A Strong Baseline for Semi-Supervised Incremental Few-Shot Learning

Linglan Zhao\textsuperscript{*1}  
llzhao@sjtu.edu.cn  
\textsuperscript{1} Dept. of Electronic Engineering  
Shanghai Jiao Tong University  
Shanghai, China  

Dashan Guo\textsuperscript{2}, Yunlu Xu\textsuperscript{†2}, Liang Qiao\textsuperscript{2}  
\{guodashan,xuyunlu,qiaoliang6\}@hikvision.com\textsuperscript{2}  
Hikvision Research Institute  
Hangzhou, China  

Zhanzhan Cheng\textsuperscript{2}, Shiliang Pu\textsuperscript{2}, Yi Niu\textsuperscript{2}  
\{chengzhanzhan,pushiliang,niuyi\}@hikvision.com  

Xiangzhong Fang\textsuperscript{1}  
xzfang@sjtu.edu.cn

Abstract

Few-shot learning (FSL) aims to learn models that generalize to novel classes with limited training samples. Recent works advance FSL towards a scenario where unlabeled examples are also available and propose semi-supervised FSL methods. Another line of methods also cares about the performance of base classes in addition to the novel ones and thus establishes the incremental FSL scenario. In this paper, we generalize the above two under a more realistic yet complex setting, named by Semi-Supervised Incremental Few-Shot Learning (S\textsuperscript{2}I-FSL). To tackle the task, we propose a novel paradigm containing two parts: (1) a well-designed meta-training algorithm for mitigating ambiguity between base and novel classes caused by unreliable pseudo labels and (2) a model adaptation mechanism to learn discriminative features for novel classes while preserving base knowledge using few labeled and all the unlabeled data. Extensive experiments on standard FSL, semi-supervised FSL, incremental FSL, and the firstly built S\textsuperscript{2}I-FSL benchmarks demonstrate the effectiveness of our proposed method.

1 Introduction

Despite the great success of deep learning models in recent years, modern deep learning approaches rely heavily on abundant labeled data which can be prohibitively expensive to acquire. In contrast, humans can quickly learn novel concepts with only a few samples. To tackle this problem, Few-Shot Learning (FSL) \cite{6, 31, 33, 38} methods are proposed to bridge the gap between supervised learning and the intriguing property of human cognition.

The standard FSL setting is shown in Fig. 1 (a) that given new tasks of limited supervision, the system turns the prior knowledge from base tasks. Recent works \cite{14, 26, 44} advance the few-shot paradigm towards a scenario where unlabeled examples are available, namely the semi-supervised FSL shown in Fig. 1 (b). Exploring extra unlabeled data along with the limited supervised ones makes sense since unlabeled data is much easier to obtain than the labeled. However, the standard or semi-supervised FSL focuses solely on recognizing few novel classes, while ignoring the ability to handle the previously learned base classes.

\textsuperscript{*} Linglan Zhao did this work when he was an intern at Hikvision Research Institute. \textsuperscript{†}: corresponding author.  
© 2021. The copyright of this document resides with its authors.  
It may be distributed unchanged freely in print or electronic forms.
To prevent forgetting prior knowledge for base tasks, **incremental FSL** [9, 23, 27, 43] is built for simultaneously regarding performance on base and novel classes, which shows some realistic significance, as shown in Fig. 1 (c). Nevertheless, there is still room for improvement as only supervised data is used which is hard to adapt to new classes.

From a more realistic perspective, besides few annotated samples, we expect to additionally utilize more easily available unlabeled data containing both base and novel class samples to overcome the performance limitation, especially the performance of novel classes. Therefore, we propose a new benchmark incorporating the above semi-supervised and incremental settings together called **Semi-Supervised Incremental Few-Shot Learning (S²I-FSL)**, which generalizes previous assumptions under a more realistic yet complex setting (Fig. 1 (d)).

Intuitively, a straightforward solution for the proposed S²I-FSL is to simply combine the existing semi-supervised and incremental FSL approaches, but it does not work well. For detail, as illustrated in Section 4.1, we utilize a simple yet efficient meta-learning-based method as our incremental FSL baseline (denoted as Meta-Inc-Baseline). When extended with unlabeled data in a semi-supervised manner by the mainstreaming methods including weights refinement [18, 19, 26], label propagation [20, 28] and FixMatch [32], unfortunately, it can not perform well. We find that novel class weights refined with incorrectly predicted pseudo-labels will overlap with base class weights and lead to ambiguity among classes in the embedding space. For example, the weight of novel class “school bus” in Fig. 1 (d) is likely to be incorrectly refined with unlabeled samples from base class “unicycle” (both have the wheel pattern), “ashcan” (similar color and texture), and “street sign” (two objects often appear in the same image), especially when only a few labeled “school bus” samples are given as in few-shot scenarios. Overall, S²I-FSL brings additional challenges in the FSL field: (1) using unlabeled data from both base and novel classes leads easily to ambiguity among classes in the embedding space, showing performance degradation; (2) unbalanced sample number of different classes in unlabeled set instead of a balanced number in each novel class makes it harder to learn the classifier. In this paper, we mainly handle the former challenge and show the comparison effect on the latter one.

Upon the above, we propose a novel S²I-FSL paradigm of two phases including two key components, a novel meta-training scheme and a well-designed model adaptation mechanism. In the meta-training phase, the training set is sampled using a mechanism to mimic the semi-supervised scenario during testing. As a result, the model learns to make use of unlabeled data for incremental learning and mitigates confusion between base and novel classes.

Figure 1: Comparisons between our proposed S²I-FSL and other FSL benchmarks. \( N_b \) and \( N \) are the numbers of base and novel classes, respectively. Best viewed in color.
classes. Considering that the feature extractor is trained on the training set, it may not contain discriminative information for recognizing novel classes unseen during training, a model adaptation scheme is proposed. Concretely, besides standard classification loss on few labeled data, a contrastive learning loss function is applied to unlabeled data to explore the novel class distribution implicitly. In addition, we also use distillation technique to preserve the classification ability on base classes.

To summarize, the contributions of this work are three-fold: (1) A Semi-Supervised Incremental Few-Shot Learning ($S^2$I-FSL) benchmark is proposed to generalize previous semi-supervised FSL and incremental FSL under a more realistic and challenging setting; (2) Technically, we propose an efficient meta-training paradigm and a model adaptation mechanism for the newly built task, which can be regarded as a baseline for future researches; (3) Extensive experiments on standard FSL, incremental FSL, semi-supervised FSL, and the $S^2$I-FSL benchmark demonstrate the effectiveness of our proposed method.

2 Related works

2.1 Few-Shot Learning

Few-shot learning (FSL) aims to learn novel categories with limited labeled examples. Previous episodic-training \cite{Vinyals2016} based FSL works can be roughly divided into 4 categories. Metric learning based approaches \cite{Sung2018, Oreshkin2018, Vinyals2016} aim at learning the similarity metric between query (i.e., test) samples and support (i.e., training) samples. The goal of initialization based FSL methods \cite{Bello2017, Snell2017, Wei2017} is to learn good model initialization for fast adaptation to current tasks. Weight generation based methods \cite{B+F18, B+F19, Snell2017} directly predict classification weights for novel classes. Moreover, hallucination based approaches \cite{Yao2018, Liu2019, Rothe2019} learn a generator for data augmentation. However, recent works \cite{B+F18, B+F19, Yin2019, Yao2018} reveal that a well-trained feature extractor, rather than other complex meta-trained modules, is the key component to achieve state-of-the-art performance. This motivates us to adopt a modified version of \cite{B+F18} as a baseline method (Meta-Inc-Baseline) for incremental FSL and $S^2$I-FSL benchmark.

2.2 Few-Shot Incremental Learning

Incremental FSL involves training a model to classify novel classes while preserving the capability of handling previously learned classes. There are two main different settings in the literature. The first setting is also called generalized (or dynamic) FSL, which contains two sequences of tasks. The first task contains abundant training samples from base classes, while the second task contains only a few novel class training samples. Methods proposed in this setting include novel class weight generation \cite{B+F19, Yin2019, B+F18, Wei2017}, attention-based regularization \cite{Han2019} and task adaptive features extraction \cite{Yao2018, Han2019}. The second setting \cite{B+F18, Yao2018, Yin2019, Li2019} shares the same core idea with the first one but considers several sequences of tasks instead
of two. We build S²I-FSL mainly based on the first setting since extra unlabeled data are not available in the second setting. In addition, the second setting uses unique data splits which makes it harder to compare to other standard/transductive/semi-supervised FSL works.

2.3 Semi-Supervised Few-Shot Learning

Semi-supervised/transductive FSL works take additional unlabeled data into account to boost the performance of FSL. Pseudo-labeling based methods and graph-based methods are two main lines of efforts. Pseudo-labeling based methods consist of prototype refinement [13, 18, 19, 26], self-training [17, 39] and entropy minimization [2, 7]. As for graph-based approaches [14, 20, 28, 41], graph models are constructed to propagate information from labeled data to unlabeled data. Among these approaches, we mainly focus on the utilization of prototype refinement for its effectiveness and extensiveness.

3 Task Formulation

In this section, we introduce the detailed settings for the proposed S²I-FSL benchmark.

Dataset Splits. For evaluating classification performance on both base and novel classes, we follow the dataset splits in incremental FSL [9, 27, 43]. Concretely, \( \mathcal{D}_{\text{base}} \) and \( \mathcal{D}_{\text{novel}} \) are two datasets for base classes \( \mathcal{C}_{\text{base}} \) and novel classes \( \mathcal{C}_{\text{novel}} \) respectively, and \( \mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \emptyset \). \( \mathcal{D}_{\text{base}} \) is comprised of \( \mathcal{D}_{\text{base/train}} \), \( \mathcal{D}_{\text{base/val}} \) and \( \mathcal{D}_{\text{base/test}} \) as base class training, validation and testing sets (\( |\mathcal{C}_{\text{base}}| = N_b \)). \( \mathcal{D}_{\text{novel}} \) consists of three splits of disjoint novel classes: \( \mathcal{D}_{\text{novel/train}} \), \( \mathcal{D}_{\text{novel/val}} \) and \( \mathcal{D}_{\text{novel/test}} \) for meta-training, validation and testing, respectively.

Training Stage. As shown in the left part of Fig. 2, the training stage is split into pre-training phase and meta-training phase. In the pre-training phase, feature extractor \( f_\theta \) and the base class weights \( W_b \) are trained on \( \mathcal{D}_{\text{base/train}} \) as in standard supervised learning. After that, in the meta-training phase, the model is further trained using both base class training set \( \mathcal{D}_{\text{base/train}} \) and novel class training set \( \mathcal{D}_{\text{novel/train}} \) in an episodic training manner [38].

Testing Stage. Following the evaluation protocol in [9, 27, 43], average accuracy on sampled incremental FSL episodes (i.e., tasks) is reported. For each test episode, \( N \) novel categories are randomly chosen from \( \mathcal{D}_{\text{novel/test}} \) (\( \mathcal{D}_{\text{novel/val}} \) for validation), which are relabeled with new labels from \( N_b + 1 \) to \( N_b + N \). After that, \( K \) per-class training samples are sampled from these \( N \) classes (\( N \)-way \( K \)-shot) to construct a support set \( \mathcal{S} \). The query set \( \mathcal{Q} \) consists of samples from both novel and base classes, i.e., additional examples from the selected \( N \) novel classes and base class test examples from \( \mathcal{D}_{\text{base/test}} \) (\( \mathcal{D}_{\text{base/val}} \) for validation). The main difference between S²I-FSL and incremental FSL is that unlabeled data is available in S²I-FSL. Two different settings are considered: a transductive setting where the unlabeled set \( \mathcal{U} \) is exactly the query set \( \mathcal{Q} \); a semi-supervised setting where \( \mathcal{U} \) contains extra unlabeled samples from both base classes and novel classes.

4 Methodology

4.1 Preliminary

Since Meta-Baseline [13] provides an effective yet strong baseline for standard FSL, we extend it for incremental setting, noted as Meta-Inc-Baseline. During the pre-training phase, the feature extractor \( f_\theta (x) \in \mathbb{R}^d \) is trained using a cosine classifier on \( \mathcal{D}_{\text{base/train}} \) as in standard supervised training. The conditional probability of sample \( x \) belonging to class \( k \) is:

\[
P(y = k \mid x; W_b) = \frac{\exp (\gamma \cos (f_\theta (x), w_k))}{\sum_j \exp (\gamma \cos (f_\theta (x), w_j))},
\]

(1)
where $W_b = [w_1, \ldots, w_{N_b}] \in \mathbb{R}^{d \times N_b}$ is the base class weights and $\gamma$ is a learnable scalar.

In the testing stage, novel class weights can be constructed as class means of the support set $S$, i.e., the novel class weights $W_n = [p_1, \ldots, p_N] \in \mathbb{R}^{d \times N}$ are comprised of prototypes of each category where $S_j = \{(x, y) \mid (x, y) \in S, y = N_b + j\}$:

$$p_j = \frac{1}{|S_j|} \sum_{(x, y) \in S_j} f_\theta(x).$$

For extending Meta-Baseline to handle both base and novel classes, the overall classification weights $W$ is constructed by concatenating the base class weights $W_b$ and novel class weights $W_n$, i.e. $W = [W_b, W_n] \in \mathbb{R}^{d \times (N_b + N)}$. Finally, the model makes predictions on query set $Q$ containing both base and novel class samples based on Eq. 1 with $W_b$ replaced with $W$.

After pre-training is a meta-training phase. Since the model is evaluated on incremental FSL tasks, instead of only optimizing $f_\theta$ and discarding $W_b$ [5], we meta-train both $f_\theta$ and $W_b$ by sampling incremental FSL tasks from $D_{\text{base/train}}$ and $D_{\text{novel/train}}$ as in Algorithm 1.

Algorithm 1 Meta-training algorithm for Meta-Inc-Baseline.

Input: pre-trained $f_\theta$ and $W_b$, learning rate for $\theta$ and $W_b$ respectively: $\eta_1$ and $\eta_2$.

Output: Meta-trained feature extractor $f_\theta$ and base class weights $W_b$.

1: \textbf{while} not done \textbf{do}
2: \{S, Q_n\} \leftarrow \text{Sample support and query set from } D_{\text{novel/train}}
3: \{Q_b\} \leftarrow \text{Sample query set from } D_{\text{base/train}}
4: Q = Q_b \cup Q_n // \text{Construct an incremental FSL episode}
5: Construct overall classification weight $W = [W_b, W_n]$ using Eq. 2
6: Get predictions $\hat{Y}_q \in \mathbb{R}_+^{n_q \times (N_b + N)}$ on $Q$ ($n_q = |Q|$) using Eq. 1 with $W_b \leftarrow W$
7: Calculate cross-entropy loss $L_{\text{cls}}(\hat{Y}_q, Y_q)$ based on $\hat{Y}_q$ and ground truth labels $Y_q$
8: Update: $\theta \leftarrow \theta - \eta_1 \nabla_\theta L_{\text{cls}}, W_b \leftarrow W_b - \eta_2 \nabla_{W_b} L_{\text{cls}}$
9: \textbf{end while}

4.2 Prototype Refinement with Fake Unlabeled Data

Upon Meta-Inc-Baseline, we further generalize to S²I-FSL in Section 4.2 and Section 4.3. Since prototype refinement using pseudo labels has been proved effective in semi-supervised methods [13, 18, 19, 26], an intuitive solution to extend Meta-Inc-Baseline to S²I-FSL is to refine the novel class weights $W_n$ with unlabeled data. Meanwhile, base class weights $W_b$ trained on abundant base class samples remain unchanged.

Concretely, predictions $\hat{Y}_u \in \mathbb{R}_+^{n_u \times (N_b + N)}$ on $U$ ($n_u = |U|$) can be obtained as in line 5-6 of Algorithm 1 similar to $\hat{Y}_q$. Since $U$ contains unlabeled samples from both base and novel classes, samples predicted as base classes should be filtered out as they are viewed as distractors in the refinement process: $\hat{Y}' = \hat{Y}_u[;, N_b : N_b + N] \in \mathbb{R}_+^{n_u \times N}$ where $[,]$ denotes slicing operation. Based on $\hat{Y}'$, novel class prototypes $\{p'_j\}_{j=1}^N$ can be re-estimated:

$$p'_j = \frac{\sum_{i=1}^{n_u} \hat{Y}'_{ij} f_\theta(x_i) + \sum_{(x_k, y_k) \in S_j} f_\theta(x_k)}{\sum_{i=1}^{n_u} \hat{Y}'_{ij} + |S_j|}.$$  

Then we progressively refine $W_n = [p_1, \ldots, p_N] \in \mathbb{R}^{d \times N}$:

$$p_j \leftarrow \alpha \cdot p'_j + (1 - \alpha) \cdot p_j,$$

where $\alpha \in (0, 1)$. After iterating the above process of Eq. 3-4 for $n_{\text{steps}}$ loops, query samples are evaluated with $W = [W_b, W_n]$ where $W_n$ is the refined novel class weights.

However, as discussed in Section 1 and verified in Section 5, this straightforward modification leads to little improvement or even obvious degradation. It is because that novel class weights refined with incorrectly predicted pseudo-labels will be confused with similar base
classes. For utilization of unlabeled samples and mitigating ambiguity, a novel meta-training algorithm with fake unlabeled examples is proposed in Algorithm 2 as a substitution for Algorithm 1. Concretely, compared to labeled support set \( S \) or query set \( Q \) whose labels are used in the loss function, the label information of sampled unlabeled set \( U \) in each training episode is not used to mimic the testing scenario. By default, \( n_{\text{steps}} = 1, \alpha = 1.0 \) in training.

Algorithm 2 Meta-training with fake unlabeled data.

\begin{algorithm}
\textbf{Input:} pre-trained \( f_\theta \) and \( W_b \), learning rate for \( \theta \) and \( W_b \): \( \eta_1 \) and \( \eta_2 \). refinement parameters: \( n_{\text{steps}} = 1, \alpha = 1.0 \).
\textbf{Output:} Meta-trained feature extractor \( f_\theta \) and base class weights \( W_b \).

1: while not done do
2: \{\( S, Q_n, U_b \}\} ← Sample support, query and unlabeled set from \( D_{\text{novel/train}} \)
3: \{\( Q_b, U_b \}\} ← Sample query and unlabeled set from \( D_{\text{base/train}} \)
4: \( Q = Q_b \cup Q_n, U = U_b \cup U_n \) // Construct an incremental FSL episode with fake unlabeled set
5: Get refined \( W_b \) using Eq. 3-4 based on \( U \), and construct classification weight \( W = [W_b, W_n] \)
6: Get predictions \( \hat{Y}_q \in \mathbb{R}_+^{n_q \times (N_b + N)} \) on \( Q (n_q = |Q|) \) using Eq. 1 with \( W_b ← W \)
7: Calculate cross-entropy loss \( L_{\text{cls}}(\hat{Y}_q, Y_q) \) based on \( \hat{Y}_q \) and ground truth labels \( Y_q \)
8: Update: \( \theta ← \theta - \eta_1 \nabla_\theta L_{\text{cls}}, W_b ← W_b - \eta_2 \nabla_{W_b} L_{\text{cls}} \)
9: end while
\end{algorithm}

### 4.3 Model Adaptation without Forgetting

As described in Section 1, feature extractor \( f_\theta \) trained on the training set may not contain discriminative information for novel classes during testing [3]. However, as observed in few-shot incremental literature [3, 43, 4], directly finetuning models on few labeled novel class samples will cause not only overfitting to the novel class but also catastrophic forgetting of the old classes. Fortunately, since additional unlabeled data is available, \{\( f_\theta, W_b \}\} can be further adapted to the current test episode \( T = \{S, Q, U\} \) in the testing stage.

As shown in the right part of Fig. 2, for learning better representations for novel classes, we explore both labeled support set \( S \) using cross-entropy loss \( L_{\text{cls}} \) for classification, and the unlabeled set \( U \) using a contrastive loss [3] \( L_{\text{ctr}} \):

\[
L_{\text{ctr}} = -\frac{1}{2B} \sum_{i,j=1}^{B} \log \frac{\exp \left( \cos \left( f_i, f_j \right) / \tau_1 \right)}{\sum_{k \neq i} \exp \left( \cos \left( f_i, f_k \right) / \tau_1 \right)},
\]

where \( B \) is the batch size to sample batches from \( U \), \( \tau_1 \) is a temperature parameter and \((f_i, f_j)\) is a positive feature pair (different augmentations from the same image) based on \( f_\theta \).

For preserving the knowledge learned from base classes, we adopt a distillation [4] loss \( L_{\text{dst}} \) on base class predictions using unlabeled data \( U \):

\[
L_{\text{dst}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{N_b} \bar{z}_{ik} \log z_{ik},
\]

where \( z_{ik} \) and \( \bar{z}_{ik} \) are \( \tau_2 \) softened predictions of the \( i \)-th sample belonging to class \( k \) from previous meta-trained parameters \{\( \hat{\theta}, W_{b} \}\} and current adapted ones \{\( \theta, W_b \)\}, respectively:

\[
\bar{z}_{ik} = \frac{\exp \left( \gamma \cos \left( f_{\hat{\theta}}(\mathbf{x}_i), \mathbf{w}_k / \tau_2 \right) \right)}{\sum_{j=1}^{N_b} \exp \left( \gamma \cos \left( f_{\hat{\theta}}(\mathbf{x}_i), \mathbf{w}_j / \tau_2 \right) \right)},
\]

\[
z_{ik} = \frac{\exp \left( \gamma \cos \left( f_{\theta}(\mathbf{x}_i), w_k / \tau_2 \right) \right)}{\sum_{j=1}^{N_b} \exp \left( \gamma \cos \left( f_{\theta}(\mathbf{x}_i), w_j / \tau_2 \right) \right)}. \tag{7}
\]

The overall loss function is formulated as:

\[
L = w_{\text{cls}} \cdot L_{\text{cls}} + w_{\text{ctr}} \cdot L_{\text{ctr}} + w_{\text{dst}} \cdot L_{\text{dst}}\]

where \( w_{\text{cls}}, w_{\text{ctr}} \) and \( w_{\text{dst}} \) (hyperparameters in Appendix C) control the tradeoff between each term.

---

1 We use the terminology fake unlabeled since a subset is randomly sampled from the supervised training set in each meta-training episode, but its labels are not used to simulate the testing scenario.
5 Experiments

5.1 Experiment Setup

Datasets. We follow the common incremental FSL setting [9, 27, 43], and conduct experiments on mini-ImageNet [38] and tiered-ImageNet [26] with 64 and 200\(^2\) base classes, respectively. More details of dataset statistics can be found in Appendix A.

Evaluation Protocol. The performance is evaluated using the average accuracy and corresponding 95% confidence interval over 600 randomly sampled episodes with a fixed \(\text{seed}\), which makes results more stable and reliable. In incremental FSL and S\(^2\)I-FSL, we report joint accuracy \(\text{Acc}_{\text{all/all}}\) (Acc. for short) which is the accuracy of classifying all the test samples to all (both base and novel) classes. To measure the performance gap between joint accuracy and individual accuracy, following [27, 43], metric \(\Delta\) is also reported. Concretely, \(\Delta = (\Delta_b + \Delta_n)/2\) is computed as the mean of performance degradation on base and novel classes. \(\Delta_b\) (\(\Delta_n\)) is the performance degradation between the accuracy of classifying base (novel) class test samples to all classes \(\text{Acc}_{b/all}\) (\(\text{Acc}_{n/all}\)) and the accuracy of classifying base (novel) class test samples to only base (novel) classes \(\text{Acc}_{b/b}\) (\(\text{Acc}_{n/n}\)):

\[
\Delta_b = \text{Acc}_{b/all} - \text{Acc}_{b/b}, \quad \Delta_n = \text{Acc}_{n/all} - \text{Acc}_{n/n}.
\] (8)

Implementation Details. In most of the experiments, ResNet12 [11, 22] (detailed in Appendix B) is adopted as the backbone \(f_B\) to have fair comparisons between other methods. Unless specified in S\(^2\)I-FSL, unlabeled base and novel class samples are maintained in equal proportion. Concretely, in each test episode, we sample 5 novel classes with 1/5 per-class training samples (5-way 1/5-shot) for support set \(S\). Query set \(Q\) contains 5 \(\times\) 15 novel class test samples and 75 examples sampled uniformly from the base class test set [27, 43]. After adopting the semi-supervised FSL [17, 18, 39] setting, unlabeled set \(U\) in S\(^2\)I-FSL contains 5 \(\times\) 30 / 5 \(\times\) 50 novel class samples and 150 / 250 base class samples for 1/5-shot setting. We leave more details of hyper-parameter settings in Appendix C.

5.2 Experimental Results

Incremental FSL. We first conduct experiments to evaluate the robustness of the proposed Meta-Inc-Baseline on the incremental FSL classification task. The upper parts of Table 1 and Table 2 show the results (denoted as Inductive) on the mini-ImageNet and tiered-ImageNet, respectively. Compared to other approaches, our Meta-Inc-Baseline is able to achieve much higher joint accuracy and diminish the gap between joint accuracy and individual accuracy. The direct performance enhancement is mainly attributed to the extension of the simple but strong baseline of Meta-baseline [6]. The detailed verification is described in Appendix D. Moreover, the effectiveness of Meta-Inc-Baseline on incremental FSL benchmark with multi sessions [26] is also verified in Appendix E.

S\(^2\)I-FSL. The proposed method is then generalized into two settings where unlabeled data is available: the Transductive setting where unlabeled set \(U\) is exactly the query set \(Q\), and the Semi-supervised setting where extra unlabeled set \(U\) is available. Besides our proposed method, extensions with prevailing semi-supervised FSL methods such as label propagation [20] (a graph model is constructed with base and novel class weights as labeled vertices and unlabeled samples as unlabeled vertices), prototype refinement [18] and semi-supervised learning approach FixMatch [32] are also evaluated for comparison.

As shown in the bottom parts of Table 1 and Table 2, label propagation [20] is not suitable for the more challenging task since it always results in lower accuracy and degrades \(\Delta\) metric.

\(^2\)Due to the dataset splits in incremental FSL [4, 34], the number of base classes is 200 instead of 351.
While FixMatch [32] can get about 0.5% performance gain in 5-shot case, the accuracy decreases significantly in 1-shot case. It is because that FixMatch relies heavily on the quality of pseudo-labels that cannot be obtained in few-shot scenarios especially the 1-shot case. Moreover, the direct combination of strong baseline and prototype refinement [18] can get a limited improvement on \( S^2 \)-ImageNet but leads to obvious performance degradation on tiered-ImageNet. Since tiered-ImageNet contains more base classes, confusion between base and novel classes caused by unreliable pseudo labels is prone to happen.

In contrast, our proposed method can obtain significant performance gains on both joint accuracy and resistance to forgetting consistently. In Transductive setting, compared to Meta-Inc-Baseline + PR, we get 5.7%/12.0% 1-shot classification performance gains and 5.3%/10.1% \( \Delta \) metric improvements on the two datasets, respectively. It is because that our proposed method can alleviate confusion between base and novel classes, and is well adapted to the current test episode to learn discriminative features for novel classes. In Semi-Supervised setting, our method outperforms Meta-Inc-Baseline by about 10% in 1-shot classification on both benchmarks, which shows the superiority of utilizing the unlabeled set.

For better understanding the challenges of \( S^2 \)-FSL, we report detailed results in Table 3. Although all the methods can improve \( \text{Acc}_{b/a} \) metric which is the only focus of standard/transductive/semi-supervised FSL methods, other competitors obtain decreased joint accuracy or only marginal overall improvement. This indicates that \( S^2 \)-FSL is a more difficult task and poses challenges that have not been solved by existing works. We emphasize that there is a trade-off between \( \text{Acc}_{b/a} \) and \( \text{Acc}_{n/a} \), and balancing between them is significant for the final result. For example, although \( \text{Acc}_{b/all} \) increases using graph models [\text{others}], \( \text{Acc}_{n/all} \) decreases dramatically which results in degraded joint accuracy. As mentioned in Section 1, there is an imbalance between base and novel classes. Since the number of base classes is larger than that of novel classes, label information from base vertices (\( N_b = 200 \)) is more likely propagated to unlabeled vertices which decreases \( \text{Acc}_{n/all} \). The marginal improvement of FixMatch [32] in 5-shot case is attributed to that the increase in \( \text{Acc}_{n/all} \) slightly outweighs the decrease in \( \text{Acc}_{b/all} \). However, this limited improvement is unstable, since it gets degraded performance in more challenging 1-shot scenarios on both datasets. As for prototype refinement [\text{others}], \( \text{Acc}_{b/all} \) degrades significantly due to the confusion between base and novel classes. Concretely, some base class unlabeled samples are incorrectly predicted.

### Table 1: Incremental FSL and \( S^2 \)-FSL results on mini-ImageNet.

<table>
<thead>
<tr>
<th>Method</th>
<th>( f_\theta )</th>
<th>Type</th>
<th>( 64+5 )-way 1-shot</th>
<th>( 64+5 )-way 5-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \text{Acc.} ) ( \Delta )</td>
<td>( \text{Acc.} ) ( \Delta )</td>
</tr>
<tr>
<td>Imprint [\text{ours}]</td>
<td>Res10</td>
<td>Incremental FSL (Inductive)</td>
<td>65.12±0.48</td>
<td>-18.32</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + Graph [\text{ours}]</td>
<td>Res12</td>
<td>( S^2 )-FSL (Transductive)</td>
<td>65.12±0.48</td>
<td>-18.32</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + FixMatch [\text{ours}]</td>
<td>Res12</td>
<td>( S^2 )-FSL (Transductive)</td>
<td>66.43±0.54</td>
<td>-8.31</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + PR [\text{ours}]</td>
<td>Res12</td>
<td>( S^2 )-FSL (Transductive)</td>
<td>65.12±0.48</td>
<td>-18.32</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + Graph [\text{ours}]</td>
<td>Res12</td>
<td>( S^2 )-FSL (Transductive)</td>
<td>66.43±0.54</td>
<td>-8.31</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + FixMatch [\text{ours}]</td>
<td>Res12</td>
<td>(Inductive)</td>
<td>65.12±0.48</td>
<td>-18.32</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + PR [\text{ours}]</td>
<td>Res12</td>
<td>(Inductive)</td>
<td>65.12±0.48</td>
<td>-18.32</td>
</tr>
<tr>
<td>Our proposed method</td>
<td>Res12</td>
<td>(Inductive)</td>
<td>65.12±0.48</td>
<td>-18.32</td>
</tr>
</tbody>
</table>

\( \pm \): ResNet12 backbone with 256 output channels as [\text{others}]. PR: directly Prototype Refinement using unlabeled data.
as novel classes and are used to refine novel weights. Base class testing samples may be predicted as novel if the corresponding weight is refined with the similar but incorrect base class unlabeled samples. For example, as discussed in Section 1, novel class weight “school bus” in Fig. 1 (d) is likely to be incorrectly refined with unlabeled samples from base class “unicycle” (same wheel pattern), “ashcan” (similar color and texture), and “street sign” (often appear in the same image) which causes confusion. By contrast, our method can boost both Accb/all and Accn/all, which accounts for the obvious improvement on joint accuracy.

### Table 2: Incremental FSL and S2I-FSL results on tiered-ImageNet.

<table>
<thead>
<tr>
<th>Method</th>
<th>fθ</th>
<th>Type</th>
<th>200+5-way 1-shot Acc.</th>
<th>Δ</th>
<th>200+5-way 5-shot Acc.</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imprint [4]</td>
<td>Res18</td>
<td>Incremental FSL (Inductive)</td>
<td>40.83±0.45</td>
<td>-22.29</td>
<td>53.87±0.48</td>
<td>-17.18</td>
</tr>
<tr>
<td>LwoF [3]</td>
<td>Res18</td>
<td></td>
<td>53.42±0.56</td>
<td>-9.59</td>
<td>63.22±0.52</td>
<td>-7.27</td>
</tr>
<tr>
<td>Attractor [4]</td>
<td>Res18</td>
<td></td>
<td>56.11±0.33</td>
<td>-6.11</td>
<td>65.52±0.31</td>
<td>-4.48</td>
</tr>
<tr>
<td>TapNet [5]</td>
<td>Res18</td>
<td></td>
<td>55.42±0.59</td>
<td>-8.09</td>
<td>65.38±0.53</td>
<td>-6.37</td>
</tr>
<tr>
<td>XtarNet [6]</td>
<td>Res18</td>
<td></td>
<td>56.57±0.60</td>
<td>-8.02</td>
<td>66.52±0.49</td>
<td>-6.31</td>
</tr>
<tr>
<td>Meta-Inc-Baseline (ours)</td>
<td>Res18</td>
<td></td>
<td>63.16±0.51</td>
<td>-7.53</td>
<td>73.28±0.40</td>
<td>-5.35</td>
</tr>
<tr>
<td>LCwoF unlim [7]</td>
<td>Res12</td>
<td></td>
<td>58.82±N/A</td>
<td>-N/A</td>
<td>66.61±N/A</td>
<td>-N/A</td>
</tr>
<tr>
<td>Meta-Inc-Baseline (ours)</td>
<td>Res12</td>
<td></td>
<td>61.92±0.50</td>
<td>-7.87</td>
<td>72.98±0.38</td>
<td>-5.39</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + Graph [8]</td>
<td>Res12</td>
<td></td>
<td>59.73±0.56</td>
<td>-11.25</td>
<td>70.18±0.41</td>
<td>-8.58</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + FixMatch [9]</td>
<td>Res12</td>
<td>S2I-FSL</td>
<td>57.49±0.59</td>
<td>-8.12</td>
<td>73.34±0.38</td>
<td>-5.46</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + PR [10]</td>
<td>Res12</td>
<td>(Transductive)</td>
<td>57.99±0.61</td>
<td>-15.58</td>
<td>72.08±0.39</td>
<td>-7.47</td>
</tr>
<tr>
<td>Our proposed method</td>
<td>Res12</td>
<td></td>
<td>70.03±0.57</td>
<td>-5.41</td>
<td>75.73±0.37</td>
<td>-4.53</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + Graph [11]</td>
<td>Res12</td>
<td></td>
<td>51.87±0.48</td>
<td>-19.71</td>
<td>65.86±0.41</td>
<td>-12.63</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + FixMatch [12]</td>
<td>Res12</td>
<td>S2I-FSL</td>
<td>56.56±0.59</td>
<td>-8.37</td>
<td>73.22±0.39</td>
<td>-5.57</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + PR [13]</td>
<td>Res12</td>
<td></td>
<td>59.35±0.58</td>
<td>-14.55</td>
<td>72.68±0.39</td>
<td>-6.93</td>
</tr>
<tr>
<td>Our proposed method</td>
<td>Res12</td>
<td></td>
<td>71.64±0.56</td>
<td>-4.97</td>
<td>76.24±0.36</td>
<td>-4.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>fθ</th>
<th>Type</th>
<th>200+5-way 5-shot Acc.</th>
<th>Δ</th>
<th>Accb/all</th>
<th>Accn/all</th>
<th>Accb/h</th>
<th>Accn/h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-Inc-Baseline (baseline)</td>
<td></td>
<td></td>
<td>72.98</td>
<td>-5.39</td>
<td>65.91</td>
<td>80.05</td>
<td>73.36</td>
<td>83.39</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + Graph [14]</td>
<td></td>
<td></td>
<td>65.86</td>
<td></td>
<td>-12.63</td>
<td></td>
<td></td>
<td>70.65</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + FixMatch [15]</td>
<td></td>
<td></td>
<td>73.22 ↑</td>
<td></td>
<td>-5.57</td>
<td>64.13 ↓</td>
<td></td>
<td>82.30 ↑</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + PR [16]</td>
<td></td>
<td></td>
<td>72.68 ↓</td>
<td></td>
<td>-6.93</td>
<td>60.32 ↓</td>
<td></td>
<td>85.04 ↑</td>
</tr>
<tr>
<td>Our proposed method</td>
<td></td>
<td></td>
<td>76.24 ↑</td>
<td></td>
<td>-4.21</td>
<td>67.58 ↑</td>
<td></td>
<td>84.90 ↑</td>
</tr>
</tbody>
</table>

*↑(↓): Certain metric increases (significantly). ↓(↑): Certain metric decreases (significantly).*

### Table 3: Detailed evaluations of S2I-FSL on tiered-ImageNet 5-shot setting.

#### Semi-Supervised/Transductive FSL

Our model trained for S2I-FSL is also directly applied to regular semi-supervised and transductive FSL benchmarks. Here, we only need to perform classification on novel classes where unlabeled data solely contains novel class samples. As shown in Table 4, our method consistently outperforms other approaches using the same or deeper backbones. It can be attributed to the proposed Algorithm 2 for effectively utilizing unlabeled samples and the model adaptation mechanism for learning discriminative features of novel classes. Experiments on tiered-ImageNet are provided in Appendix F.

<table>
<thead>
<tr>
<th>Method</th>
<th>fθ</th>
<th>Transductive FSL 1-shot</th>
<th>Transductive FSL 5-shot</th>
<th>Semi-supervised FSL 1-shot</th>
<th>Semi-supervised FSL 5-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPN [26]</td>
<td>Conv4</td>
<td>55.51±0.86</td>
<td>69.86±0.65</td>
<td>Soft k-M [20]</td>
<td>Conv4</td>
</tr>
<tr>
<td>TEAM [24]</td>
<td>Res12</td>
<td>60.07±N/A</td>
<td>75.90±N/A</td>
<td>LST [20]</td>
<td>Res12</td>
</tr>
<tr>
<td>BD-CSPN [20]</td>
<td>Res12</td>
<td>65.94±N/A</td>
<td>79.23±N/A</td>
<td>LR+ICI [39]</td>
<td>Res12</td>
</tr>
<tr>
<td>LR+ICI [39]</td>
<td>Res12</td>
<td>68.70±1.10</td>
<td>79.69±0.66</td>
<td>EPNet [32]</td>
<td>Res12</td>
</tr>
<tr>
<td>DPGN [32]</td>
<td>Res12</td>
<td>67.77±0.32</td>
<td>84.60±0.43</td>
<td>TransMatch [44]</td>
<td>WRN28</td>
</tr>
<tr>
<td>Completion [35]</td>
<td>Res12</td>
<td>73.13±0.82</td>
<td>82.06±0.54</td>
<td>ICA+MSP [32]</td>
<td>DenseNet</td>
</tr>
<tr>
<td>Our proposed</td>
<td>Res12</td>
<td>77.37±0.85</td>
<td>86.02±0.41</td>
<td></td>
<td>Our proposed</td>
</tr>
</tbody>
</table>

### Table 4: Regular semi-supervised/transductive FSL results on mini-ImageNet.
5.3 Ablation Studies

Effectiveness of each component. We present detailed results of one-by-one employment of different components in Table 5. As discussed in Table 3, directly applying prototype refinement results in decreased joint accuracy and $\Delta$ metric, which is caused by the confusion between base and novel classes. With the help of our proposed meta-training Algorithm 2, the severe confusion is alleviated and about 11.0%/3.1% absolute performance gains are obtained in 1/5-shot cases. Moreover, model adaptation mechanism that learns discriminative features for novel classes yields another 1.3%/0.4% improvement.

<table>
<thead>
<tr>
<th>Method (Semi-supervised)</th>
<th>Proto refine</th>
<th>Fake unlabel</th>
<th>Model adapt</th>
<th>200+5-way 1-shot Acc.</th>
<th>$\Delta$</th>
<th>Accu/n</th>
<th>200+5-way 5-shot Acc.</th>
<th>$\Delta$</th>
<th>Accu/n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-Inc-Baseline</td>
<td></td>
<td></td>
<td>✓</td>
<td>61.92</td>
<td>-7.87</td>
<td>67.79</td>
<td>72.98</td>
<td>-5.39</td>
<td>83.39</td>
</tr>
<tr>
<td>Meta-Inc-Baseline + PR</td>
<td>✓</td>
<td></td>
<td></td>
<td>59.35</td>
<td>-14.55</td>
<td>75.41</td>
<td>72.68</td>
<td>-6.93</td>
<td>85.86</td>
</tr>
<tr>
<td>Our method w/o adaptation</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>70.34</td>
<td>-5.90</td>
<td>79.82</td>
<td>75.82</td>
<td>-4.49</td>
<td>87.21</td>
</tr>
<tr>
<td>Our full method</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>71.64</td>
<td>-4.97</td>
<td>80.91</td>
<td>76.24</td>
<td>-4.21</td>
<td>87.82</td>
</tr>
</tbody>
</table>

Table 5: Ablation study of each component proposed for S$^2$I-FSL on tiered-ImageNet.

Visualization results. The above observations are also confirmed by t-SNE visualizations. Fig. 3 shows a 200+5-way 1-shot semi-supervised incremental test episode on tiered-ImageNet without (left) and with (right) our proposed components. For clarity, 5 base classes and 5 novel classes are randomly chosen and features of 30 per-class test samples are considered. The blue novel class prototype (triangle) which was confused with green and orange base classes is more separable from them using Algorithm 2, and novel class features (blue and cyan, magenta and black) become more distinguishable after model adaptation.

Studies on unlabeled set U. To observe the performance on the second challenge of S$^2$I-FSL mentioned in Section 1, we vary the sample ratio in U as shown in Fig. 4. The accuracy decreases when fewer novel class samples are included in U, since the more unbalanced samples between base and novel classes, the more challenging the task becomes. Compared to Meta-Inc-Baseline + PR, our method is more robust to changes in sample ratio, as less degradation ($\delta$) is observed. More comparison and evaluation details are in Appendix G.

6 Conclusion

In this paper, a Semi-Supervised Incremental Few-Shot Learning (S$^2$I-FSL) benchmark is proposed to generalize previous semi-supervised FSL and incremental FSL under a more realistic and challenging setting. To solve S$^2$I-FSL, we introduce an efficient meta-baseline for incremental FSL. Based on that, a novel meta-training paradigm and a model adaptation scheme are proposed to fully explore unlabeled data. Extensive experiments on incremental FSL, semi-supervised FSL, and S$^2$I-FSL demonstrate the effectiveness of our method.
References


