ResNet12	62.53	79.77	-	-
MetaOpt [16]	CVPR,2019	ResNet12	62.64	78.63	65.99	81.56
TEAM [27]	ICCV,2019	ResNet12	60.07	75.90	-	-
DWC [19]	ICCV,2019	ResNet12	63.73	81.19	70.44	85.43
S2M2 [14]	WACV,2020	WRN	64.93	83.18	73.71	88.59
Fine-tuning [47]	ICLR,2020	WRN	65.73	78.40	73.34	85.50
DSN-MR [48]	CVPR,2020	ResNet12	64.60	79.51	67.39	82.85
MABAS [49]	ECCV,2020	ResNet12	64.21	81.01	-	-
DivCoop [39]	ECCV,2020	ResNet12	64.14	81.23	-	-
HGNN [50]	TCSVT,2021	4CONV	60.03	79.64	64.32	83.34
URT [40]	ICLR,2021	ResNet12	72.23	83.35	80.30	88.63
DC [32]	ICLR,2021	WRN	68.57	82.88	78.19	89.90
BOIL [51]	ICLR,2021	ResNet12	66.80	79.26	80.79	87.92
MELR [52]	ICLR,2021	ResNet12	67.40	83.40	72.14	87.01
ODE [53]	CVPR,2021	ResNet12	67.76	82.71	71.89	85.96
MDFM (2-Dec)	**Ours**	ResNet12	74.91	83.88	84.17	89.95
MDFM (4-Dec)	**Ours**	ResNet12	**75.70**	**86.04**	**84.57**	**90.94**

TABLE III: The 5-way supervised few-shot classification accuracies on CIFAR-FS and FC100 with 95% confidence intervals over 600 episodes. "Dec" is the abbreviation of "Decision".

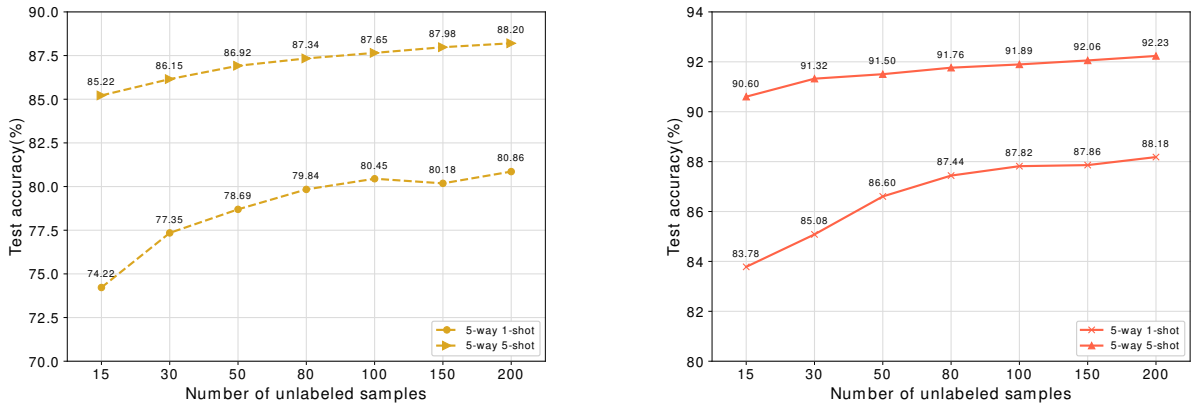
Method	Venue	Backbone	CIFAR-FS		FC100	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
ProtoNet [15]	NeurIPS,2017	4CONV	55.50	72.00	35.30	48.600
MAML [21]	ICML,2018	4CONV	58.90	71.50	-	-
RelationNet [42]	CVPR,2018	4CONV	55.00	69.30	-	-
TADAM [25]	NeurIPS,2018	ResNet12	-	-	40.10	56.10
DenseCls [38]	CVPR,2019	ResNet12	-	-	42.04	57.63
MetaOpt [16]	CVPR,2019	ResNet12	72.00	84.20	41.10	55.50
TEAM [27]	ICCV,2019	ResNet12	70.43	81.25	-	-
MABAS [49]	ECCV,2020	ResNet12	73.24	85.65	41.74	57.11
Fine-tuning [47]	ICLR,2020	WRN	76.58	85.79	43.16	57.57
DSN-MR [48]	CVPR,2020	ResNet12	75.60	86.20	-	-
MDFM (2-Dec)	**Ours**	ResNet12	81.68	88.57	46.97	59.96
MDFM (4-Dec)	**Ours**	ResNet12	**82.20**	**89.75**	**48.69**	**63.58**

A. Datasets

We carry out experiments on five benchmark datasets, including mini-ImageNet [56], tiered-ImageNet [54], CIFAR-FS [57], FC100 [25], and CUB [58]. Both mini-ImageNet and tiered-ImageNet are the subsets of ImageNet dataset [59]. mini-ImageNet consists of 100 classes and tiered-ImageNet

TABLE IV: The 5-way semi-supervised few-shot classification accuracies on mini-ImageNet and tiered-ImageNet with 95% confidence intervals over 600 episodes. "Dec" is the abbreviation of "Decision".

Method	Venue	Backbone	mini-ImageNet		tiered-ImageNet	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MSkM [54]	ICLR,2018	4CONV	50.41	63.39	49.04	62.96
TPN [17]	ICLR,2019	4CONV	55.51	69.86	59.91	73.30
LST [30]	NeurIPS,2019	ResNet12	70.10	78.70	77.70	85.20
EPNet [13]	ECCV,2020	ResNet12	75.36	84.07	81.79	88.45
TransMatch [55]	CVPR,2020	WRN	63.02	81.19	-	-
ICI [22]	CVPR,2020	ResNet12	71.41	81.12	85.44	89.12
MDFM (2-Dec)	**Ours**	ResNet12	**80.45**	87.65	**87.82**	91.90
MDFM (4-Dec)	**Ours**	ResNet12	80.42	**88.09**	87.72	**92.08**



(a) mini-ImageNet (b) tiered-ImageNet

Fig. 3: The comparison results of semi-supervised few-shot classification with varied unlabeled samples on mini-ImageNet and tiered-ImageNet.

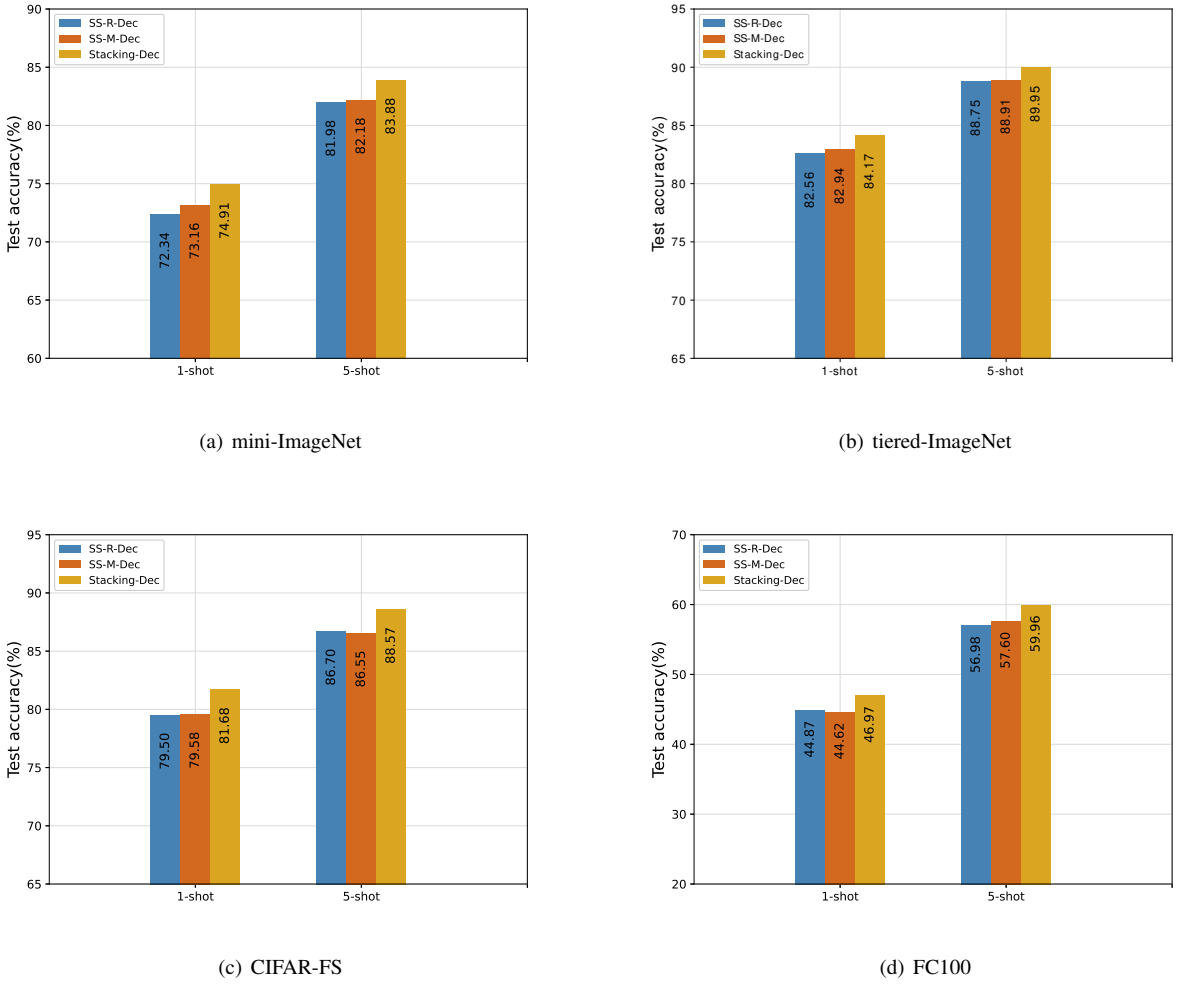
contains 608 classes. For both datasets, the number of images for each class is 600 and the size of each image is 84×84. We follow standard split as [22], select 64 classes as the base set, 16 classes as the validation set and 20 classes as the novel set for mini-ImageNet, and select 351 classes as the base set, 97 classes as the validation set and 160 classes as the novel set for tiered-ImageNet. Both CIFAR-FS and FC100 are the subsets of CIFAR-100 dataset [60], and consist of 100 classes. We follow the split introduced in [57] to divide CIFAR-FS into 64 classes as base set, 16 classes as validation set, 20 classes as novel set, and divide FC100 into 60 classes as base set, 20 classes as validation set, 20 classes as novel set. All the image size is 32×32. CUB totally includes $11,788$ images with 200 categories. We follow the setting in ICI [22] to split it into 100 classes as base set, 50 classes as validation set and 50 classes as novel set. The images are cropped into 84×84.

B. Implementation Details

In this paper, all the FEMs on different views adopt the ResNet12 [61] backbone, consisting of four residual blocks (3×3 convolution layer, batch normalization layer, LeakyReLU layer), four 2×2 max pooling layers, and four dropout layers. We adopt stochastic gradient descent (SGD) optimizer with Nesterov momentum (0.9) for the optimizer. For the parameter η in Equation (4), we fix it to 0.5 for convenience. We set the training epochs to 120 and test over 600 episodes with 15 query samples per class for all the models. Since that our method has decoupled the learning of representations and classifiers for the FSL, we have opportunities to deal with the extracted feature embeddings before classification. To this end, we introduce subspace transformation methods to strengthen the discrimination of the feature. Specifically, for the mini-ImageNet and tiered-ImageNet, we use Laplacian Eigenmap (LE) [62], and for the CIFAR-FS and FC100, we use the Principal Component Analysis (PCA) [63]. Besides, we choose the Logistic Regression (LR) classifier with the default implementation of scikit-learn [64] and have no fine-tuning process when classifying the novel data. For other settings, such as the data augmentation, the number of filters, we follow the ICI [22].

TABLE V: Results of Multi-Decision Fusing with the supervised setting on mini-ImageNet on 5-way 1-shot case. "Dec" is the abbreviation of "Decision".

Decisions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Std-Dec	✓				✓	✓	✓				✓	✓	✓		✓
Meta-Dec		✓			✓			✓	✓		✓	✓		✓	✓
SS-R-Dec			✓			✓		✓		✓	✓		✓	✓	✓
SS-M-Dec				✓			✓		✓	✓		✓	✓	✓	✓
ACC	68.2	65.8	72.3	72.6	71.7	73.3	73.5	73.3	73.4	74.9	74.1	73.8	75.2	75.1	75.7



(a) mini-ImageNet

(b) tiered-ImageNet

(c) CIFAR-FS

(d) FC100

Fig. 4: Ablation studies to show the performances of different decisions on the supervised setting.

C. Experimental Results

We compare the proposed MDFM with several state-of-the-art methods, the results are listed in Table II, III, IV. Here, we list some observations.

i) First, we look at the supervised results from Table II, III. Obviously, our MDFM has far surpassed other approaches, especially on 5-way 1-shot case, at least 3.5%, 3.8%, 5.6% and 5.5% on mini-ImageNet, tiered-ImageNet, CIFAR-FS, FC100 datasets. The performances of our MDFM on the 5-way 1-shot case are even better than many other methods on the 5-way 5-shot case. And on the 5-way 5-shot case, the MDFM also exceeds others at least 2.6%, 1.0%, 3.6% and 6.0% on mini-ImageNet, tiered-ImageNet, CIFAR-FS, FC100 datasets.

ii) Next, we compare our MDFM with other recently proposed multi-view based methods, including DenseCls [38], DWC [19], DivCoop [39], URT [40]. Obviously, our method outperforms them at least 3.5% on the 5-way 1-shot case and at least 2.3% on the 5-way 5-shot case.

iii) Then, we introduce a self-training strategy to extend our MDFM to the semi-supervised setting. Compared the results of MDFM in Table II and Table IV, we find that the unlabeled samples are really helpful to improve the performance for

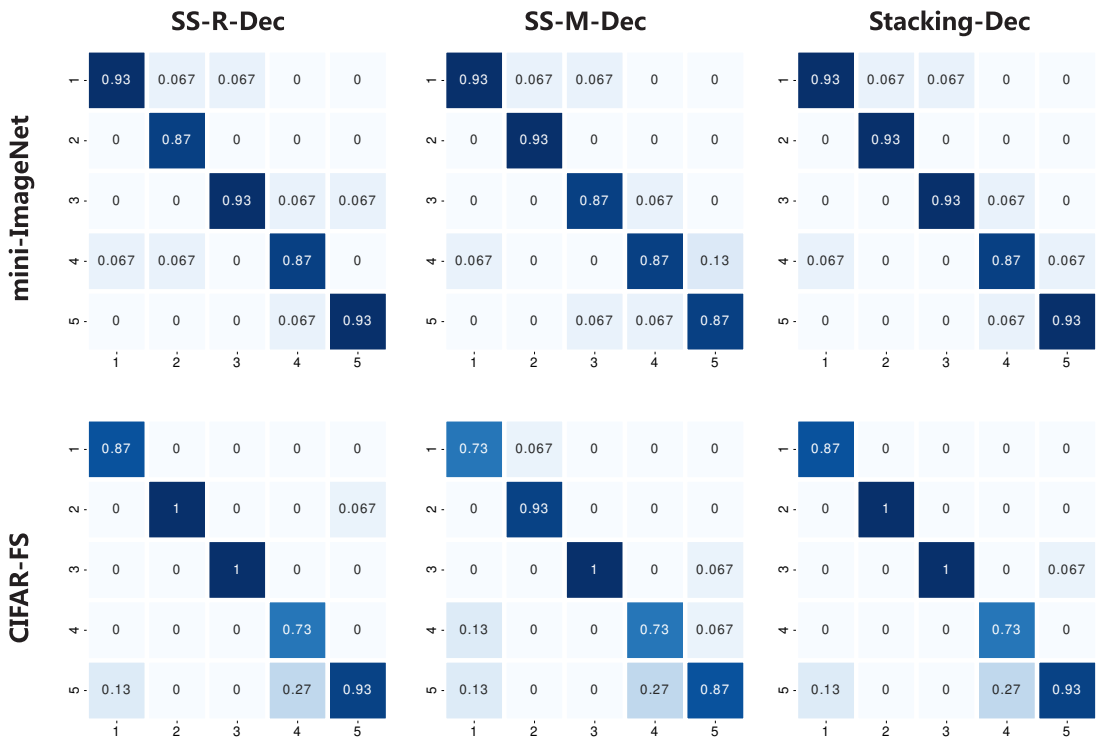


Fig. 5: Ablation studies to show the samples' confusion matrices of one episode on the supervised setting.

TABLE VI: Comparison results with fixed weights on 5-way few-shot case. (a, b) denotes that the SS-R-Dec's weight is "a", and SS-M-Dec's weight is "b". Our method exploits the designed weighting mechanism to update the weights automatically for each episode.

Weight	mini-ImageNet		tiered-ImageNet	
	1-shot	5-shot	1-shot	5-shot
(0.1, 0.9)	73.9	81.3	83.5	88.4
(0.3, 0.7)	73.3	83.4	83.1	89.2
(0.5, 0.5)	73.2	83.6	83.2	89.7
(0.7, 0.3)	73.4	82.7	83.6	89.2
(0.9, 0.1)	72.7	82.8	83.1	88.7
MDFM	**74.9**	**83.9**	**84.2**	**90.0**

FSL. Besides, from Table IV, compared our MDFM with other semi-supervised methods, our method (use 100 unlabeled samples) also achieves excellent performances. We see that MDFM has significant improvements of at least 5.1% and 4.0% on mini-ImageNet with 5-way 1-shot and 5-way 5-shot case, 2.4% and 3.0% on tiered-ImageNet with 5-way 1-shot and 5-way 5-shot case. In addition, for semi-supervised methods, the final results are influenced by the number of employed unlabeled samples. Thus, we use the mini-ImageNet and tiered-ImageNet as examples to observe the impact and list the results in Figure 3. The x-axis denotes the number of unlabeled samples. With the increase of unlabeled samples, the proposed method has became more effective. And the results start to saturate after 100 unlabeled samples.

iv) From Table II, III, IV, we find that the 4-Dec based results are better than the 2-Dec based results in most cases, but in Table IV mini-ImageNet 5-way 1-shot case and tiered-ImageNet 5-way 1-shot case, the conclusion is inverse. The reason is that introducing a large number of unlabeled data is helpful to calibrate the feature's distribution and weak the impact of our multi-decision fusion.

D. Ablation Studies

In this paper, we propose a multi-decision fusing method for few-shot learning. It is interesting to know the influence of *fusing*. All the experiments are conducted on supervised setting.

i) We list the results of only using one view's decision in Figure 4 and compare them with the fusing-view result (two view). From this figure, we can see that the performance of fusing-view improves significantly compared with the single-view, especially on the 1-shot case. It has demonstrated the efficiency of our *fusing* to some extent.

ii) To further evaluate the impact of the *fusing*, we illustrate the experimental results of each class on mini-ImageNet and CIFAR-FS datasets. Specifically, we randomly select one episode (include 5 classes) on a 5-shot case and show the corresponding confusion matrix of each view in Figure 5. Obviously, for a certain class, we obtain different results from different views, while the proposed MDFM at least achieves similar performance as the best result of single-view. Thus, as the number of categories increases, the proposed method can naturally obtain more favorable results.

iii) As described in Section IV-D, the proposed method is capable of fusing multiple decisions. But the reported results in Table II, III, IV only list the results of two ways (e.g., fuse 2-Dec and 4-Dec) for convenience. To further evaluate the proposed method, we carry out an experiment to show

TABLE VII: Comparison in cross-domain dataset scenario. Our MDFM is on supervised setting. $(\cdot)^\flat$ and $(\cdot)^\sharp$ indicate the reported results come from [65] and [14].

Method	mini-ImageNet ⟶ CUB	
	5-way 1-shot	5-way 5-shot
Baseline$^\flat$ [43]	-	53.1
MatchNet$^\flat$ [56]	-	53.1
MAML$^\flat$ [21]	-	51.3
ProtoNet$^\flat$ [15]	-	62.0
RelationNet$^\flat$ [42]	-	57.7
GNN$^\flat$ [66]	-	66.9
Neg-Cosine$^\flat$ [67]	-	67.0
LaplacianShot$^\flat$ [68]	-	66.3
TIM-GD$^\flat$ [65]	-	71.0
MetaOpt$^\sharp$ [57]	44.79	64.98
Manifold Mixup$^\sharp$ [69]	46.21	66.03
S2M2$^\sharp$ [14]	48.24	70.44
MDFM	**60.56**	**78.30**

the performances of more kinds of fusing ways with the supervised setting on mini-ImageNet. The results are listed in Table V.

iv) In theory, the proposed method can automatically assert weight to each decision through Equation (6). Thus, the ideal result is that: The more decisions are fused, the more choices are owned, and the better the results are obtained. From this table, we find that the conclusion is satisfactory. For example, 5's ACC is higher than 1's and 2's; 11's ACC is higher than 1's, 2's, 5's, 6's, and 8's; 15's ACC is higher than all the others. Besides, we find that if the to-be-fused decisions have similar performances, the final fusing result may have significant improvement, such as (1, 2, 5), (3, 4, 10).

v) From Table V, we find that the Std-Dec has outperformed many classical methods shown in Table II. There are two main reasons: On the one hand, the adopted standard FEM (Std-FEM) is captured from the ICI-based FEM [22], which is a very strong FEM and enables the extracted features to have sufficient discrimination. On the other hand, in the meta-test phase, we make process to the extracted feature embeddings, e.g., subspace transformation. The details are shown in Section V-B. This operation can further enhance the discrimination of the data, which is very helpful for classification.

vi) In addition, it's not hard to find that with the fused views increase, the performance becomes saturated. This is because the success of our approach depends largely on the diversity and complementarity of different views. As views increase, the more complete the model becomes, so the contribution of newly introduced views decreases.

vii) In order to reasonably integrate multi-view decisions, this paper proposes an weighting mechanism to dynamic weight different views of features, but how it works? Here, we design an experiment to compare the results with fixed weights to our method, which is listed in Table VI. The results show that the updated weights are more reasonable for our method and the weighting mechanism is crucial.

E. Cross-Domain Few-Shot Learning

After introducing multi-view information for selecting appropriate features for different categories, we believe that the MDFM is an extremely robust method in practical scenarios. To this end, we evaluate the proposed method with the supervised setting on a cross-domain dataset: e.g., mini-ImageNet ⟶ CUB. The results are reported in Table VII. Obviously, compared to the state-of-the-art method, we have a significant improvement at least 12.3% on 1-shot case and 7.3% on 5-shot case. Thus, the proposed MDFM would be powerful in real practice.

VI. Conclusion

Few-shot learning (FSL) based tasks have a fundamental problem, e.g., distribution shift problem. To address this challenge, we propose Multi-Decision Fusing Model (MDFM), which introduces multiple decisions to strengthen the FSL based model's efficacy and robustness. MDFM is a simple non-parametric method that can directly apply to the existing FEMs. Experimental results have demonstrated the effectiveness of MDFM. In our future work, it would be interesting to consider other fusing ways for FSL.

Acknowledgment

The paper was supported by the National Natural Science Foundation of China (Grant No. 62072468), the Natural Science Foundation of Shandong Province, China (Grant No. ZR2019MF073), the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China) (Grant No. 20CX05001A), the Graduate Innovation Project of China University of Petroleum (East China) YCX2021117, and the Graduate Innovation Project of China University of Petroleum (East China) YCX2021123.

References

[1] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.
[2] S. Shao, R. Xu, W. Liu, B.-D. Liu, and Y.-J. Wang, "Label embedded dictionary learning for image classification," *Neurocomputing*, vol. 385, pp. 122–131, 2020.
[3] Y. Yao, J. Deng, X. Chen, C. Gong, J. Wu, and J. Yang, "Deep discriminative cnn with temporal ensembling for ambiguously-labeled image classification," in *AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 669–12 676.
[4] L. Wang, B. Fan, Z. Guo, Y. Zhao, R. Zhang, R. Li, and W. Gong, "Dense-scale feature learning in person re-identification," in *Asian Conference on Computer Vision*, 2020.
[5] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2018.
[6] B. Fan, L. Wang, R. Zhang, Z. Guo, Y. Zhao, R. Li, and W. Gong, "Contextual multi-scale feature learning for person re-identification," in *ACM International Conference on Multimedia*, 2020, pp. 655–663.
[7] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.
[8] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.
[9] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[10] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3674–3683.

[11] W. Zhang, C. Ma, Q. Wu, and X. Yang, "Language-guided navigation via cross-modal grounding and alternate adversarial learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[12] A. Majumdar, A. Shrivastava, S. Lee, P. Anderson, D. Parikh, and D. Batra, "Improving vision-and-language navigation with image-text pairs from the web," in *European Conference on Computer Vision*. Springer, 2020, pp. 259–274.

[13] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *European Conference on Computer Vision*, 2020.

[14] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, and V. N. Balasubramanian, "Charting the right manifold: Manifold mixup for few-shot learning," in *IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2218–2227.

[15] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Neural Information Processing Systems*, 2017, pp. 4077–4087.

[16] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 657–10 665.

[17] 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 *International Conference on Learning Representations*, 2019.

[18] L. Yang, L. Li, Z. Zhang, X. Zhou, E. Zhou, and Y. Liu, "Dpgn: Distribution propagation graph network for few-shot learning," in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 390–13 399.

[19] N. Dvornik, C. Schmid, and J. Mairal, "Diversity with cooperation: Ensemble methods for few-shot classification," in *International Conference on Computer Vision*, 2019, pp. 3723–3731.

[20] M. N. Rizve, S. Khan, F. S. Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," in *Conference on Computer Vision and Pattern Recognition*, 2021.

[21] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, 2017, pp. 1126–1135.

[22] Y. Wang, C. Xu, C. Liu, L. Zhang, and Y. Fu, "Instance credibility inference for few-shot learning," in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 836–12 845.

[23] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[24] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," in *International Conference on Learning Representations*, 2019.

[25] B. Oreshkin, P. R. López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Neural Information Processing Systems*, 2018, pp. 721–731.

[26] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1091–1102, 2020.

[27] L. Qiao, Y. Shi, J. Li, Y. Wang, T. Huang, and Y. Tian, "Transductive episodic-wise adaptive metric for few-shot learning," in *International Conference on Computer Vision*, 2019, pp. 3603–3612.

[28] S. X. Hu, P. G. Moreno, Y. Xiao, X. Shen, G. Obozinski, N. D. Lawrence, and A. Damianou, "Empirical bayes transductive meta-learning with synthetic gradients," in *International Conference on Learning Representations*, 2020.

[29] C. Zhang, C. Li, and J. Cheng, "Few-shot visual classification using image pairs with binary transformation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2867–2871, 2019.

[30] X. Li, Q. Sun, Y. Liu, Q. Zhou, S. Zheng, T.-S. Chua, and B. Schiele, "Learning to self-train for semi-supervised few-shot classification," in *Neural Information Processing Systems*, vol. 32, 2019, pp. 10 276–10 286.

[31] S. Shao, L. Xing, Y. Wang, R. Xu, C. Zhao, Y.-J. Wang, and B.-D. Liu, "Mhfc: Multi-head feature collaboration for few-shot learning," in *ACM International Conference on Multimedia*, 2021.

[32] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," in *International Conference on Learning Representations*, 2021.

[33] W. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview hessian discriminative sparse coding for image annotation," *Computer Vision and Image Understanding*, vol. 118, pp. 50–60, 2014.

[34] B.-D. Liu, J. Meng, W.-Y. Xie, S. Shao, Y. Li, and Y. Wang, "Weighted spatial pyramid matching collaborative representation for remote-sensing-image scene classification," *Remote Sensing*, vol. 11, no. 5, p. 518, 2019.

[35] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Knowledge Discovery and Data Mining*, 2010, pp. 1–4.

[36] M. Liu, Y. Gao, P.-T. Yap, and D. Shen, "Multi-hypergraph learning for incomplete multimodality data," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1197–1208, 2017.

[37] Z. Zhang, H. Lin, X. Zhao, R. Ji, and Y. Gao, "Inductive multi-hypergraph learning and its application on view-based 3d object classification," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 5957–5968, 2018.

[38] Y. Lifchitz, Y. Avrithis, S. Picard, and A. Bursuc, "Dense classification and implanting for few-shot learning," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9258–9267.

[39] N. Dvornik, C. Schmid, and J. Mairal, "Selecting relevant features from a multi-domain representation for few-shot classification," in *European Conference on Computer Vision*. Springer, 2020, pp. 769–786.

[40] L. Liu, W. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," in *International Conference on Learning Representations*, 2021.

[41] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *International Conference on Machine Learning*, 2007, pp. 759–766.

[42] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[43] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *International Conference on Learning Representations*, 2019.

[44] C. Xing, N. Rostamzadeh, B. Oreshkin, and P. O. Pinheiro, "Adaptive cross-modal few-shot learning," in *Neural Information Processing Systems*, 2019, pp. 4847–4857.

[45] S. W. Yoon, J. Seo, and J. Moon, "Tapnet: Neural network augmented with task-adaptive projection for few-shot learning," in *International Conference on Machine Learning*, 2019, pp. 7115–7123.

[46] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding task-relevant features for few-shot learning by category traversal," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1–10.

[47] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *International Conference on Learning Representations*, 2020.

[48] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4136–4145.

[49] J. Kim, H. Kim, and G. Kim, "Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning," *European Conference on Computer Vision*, pp. 599–617, 2020.

[50] C. Chen, K. Li, W. Wei, J. T. Zhou, and Z. Zeng, "Hierarchical graph neural networks for few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[51] J. Oh, H. Yoo, C. Kim, and S.-Y. Yun, "Boil: Towards representation change for few-shot learning," in *International Conference on Learning Representations*, 2021.

[52] N. Fei, Z. Lu, T. Xiang, and S. Huang, "Melr: Meta-learning via modeling episode-level relationships for few-shot learning," in *International Conference on Learning Representations*, 2021.

[53] C. Xu, C. Liu, L. Zhang, C. Wang, J. Li, F. Huang, X. Xue, and Y. Fu, "Learning dynamic alignment via meta-filter for few-shot learning," in *Conference on Computer Vision and Pattern Recognition*, 2021.

[54] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *International Conference on Learning Representations*, 2018.

[55] Z. Yu, L. Chen, Z. Cheng, and J. Luo, "Transmatch: A transfer-learning scheme for semi-supervised few-shot learning," in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 856–12 864.

[56] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Neural Information Processing Systems*, vol. 29, 2016, pp. 3630–3638.

[57] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," in *International Conference on Learning Representations*, 2019.
[58] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.
[59] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[60] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Computer Science Department, University of Toronto*, 2009.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[62] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Neural Information Processing Systems*, 2002, pp. 585–591.
[63] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.
[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[65] M. Boudiaf, Z. I. Masud, J. Rony, J. Dolz, P. Piantanida, and I. B. Ayed, "Transductive information maximization for few-shot learning," in *Neural Information Processing Systems*, 2020.
[66] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, "Cross-domain few-shot classification via learned feature-wise transformation," in *International Conference on Learning Representations*, 2020.
[67] 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 *European Conference on Computer Vision*, 2020, pp. 438–455.
[68] I. Ziko, J. Dolz, E. Granger, and I. B. Ayed, "Laplacian regularized few-shot learning," in *International Conference on Machine Learning*, 2020, pp. 11 660–11 670.
[69] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International Conference on Machine Learning*, 2019, pp. 6438–6447.