Pseudo-loss Confidence Metric for Semi-supervised Few-shot Learning

Kai Huang, Jie Geng*, Wen Jiang*, Xinyang Deng, Zhe Xu
Northwestern Polytechnical University
KaiHuangk@mail.nwpu.edu.cn, gengjie, jiangwen, xinyang.deng@nwpu.edu.cn
alan.xu@mail.nwpu.edu.cn

Abstract

Semi-supervised few-shot learning is developed to train a classifier that can adapt to new tasks with limited labeled data and a fixed quantity of unlabeled data. Most semi-supervised few-shot learning methods select pseudo-labeled data of unlabeled set by task-specific confidence estimation. This work presents a task-unified confidence estimation approach for semi-supervised few-shot learning, named pseudo-loss confidence metric (PLCM). It measures the data credibility by the loss distribution of pseudo-labels, which is synthetical considered multi-tasks. Specifically, pseudo-labeled data of different tasks are mapped to a unified metric space by mean of the pseudo-loss model, making it possible to learn the prior pseudo-loss distribution. Then, confidence of pseudo-labeled data is estimated according to the distribution component confidence of its pseudo-loss. Thus highly reliable pseudo-labeled data are selected to strengthen the classifier. Moreover, to overcome the pseudo-loss distribution shift and improve the effectiveness of classifier, we advance the multi-step training strategy coordinated with the class balance measures of class-apart selection and class weight. Experimental results on four popular benchmark datasets demonstrate that the proposed approach can effectively select pseudo-labeled data and achieve the state-of-the-art performance.

1. Introduction

Deep learning has made great strides in many visual recognition tasks, and its outstanding performances even exceed human being in some scenarios [7]. However, it always relies on numerous labeled data which may be a heavy burden of data collection and maintenance in reality [35]. How to get rid of the limitation of labeled samples and learn a novel category only with one or few labeled samples is the core of few-shot learning. Since few-shot learning has great significance for its extensive applications on artificial intelligence, it aroused a growing academic interest in recent years.

As a typical transfer learning method, fine-tuning [5] is the preliminary exploration for transferring accumulated experience to new tasks. However, it is hard to perform the domain adaptation with only few training data, in which limited samples cannot represent the distribution of its class [26]. Episodes-based training strategy [6][39] clarifies the few-shot learning problem and has been the foundation of majority few-shot learning methods. Particularly, each episode learns a specific classification task, in which only a few samples per class are available for training. Performances are calculated on a series of episodes data for testing the ability of rapidly adapting to new tasks. Meta-based learning methods [6] [31] adopt the meta-learner to improve the capability of acclimatizing themselves to different tasks. Metric learning methods [11][12] attempt to find more effective distance metrics from numerous episodic tasks. Utilizing the unified metrics formula, the class distribution is more distinctive in metrics space.

More recently, there has been extensive research on semi-supervised few-shot learning (SSFSL), aiming to improve the model by utilizing a certain amount of unlabeled data. Predicting pseudo-labels of unlabeled samples and selecting high-confidence data for iterative training is a direct and valid way for SSFSL [20][40]. However, the task-specific confidence inference of pseudo-label suffers from lacking of adequate instances support in single task. To address this problem, we propose a task-free credibility estimation approach to select the credible pseudo-labeled data, by means of building a unitive confidence metric space.

In this paper, we focus on constructing the reliability estimation of pseudo-labeled samples and proposing a novel semi-supervised few-shot learning approach called pseudo-loss confidence metric (PLCM). The full procedure is showed in Algorithm 1 and illustrated in Figure 1. We first map pseudo-labeled data to pseudo-loss space by utilizing the pseudo-loss model, which can reflect the acceptance of current classifier to the unlabeled data with its pseudo-label. In general, classifier tends to give decep-

[*] Corresponding authors.

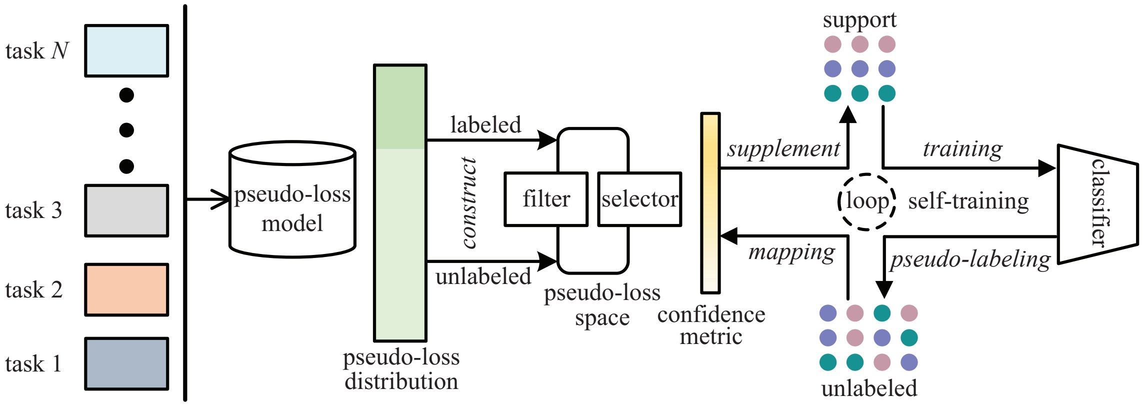


Figure 1. The overview of our proposed framework. On training process, we construct the pseudo-loss space according to the pseudo-loss model for multi-tasks. Then, a selector and a filter are developed to perform confidence metric for unselected and selected pseudo-labeled data respectively. Finally, the self-training strategy is adopted to fit the mixed set on evaluating process.

tive prediction results if the sample is hard to understand. With the undependable prediction, it is extremely possible to generate noisy pseudo-labels which cause heavier loss than correct pseudo-labels do. Based on this, we set up the semi-supervised Gaussian mixture model (ss-GMM) to fit the pseudo-loss distribution and allocate the credibility of pseudo-label according to learned distribution. Different from other samples selection-based SSFSL methods, our pseudo-loss confidence metric is based on the statistics of multi-episodes tasks, concentrating on generality and unity. Once the fitting is finished on training process, we estimate the credibility of pseudo-labels on evaluation process only through swift reasoning without any extra training.

The main contributions of this work are summarized as: 1) We present a novel pseudo-labeled data reliability estimation approach for semi-supervised few-shot learning, dubbed as pseudo-loss confidence metric (PLCM). Different from the previous work, we assess the confidence of pseudo-labeled data in a unified pseudo-loss metric space instead of apart between different tasks. 2) We devise the multi-step training strategy that learns a more flexible pseudo-loss distribution to follow the training of classifier, which offers stabler confidence metric. 3) Experimental results on four widely popular benchmark datasets for few-shot learning demonstrate that the proposed method achieves higher performance compared with other state-of-the-art methods.

1.1. Related Work

Few-shot Learning. The existing few-shot learning methods can summarize as three aspects: (1) Metric learning methods pay more attention to model the distance metrics to better discriminate classes. Matching Networks [39] train a fixed liner classifier with the distance of support set and query set in the embedding space. Prototypical Networks [32] search the prototype of different classes with a learned mapping function. (2) Meta learning methods aim to obtain a universal model which can rapidly adapt to new tasks. MAML [6] optimizes model parameters according to the gradients of multi-tasks, making it possible to fit new tasks with a few steps. MetaOptNet [17] learns the feature embedding by a linear classifier which is maintained as a convex learning problem. (3) Graph network methods explore the label structure or embedding structure between the samples of support set and query set. TPN [22] achieves label propagation from labeled instances to unlabeled test instances with a graph construction model. DPGN [43] combine the distribution-level relations and instance-level relations with a dual complete graph network.

Semi-supervised Learning. Semi-supervised learning (SSL) is developed on the condition that few labeled data and abundant unlabeled data are available, hoping to obtain the similar or even same performance as supervised learning. The existing SSL methods can be roughly summarized into three categories: (1) Self-training is the most widely used semi-supervised method because of its simplicity and effectiveness. Pseudo-labeling [16] provided by the most confident class of base classifier is a typical self-training approach. Furthermore, co-training [41] tries to understand data with multiple views for solving the accumulative error problem that occurs in self-training. (2) Consistency regularization methods focus on improving the robustness of model and keeping the label distribution even if images are noisy. π-model [15] introduces image augmentation as input noise and regularizes itself with the extra consistency loss. Mean Teacher [38] regularizes the model by adopting the exponential moving average of parameters. (3) Mix methods try to combine the current dominant approaches such as self-training and consistency regularization so as to obtain a unified framework of SSL. MixMatch method [3] mixes labeled and unlabeled data with MixUp and proposes a unified loss combined consistency regularization and entropy minimization. FixMatch method [33] demonstrates a simple but powerful model with the help of consistency regularization and pseudo-labeling.

2. Methodology

2.1. Problem Formulation

Definition. The goal of semi-supervised few-shot learning is to adapt to the task with only a few labeled data and a certain number of unlabeled data acquired. Specifically, for a S-shot W-way Q-query U-unlabeled task, S labeled samples from each of W classes comprise the support set $\mathcal{S}$, $Q\times W$ samples as unseen datapoints for evaluation make up the query set $\mathcal{Q}$, and U unlabeled samples from each of W classes constitute the unlabeled set $\mathcal{U}$. The model needs to classify the query set $\mathcal{Q}$ with only a few labeled samples of the support set $\mathcal{S}$ available, assist with a fixed number of unlabeled samples of the unlabeled set $\mathcal{U}$.

Training Process. Given a dataset $\mathcal{D}_{train}$ with a set of classes $\mathcal{C}_{train}$, $\mathcal{D}_{train}$ consists of a labeled sub-set $D_l = \{(I, y), y \in \mathcal{C}_{train}\}$ and an unlabeled sub-set $\mathcal{D}_{unl} = \{I_{unl}\}$ with the same classes set. We can sample many SS-FSL tasks with episodes [37] to train the model. In each episode, W classes are confirmed by randomly selecting from $\mathcal{C}_{train}$. (S + Q) labeled samples of each class are sampled from sub-set D_l to form support set and query set. U unlabeled samples per class from sub-set $\mathcal{D}_{unl}$ construct unlabeled set. Training is conducted by incessantly feeding support set, unlabeled set and query set with different episodes to the model.

Evaluation Process. Given another dataset $\mathcal{D}_{test}$ with a set of novel classes $\mathcal{C}_{test}$. Just like the training process, we evaluate the model with episodic tasks. Once an under-evaluated task is sampled from $\mathcal{D}_{test}$, the model should quickly adapt to it with the help of support set and unlabeled set, and then is tested on query set. The final classification performance is reported by averaging the results on query set with a series of episodic tasks.

2.2. Pseudo-loss Model for SSFSL

Loss model is usually applicable to the learning of data with noisy labels [2][18]. Noisy samples always have higher loss during the early training, making it possible to apply the mixture models to distinguish between clean samples and noisy samples from the loss distribution. Inspired by it, we extend the loss model to semi-supervised few-shot learning, which aims to identify reliable pseudo-labeled samples of unlabeled set so as to augment support set by learning its pseudo-loss distribution. Formally, consider $\mathcal{S} = \{(I_s, y_s), y_s \in \mathcal{C}_{train}\}$ as support set and $\mathcal{U} = \{(I_u)\}$ represents unlabeled set. At first support set is used to help classifier to adapt to this specific task. Let θ_s represents the parameter of classifier warmed up by support set, Thus we obtain the pseudo-labels of unlabeled set with the classifier:

$$y_p^u = \arg\ \max(P_c(I_u;\theta_s)) \ , \tag{1}$$

when P_c is the softmax output of classifier. The pseudo-loss of unlabeled set between the predictions of classifier with parameter θ_s and pseudo-label is formulated as:

$$L(\mathcal{U}|\theta_s) = \{-y_p^u \log(P_c(I_u;\theta_s)),\ I_u \in \mathcal{U}\} \ , \tag{2}$$

Similarly, the sample I_u with noisy pseudo-label often has higher pseudo-loss than that with clean pseudo-label. Therefore, it is possible to distinguish the unlabeled data with clean pseudo-labels by conducting confidence metric for pseudo-loss.

2.3. Pseudo-loss Confidence Metric

Pseudo-loss Space. Specifically, noisy pseudo-labeled data and clean pseudo-labeled data tend to have different pseudo-loss distributions. To a certain extent, each pseudo-loss component approximately follows normal distribution. Given that learned data normally have lower loss than unseen data, we divide pseudo-labeled samples into two branches: unselected set and selected set. Inevitably, both sets have noisy pseudo-labeled data and clean pseudo-labeled data. Therefore, the pseudo-loss space made up of unlabeled set consists of four pseudo-losses:

$$P(\mathcal{U}) = \{\ell | \ell \in L(\mathcal{U}|\theta_s, F_L, F_I)\}, \tag{3}$$

where F_L indicates whether its pseudo-label is clean or noisy, and F_I indicates whether this sample has been selected or not. On training process, we introduce labeled pseudo-loss instances by combining the pseudo-labels and the ground truth of query set, helping to learn the pseudo-loss distributions as supervised information.

Selector and Filter. Selector is designed to learn the pseudo-loss distributions of unselected set and recognize clean labeled data. Filter serves to screen noisy pseudo-labeled data from selected set which is mistakenly chosen by the selector. Thus, four-component ss-GMM is built to fit pseudo-loss distribution. Two components are for the selector and the other two are for the filter. The probability density function of the pseudo-loss ℓ_I with K components Gaussian mixture model is:

$$p(l_I) = \sum_{k=1}^{K} \pi_k g_k(\ell_I;\ \mu_k, \Sigma_k) \ , \tag{4}$$

where π_k indicates the weight of kth Gaussian component subject to $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. For a given pseudo-loss instance ℓ_I of sample I, $\pi_k g_k$ means the confidence of ℓ_I classified into *k*th Gaussian component. Suppose pseudo-loss instances are collected by N episodes, labeled pseudo-loss instances coming from query set of N episodes is $\mathcal{D}_L$, unlabeled pseudo-loss of unlabeled set indicates as $\mathcal{D}_U$. We then maximize the log-likelihood of both the labeled and unlabeled pseudo-loss instances to seek maximum likelihood estimate (MLE) of ss-GMM model:

$$\hat{\beta} = \arg\max_{\beta} \left[\log p \left(\{ \mathcal{D}_L, \mathcal{D}_U \} \mid \beta \right) \right], \tag{5}$$

where β indicates $\{(\pi_i, \mu_i, \Sigma_i) \mid 1 \leq i \leq K\}$, the log likelihood is further expressed as:

$$\begin{aligned} \log p \left(\mathcal{D}_L, \mathcal{D}_U | \beta \right) = \lambda \sum_{i=1}^{|\mathcal{D}_L|} \log p \left(y_L^i \mid \beta \right) p \left(\ell_L^i \mid y_L^i, \beta \right) \\ + (1 - \lambda) \sum_{i=1}^{|\mathcal{D}_U|} \log \, p \left(\ell_U^i \mid \beta \right), \end{aligned} \tag{6}$$

where λ is a weight coefficient introduced in [42] to balance the labeled and unlabeled information for parameters estimation, and can be calculated with $|\mathcal{D}_L| \, / \, (|\mathcal{D}_L| + |\mathcal{D}_U|)$. Since it is hard to solve the MLE analytically, EM algorithm is used to find a locally optimal solution with iterative procedures. In the E-step, the posterior probability of ℓ_U^i belongs to *k*th Gaussian component $p \left(g_k \mid \ell_U^i \right)$ obtained with the current ss-GMM, and the estimated parameters of the ss-GMM are updated in M-step.

Confidence Metric. After confirming the pseudo-loss distributions with the ss-GMM, we perform the confidence metric of pseudo-labeled data according to the posterior probability $p \left(g_c \mid \ell_U^i \right)$, where g_c is the Gaussian component with clean pseudo-label. The pseudo-labeled data with high posterior probability is more likely to own the clean pseudo-label and can be selected as authentic data for expanding support set. The classifier is then re-trained with the mixed data set:

$$\mathcal{S}_{mix} = \left\{ \left(I_S^i, y_S^i \right) \right\}_{i=1}^{|\mathcal{S}|} \cup \left\{ \left(I_{\mathcal{U}_s}^i, y_p^i \right) \right\}_{i=1}^{|\mathcal{U}_s|}, \tag{7}$$

where $\mathcal{U}_s$ denotes the selected samples of unlabeled set and y_p^i is the pseudo-label of sample $I_{\mathcal{U}_s}^i$.

Self-training. Furthermore, in order to dig out the available information of unlabeled set as thoroughly as possible, we update the pseudo-labels of unlabeled set and re-calculate its pseudo-loss distribution through re-training the classifier. In general, augmented training set is support to help classifier offers more reliable pseudo-labels. More samples of unlabeled set that meet the requirement are selected to enlarge the mixed set. The re-training and re-selecting are iterated until the number of selected instances and the predicted pseudo-labels keep stable.

2.4. Multi-step Strategy for Instance Selection

During self-training, the pseudo-loss distribution of unlabeled set changes along with the times of re-selecting and re-training. Conflict may arise if only the primary ss-GMM is used without any treatment. Figure 2 gives an example with the normalized loss instance $\ell = 0.53$. Classifier with only one time of re-selecting and re-training decides this instance as incredible data (see Figure 2a), while it actually more likely to be a reliable data for the classifier with five times re-selecting and re-training (see Figure 2b). Confidence metric turns dubious when the learned pseudo-loss distribution cannot follow the training of classifier.

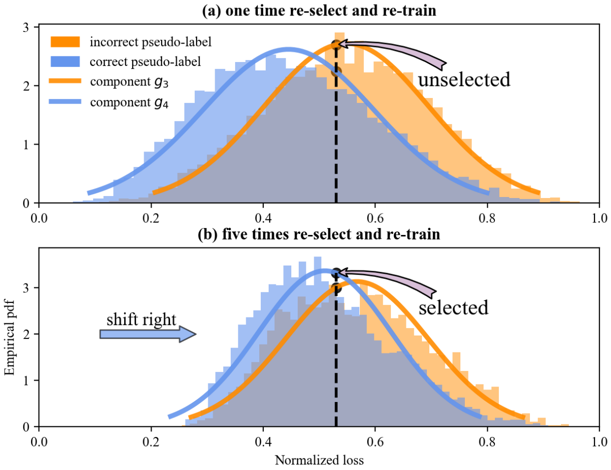


Figure 2. Comparison of pseudo-loss distribution with different times re-selecting and re-training. (a) shows re-selecting with unlabeled set and re-training with mix data only one time; (b) shows re-selecting and re-training for five times each episode task. For convenience, the figure shows only the pseudo-losses belonging to the Gaussian component g_3 and g_4.

To address this issue, we adjust the ss-GMM's parameters with the multi-step training strategy to catch the alteration of pseudo-loss distribution. We fit a ss-GMM group with each item corresponding to the classifier trained with different number of iterations. More specifically, the training process is divided into T number of steps, which corresponds to T times of re-selecting and re-training for each episode task. Each step is in charge of one ss-GMM, G_t represents the ss-GMM's parameters learned in *t*th step. The pseudo-loss instances for fitting G_t are calculated by the classifier re-selected and re-trained for t times. The reliable pseudo-labeled samples of *k*th ($k \leq t$) time is re-selected according to G_k, which is generated in *k*th step.

2.5. Class Balance

Generally, the supplementary data selected from unlabeled set are imbalanced between different classes, which causes volatile performance of classifier [19]. The situation gets worse if exiguous training data available, making the decision boundary unstable and tending to misclassify the samples of minority classes into majority classes. In particular, two serviceable class balance approaches are employed in our method to solve this problem.

Class-apart Selection. The imbalance of selected pseudo-labeled data is the cause leading to class imbalance. For this reason, we propose keeping the selected samples roughly equal per class with the class-apart selection. Specifically, the unlabeled samples are selected instead of the whole unlabeled set, considering that the confidence of pseudo-labels is separated from different classes.

Class Weight. We also adopt class weight [13] to the loss of classifier to balance the imbalanced data when the class-apart selection effect fades as alternative samples decreased. The major classes will get fewer weights and the samples of minor classes gain more attention in loss backward by applying the class weight.

3. Experiments

3.1. Datasets and Setups

3.1.1 Dataset

mini-ImageNet consists of 100 classes with 600 samples of size 84×84 per class, which are selected from ILSVRC-2012 [30]. Following [39], these classes are randomly divided into 64 classes for training, 16 classes for validation and 20 classes for evaluation.

tiered-ImageNet is a larger subset of ILSVRC-2012 [30] which contains 608 classes grouped into 34 higher-level category nodes with the hierarchical structure made by human beings. Following [28], we split these category notes into 20 (351 classes), 6 (97 classes), and 8 (160 classes) for training, validation and evaluation respectively. All images have the size 84×84.

CIFAR-FS is a variant of CIFAR-100 [14] with low-resolution. It has 100 object classes and each of them contains 600 samples of 32×32 color images. Following [4], dataset is partitioned into 64, 16 and 20 classes for training, validation and evaluation respectively.

FC100 is also based on the dataset CIFAR-100 [14] which provide more challenging scenario with low-resolution and super-classes. Following [25], the 100 classes are split into 20 super-classes, 12 super-classes (60 classes) for training, 4 super-classes (20 classes) for validation and 4 super-classes (20 classes) for evaluation.

3.1.2 Experimental setup

Network Architectures. For fair comparison, we adopt the ResNet-12 [8] as the feature extractor which consist of four residual blocks, each block has three 3×3 convolutional layers followed by a BatchNorm layer and a LeakyReLu activation. In addition, a 2×2 max-pooling layer is applied to reduce the size of output at the end of each block. Following [17], we utilize the Dropout [34] to prevent the overfitting. 10% output is dropped randomly in the first two blocks. At the end of final block, a mean-pooling layer is applied to refine the feature embedding of input images. The base classifier is Logistic Regression with L2 regularization.

Algorithm 1: PLCM Training Process

```
Input: dataset $\mathcal{D}_{train}$, step $T$, iteration $N$
Output: ss-GMM group $G$
for $t = 1$ to $T$ do
    Initialize: $\beta^t = \{\pi, \mu, \Sigma\}^t$ with Bayes estimation
    for $n = 1$ to $N$ do
        Sample an episodic task from $\mathcal{D}_{train}$
        iterative train the classifier $P_c$ with $S_{mix}$
        $\bar{y}_u, \bar{y}_q \leftarrow \arg\ \max(P_c(u, q; \theta_s))$
        $\bar{\ell}_u, \bar{\ell}_q \leftarrow -\bar{y}_{u,q} \log(P_c(u, q; \theta_s))$
        $y_{\bar{\ell}_q} \leftarrow L(\bar{y}_q, y_q | F_L, F_I)$
    end
    while $\beta^t$ is not converged do
        $\gamma_{uk} = p\left(g_k \mid \bar{\ell}_u\right) \leftarrow \frac{\pi_k g_k\left(\bar{\ell}_u ; \mu_k, \Sigma_k\right)}{\sum_{k=1}^K \pi_k g_k\left(\bar{\ell}_u ; \mu_k, \Sigma_k\right)}$
        $\gamma_{qk} = p\left(g_k \mid \bar{\ell}_q\right) \leftarrow 1$ if $y_{\bar{\ell}_q} == k$ else $0$
        $\pi_k \leftarrow \frac{\lambda \sum_i \gamma_{qk}^i + (1-\lambda) \sum_j \gamma_{uk}^j}{\lambda |q| + (1-\lambda)|u|}$
        $\mu_k \leftarrow \frac{\lambda \sum_i \gamma_{qk}^i \bar{\ell}_q^i + (1-\lambda) \sum_j \gamma_{uk}^j \bar{\ell}_u^j}{\lambda \sum_i \gamma_{qk}^i + (1-\lambda) \sum_j \gamma_{uk}^j}$
        $\Sigma_k \leftarrow \frac{\lambda \sum_i \gamma_{qk}^i \bar{\ell}_q^i \left(\bar{\ell}_q^i - \mu_k\right)^2 + (1-\lambda) \sum_j \gamma_{uk}^j \bar{\ell}_u^j \left(\bar{\ell}_u^j - \mu_k\right)^2}{\lambda \sum_i \gamma_{qk}^i + (1-\lambda) \sum_j \gamma_{uk}^j}$
    end
    $G^t \leftarrow \beta^t$
end
```

Hyperparameters. For the training of feature extraction network, the base learning rate is set to 0.1 initially and decay 10 times every 30 epochs with the total 120 epochs. The number of episodes for collecting pseudo-loss instances for the ss-GMM fitting is set to 600. We conduct 10 steps re-selecting and re-training for instances selection on evaluating process.

Comparing Methods. We compare our algorithm with other method mainly in the three aspects. (1) In terms of **basic semi-supervised setting**, we compare our method with the recently generalized SSFSL alternatives: TPN [22], TransMatch [44], LST [20], EPNet [29] and ICI [40]. Since the number of unlabeled samples is a key factor for the semi-supervised few-shot learning, we report our comparison results under the same semi-supervised condition. Following [20], the experiments on 5-way 1-shot use 30 unlabeled data each class and 50 is for 5-way 5-shot. Meanwhile, we conduct the transductive setting experiments for validating the effectiveness of our framework. The comparison includes the SSFSL methods applied in the transductive setting [22] [40] and other current TFSL approaches [9][10][27] . (2) In terms of **distraction semi-supervised setting**, we compare our method with other SSFL meth-

Table 1. The 5-way few-shot classification test accuracies with 95% confidence intervals over 600 episodes on mini-ImageNet and tiered-ImageNet. † indicates that it is implemented by public code. **In.** means the methods tested in inductive setting, **Tran.** denotes transductive setting with 15 query and **Semi.(30/50)** is semi-supervised setting with 30 unlabel for 5-way 1-shot and 50 unlabel for 5-way 5-shot.

Setting	Method	Backbone	mini-ImageNet		tiered-ImageNet	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
In.	MatchingNet[39]	4 CONV	43.56±0.84%	55.31±0.73%	-	-
	MAML[6]	4 CONV	48.70±1.84%	63.11±0.92%	51.67±1.81%	70.30±1.75%
	ProtoNet[32]	4 CONV	49.42±0.78%	68.20±0.66%	53.31±0.89%	72.69±0.74%
	LEO[31]	WRN-28-10	61.76±0.08%	77.59±0.12%	66.33±0.05%	81.44±0.09%
	DeepEMD[45]	ResNet-12	65.91±0.82%	82.41±0.56%	71.16±0.87%	86.03±0.58%
Tran.	SIB[10]	ResNet-12	70.00±0.60%	79.20±0.40%	72.90±0.65%	82.80±0.37%
	CAN+T[9]	ResNet-12	67.19±0.55%	80.64±0.35%	73.21±0.58%	84.93±0.38%
	BD-CSPN[21]	WRN-28-10	70.31±0.93%	81.89±0.60%	78.74±0.95%	86.92±0.63%
	E^{3}BM[23]	WRN-28-10	**71.40±0.50%**	81.20±0.40%	75.60±0.60%	84.30±0.40%
Semi. → Tran.	TPN[22]	4 CONV	55.51±0.86%	69.86±0.65%	59.91±0.94%	73.30±0.75%
	EPNet[29]	ResNet-12	66.50±0.89%	81.06±0.60%	76.53±0.87%	87.32±0.64%
	ICI[40]	ResNet-12	66.80±1.10%	79.26±0.68%	80.79±1.11%	87.92±0.69%
	PLCM (ours)	ResNet-12	70.92±1.03%	**82.74±0.55%**	**82.61±1.08%**	**89.47±0.56%**
Semi.(30/50)	TPN[22]	4 CONV	52.78±0.27%	66.42±0.21%	55.74±0.23%	71.01±0.17%
	TransMatch[44]	WRN-28-10	60.02±1.02%	79.30±0.59%	72.19±1.27%	82.12±0.92%
	LST[20]	ResNet-12	70.01±1.90%	78.70±0.80%	77.70±1.60%	85.20±0.80%
	EPNet†[29]	ResNet-12	70.50±1.32%	80.20±0.77%	75.90±1.18%	82.11±0.62%
	ICI[40]	ResNet-12	69.66±1.13%	80.11±0.72%	84.01±1.03%	89.00±0.67%
	PLCM (ours)	ResNet-12	**72.06±1.08%**	**83.71±0.63%**	**84.78±0.96%**	**90.11±0.57%**

Table 2. The 5-way few-shot classification test accuracies with 95% confidence intervals on CIFAR-FS. * denotes that it is reported in [4]. the best-performing result is highlighted.

Method	**Backbone**	**5-way 1-shot**	**5-way 5-shot**
ProtoNet*[32]	4 CONV	55.50±0.70%	72.00±0.60%
MAML*[6]	4 CONV	58.90±1.90%	71.50±1.00%
R2D2[4]	4 CONV	65.30±0.20%	79.40±0.10%
TEAM[27]	ResNet-12	70.43±1.03%	81.25±0.92%
MetaOptNet[17]	ResNet-12	72.00±0.70%	84.20±0.50%
ICI[40]	ResNet-12	76.51±1.22%	84.32±0.70%
PLCM (ours)	ResNet-12	**77.62±1.15%**	**86.13±0.67%**

Table 3. The 5-way few-shot classification test accuracies with 95% confidence intervals on FC100. * denotes that it is reported in [17]. the best-performing result is highlighted.

Method	**Backbone**	**5-way 1-shot**	**5-way 5-shot**
ProtoNet*[32]	4 CONV	37.50±0.60%	52.50±0.60%
TADAM[25]	4 CONV	40.10±0.40%	56.10±0.40%
MetaOptNet[17]	ResNet-12	41.10±0.60%	55.50±0.60%
MatchNet[39]	ResNet-12	43.88±0.75%	57.05±0.71%
SIB[10]	WRN-28-10	45.20±0.81%	55.90±0.74%
MTL[36]	ResNet-12	45.10±1.80%	57.60±0.90%
E^{3}BM[24]	WRN-28-10	46.00±0.60%	57.10±0.40%
Centroid[1]	ResNet-18	45.83±0.48%	59.74±0.56%
PLCM (ours)	ResNet-12	**48.35±1.00%**	**62.75±0.82%**

ods under more realistic conditions, in which unlabeled set contains distractive classes that are excluded in support set [20][28]. Since few researchers pay attention to it, we report part of the results neatened by LST [20], and remaining results are performed by the public codes [29][40]. Following [28], we test our method with both the 5-way 1-shot 5-unlabeled and 5-way 5-shot 20-unlabled, and use 5 distracting classes with the same samples of unlabeled set. (3) In terms of **variety-unlabeled semi-supervised setting**, we compare our method with the condition that the number of unlabeled samples in each class are different, which are set to 15, 30, 50, 80 and 100 respectively. The comparison results are mainly from LST [20], ICI [40], EPNet [29] and their public codes.

3.2. Experiment Results

Basic Semi-supervised Few-shot Setting. We compare our method with several current approaches on mini-ImageNet, tiered-ImageNet, CIFAR-FS and FC100. From Table 1, 2 and 3, we obtain the following conclusion: (1) The proposed PLCM shows substantial gains compared with other existing SSFSL methods and achieves the state-of-the-art performance of all few-shot settings and datasets. (2) Our method is superior to other existing semi-supervised few-

Table 4. The 5-way few-shot classification test accuracies with distraction semi-supervised setting. * denotes that it is carried out with their public codes. The best-performing result is highlighted.

Method	mini		tiered	
	1 shot	5-shot	1-shot	5-shot
MS k-Means[22]	49.0%	63.0%	51.4%	69.1%
TPN[28]	50.4%	64.9%	53.5%	69.9%
TPN with MTL[20]	61.3%	72.4%	71.5%	82.7%
LST[20]	64.1%	77.4%	73.5%	83.4%
EPNet*[29]	64.7%	76.8%	72.2%	82.1%
ICI*[40]	65.4%	75.1%	75.4%	82.5%
PLCM	**68.5%**	**80.2%**	**79.1%**	**87.8%**

Table 5. We report the compared result of our model without or with the selected filter. For convenience, only 3 to 10 steps are showed this table. The former results of "·/·" are obtained by model without selected filter and the latter takes the selected filter.

Step	3	4	5	6
error rate (%)	2.4/3.7	4.7/5.2	5.2/5.8	9.4/9.1
Accuracy (%)	80.5/80.1	82.0/81.2	82.2/81.1	81.7/81.8

Step	7	8	9	10
error rate (%)	13.2/10.7	16.0/12.6	18.1/12.8	20.4/13.2
Accuracy (%)	81.4/82.3	80.3/82.9	79.1/83.3	78.8/83.7

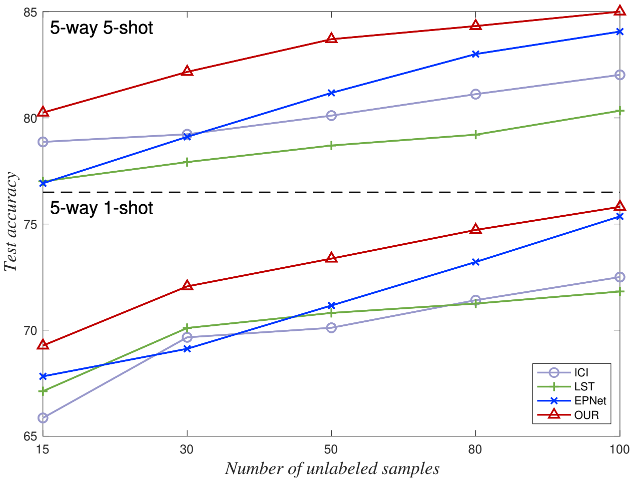


Figure 3. The comparison results of semi-supervised few-shot classification with varied unlabeled samples on mini-ImageNet.

shot learning methods with the same settings, especially for credibility samples selection based SSFSL methods such as LST [20] and ICI [40]. The outstanding performances indicate that the confidence estimate of pseudo-labeled data in our method is more precise and can take advantage of unlabeled information more effectively. (3) The experimental results on transductive setting indicate that our method also achieves competitive performance which further shows the effectivity and robustness under the condition of lacking both labeled data and unlabeled data.

Table 6. Several class balance results with base SSFSL setting and distractive SSFSL setting on mini-ImageNet. 'CW' denotes class weight and 'CSS' denotes class-apart selection. Our method uses both class weight and class-apart selection and both LST and ICI methods already use class-apart selection.

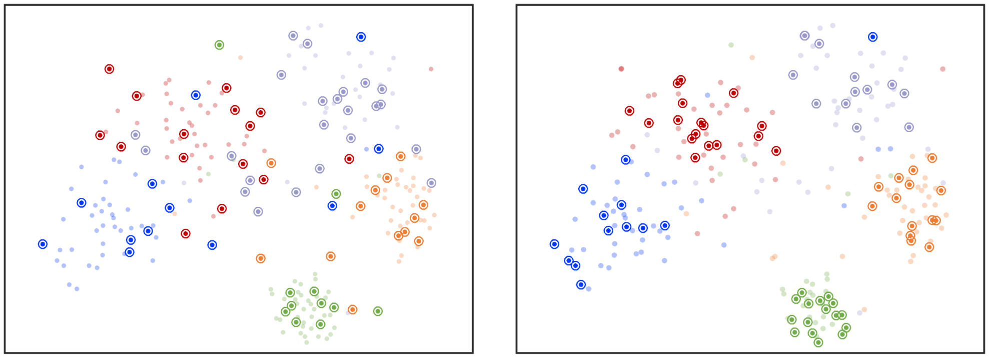
Setting	Method	Base SSFL		Distrative SSFL	
		1 shot	5-shot	1-shot	5-shot
+cw	LST[20]	65.58	70.43	58.27	70.86
	ICI[40]	64.42	73.81	57.29	69.29
	PLCM	**68.28**	**78.19**	**62.81**	**74.45**
+css	LST[20]	70.01	78.70	64.12	77.39
	ICI[40]	69.66	80.11	65.37	75.11
	PLCM	**71.76**	**83.03**	**67.73**	**79.60**
+css & +cw	LST[20]	70.55	79.11	64.82	77.95
	ICI[40]	70.07	80.60	65.91	75.57
	PLCM	**72.06**	**83.71**	**68.50**	**80.21**

Distractive Semi-supervised Few-shot Setting. In reality, it is hard to get the clean unlabeled set without mixing any data of other classes. In order to show the adaptability of our method, we compare the PLCM with several other SSFSL methods on distractive semi-supervised few-shot setting. The comparison results are presented in Table 4. It is clear that our approach is more effective than other existing SSFSL methods and achieve the highest accuracy in all distractive semi-supervised few-shot classification settings.

Variety-unlabeled Semi-supervised Few-shot Setting. To validate the robustness of our framework under variety-unlabeled semi-supervised few-shot setting, we conduct the 5-way 1-shot and 5-shot experiments on mini-ImageNet with different number of unlabeled samples. Figure 3 shows the variation of test accuracy as the number of unlabeled samples increases. Obviously, our method performs the best in all variety-unlabeled semi-supervised few-shot settings. Compared to other SSFSL methods which consider only samples selection on specific tasks, PLCM utilizes the pseudo-loss distribution of multi-tasks to establish a unitary selection mechanism in a more sufficient and steady way.

3.3. Ablation Studies

Effectiveness of PLCM. Figure 4 visualizes the samples selection process for one task. We compare our method's selection performance with that of the hard selection approach adopted by the LST method, which only picks out the pseudo-labeled samples with high prediction scores. It is obvious that the samples selected with our method are more centralized and distinct than the LST method. Figure 4 (a) shows that the selection with the LST method is disheveled, and easier to select the samples away from their class cluster. Since PLCM selects pseudo-labeled samples according to the unitary pseudo-loss space, we can pick out



(a)

(b)

Figure 4. The t-SNE visualization of samples selection for 5-way 5-shot 50-unlabeled task. The points with different colors indicate unlabeled samples with different classes and the circled points mean selected samples. For convenience, we only show the samples selection on third loop (75 unlabeled samples are selected approximately) by hard selection method and our PLCM selection method.

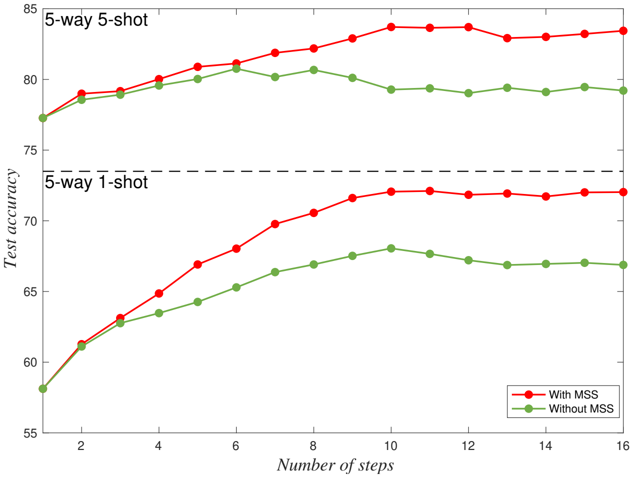


Figure 5. The results of model with or without multi-step strategy on mini-ImageNet. 'MSS' denotes multi-step strategy.

more credible supplementary data to boost the robustness of decision boundary.

Impact of Selected Filter. We aim to collect the samples with high reliable pseudo-labels by selector. However, classifier may also adapt to wrong pseudo-labeled samples once the performance of selector is poor. The pseudo-loss of these pseudo-samples will be more similar to correct labeling samples, which puzzle the selector further. Therefore, the filter works as an important component of PLCM by picking out the incorrect information from selected data. As showed in Table 5, it is obvious that the filter can slow down the increase of the error ratio of selected samples during multi-step training. Profiting from cleaner supplementary information, the model with the filter achieve better results.

Impact of Class Balance. We further analyze the effect of class balance approaches in SSFSL. Table 6 shows that: (1) the accuracy descends down when we weaken the class balance measures and our method still achieves competitive performance. (2) Both class-apart selection and class weight pay a positive role not only in our method but also other SSFSL methods, proving its importance for SSFSL. (3) Since class-apart selection roughly keeps class balance of selected samples and class weight offers more detailed attention to loss, our model reduces the influence of class imbalance to a larger degree and gains higher accuracies.

Impact of Multi-step Training Strategy. Figure 5 shows the influence of the multi-step training strategy in our method on mini-ImageNet. It is clear that the performance of the multi-step strategy is superior to the normal loop strategy with the increase of step or loop. Owing to the self-adaptive pseudo-loss distribution fitting by the multi-step training strategy, our confidence metric synchronizes the classifier and works more effectively and precisely.

4. Conclusions

In this paper, we propose a task-unified pseudo-loss confidence metric for semi-supervised few-shot learning. It can effectively estimate the quality of pseudo-labeled data and exploit useful unlabeled data to enhance the training of classifier. A unified pseudo-loss space by jointing multi-tasks is constructed for confidence metric, which is proved to select superior pseudo-labeled data. Further, multi-step training strategy is able to learn more credible pseudo-loss distribution to follow the training of classifier, which contributes to the confidence metric. Extensive experiments on four few-shot settings, including transductive setting, base, distractive and variety-unlabeled semi-supervised setting, demonstrate that our method outperforms other algorithms.

Acknowledgments. This research was supported in part by the National Natural Science Foundation of China under Grant 61901376 and Shaanxi key research and development program under Grant 2021ZDLGY01-04.

References

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Associative alignment for few-shot image classification. In *ECCV*, volume 12350, pages 18–35, 2020. 6

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, pages 312–321, 2019. 3

[3] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, pages 5050–5060, 2019. 3

[4] Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *ICLR*, 2019. 5, 6

[5] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019. 1

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 1, 2, 6

[7] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 1

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[9] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *NeurIPS*, pages 4003–4014, 2019. 5, 6

[10] Shell Xu Hu, Pablo G Moreno, Yang Xiao, Xi Shen, Guillaume Obozinski, Neil D Lawrence, and Andreas Damianou. Empirical bayes transductive meta-learning with synthetic gradients. In *ICLR*, 2020. 5, 6

[11] 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*, pages 1–1, 2020. 1

[12] L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, and A. M. Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *CVPR*, pages 5192–5201, 2019. 1

[13] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001. 5

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[15] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 2

[16] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML*, 2013. 2

[17] K. Lee, S. Maji, A. Ravichandran, and S. Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, pages 10649–10657, 2019. 2, 5, 6

[18] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020. 3

[19] Shoushan Li, Zhongqing Wang, Guodong Zhou, and Sophia Yat Mei Lee. Semi-supervised learning for imbalanced sentiment classification. In *IJCAI*, pages 1826–1831, 2011. 4

[20] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *NeurIPS*, 2019. 1, 5, 6, 7

[21] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In *ECCV*, pages 741–756. Springer, 2020. 6

[22] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *ICLR*, 2019. 2, 5, 6, 7

[23] Yaoyao Liu, Bernt Schiele, and Qianru Sun. An ensemble of epoch-wise empirical bayes for few-shot learning. In *ECCV*, pages 404–421. Springer, 2020. 6

[24] Yaoyao Liu, Bernt Schiele, and Qianru Sun. An ensemble of epoch-wise empirical bayes for few-shot learning. In *ECCV*, volume 12361, pages 404–421. Springer, 2020. 6

[25] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, pages 721–731, 2018. 5, 6

[26] N. Patricia and B. Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *CVPR*, pages 1442–1449, 2014. 1

[27] L. Qiao, Y. Shi, J. Li, Y. Tian, T. Huang, and Y. Wang. Transductive episodic-wise adaptive metric for few-shot learning. In *ICCV*, pages 3602–3611, 2019. 5, 6

[28] Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018. 5, 6, 7

[29] Pau Rodríguez, Issam Laradji, Alexandre Drouin, and Alexandre Lacoste. Embedding propagation: Smoother manifold for few-shot classification. In *ECCV*, 2020. 5, 6, 7

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[31] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019. 1, 6

[32] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4080–4090, 2017. 2, 6

[33] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 3

[34] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 5

[35] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852, 2017. 1

[36] Q. Sun, Y. Liu, T. Chua, and B. Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, pages 403–412, 2019. 6

[37] 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 *CVPR*, pages 1199–1208, 2018. 3

[38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, pages 1195–1204, 2017. 2

[39] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016. 1, 2, 5, 6

[40] Y. Wang, C. Xu, C. Liu, L. Zhang, and Y. Fu. Instance credibility inference for few-shot learning. In *CVPR*, pages 12833–12842, 2020. 1, 5, 6, 7

[41] Q. Xie, M. T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10684–10695, 2020. 2

[42] H. Yan, J. Zhou, and C. K. Pang. Gaussian mixture model using semisupervised learning for probabilistic fault diagnosis under new data categories. *IEEE Transactions on Instrumentation and Measurement*, 66(4):723–733, 2017. 4

[43] L. Yang, L. Li, Z. Zhang, X. Zhou, E. Zhou, and Y. Liu. Dpgn: Distribution propagation graph network for few-shot learning. In *CVPR*, pages 13387–13396, 2020. 2

[44] Z. Yu, L. Chen, Z. Cheng, and J. Luo. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *CVPR*, pages 12853–12861, 2020. 5, 6

[45] C. Zhang, Y. Cai, G. Lin, and C. Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*, pages 12200–12210, 2020. 6