SEMI-SUPERVISED FEW-SHOT LEARNING WITH PSEUDO LABEL REFINEMENT

*Pan Li, Guile Wu, Shaogang Gong and Xu Lan**

Queen Mary University of London
{pan.li, guile.wu, s.gong, x.lan}@qmul.ac.uk

ABSTRACT

Few-shot classification aims at recognising novel categories with very limited labelled samples. Although substantial achievements have been obtained, few-shot classification remains challenging due to the scarcity of labelled examples. Recent studies resort to leveraging unlabelled data to expand the training set using pseudo labelling, but this strategy often yields significant label noise. In this work, we introduce a new baseline method for semi-supervised few-shot learning by iterative pseudo label refinement to reduce noise. Then, we investigate the label noise propagation problem and improve the baseline with a denoising network to learn distributions of clean and noisy pseudo-labelled examples via a mixture model. This helps to estimate confidence values of pseudo labelled examples and to select the reliable ones with less noise for iteratively refining a few-shot classifier. Extensive experiments on three widely used benchmarks, miniImagenet, tieredImagenet and CIFAR-FS, show the superiority of the proposed methods over the state-of-the-art methods.

***Index Terms*—** Semi-Supervised Few-Shot Learning, Pseudo Label Refinement, Mixture Model

1. INTRODUCTION

Few-shot classification is a challenging task aiming at recognising novel classes with limited labelled data. Conventional deep neural networks often fail in this task because they contain lots of model parameters which lead to overfitting to the scarce labelled data. To solve this problem, many few-shot learning solutions have been proposed recently [1, 2, 3, 4]. A general pipeline is training a recognition model with sufficient labelled data from base categories and then fine-tuning a new classifier for novel categories. However, due to the scarcity of labelled examples from new classes, traditional few-shot classification methods usually yield inferior performance.

To alleviate this drawback, some studies [5, 6, 7] resort to semi-supervised few-shot learning (SS-FSL) by leveraging additional unlabelled data from novel classes. Contemporary SS-FSL approaches mainly follow a meta-learning pipeline and use pseudo label estimation (*e.g.* soft k-means clustering with masking [5], label propagation [6] and self-training with hard and soft pseudo labels [7]) to leverage both scarce labelled data and abundant unlabelled data for learning a meta-learner. However, these methods require to mimic SS-FSL tasks during meta-training and meta-testing stages, resulting in sophisticated episodic learning processes and poor extension ability. On the other hand, recent study [8] adopts a transfer-learning pipeline by pre-training a feature extractor, imprinting classifier weights for novel classes and updating the model with an off-the-self semi-supervised learning method. But such a simple combination with off-the-self semi-supervised methods without careful adjustments usually results in sub-optimal performance for SS-FSL.

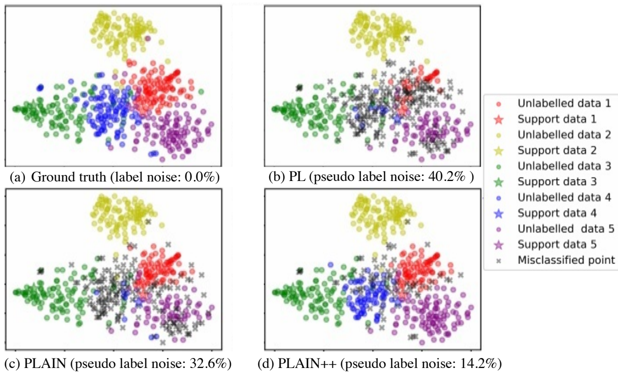

Fig. 1. Visualisation of embeddings of 5-way 1-shot tasks with 100 unlabelled data per class on miniImagenet. (a) shows distributions of support exemplars and unlabelled data with ground truth labels, whilst (b), (c), (d) show distributions of support exemplars and unlabelled data with pseudo labels estimated by pseudo-labelling (PL), PLAIN and PLAIN++. Round points, stars and black crosses represent unlabelled data, support exemplars and misclassified points, respectively.

In this work, we introduce a simple baseline method for SS-FSL by modifying a transfer-learning framework with **P**seudo **LA**bel ref**IN**ement (**PLAIN**). Pseudo labelling [9] is one of the key techniques for assigning labels of unlabelled samples in novel classes. A common practice is to estimate

∗ Corresponding author. This work was supported by Vision Semantics Limited, the Alan Turing Institute Turing Fellowship, the China Scholarship Council, and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

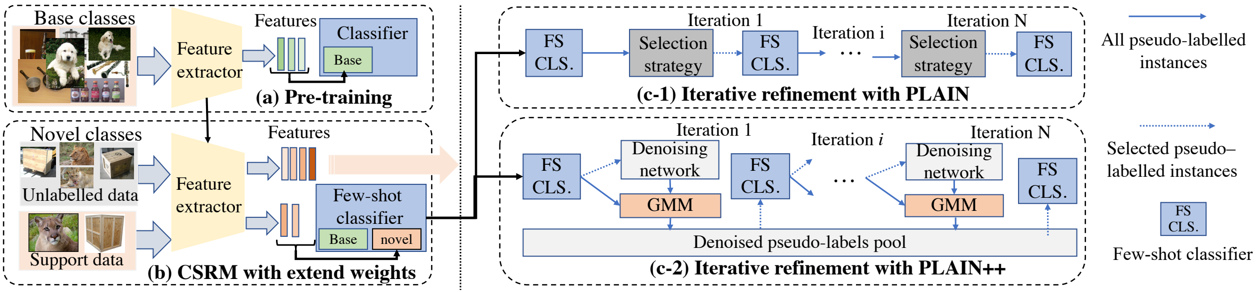

Fig. 2. The overall framework of PLAIN and PLAIN++ for semi-supervised few-shot learning. The baseline method PLAIN consists of (a), (b) and (c-1), while PLAIN++ contains (a), (b) and (c-2).

pseudo labels for unlabelled data with an initial classifier and then update the classifier with pseudo labelled data. However, this approach is usually affected by noisy pseudo labels, leading to inaccurate prediction (*e.g.* blue points in Fig. 1 (b)). Thus, in this work, we develop a method called PLAIN for SS-FSL by integrating iterative self-training with reliable pseudo-label selection into a transfer learning framework. As shown in Fig. 2 (a), (b) and (c-1), we pre-train a feature extractor, fine-tune a cosine similarity based recognition model with classification weights for novel classes, and then iteratively refine pseudo labels to learn a classifier without elaborately sampling meta tasks or adopting off-the-self semi-supervised learning methods. This baseline method is simple but can effectively refine reliable pseudo labels (*e.g.* red and blue points in Fig. 1 (c)) for learning a few-shot classifier.

Given that pseudo labels are iteratively updated using a fixed feature extractor in PLAIN, it is inevitable that noisy pseudo labels produced by the bias of the feature extractor will be easily amplified in the refinement process, causing the *label noise propagation* problem [10]. To further address this problem, we improve PLAIN with a denoising network to reduce pseudo label noise via adapting knowledge on novel classes and a Gaussian Mixture Model (GMM) to learn distributions of clean and noisy pseudo-labels for obtaining reliable pseudo-labelled instances, resulting in an advanced SS-FSL method called PLAIN++. As shown in Fig. 2 (c-2), compared with PLAIN, PLAIN++ requires to train a denoising network using pseudo labelled examples with high confidence. We use this denoising network to evaluate confidence values of pseudo labels with GMM, which models distributions of clean and noisy pseudo labelled examples, so that we can select η percentage of pseudo labels to update the few-shot classifier. This process is alternately performed until the pre-defined number of iterations. Thus, PLAIN++ can help to estimate the confidence values of pseudo labelled examples and alleviate pseudo label noise (*e.g.* Fig. 1 (d)) by pseudo-labelled examples selection in each iterative step.

Our **contributions** are: **(1)** We introduce a simple yet effective baseline (PLAIN) for SS-FSL. Although it uses some basic ideas of existing methods (*e.g.* pseudo labelling), it is a new formulation achieving competitive performance against existing complex SS-FSL methods. **(2)** We discuss the label noise propagation issue and further propose PLAIN++ with a denoising network and a mixture model. **(3)** Extensive experiments on three widely used benchmarks (miniImagenet [11], tieredImagenet [5] and CIFAR-FS [12]) show the superiority of PLAIN and PLAIN++ over the state-of-the-art methods.

2. RELATED WORK

Few-Shot Classification can be categorised as metric-based and gradient-based methods. Metric-based methods [11, 1] focus on learning a generalised feature space where data from the same class can be easily distinguished from those from different classes using a distance metric, whilst gradient-based methods [3, 13] use a meta-learner as an optimiser for learning to learn model's meta parameters. But these methods usually suffer from intrinsic drawback brought by limited labelled data, and therefore achieves inferior performance.
Semi-Supervised Few-Shot Learning (SS-FSL) mostly follow a meta-learning pipeline and estimate pseudo labels for unlabelled data to update classifier. Ren *et al.* [5] propose to extend ProtoNet [1] for SS-FSL by adopting soft k-means to estimate pseudo labels for unlabelled data. Li *et al.* [7] propose a learning to self-train (LST) method to meta-learn a soft weight network for unlabelled data. However, these methods show poor extension ability for dynamically recognising novel classes and require episodic training. Recently, TransMatch [8] uses a transfer-learning framework for SS-FSL by learning a cosine similarity based recognition model without episodic training, but it does not consider pseudo label noise for unlabelled data, resulting in sub-optimal performance.
Semi-Supervised Learning aims to leverage unlabelled data to learn a model that better fits underlying data distributions. Conventional solutions (*e.g.* consistency regularisation [14] and entropy minimisation [15]) have shown promising performance for semi-supervised learning but they cannot be readily used in SS-FSL because of the scarcity of labelled examples.

3. METHODOLOGY

Problem Definition. Suppose we have a large-scale dataset D_b which contains sufficient labelled examples from base classes in C_b and a small-scale dataset D_n which has only a few labelled examples and some unlabelled examples from novel classes in C_n, where C_n is disjoint from C_b. The aim of SS-FSL is to learn a classifier for recognising novel classes using both few labelled examples and unlabelled examples in D_n and labelled examples in D_b as auxiliary data. Generally, a small support set of N classes with K labelled exemplars per class is sampled from D_n, resulting to a N-way K-shot problem. Besides, additional R unlabelled images are sampled from each of the N novel classes or distractor classes.

3.1. PLAIN: A Baseline Method for SS-FSL

As shown in Fig. 2, there are three steps in PLAIN: (a) Pre-training, (b) Cosine similarity based recognition model with extended weights, and (c-1) Iterative pseudo label refinement.

Pre-Training. We learn a Cosine-Similarity based Recognition Model (CSRM) [16, 17] $f_{(\theta,W)}$, which includes a feature extractor Φ_θ and a classifier $\sigma(\Phi_\theta|W)$ with classification weights $W = W_b$ for base categories, on a base training dataset $D_b = \bigcup_{b=1}^{C_b} \{x_{b,i}\}_{i=1}^{N_b}$ with C_b categories. We optimise this model using cross entropy loss, which is formulated as: $\frac{1}{C_b}\sum_{b=1}^{C_b} \frac{1}{N_b}\sum_{i=1}^{N_b} loss(x_{b,i}, b)$, where $loss(x_{b,i}, b) = -log(p_b)$ and p_b is the probability of $x_{b,i}$ over the b-th category. Then, we evaluate this model on a validation set to get a feature extractor Φ_{θ^*} with the best generalisation.

Cosine Similarity based Recognition Model with Extended Weights. After pre-training, we get a CSRM $f_{(\theta^*,W_b)}$ and extend its classification weights as $W_e = W_b \bigcup W_n$, where W_n are the classification weights for novel categories. Specifically, suppose there are $N(N \geqslant 1)$ support exemplars $x^i_{sup.}\{i = 1, ..., N\}$ per class, we infer classification weights $W_{sup.}$ of a class by averaging feature vectors of training exemplars from that class: $W_{sup.} = \frac{1}{N}\sum_{i=1}^{N} \Phi_{\theta^*}(x^i_{sup})$. Then, we normalise the weight vectors to unit length $W_n = \frac{W_{sup.}}{||W_{sup.}||}$ and concatenate the weights for base and novel categories to get classification weights $W_e = W_b \bigcup W_n$, resulting in an extended CSRM $f_{(\theta^*,W_e)}$ for recognising both base and novel categories. In this work, we use the CSRM $f_{(\theta^*,W_n)}$ as the few-shot classifier to recognise novel categories.

Iterative Pseudo Label Refinement. After the first two steps, we use pseudo label refinement with iterative self-training to learn a classifier with unlabelled data. Specifically, we use the few-shot classifier $f_{(\theta^*,W_n)}$ to estimate pseudo labels for unlabelled data from C_n based on their probability, and then we can select pseudo labels with high prediction confidence to fine-tune the classification weights W_n of $f_{(\theta^*,W_n)}$ by averaging the feature embeddings of support and selected pseudo-labelled instances. As shown in Fig. 2(c-1), this process is iteratively performed to remit the label noise and gradually improving the classifier. We summarise the training process of PLAIN in Algorithm 1 in the supplementary material. Code will be available in `https://github.com/panli93/SSFSL_PLAIN`.

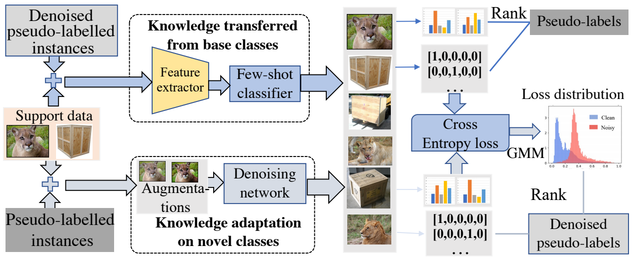

Fig. 3. The pipeline of iterative pseudo label refinement with pseudo label denoising in the proposed PLAIN++.

3.2. PLAIN++ for Resolving Label Nosie Propagation

Label Noise Propagation. During the iterative refinement process, once a sample is assigned with an incorrect label (*e.g.* blue points in Fig. 1), it might suffer from incorrect prediction in the subsequent iterations and be assigned with higher confidence value. This causes the *label noise propagation* issue (a.k.a. the confirmation bias problem [18]). To address this issue, we design a denoising network to learn reliable knowledge from novel classes for reducing bias derived from base classes and use a Gaussian Mixture Model (GMM) to model loss distributions of pseudo labels and penalise noisy pseudo labels for reducing accumulated label noise.

Denoising Network. As shown in Fig. 2 (a), (b) and (c-2), PLAIN++ consists of three steps, in which the first two steps are the same as PLAIN, whilst the third step is improved with a pseudo label denoising process. The pipeline of iterative pseudo label refinement with pseudo label denoising is depicted in Fig. 3. With the pseudo labels L_{pl} assigned by the few-shot classifier $f_{(\theta^*,W_n)}$, we select reliable pseudo-labelled instances D^{pl}_{select} and support data $D_{sup.}$ to train a denoising network. Generally, we use ξ percentage of pseudo-labelled instances with high confidence per-class from $D_{unl.}$ as D^{pl}_{select} and perform two different random data augmentations, *i.e.* weak augmentation (random crop and random flip) and strong augmentation (RandAugment [15] using three different items for augmentation with magnitude 10), on these instances and support data to generate augmented images X_w and X_r. Since pseudo labels with high confidence usually contain less noise and random augmented data contain potential transformations of instances, so they can be used to learn richer data distributions of novel classes. Then, to train a denoising network with these data, we use a cross-entropy loss L_{CE} for classification and employ a distillation loss L_{KD} [19, 20] to learn soft data distributions, which helps to improve the generalisation of the denoising network for remitting pseudo label noise. Here, L_{KD} is formulated as $L_{KD} = L_{KL}(p_w||p_r) + L_{KL}(p_r||p_w)$, where $L_{KL}(x||y)$ is a loss metric with Kullback-Leibler (KL) divergence.

Denoising Pseudo Labels with GMM. During the network training process, noisy labels often take longer to learn than clean labels, so noisy pseudo-labelled examples will produce higher losses at the early stage. This provides us a chance to distinguish clean and noisy samples based on their loss distributions [21]. To this end, we use a two-component GMM (J=2, $l \sim N(\mu_j, \sum_j)$) to model loss distributions. For each pseudo-labelled sample, the mixture model estimates a confidence value for the pseudo label according to the corresponding loss and penalises samples that do not satisfy the clean label distribution, avoiding assigning high confidence to incorrect prediction instances in the next iterations. Specifically, with trained denoising network, we first get denoised pseudo labels L_{dpl} for $D_{unl.} \bigcup D_{sup.}$ and calculate the loss l between predictions of the denoising network and original pseudo labels L_{pl}. Then, we fit GMM with l using the expectation-maximization algorithm [22] and compute a confidence value w_i of each sample based on the posterior probability $p(g|l^i)$, where g is the gaussian component with a smaller loss. With the few-shot classifier $f_{(\theta^*, W_n)}$ and the denoising network, we obtain two types of pseudo labels, *i.e.* L_{pl} and L_{dpl} for a given sample. The confidence values produced by GMM helps to select reliable pseudo-labels from L_{pl} or L_{dpl}. Since L_{dpl} is assigned by the denoising network totally trained on novel classes, we adopt the selected L_{dpl} to refine the few-shot classifier $f_{(\theta^*, W_n)}$, which prevents the label noise of L_{pl} from being amplified during iterative refinement.

Besides, to further improve the quality of selected denoised pseudo-labels, we employ weak and strong (RandAugment [23]) methods to transform instances. Thus, we have two predictions with a confidence value for each sample, *i.e.* p_w with w_{iw} and p_r with w_{ir}. Then, we update the denoised pseudo-label pool by aligning two predictions and select η percentage of reliable instances D^{dpl}_{select} with high confidence values $w_{iw} + w_{ir}$. After that the classification weights W_n of few-shot classifier $f_{(\theta^*, W_n)}$ are updated by averaging feature embeddings of D^{dpl}_{select} and $D_{sup.}$. By iteratively refining $f_{(\theta^*, W_n)}$ and the denoising network with pseudo-labelled instances (D^{dpl}_{select} and D^{pl}_{select}), the label noise propagation problem derived from self-training is gradually reduced. We summarise the training process of PLAIN++ in Algorithm 2 in the supplementary material.

4. EXPERIMENTS

Datasets. (1) *mini*Imagenet [11] is a subset of the ILSVRC-12 dataset [26], containing 100 classes with 600 images per class. Following [13], we used 64, 16 and 20 classes as base, validation and novel set. **(2) *tiered*ImageNet** [5] is a larger subset of ILSVRC-12 with 608 classes, which are semantically grouped into 34 broader categories. Following [5], we used 20, 6, 8 categories as base, validation and novel set. **(3) *CIFAR-FS*** [12] is a subset of CIFAR100 and includes 100 classes with 600 low-resolution images per class. Following [12], we used 64, 16, 20 classes as base, validation and novel set. For each dataset, we resized all images to 84×84. We used the base set to pre-train a feature extractor and selected a feature extractor with the best performance on validation set. We randomly selected 600 tasks from the novel set, where each task has K support labelled data, 15 query data and R unlabelled data per-class from N novel categories.

Implementation Details. Following [25], we used ResNet-12 as the backbone for pre-training a feature extractor. ResNet-12 contains 4 residual blocks, where each block has three 3×3 convolutional layers and every convolutional layer is followed by a BatchNorm layer and a LeakyReLU activation with 0.1. We employed dropout in each block and applied a 2×2 max-pooling layer at the end of each residual block. We used SGD with momentum 0.9 and L2 weight decay 5e-4 as the optimiser. We set the initial learning rate to 0.1 and trained the model with 30 epochs for CIFAR-FS, 60 epochs for other datasets. In each epoch, we randomly selected 8000 batches with size 32. As for the denoising network, we adopt ResNet-10 with 4 blocks as the backbone. Each block of ResNet-10 consists of three 3×3 convolution layers, where each convolutional layer is followed by a BatchNorm layer. We used SGD with momentum 0.9 and weight decay 5e-4 as the optimiser. The batch size and learning rate were set to 64 and 5e-3, respectively. We set the number of iterations M as 15 when $\eta > 50\%$, otherwise set M as 10, and set the epochs T_e for training denoising network to 12 for warming up the network in the first iteration and 6 in the remaining iterations. For each SS-FSL task, we used $R_{unl.} = 100$ unlabelled samples per-class and maximumly selected $\eta * R_{unl.}$ pseudo-labelled instances per-class to update CSRM $f_{(\theta^*, W_n)}$ ($\eta = \{50\%, 100\%\}$). For training the denoising network, we set ξ to 60%/80% for the 1/5 shot setting in all experiments.

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

In Table 1, we compared the proposed methods with 10 state-of-the-art approaches. From Table 1, we see that: (1) Compared with state-of-the-art methods, PLAIN achieves competitive performance though it is simple, which shows effectiveness of this baseline; (2) With pseudo label denoising for resolving label noise propagation, PLAIN++ further improves PLAIN outperforming state-of-the-art methods on *mini*Imagenet and CIFAR-FS and is on par with ICI on *tiered*Imagenet; (3) With more pseudo-labelled instances, the performance of PLAIN and PLAIN++ gradually improve.

4.2. Ablation Study

Components Analysis. To verify the effectiveness of each component in PLAIN and PLAIN++, we conducted experiments with CSRM($f_{(\theta^*, W_n)}$), CSRM with pseudo label (PL), PLAIN (full model), PLAIN with GMM (partial PLAIN++ model), PLAIN with weak and strong augmented (WSA) images and GMM (full PLAIN++ model). As shown in Table 2, PLAIN with pseudo label refinement achieves substantial improvement compared with CSRM w/ or w/o pseudo labels in

Table 1. Mean classification accuracies of the 5-way 1/5-shot tasks on *mini*ImageNet, *tiered*ImageNet and CIFAR-FS with 95% confidence interval.(a/b) represents selecting maximum $a = \eta * R_{unl}$ pseudo labelled instances per-class from $b = R_{unl}$ unlabelled data per-class in 5-way 1/5-shot learning. **Bold** and <u>Underline</u> are the best and second best results, respectively.

	Method	Venue	Backbone	*mini*Imagenet		*tiered*Imagenet		*CIFAR-FS*	
				1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Supervised FSL	ProtoNet [1]	NeurIPS16	Conv4-64	49.42±0.78	68.20±0.66	-	-	72.20	83.50
	Dynamic [16]	CVPR2018	Conv4-64	56.20±0.86	72.81±0.62	-	-	-	-
	Imprinting [17]	CVPR2018	ResNet12	58.68±0.81	76.06±0.59	-	-	-	-
	DSN-MR [24]	CVPR2020	ResNet12	64.60±0.72	79.51±0.50	67.39±0.82	82.85±0.56	75.6±0.9	86.2±0.6
Meta-learning based SS-FSL	MS k-means [5]	ICLR2018	Conv4-64	50.41±0.31	64.39±0.24	52.4	69.9	-	-
	TPN-semi [6]	ICLR2019	Conv4-64	52.78±0.27	66.42±0.21	55.7	71.00	-	-
	semi-DSN [24]	CVPR2020	Conv4-64	53.01±0.82	69.12±0.62	54.06±0.96	72.07±0.69	-	-
	LST [7]	NeurIPS19	ResNet12	70.1±1.9	78.7±0.8	77.7	85.2	-	-
Transfer-learning based SS-FSL	TransMatch [8]	CVPR2020	WRN/28/10	63.02±1.07	81.19±0.59	-	-	-	-
	ICI [25]	CVPR2020	ResNet12	71.41	81.12	**85.44**	**89.12**	78.07	84.79
	PLAIN(80/80)	Ours	ResNet12	72.42±2.11	80.88±1.17	82.69±1.84	88.20±1.02	84.93±1.77	87.98±1.15
	PLAIN (50/100)	Ours	ResNet12	72.06±1.94	79.75±1.49	82.40±1.85	87.29±1.13	83.47±1.61	87.42±0.97
	PLAIN (100/100)	Ours	ResNet12	72.84±2.20	81.01±1.10	82.32±2.19	88.17±1.34	84.32±1.63	88.35±1.06
	PLAIN++ (80 /80)	Ours	ResNet12	73.18±2.19	<u>81.77±1.11</u>	82.80±1.86	88.26±1.01	**85.64±1.72**	88.18±1.15
	PLAIN++ (50/100)	Ours	ResNet12	<u>73.88±1.98</u>	81.73±1.13	82.62±1.93	87.99±1.20	84.50±1.67	<u>88.37±1.04</u>
	PLAIN++ (100/100)	Ours	ResNet12	**74.38±2.06**	**82.02±1.08**	<u>82.91±2.09</u>	<u>88.29±1.25</u>	<u>85.21±1.62</u>	**88.78±1.01**

Table 2. Component effectiveness analysis on *mini*ImageNet with ResNet12 (mean accuracies (%) with 95% confidence interval, 5-way 1/5-shot). We set $R_{unl.} = 100, \eta = 50\%$.

Method	1-shot	5-shot
CSRM	60.06	75.88
CSRM + PL	68.66±1.55	80.58±1.56
PLAIN	72.05±1.94	79.75±1.49
PLAIN+GMM	73.65±1.91	81.58±1.09
PLAIN+GMM+WSA	**73.88±1.98**	**81.73±1.13**

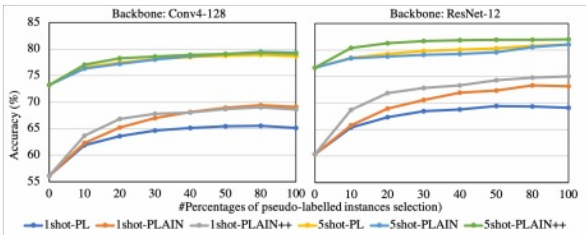

Fig. 4. Comparisons of CSRM+pseudo label (PL), PLAIN, PLAIN++ on *mini*Imagenet using different instance percentages η and different backbones (Conv4-128 and ResNet-12).

the 1-shot setting. Although the improvement on 5-shot learning is not obvious for iterative pseudo label refinement, this can be attributed to the label noise propagation problem. This problem can be solved by the proposed GMM and WSA. As shown in Table 2, PLAIN+GMM performs significantly better than CSRM+PL and PLAIN in both 1/5-shot setting, while the PLAIN+GMM+WSA further improves the performance.

Effect of Different Backbones and Percentage of Selected Pseudo Labelled Instances. In Fig. 4, we reported results of CSRM+PL, PLAIN, PLAIN++ with different η on *mini*Imagenet using ResNet12 and Conv4-128 [16]. We set η=$\{0, 10, 20, 30, 40, 50, 80, 100\}$% and $R_{unl.}$=100. We can see that with a deeper network as the backbone, all compared

Table 3. Accuracies of various methods with $R_{unl.} = \{0, 15, 50, 100, 150, 200\}$ on *mini*Imagenet with ResNet12. We both set $\eta = 50\%$ in all settings.

$R_{unl.}$	Trans.		PLAIN		PLAIN++	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
0	58.68	76.06	60.06	75.88	60.06	75.88
15	-	-	64.67	78.30	64.63	78.58
50	61.21	79.30	70.00	79.94	70.19	81.19
80	-	-	71.76	79.78	73.16	81.41
100	**63.02**	81.19	72.05	79.82	73.88	82.19
150	-	-	**73.06**	**80.09**	74.71	**82.39**
200	62.93	**82.24**	72.50	79.14	**74.91**	81.77

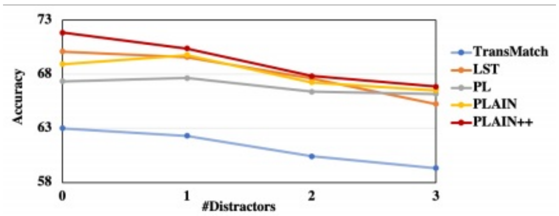

Fig. 5. Accuracies of 5-way 1-shot tasks on *mini*Imagenet under various distractors with ResNet12. We set η=20% and $R_{unl.}$=100 in CSRM+PL, PLAIN and PLAIN++.

methods improve their performance, where PLAIN++ performs the best and PLAIN performs the second-best. Besides, with different η, PLAIN++ still performs the best overall.

Effect of Number of Unlabelled Examples Per Class As shown in Table 3, when more unlabelled data per class are available, the performance of all compared methods improves, among which PLAIN++ achieves the best performance, which shows the scalability of our methods.

Robustness against Distractor Classes. Following [7, 8], we mixed the original unlabelled data with the same number of samples per-class randomly selected from other categories in the test set as the distractors and further evaluated our methods in SS-FSL with 1/2/3 distractor classes. As shown in Fig.5, when distractors are included, accuracies of all compared methods decrease, but PLAIN++ and PLAIN still perform competitive against the other methods.

5. CONCLUSIONS

In this work, we introduced a simple yet effective baseline method (**P**seudo **LA**bel ref**IN**ement, PLAIN) for SS-FSL to iteratively refine pseudo labels for learning a new classifier for novel categories. Then, we discussed the label noise propagation problem and proposed PLAIN++ by improving PLAIN with a denoising network for generating deniosed pseudo-labels and a mixture model for learning distributions of clean and noisy pseudo-labelled examples to select reliable pseudo-labelled instances with less noise. We conducted extensive experiments on *mini*Imagenet, *tiered*Imagenet and CIFAR-FS. Experimental results show the effectiveness of PLAIN and PLAIN++ over the state-of-the-art SS-FSL methods.

6. REFERENCES

[1] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017.

[2] Chenrui Zhang and Yuxin Peng, "Visual data synthesis via gan for zero-shot video classification," in *IJCAI*, 2018.

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.

[4] 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.

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

[6] 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.

[7] 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.

[8] Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo, "Transmatch: A transfer-learning scheme for semi-supervised few-shot learning," in *CVPR*, 2020.

[9] Dong-Hyun Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshops*, 2013, vol. 3.

[10] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, 2017.

[11] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., "Matching networks for one shot learning," in *NeurIPS*, 2016.

[12] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi, "Meta-learning with differentiable closed-form solvers," in *ICLR*, 2019.

[13] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2016.

[14] Samuli Laine and Timo Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, 2017.

[15] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *NeurIPS*. 2019.

[16] Spyros Gidaris and Nikos Komodakis, "Dynamic few-shot visual learning without forgetting," in *CVPR*, 2018.

[17] Hang Qi, Matthew Brown, and David G Lowe, "Low-shot learning with imprinted weights," in *CVPR*, 2018.

[18] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *IJCNN*, 2020.

[19] Xu Lan, Xiatian Zhu, and Shaogang Gong, "Knowledge distillation by on-the-fly native ensemble," in *NeurIPS*. 2018.

[20] Guile Wu and Shaogang Gong, "Peer collaborative learning for online knowledge distillation," in *AAAI*, 2021.

[21] Junnan Li, Richard Socher, and Steven CH Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *ICLR*, 2020.

[22] Haim Permuter, Joseph Francos, and Ian Jermyn, "A study of gaussian mixture models of color and texture features for image classification and segmentation," *PR*, vol. 39, no. 4, 2006.

[23] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *CVPR Workshops*, 2020.

[24] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi, "Adaptive subspaces for few-shot learning," in *CVPR*, 2020.

[25] Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu, "Instance credibility inference for few-shot learning," in *CVPR*, 2020.

[26] 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," *IJCV*, vol. 115, no. 3, 2015.