

# Meta-Transfer Learning for Few-Shot Learning

Qianru Sun<sup>1,3\*</sup> Yaoyao Liu<sup>2\*</sup> Tat-Seng Chua<sup>1</sup> Bernt Schiele<sup>3</sup>

<sup>1</sup>National University of Singapore <sup>2</sup>Tianjin University<sup>†</sup>

<sup>3</sup>Max Planck Institute for Informatics, Saarland Informatics Campus

{qsun, schiele}@mpi-inf.mpg.de

liuyaoyao@tju.edu.cn {dcssq, dcscts}@nus.edu.sg

## Abstract

Meta-learning has been proposed as a framework to address the challenging few-shot learning setting. The key idea is to leverage a large number of similar few-shot tasks in order to learn how to adapt a base-learner to a new task for which only a few labeled samples are available. As deep neural networks (DNNs) tend to overfit using a few samples only, meta-learning typically uses shallow neural networks (SNNs), thus limiting its effectiveness. In this paper we propose a novel few-shot learning method called **meta-transfer learning (MTL)** which learns to adapt a **deep NN** for **few shot learning tasks**. Specifically, meta refers to training multiple tasks, and transfer is achieved by learning scaling and shifting functions of DNN weights for each task. In addition, we introduce the **hard task (HT) meta-batch** scheme as an effective learning curriculum for MTL. We conduct experiments using (5-class, 1-shot) and (5-class, 5-shot) recognition tasks on two challenging few-shot learning benchmarks: minImageNet and Fewshot-CIFAR100. Extensive comparisons to related works validate that our **meta-transfer learning** approach trained with the proposed **HT meta-batch** scheme achieves top performance. An ablation study also shows that both components contribute to fast convergence and high accuracy<sup>1</sup>.

## 1. Introduction

While deep learning systems have achieved great performance when sufficient amounts of labeled data are available [17, 46, 58], there has been growing interest in reducing the required amount of data. Few-shot learning tasks have been defined for this purpose. The aim is to learn new concepts from few labeled examples, e.g. 1-shot learn-

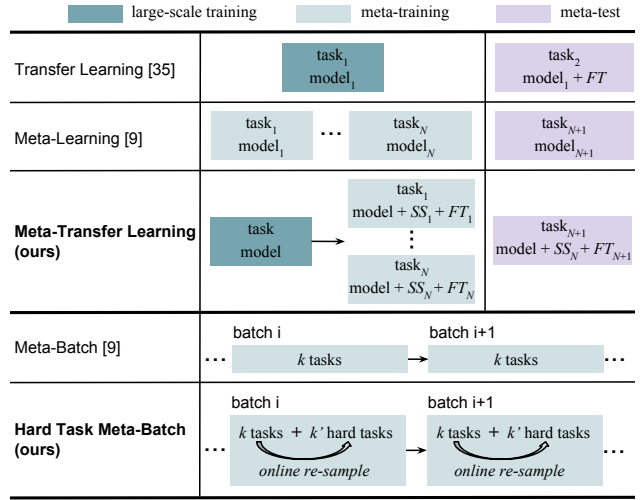


Figure 1. Meta-transfer learning (MTL) is our meta-learning paradigm and hard task (HT) meta-batch is our training strategy. The upper three rows show the differences between MTL and related methods, transfer-learning [35] and meta-learning [9]. The bottom rows compare HT meta-batch with the conventional meta-batch [9]. FT stands for fine-tuning a classifier. SS represents the Scaling and Shifting operations in our MTL method.

ing [25]. While humans tend to be highly effective in this context, often grasping the essential connection between new concepts and their own knowledge and experience, it remains challenging for machine learning approaches. E.g., on the CIFAR-100 dataset, a state-of-the-art method [34] achieves only 40.1% accuracy for 1-shot learning, compared to 75.7% for the all-class fully supervised case [6].

Few-shot learning methods can be roughly categorized into two classes: data augmentation and task-based meta-learning. Data augmentation is a classic technique to increase the amount of available data and thus also useful for few-shot learning [21]. Several methods propose to learn a data generator e.g. conditioned on Gaussian noise [29, 44, 54]. However, the generation models often

\*Equal contribution.

<sup>†</sup>Yaoyao Liu did this work during his internship at NUS.

<sup>1</sup>Code: <https://github.com/yaoyao-liu/meta-transfer-learning>

underperform when trained on few-shot data [1]. An alternative is to merge data from multiple tasks which, however, is not effective due to variances of the data across tasks [54].

In contrast to data-augmentation methods, meta-learning is a task-level learning method [2, 33, 52]. Meta-learning aims to accumulate experience from learning multiple tasks [9, 13, 31, 39, 48], while base-learning focuses on modeling the data distribution of a single task. A state-of-the-art representative of this, namely Model-Agnostic Meta-Learning (MAML), learns to search for the optimal initialization state to fast adapt a base-learner to a new task [9]. Its task-agnostic property makes it possible to generalize to few-shot supervised learning as well as unsupervised reinforcement learning [10, 13]. However, in our view, there are two main limitations of this type of approaches limiting their effectiveness: i) these methods usually require a large number of similar tasks for meta-training which is costly; and ii) each task is typically modeled by a low-complexity base learner (such as a shallow neural network) to avoid model overfitting, thus being unable to use deeper and more powerful architectures. For example, for the mini-ImageNet dataset [53], MAML uses a *shallow* CNN with only 4 CONV layers and its optimal performance was obtained learning on 240k tasks.

In this paper, we propose a novel meta-learning method called **meta-transfer learning (MTL)** leveraging the advantages of both transfer and meta learning (see conceptual comparison of related methods in Figure 1). In a nutshell, MTL is a novel learning method that helps deep neural nets converge faster while reducing the probability to overfit when using few labeled training data only. In particular, “transfer” means that DNN weights trained on large-scale data can be used in other tasks by two light-weight neuron operations: *Scaling* and *Shifting* (SS), i.e.  $\alpha X + \beta$ . “Meta” means that the parameters of these operations can be viewed as hyper-parameters trained on few-shot learning tasks [26, 31]. Large-scale trained DNN weights offer a good initialization, enabling fast convergence of meta-transfer learning with fewer tasks, e.g. only 8k tasks for miniImageNet [53], 30 times fewer than MAML [9]. Light-weight operations on DNN neurons have less parameters to learn, e.g. less than  $\frac{2}{49}$  if considering neurons of size  $7 \times 7$  ( $\frac{1}{49}$  for  $\alpha$  and  $< \frac{1}{49}$  for  $\beta$ ), reducing the chance of overfitting. In addition, these operations keep those trained DNN weights unchanged, and thus avoid the problem of “catastrophic forgetting” which means forgetting general patterns when adapting to a specific task [27, 28].

The second main contribution of this paper is an effective meta-training curriculum. Curriculum learning [3] and hard negative mining [47] both suggest that faster convergence and stronger performance can be achieved by a better arrangement of training data. Inspired by these ideas, we design our **hard task (HT) meta-batch** strategy to offer a

challenging but effective learning curriculum. As shown in the bottom rows of Figure 1, a conventional meta-batch contains a number of random tasks [9], but our HT meta-batch online re-samples harder ones according to past failure tasks with lowest validation accuracy.

Our overall contribution is thus three-fold: i) we propose a novel **MTL** method that learns to transfer large-scale pre-trained DNN weights for solving few-shot learning tasks; ii) we propose a novel **HT meta-batch** learning strategy that forces meta-transfer to “grow faster and stronger through hardship”; and iii) we conduct extensive experiments on two few-shot learning benchmarks, namely miniImageNet [53] and Fewshot-CIFAR100 (FC100) [34], and achieve the state-of-the-art performance.

## 2. Related work

**Few-shot learning** Research literature on few-shot learning exhibits great diversity. In this section, we focus on methods using the supervised meta-learning paradigm [9, 12, 52] most relevant to ours and compared to in the experiments. We can divide these methods into three categories. 1) *Metric learning* methods [48, 51, 53] learn a similarity space in which learning is efficient for few-shot examples. 2) *Memory network* methods [30, 31, 34, 42] learn to store “experience” when learning seen tasks and then generalize that to unseen tasks. 3) *Gradient descent* based methods [9, 13, 24, 39, 60] have a specific *meta-learner* that learns to adapt a specific *base-learner* (to few-shot examples) through different tasks. E.g. MAML [9] uses a meta-learner that learns to effectively initialize a base-learner for a new learning task. Meta-learner optimization is done by gradient descent using the validation loss of the base-learner. Our method is closely related. An important difference is that our MTL approach leverages transfer learning and benefits from referencing neuron knowledge in pre-trained deep nets. Although MAML can start from a pre-trained network, its element-wise fine-tuning makes it hard to learn deep nets without overfitting (validated in our experiments).

**Transfer learning** *What* and *how* to transfer are key issues to be addressed in transfer learning, as different methods are applied to different source-target domains and bridge different transfer knowledge [35, 55, 57, 59]. For deep models, a powerful transfer method is adapting a pre-trained model for a new task, often called *fine-tuning (FT)*. Models pre-trained on large-scale datasets have proven to generalize better than randomly initialized ones [8]. Another popular transfer method is taking pre-trained networks as backbone and adding high-level functions, e.g. for object detection and recognition [18, 49, 50] and image segmentation [5, 16]. Our meta-transfer learning leverages the idea of transferring pre-trained weights and aims to meta-learn how to effectively transfer. In this paper, large-scale trained DNN weights are *what* to transfer, and the operations of

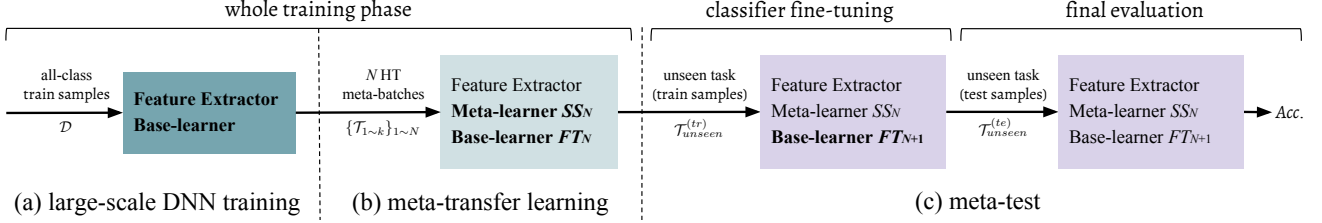


Figure 2. The pipeline of our proposed few-shot learning method, including three phases: (a) DNN training on large-scale data, i.e. using all training datapoints (Section 4.1); (b) Meta-transfer learning (MTL) that learns the parameters of *Scaling* and *Shifting* (*SS*), based on the pre-trained feature extractor (Section 4.2). Learning is scheduled by the proposed HT meta-batch (Section 4.3); and (c) meta-test is done for an unseen task which consists of a base-learner (classifier) *Fine-Tuning* (*FT*) stage and a final evaluation stage, described in the last paragraph in Section 3. Input data are along with arrows. Modules with names in bold get updated at corresponding phases. Specifically, *SS* parameters are learned by meta-training but fixed during meta-test. Base-learner parameters are optimized for every task.

*Scaling* and *Shifting* indicate *how* to transfer. Similar operations have been used to modulating the per-feature-map distribution of activations for visual reasoning [37].

Some few-shot learning methods have been proposed to use pre-trained weights as initialization [20, 30, 38, 41, 45]. Typically, weights are fine-tuned for each task, while we learn a meta-transfer learner through all tasks, which is different in terms of the underlying learning paradigm.

**Curriculum learning & Hard sample mining** Curriculum learning was proposed by Bengio *et al.* [3] and is popular for multi-task learning [14, 36, 43, 56]. They showed that instead of observing samples at random it is better to organize samples in a meaningful way so that fast convergence, effective learning and better generalization can be achieved. Pentina *et al.* [36] use adaptive SVM classifiers to evaluate task difficulty for later organization. Differently, our MTL method does task evaluation online at the phase of episode test, without needing any auxiliary model.

Hard sample mining was proposed by Shrivastava *et al.* [47] for object detection. It treats image proposals overlapped with ground truth as hard negative samples. Training on more confusing data enables the model to achieve higher robustness and better performance [4, 7, 15]. Inspired by this, we sample harder tasks online and make our MTL learner “grow faster and stronger through more hardness”. In our experiments, we show that this can be generalized to enhance other meta-learning methods, e.g. MAML [9].

### 3. Preliminary

We introduce the problem setup and notations of meta-learning, following related work [9, 34, 39, 53].

**Meta-learning** consists of two phases: meta-train and meta-test. A meta-training example is a classification task  $\mathcal{T}$  sampled from a distribution  $p(\mathcal{T})$ .  $\mathcal{T}$  is called episode, including a training split  $\mathcal{T}^{(tr)}$  to optimize the base-learner, and a test split  $\mathcal{T}^{(te)}$  to optimize the meta-learner. In particular, meta-training aims to learn from a number of episodes  $\{\mathcal{T}\}$  sampled from  $p(\mathcal{T})$ . An unseen task  $\mathcal{T}_{unseen}$  in meta-test will start from that experience of the meta-learner and

adapt the base-learner. The final evaluation is done by testing a set of unseen datapoints  $\mathcal{T}_{unseen}^{(te)}$ .

**Meta-training phase.** This phase aims to learn a meta-learner from multiple episodes. In each episode, meta-training has a two-stage optimization. Stage-1 is called base-learning, where the cross-entropy loss is used to optimize the parameters of the base-learner. Stage-2 contains a feed-forward test on episode test datapoints. The test loss is used to optimize the parameters of the meta-learner. Specifically, given an episode  $\mathcal{T} \in p(\mathcal{T})$ , the base-learner  $\theta_{\mathcal{T}}$  is learned from episode training data  $\mathcal{T}^{(tr)}$  and its corresponding loss  $\mathcal{L}_{\mathcal{T}}(\theta_{\mathcal{T}}, \mathcal{T}^{(tr)})$ . After optimizing this loss, the base-learner has parameters  $\hat{\theta}_{\mathcal{T}}$ . Then, the meta-learner is updated using test loss  $\mathcal{L}_{\mathcal{T}}(\hat{\theta}_{\mathcal{T}}, \mathcal{T}^{(te)})$ . After meta-training on all episodes, the meta-learner is optimized by test losses  $\{\mathcal{L}_{\mathcal{T}}(\hat{\theta}_{\mathcal{T}}, \mathcal{T}^{(te)})\}_{\mathcal{T} \in p(\mathcal{T})}$ . Therefore, the number of meta-learner updates equals to the number of episodes.

**Meta-test phase.** This phase aims to test the performance of the trained meta-learner for fast adaptation to unseen task. Given  $\mathcal{T}_{unseen}$ , the meta-learner  $\hat{\theta}_{\mathcal{T}}$  teaches the base-learner  $\theta_{\mathcal{T}_{unseen}}$  to adapt to the objective of  $\mathcal{T}_{unseen}$  by some means, e.g. through initialization [9]. Then, the test result on  $\mathcal{T}_{unseen}^{(te)}$  is used to evaluate the meta-learning approach. If there are multiple unseen tasks  $\{\mathcal{T}_{unseen}\}$ , the average result on  $\{\mathcal{T}_{unseen}^{(te)}\}$  will be the final evaluation.

### 4. Methodology

As shown in Figure 2, our method consists of three phases. First, we train a DNN on large-scale data, e.g. on miniImageNet (64-class, 600-shot) [53], and then fix the low-level layers as Feature Extractor (Section 4.1). Second, in the meta-transfer learning phase, MTL learns the *Scaling* and *Shifting* (*SS*) parameters for the Feature Extractor neurons, enabling fast adaptation to few-shot tasks (Section 4.2). For improved overall learning, we use our HT meta-batch strategy (Section 4.3). The training steps are detailed in Algorithm 1 in Section 4.4. Finally, the typical meta-test phase is performed, as introduced in Section 3.

#### 4.1. DNN training on large-scale data

This phase is similar to the classic pre-training stage as, e.g., pre-training on Imagenet for object recognition [40]. Here, we do not consider data/domain adaptation from other datasets, and pre-train on readily available data of few-shot learning benchmarks, allowing for fair comparison with other few-shot learning methods. Specifically, for a particular few-shot dataset, we merge all-class data  $\mathcal{D}$  for pre-training. For instance, for miniImageNet [53], there are totally 64 classes in the training split of  $\mathcal{D}$  and each class contains 600 samples used to pre-train a 64-class classifier.

We first randomly initialize a feature extractor  $\Theta$  (e.g. CONV layers in ResNets [17]) and a classifier  $\theta$  (e.g. the last FC layer in ResNets [17]), and then optimize them by gradient descent as follows,

$$[\Theta; \theta] =: [\Theta; \theta] - \alpha \nabla \mathcal{L}_{\mathcal{D}}([\Theta; \theta]), \quad (1)$$

where  $\mathcal{L}$  denotes the following empirical loss,

$$\mathcal{L}_{\mathcal{D}}([\Theta; \theta]) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} l(f_{[\Theta; \theta]}(x), y), \quad (2)$$

e.g. cross-entropy loss, and  $\alpha$  denotes the learning rate. In this phase, the feature extractor  $\Theta$  is learned. It will be frozen in the following meta-training and meta-test phases, as shown in Figure 2. The learned classifier  $\theta$  will be discarded, because subsequent few-shot tasks contain different classification objectives, e.g. 5-class instead of 64-class classification for miniImageNet [53].

#### 4.2. Meta-transfer learning (MTL)

As shown in Figure 2(b), our proposed meta-transfer learning (MTL) method optimizes the meta operations *Scaling* and *Shifting* (SS) through HT meta-batch training (Section 4.3). Figure 3 visualizes the difference of updating through SS and FT. SS operations, denoted as  $\Phi_{S_1}$  and  $\Phi_{S_2}$ , do not change the frozen neuron weights of  $\Theta$  during learning, while FT updates the complete  $\Theta$ .

In the following, we detail the SS operations. Given a task  $\mathcal{T}$ , the loss of  $\mathcal{T}^{(tr)}$  is used to optimize the current base-learner (classifier)  $\theta'$  by gradient descent:

$$\theta' \leftarrow \theta - \beta \nabla_{\theta} \mathcal{L}_{\mathcal{T}^{(tr)}}([\Theta; \theta'], \Phi_{S_{\{1,2\}}}), \quad (3)$$

which is different to Eq. 1, as we do not update  $\Theta$ . Note that here  $\theta$  is different to the one from the previous phase, the large-scale classifier  $\theta$  in Eq. 1. This  $\theta$  concerns only a few of classes, e.g. 5 classes, to classify each time in a novel few-shot setting.  $\theta'$  corresponds to a temporal classifier only working in the current task, initialized by the  $\theta$  optimized for the previous task (see Eq. 5).

$\Phi_{S_1}$  is initialized by ones and  $\Phi_{S_2}$  by zeros. Then, they are optimized by the test loss of  $\mathcal{T}^{(te)}$  as follows,

$$\Phi_{S_i} =: \Phi_{S_i} - \gamma \nabla_{\Phi_{S_i}} \mathcal{L}_{\mathcal{T}^{(te)}}([\Theta; \theta'], \Phi_{S_{\{1,2\}}}). \quad (4)$$

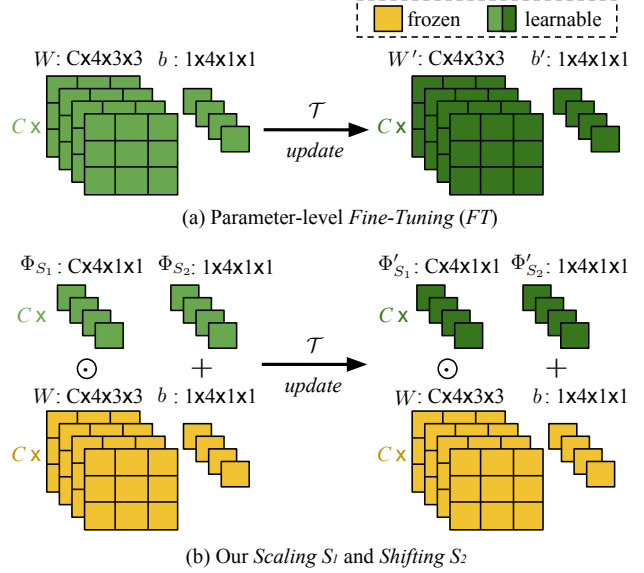


Figure 3. (a) Parameter-level *Fine-Tuning* (FT) is a conventional meta-training operation, e.g. in MAML [9]. Its update works for all neuron parameters,  $W$  and  $b$ . (b) Our neuron-level *Scaling* and *Shifting* (SS) operations in MTL. They reduce the number of learning parameters and avoid overfitting problems. In addition, they keep large-scale trained parameters (in yellow) frozen, preventing ‘catastrophic forgetting’ [27, 28].

In this step,  $\theta$  is updated with the same learning rate  $\gamma$  as in Eq. 4,

$$\theta =: \theta - \gamma \nabla_{\theta} \mathcal{L}_{\mathcal{T}^{(te)}}([\Theta; \theta'], \Phi_{S_{\{1,2\}}}). \quad (5)$$

Re-linking to Eq. 3, we note that the above  $\theta'$  comes from the last epoch of base-learning on  $\mathcal{T}^{(tr)}$ .

Next, we describe how we apply  $\Phi_{S_{\{1,2\}}}$  to the frozen neurons as shown in Figure 3(b). Given the trained  $\Theta$ , for its  $l$ -th layer containing  $K$  neurons, we have  $K$  pairs of parameters, respectively as weight and bias, denoted as  $\{(W_{i,k}, b_{i,k})\}$ . Note that the neuron location  $l, k$  will be omitted for readability. Based on MTL, we learn  $K$  pairs of scalars  $\{\Phi_{S_{\{1,2\}}}\}$ . Assuming  $X$  is input, we apply  $\{\Phi_{S_{\{1,2\}}}\}$  to  $(W, b)$  as

$$SS(X; W, b; \Phi_{S_{\{1,2\}}}) = (W \odot \Phi_{S_1})X + (b + \Phi_{S_2}), \quad (6)$$

where  $\odot$  denotes the element-wise multiplication.

Taking Figure 3(b) as an example of a single  $3 \times 3$  filter, after SS operations, this filter is scaled by  $\Phi_{S_1}$  then the feature maps after convolutions are shifted by  $\Phi_{S_2}$  in addition to the original bias  $b$ . Detailed steps of SS are given in Algorithm 2 in Section 4.4.

Figure 3(a) shows a typical parameter-level *Fine-Tuning* (FT) operation, which is in the meta optimization phase of our related work MAML [9]. It is obvious that FT updates the complete values of  $W$  and  $b$ , and has a large number of



parameters, and our SS reduces this number to below  $\frac{2}{9}$  in the example of the figure.

In summary, SS can benefit MTL in three aspects. 1) It starts from a strong initialization based on a large-scale trained DNN, yielding fast convergence for MTL. 2) It does not change DNN weights, thereby avoiding the problem of “catastrophic forgetting” [27, 28] when learning specific tasks in MTL. 3) It is light-weight, reducing the chance of overfitting of MTL in few-shot scenarios.

### 4.3. Hard task (HT) meta-batch

In this section, we introduce a method to schedule hard tasks in meta-training batches. The conventional meta-batch is composed of randomly sampled tasks, where the randomness implies random difficulties [9]. In our meta-training pipeline, we intentionally pick up failure cases in each task and re-compose their data to be harder tasks for adverse re-training. We aim to force our meta-learner to “grow up through hardness”.

**Pipeline.** Each task  $\mathcal{T}$  has two splits,  $\mathcal{T}^{(tr)}$  and  $\mathcal{T}^{(te)}$ , for base-learning and test, respectively. As shown in Algorithm 2 line 2-5, base-learner is optimized by the loss of  $\mathcal{T}^{(tr)}$  (in multiple epochs). SS parameters are then optimized by the loss of  $\mathcal{T}^{(te)}$  once. We can also get the recognition accuracy of  $\mathcal{T}^{(te)}$  for  $M$  classes. Then, we choose the lowest accuracy  $Acc_m$  to determine the most difficult class- $m$  (also called failure class) in the current task.

After obtaining all failure classes (indexed by  $\{m\}$ ) from  $k$  tasks in current meta-batch  $\{\mathcal{T}_{1 \sim k}\}$ , we re-sample tasks from their data. Specifically, we assume  $p(\mathcal{T}|\{m\})$  is the task distribution, we sample a “harder” task  $\mathcal{T}^{hard} \in p(\mathcal{T}|\{m\})$ . Two important details are given below.

**Choosing hard class- $m$ .** We choose the failure class- $m$  from each task by ranking the class-level accuracies instead of fixing a threshold. In a dynamic online setting as ours, it is more sensible to choose the hardest cases based on ranking rather than fixing a threshold ahead of time.

**Two methods of hard tasking using  $\{m\}$ .** Chosen  $\{m\}$ , we can re-sample tasks  $\mathcal{T}^{hard}$  by (1) directly using the samples of class- $m$  in the current task  $\mathcal{T}$ , or (2) indirectly using the label of class- $m$  to sample new samples of that class. In fact, setting (2) considers to include more data variance of class- $m$  and it works better than setting (1) in general.

### 4.4. Algorithm

Algorithm 1 summarizes the training process of two main stages: large-scale DNN training (line 1-5) and meta-transfer learning (line 6-22). HT meta-batch re-sampling and continuous training phases are shown in lines 16-20, for which the failure classes are returned by Algorithm 2, see line 14. Algorithm 2 presents the learning process on a single task that includes episode training (lines 2-5) and episode test, i.e. meta-level update (lines 6). In lines 7-11,

the recognition rates of all test classes are computed and returned to Algorithm 1 (line 14) for hard task sampling.

---

#### Algorithm 1: Meta-transfer learning (MTL)

---

**Input:** Task distribution  $p(\mathcal{T})$  and corresponding dataset  $\mathcal{D}$ , learning rates  $\alpha, \beta$  and  $\gamma$

**Output:** Feature extractor  $\Theta$ , base learner  $\theta$ , SS parameters  $\Phi_{S_{\{1,2\}}}$

```

1 Randomly initialize  $\Theta$  and  $\theta$ ;
2 for samples in  $\mathcal{D}$  do
3   Evaluate  $\mathcal{L}_{\mathcal{D}}([\Theta; \theta])$  by Eq. 2;
4   Optimize  $\Theta$  and  $\theta$  by Eq. 1;
5 end
6 Initialize  $\Phi_{S_1}$  by ones, initialize  $\Phi_{S_2}$  by zeros;
7 Reset and re-initialize  $\theta$  for few-shot tasks;
8 for meta-batches do
9   Randomly sample tasks  $\{\mathcal{T}\}$  from  $p(\mathcal{T})$ ;
10  while not done do
11    Sample task  $\mathcal{T}_i \in \{\mathcal{T}\}$ ;
12    Optimize  $\Phi_{S_{\{1,2\}}}$  and  $\theta$  with  $\mathcal{T}_i$  by
      Algorithm 2;
13    Get the returned class- $m$  then add it to  $\{m\}$ ;
14  end
15  Sample hard tasks  $\{\mathcal{T}^{hard}\}$  from  $\subseteq p(\mathcal{T}|\{m\})$ ;
16  while not done do
17    Sample task  $\mathcal{T}_j^{hard} \in \{\mathcal{T}^{hard}\}$ ;
18    Optimize  $\Phi_{S_{\{1,2\}}}$  and  $\theta$  with  $\mathcal{T}_j^{hard}$  by
      Algorithm 2;
19  end
20  Empty  $\{m\}$ .
21 end

```

---



---

#### Algorithm 2: Detail learning steps within a task $\mathcal{T}$

---

**Input:**  $\mathcal{T}$ , learning rates  $\beta$  and  $\gamma$ , feature extractor  $\Theta$ , base learner  $\theta$ , SS parameters  $\Phi_{S_{\{1,2\}}}$

**Output:** Updated  $\theta$  and  $\Phi_{S_{\{1,2\}}}$ , the worst classified class- $m$  in  $\mathcal{T}$

```

1 Sample  $\mathcal{T}^{(tr)}$  and  $\mathcal{T}^{(te)}$  from  $\mathcal{T}$ ;
2 for samples in  $\mathcal{T}^{(tr)}$  do
3   Evaluate  $\mathcal{L}_{\mathcal{T}^{(tr)}}$ ;
4   Optimize  $\theta'$  by Eq. 3;
5 end
6 Optimize  $\Phi_{S_{\{1,2\}}}$  and  $\theta$  by Eq. 4 and Eq. 5;
7 while not done do
8   Sample class- $c$  in  $\mathcal{T}^{(te)}$ ;
9   Compute  $Acc_c$  for  $\mathcal{T}^{(te)}$ ;
10 end
11 Return class- $m$  with the lowest accuracy  $Acc_m$ .

```

---

## 5. Experiments

We evaluate the proposed **MTL** and **HT meta-batch** in terms of few-shot recognition accuracy and model convergence speed. Below we describe the datasets and detailed settings, followed by an ablation study and a comparison to state-of-the-art methods.

### 5.1. Datasets and implementation details

We conduct few-shot learning experiments on two benchmarks, miniImageNet [53] and Fewshot-CIFAR100 (FC100) [34]. miniImageNet is widely used in related works [9, 11, 13, 32, 39]. FC100 is newly proposed in [34] and is more challenging in terms of lower image resolution and stricter training-test splits than miniImageNet.

**miniImageNet** was proposed by Vinyals *et al.* [53] for few-shot learning evaluation. Its complexity is high due to the use of ImageNet images, but requires less resource and infrastructure than running on the full ImageNet dataset [40]. In total, there are 100 classes with 600 samples of  $84 \times 84$  color images per class. These 100 classes are divided into 64, 16, and 20 classes respectively for sampling tasks for meta-training, meta-validation and meta-test, following related works [9, 11, 13, 32, 39].

**Fewshot-CIFAR100 (FC100)** is based on the popular object classification dataset CIFAR100 [23]. The splits were proposed by [34] (Please check details in the supplementary). It offers a more challenging scenario with lower image resolution and more challenging meta-training/test splits that are separated according to object super-classes. It contains 100 object classes and each class has 600 samples of  $32 \times 32$  color images. The 100 classes belong to 20 super-classes. Meta-training data are from 60 classes belonging to 12 super-classes. Meta-validation and meta-test sets contain 20 classes belonging to 4 super-classes, respectively. These splits accord to super-classes, thus minimize the information overlap between training and val/test tasks.

The following settings are used on both datasets. We train a large-scale DNN with all training datapoints (Section 4.1) and stop this training after  $10k$  iterations. We use the same task sampling method as related works [9, 39]. Specifically, 1) we consider the 5-class classification and 2) we sample 5-class, 1-shot (5-shot or 10-shot) episodes to contain 1 (5 or 10) samples for train episode, and 15 (uniform) samples for episode test. Note that in the state-of-the-art work [34], 32 and 64 samples are respectively used in 5-shot and 10-shot settings for episode test. In total, we sample  $8k$  tasks for meta-training (same for w/ and w/o HT meta-batch), and respectively sample 600 random tasks for meta-validation and meta-test. Please check the supplementary document (or [GitHub repository](#)) for other implementation details, e.g. learning rate and dropout rate.

**Network architecture.** We present the details for the Fea-

ture Extractor  $\Theta$ , MTL meta-learner with *Scaling*  $\Phi_{S_1}$  and *Shifting*  $\Phi_{S_2}$ , and MTL base-learner (classifier)  $\theta$ .

**The architecture of  $\Theta$**  have two options, ResNet-12 and 4CONV, commonly used in related works [9, 30, 32, 34, 39, 53]. **4CONV** consists of 4 layers with  $3 \times 3$  convolutions and 32 filters, followed by batch normalization (BN) [19], a ReLU nonlinearity, and  $2 \times 2$  max-pooling. **ResNet-12** is more popular in recent works [11, 30, 32, 34]. It contains 4 residual blocks and each block has 3 CONV layers with  $3 \times 3$  kernels. At the end of each residual block, a  $2 \times 2$  max-pooling layer is applied. The number of filters starts from 64 and is doubled every next block. Following 4 blocks, there is a mean-pooling layer to compress the output feature maps to a feature embedding. **The difference** between using 4CONV and using ResNet-12 in our methods is that ResNet-12 MTL sees the large-scale data training, but 4CONV MTL is learned from scratch because of its poor performance for large-scale data training (see results in the supplementary). Therefore, we emphasize the experiments of using ResNet-12 MTL for its superior performance. **The architectures of  $\Phi_{S_1}$  and  $\Phi_{S_2}$**  are generated according to the architecture of  $\Theta$ , as introduced in Section 4.2. That is when using ResNet-12 in MTL,  $\Phi_{S_1}$  and  $\Phi_{S_2}$  also have 12 layers, respectively. **The architecture of  $\theta$**  is an FC layer. We empirically find that a single FC layer is faster to train and more effective for classification than multiple layers. (see comparisons in the supplementary).

### 5.2. Ablation study setting

In order to show the effectiveness of our approach, we design some ablative settings: two baselines without meta-learning but more classic learning, three baselines of *Fine-Tuning (FT)* on smaller number of parameters (Table 1), and two MAML variants using our deeper pre-trained model and HT meta-batch (Table 2 and Table 3). Note that the alternative meta-learning operation to *SS* is the *FT* used in MAML. Some bullet names are explained as follows.

**update  $[\Theta; \theta]$  (or  $\theta$ ).** There is no meta-training phase. During test phase, each task has its whole model  $[\Theta; \theta]$  (or the classifier  $\theta$ ) updated on  $\mathcal{T}^{(tr)}$ , and then tested on  $\mathcal{T}^{(te)}$ .

**FT  $[\Theta_4; \theta]$  (or  $\theta$ ).** These are straight-forward ways to define a smaller set of meta-learner parameters than MAML. We can freeze low-level pre-trained layers and meta-learn the classifier layer  $\theta$  with (or without) high-level CONV layer  $\Theta_4$  that is the 4th residual block of ResNet-12.

### 5.3. Results and analysis

Table 1, Table 2 and Table 3 present the overall results on miniImageNet and FC100 datasets. Extensive comparisons are done with ablative methods and the state-of-the-arts. Note that tables present the highest accuracies for which the iterations were chosen by validation. For the miniIma-

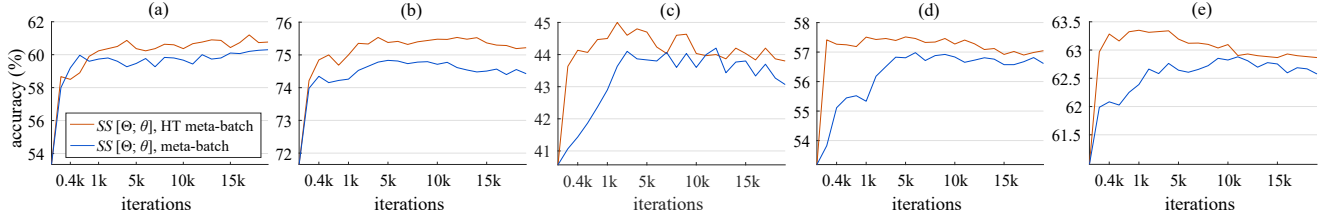


Figure 4. (a)(b) show the results of 1-shot and 5-shot on miniImageNet; (c)(d)(e) show the results of 1-shot, 5-shot and 10-shot on FC100.

geNet, iterations for 1-shot and 5-shot are at 17k and 14k, respectively. For the FC100, iterations are all at 1k. Figure 4 shows the performance gap between *with* and *without* HT meta-batch in terms of accuracy and converging speed.

**Result overview on miniImageNet.** In Table 2, we can see that the proposed MTL with  $SS [\Theta; \theta]$ , HT meta-batch and ResNet-12(pre) achieves the best few-shot classification performance with 61.2% for (5-class, 1-shot). Besides, it tackles the (5-class, 5-shot) tasks with an accuracy of 75.5% that is comparable to the state-of-the-art results, i.e. 76.7%, reported by TADAM [34] whose model used 72 additional FC layers in the ResNet-12 arch. In terms of the network arch, it is obvious that models using ResNet-12 (pre) outperforms those using 4CONV by large margins, e.g. 4CONV models have the best 1-shot result with 50.44% [51] which is 10.8% lower than our best.

**Result overview on FC100.** In Table 3, we give the results of TADAM using their reported numbers in the paper [34]. We used the public code of MAML [9] to get its results for this new dataset. Comparing these methods, we can see that MTL consistently outperforms MAML by large margins, i.e. around 7% in all tasks; and surpasses TADAM by a relatively larger number of 5% for 1-shot, and with 1.5% and 1.8% respectively for 5-shot and 10-shot tasks.

**MTL vs. No meta-learning.** Table 1 shows the results of *No meta-learning* on the top block. Compared to these, our approach achieves significantly better performance even *without* HT meta-batch, e.g. the largest margins are 10.2% for 1-shot and 8.6% for 5-shot on miniImageNet. This validates the effectiveness of our meta-learning method for tackling few-shot learning problems. Between two *No*

*meta-learning* methods, we can see that updating both feature extractor  $\Theta$  and classifier  $\theta$  is inferior to updating  $\theta$  only, e.g. around 5% reduction on miniImageNet 1-shot. One reason is that in few-shot settings, there are too many parameters to optimize with little data. This supports our motivation to learn only  $\theta$  during base-learning.

**Performance effects of MTL components.** MTL with full components,  $SS [\Theta; \theta]$ , HT meta-batch and ResNet-12(pre), achieves the best performances for all few-shot settings on both datasets, see Table 2 and Table 3. We can conclude that our large-scale network training on deep CNN significantly boost the few-shot learning performance. This is an important gain brought by the transfer learning idea in our MTL approach. It is interesting to note that this gain on FC100 is not as large as for miniImageNet: only 1.7%, 1.0% and 4.0%. The possible reason is that FC100 tasks for meta-train and meta-test are clearly split according to super-classes. The data domain gap is larger than that for miniImageNet, which makes transfer more difficult.

HT meta-batch and ResNet-12(pre) in our approach can be generalized to other meta-learning models. MAML 4CONV with HT meta-batch gains averagely 1% on two datasets. When changing 4CONV by deep ResNet-12 (pre) it achieves significant improvements, e.g. 10% and 9% on miniImageNet. Compared to MAML variants, our MTL results are consistently higher, e.g. 2.5% ~ 3.3% on FC100. People may argue that MAML fine-tuning(*FT*) all network parameters is likely to overfit to few-shot data. In the middle block of Table 1, we show the ablation study of freezing low-level pre-trained layers and meta-learn only the high-level layers (e.g. the 4-th residual block of ResNet-12) by the *FT* operations of MAML. These all yield inferior performances than using our *SS*. An additional observation is that *SS\** performs consistently better than *FT\**.

**Speed of convergence of MTL.** MAML [9] used 240k tasks to achieve the best performance on miniImageNet. Impressively, our MTL methods used only 8k tasks, see Figure 4(a)(b) (note that each iteration contains 2 tasks). This advantage is more obvious for FC100 on which MTL methods need at most 2k tasks, Figure 4(c)(d)(e). We attest this to two reasons. First, MTL starts from the pre-trained ResNet-12. And second, *SS* (in MTL) needs to learn only  $< \frac{2}{9}$  parameters of the number of *FT* (in MAML) when using ResNet-12.

	miniImageNet		FC100		
	1 (shot)	5	1	5	10
<i>update</i> $[\Theta; \theta]$	45.3	64.6	38.4	52.6	58.6
<i>update</i> $\theta$	50.0	66.7	39.3	51.8	61.0
<i>FT</i> $\theta$	55.9	71.4	41.6	54.9	61.1
<i>FT</i> $[\Theta_4; \theta]$	57.2	71.6	40.9	54.3	61.3
<i>FT</i> $[\Theta; \theta]$	58.3	71.6	41.6	54.4	61.2
<i>SS</i> $[\Theta_4; \theta]$	59.2	73.1	42.4	55.1	61.6
<i>SS</i> $[\Theta; \theta]$ (Ours)	60.2	74.3	43.6	55.4	62.4

Table 1. Classification accuracy (%) using ablative models, on two datasets. “meta-batch” and “ResNet-12(pre)” are used.

Few-shot learning method		Feature extractor	1-shot	5-shot
<i>Data augmentation</i>	Adv. ResNet, [29]	WRN-40 (pre)	55.2	69.6
	Delta-encoder, [44]	VGG-16 (pre)	58.7	73.6
<i>Metric learning</i>	MatchingNets, [53]	4 CONV	43.44 $\pm$ 0.77	55.31 $\pm$ 0.73
	ProtoNets, [48]	4 CONV	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66
	RelationNets, [51]	4 CONV	50.44 $\pm$ 0.82	65.32 $\pm$ 0.70
<i>Memory network</i>	MetaNetworks, [31]	5 CONV	49.21 $\pm$ 0.96	–
	SNAIL, [30]	ResNet-12 (pre) <sup>◊</sup>	55.71 $\pm$ 0.99	68.88 $\pm$ 0.92
	TADAM, [34]	ResNet-12 (pre) <sup>†</sup>	58.5 $\pm$ 0.3	<b>76.7 <math>\pm</math> 0.3</b>
<i>Gradient descent</i>	MAML, [9]	4 CONV	48.70 $\pm$ 1.75	63.11 $\pm$ 0.92
	Meta-LSTM, [39]	4 CONV	43.56 $\pm$ 0.84	60.60 $\pm$ 0.71
	Hier.Bayes, [13]	4 CONV	49.40 $\pm$ 1.830	–
	BilevelProgram, [11]	ResNet-12 <sup>◊</sup>	50.54 $\pm$ 0.85	64.53 $\pm$ 0.68
	MetaGAN, [60]	ResNet-12	52.71 $\pm$ 0.64	68.63 $\pm$ 0.67
	adaResNet, [32]	ResNet-12 <sup>‡</sup>	56.88 $\pm$ 0.62	71.94 $\pm$ 0.57
MAML, HT	FT $[\Theta; \theta]$ , HT meta-batch	4 CONV	49.1 $\pm$ 1.9	64.1 $\pm$ 0.9
MAML deep, HT	FT $[\Theta; \theta]$ , HT meta-batch	ResNet-12 (pre)	59.1 $\pm$ 1.9	73.1 $\pm$ 0.9
<b>MTL (Ours)</b>	SS $[\Theta; \theta]$ , meta-batch	ResNet-12 (pre)	60.2 $\pm$ 1.8	74.3 $\pm$ 0.9
	SS $[\Theta; \theta]$ , HT meta-batch	ResNet-12 (pre)	<b>61.2 <math>\pm</math> 1.8</b>	<b>75.5 <math>\pm</math> 0.8</b>

◊Additional 2 convolutional layers    ‡Additional 1 convolutional layer    †Additional 72 fully connected layers

Table 2. The 5-way, 1-shot and 5-shot classification accuracy (%) on miniImageNet dataset. “pre” means pre-trained for a single classification task using all training datapoints.

Few-shot learning method		Feature extractor	1-shot	5-shot	10-shot
<i>Gradient descent</i>	MAML, [9] <sup>‡</sup>	4 CONV	38.1 $\pm$ 1.7	50.4 $\pm$ 1.0	56.2 $\pm$ 0.8
<i>Memory network</i>	TADAM, [34]	ResNet-12 (pre) <sup>†</sup>	40.1 $\pm$ 0.4	56.1 $\pm$ 0.4	61.6 $\pm$ 0.5
MAML, HT	FT $[\Theta; \theta]$ , HT meta-batch	4 CONV	39.9 $\pm$ 1.8	51.7 $\pm$ 0.9	57.2 $\pm$ 0.8
MAML deep, HT	FT $[\Theta; \theta]$ , HT meta-batch	ResNet-12 (pre)	41.8 $\pm$ 1.9	55.1 $\pm$ 0.9	61.9 $\pm$ 0.8
<b>MTL (Ours)</b>	SS $[\Theta; \theta]$ , meta-batch	ResNet-12 (pre)	43.6 $\pm$ 1.8	55.4 $\pm$ 0.9	62.4 $\pm$ 0.8
	SS $[\Theta; \theta]$ , HT meta-batch	ResNet-12 (pre)	<b>45.1 <math>\pm</math> 1.8</b>	<b>57.6 <math>\pm</math> 0.9</b>	<b>63.4 <math>\pm</math> 0.8</b>

†Additional 72 fully connected layers    ‡Our implementation using the public code of MAML.

Table 3. The 5-way with 1-shot, 5-shot and 10-shot classification accuracy (%) on Fewshot-CIFAR100 (FC100) dataset. “pre” means pre-trained for a single classification task using all training datapoints.

**Speed of convergence of HT meta-batch.** Figure 4 shows 1) MTL with HT meta-batch consistently achieves higher performances than MTL with the conventional meta-batch [9], in terms of the recognition accuracy in all settings; and 2) it is impressive that MTL with HT meta-batch achieves top performances early, after *e.g.* about  $2k$  iterations for 1-shot,  $1k$  for 5-shot and  $1k$  for 10-shot, on the more challenging dataset – FC100.

## 6. Conclusions

In this paper, we show that our novel MTL trained with HT meta-batch learning curriculum achieves the top performance for tackling few-shot learning problems. The key operations of MTL on pre-trained DNN neurons proved highly efficient for adapting learning experience to the unseen task. The superiority was particularly achieved in the extreme 1-

shot cases on two challenging benchmarks – miniImageNet and FC100. In terms of learning scheme, HT meta-batch showed consistently good performance for all baselines and ablative models. On the more challenging FC100 benchmark, it showed to be particularly helpful for boosting convergence speed. This design is independent from any specific model and could be generalized well whenever the hardness of task is easy to evaluate in online iterations.

## Acknowledgments

This research is part of NExT research which is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IRC@SG Funding Initiative. It is also partially supported by German Research Foundation (DFG CRC 1223), and National Natural Science Foundation of China (61772359).



## References

- [1] S. Bartunov and D. P. Vetrov. Few-shot generative modelling with generative matching networks. In *AISTATS*, 2018. 2
- [2] S. Bengio, Y. Bengio, J. Cloutier, and J. Gecsei. On the optimization of a synaptic learning rule. In *Optimality in Artificial and Biological Neural Networks*, pages 6–8. Univ. of Texas, 1992. 2
- [3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009. 2, 3
- [4] O. Canévet and F. Fleuret. Large scale hard sample mining with monte carlo tree search. In *CVPR*, 2016. 3
- [5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 2
- [6] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*, 2016. 1
- [7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 3
- [8] D. Erhan, Y. Bengio, A. C. Courville, P. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11:625–660, 2010. 2
- [9] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 1, 2, 3, 4, 5, 6, 7, 8, 11
- [10] C. Finn, K. Xu, and S. Levine. Probabilistic model-agnostic meta-learning. In *NeurIPS*, 2018. 2
- [11] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, 2018. 6, 8
- [12] H. E. Geoffrey and P. C. David. Using fast weights to deblur old memories. In *CogSci*, 1987. 2
- [13] E. Grant, C. Finn, S. Levine, T. Darrell, and T. L. Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *ICLR*, 2018. 2, 6, 8
- [14] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. Automated curriculum learning for neural networks. In *ICML*, 2017. 3
- [15] B. Harwood, V. Kumar, G. Carneiro, I. Reid, and T. Drummond. Smart mining for deep metric learning. In *ICCV*, 2017. 3
- [16] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *ICCV*, 2017. 2
- [17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 4
- [18] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017. 2
- [19] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 6
- [20] R. Keshari, M. Vatsa, R. Singh, and A. Noore. Learning structure and strength of CNN filters for small sample size training. In *CVPR*, 2018. 3
- [21] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. *arXiv*, 1703.09554, 2017. 1
- [22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 1412.6980, 2014. 11
- [23] A. Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2009. 6
- [24] Y. Lee and S. Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *ICML*, 2018. 2
- [25] F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, 2006. 1
- [26] Z. Li, F. Zhou, F. Chen, and H. Li. Meta-sgd: Learning to learn quickly for few shot learning. In *ICML*, 2018. 2
- [27] D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. In *NIPS*, 2017. 2, 4, 5
- [28] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 3–17, 1989. 2, 4, 5
- [29] A. Mehrotra and A. Dukkipati. Generative adversarial residual pairwise networks for one shot learning. *arXiv*, 1703.08033, 2017. 1, 8
- [30] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. Snail: A simple neural attentive meta-learner. In *ICLR*, 2018. 2, 3, 6, 8
- [31] T. Munkhdalai and H. Yu. Meta networks. In *ICML*, 2017. 2, 8
- [32] T. Munkhdalai, X. Yuan, S. Mehri, and A. Trischler. Rapid adaptation with conditionally shifted neurons. In *ICML*, 2018. 6, 8
- [33] D. K. Naik and R. Mammone. Meta-neural networks that learn by learning. In *IJCNN*, 1992. 2
- [34] B. N. Oreshkin, P. Rodríguez, and A. Lacoste. TADAM: task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018. 1, 2, 3, 6, 7, 8, 11
- [35] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Networks*, 22(2):199–210, 2011. 1, 2
- [36] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. In *CVPR*, 2015. 3
- [37] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 3
- [38] S. Qiao, C. Liu, W. Shen, and A. L. Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018. 3
- [39] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 2, 3, 6, 8
- [40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4, 6

- [41] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019. 3
- [42] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. P. Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016. 2
- [43] N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris. Curriculum learning for multi-task classification of visual attributes. In *ICCV Workshops*, 2017. 3
- [44] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, R. S. Feris, A. Kumar, R. Giryes, and A. M. Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *NeurIPS*, 2018. 1, 8
- [45] T. R. Scott, K. Ridgeway, and M. C. Mozer. Adapted deep embeddings: A synthesis of methods for k-shot inductive transfer learning. In *NeurIPS*, 2018. 3
- [46] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. 1
- [47] A. Shrivastava, A. Gupta, and R. B. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 2, 3
- [48] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017. 2, 8
- [49] Q. Sun, L. Ma, S. Joon Oh, L. Van Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. In *CVPR*, 2018. 2
- [50] Q. Sun, B. Schiele, and M. Fritz. A domain based approach to social relation recognition. In *CVPR*, 2017. 2
- [51] 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*, 2018. 2, 7, 8
- [52] S. Thrun and L. Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pages 3–17. Springer, 1998. 2
- [53] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In *NIPS*, 2016. 2, 3, 4, 6, 8
- [54] Y. Wang, R. B. Girshick, M. Hebert, and B. Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018. 1, 2
- [55] Y. Wei, Y. Zhang, J. Huang, and Q. Yang. Transfer learning via learning to transfer. In *ICML*, 2018. 2
- [56] D. Weinshall, G. Cohen, and D. Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *ICML*, 2018. 3
- [57] J. Yang, R. Yan, and A. G. Hauptmann. Adapting SVM classifiers to data with shifted distributions. In *ICDM Workshops*, 2007. 2
- [58] L. Yann, B. Yoshua, and H. Geoffrey. Deep learning. *Nature*, 521(7553):436, 2015. 1
- [59] A. R. Zamir, A. Sax, W. B. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 2
- [60] R. Zhang, T. Che, Z. Grahahramani, Y. Bengio, and Y. Song. Metagan: An adversarial approach to few-shot learning. In *NeurIPS*, 2018. 2, 8

## Supplementary materials

These materials include the details of network architecture (§A), implementation (§B), FC100 dataset splits (§C), standard variance analysis (§D), additional ablation results (§E), and some interpretation of our meta-learned model (§F). In addition, our open-source code is on GitHub<sup>2</sup>.

### A. Network architectures

In Figure S1, we present the 4CONV architecture for feature extractor  $\Theta$ , as illustrated in Section 5.1 “Network architecture” of the main paper.

In Figure S2, we present the other architecture – ResNet-12. Figure S2(a) shows the details of a single residual block and Figure S2(b) shows the whole network consisting of four residual blocks and a mean-pooling layer.

The input of  $\Theta$  is the 3-channel RGB image, and the output is the 512-dimensional feature vector.  $\alpha = 0.1$  is set for all leakyReLU activation functions in ResNet-12.

### B. Implementation details

For the **phase of DNN training on large-scale data**, the model is trained by Adam optimizer [22]. Its learning rate is initialized as 0.001, and decays to its half every  $5k$  iterations until it is lower than 0.0001. We set the keep probability of the dropout as 0.9 and batch-size as 64. The pre-training stops after  $10k$  iterations. Note that for the hyperparameter selection, we randomly choose 550 samples each class as the training set, and the rest as validation. After the grid search of hyperparameters, we fix them and mix up all samples (64 classes, 600 samples each class), in order to do the final pre-training. Besides, these pre-training samples are augmented with horizontal flip.

For the **meta-train phase**, we sample 5-class, 1-shot (5-shot or 10-shot) episodes to contain 1 (5 or 10) sample(s) for episode training, and 15 samples for episode test uniformly, following the setting of MAML [9]. The base-learner  $\theta$  is optimized by batch gradient descent with the learning rate of 0.01. It gets updated with 20 and 60 epochs respectively for 1-shot and 5-shot tasks on the miniImageNet dataset, and 20 epochs for all tasks on the FC100 dataset. The meta-learner, *i.e.*, the parameters of the  $\mathcal{SS}$  operations, is optimized by Adam optimizer [22]. Its learning rate is initialized as 0.001, and decays to the half every  $1k$  iterations until 0.0001. The size of meta-batch is set to 2 (tasks) due to the memory limit.

Using our **HT meta-batch strategy**, hard tasks are sampled every time after running 10 meta-batches, *i.e.*, the failure classes used for sampling hard tasks are from 20 tasks. The number of hard task is selected for different settings by validation: 10 and 4 hard tasks respectively for the 1-shot

and 5-shot experiments on the miniImageNet dataset; and respectively 20, 10 and 4 hard tasks for the 1-shot, 5-shot and 10-shot experiments on the FC100 dataset.

For the **meta-test phase**, we sample 5-class, 1-shot (5-shot or 10-shot) episodes and each episode contains 1 (5 or 10) sample(s) for both episode train and episode test. On each dataset, we sample 600 meta-test tasks. All these settings are exactly the same as MAML [9].

### C. Super-class splits on FC100

In this section, we show the details of the FC100 splits according to the super-class labels, same with TADAM [34].

**Training split** super-class indexes: 1, 2, 3, 4, 5, 6, 9, 10, 15, 17, 18, 19; and corresponding labels: fish, flowers, food\_containers, fruit\_and\_vegetables, household\_electrical\_devices, household\_furniture, large\_man-made\_outdoor\_things, large\_natural\_outdoor\_scenes, reptiles, trees, vehicles\_1, vehicles\_2.

**Validation split** super-class indexes: 8, 11, 13, 16; and corresponding labels: large\_carnivores, large\_omnivores\_and\_herbivores, non-insect\_invertebrates, small\_mammals.

**Test split** super-class indexes: 0, 7, 12, 14; and corresponding labels: aquatic\_mammals, insects, medium\_mammals, people.

An episode (task) is independently sampled from a corresponding split, *e.g.* a meta-train episode contains 5 classes that can only be belonging to the 12 super-classes in the training split. Therefore, there is no fine-grained information overlap between meta-train and meta-test tasks.

### D. Standard variance analysis

The final accuracy results reported in our main paper are the mean values and standard variances of the results of 600 meta-test tasks. The standard variance is affected by the number of episode test samples. As introduced in §B, we use the same setting as MAML [9] which used a smaller number of samples for episode test (1 sample for 1-shot episode test and 5 samples for 5-shot), making the result variance higher. Other works that used more samples for episode test got lower variances, *e.g.*, TADAM [34] used 100 samples and its variances are about  $\frac{1}{6}$  and  $\frac{1}{3}$  of MAML’s respectively for miniImageNet 1-shot and 5-shot.

In order to have a fair comparison with TADAM in terms of this issue, we supplement the experiments using 100 episode test samples at the meta-test. We get the new confidence intervals (using our method: MTL w/o HT meta-batch) as 0.71% (0.3% for TADAM) and 0.54% (0.3% for TADAM) respectively for 1-shot and 5-shot on the miniImageNet dataset, and 0.70% (0.4% for TADAM), 0.63%

<sup>2</sup><https://github.com/yaoyao-liu/meta-transfer-learning>

(0.4% for TADAM) and 0.58% (0.5% for TADAM) respectively for 1-shot, 5-shot and 10-shot on the FC100 dataset.

## E. Additional ablation study

We supplement the results in Table S1, for the comparisons mentioned in Section 5.1 of main paper. Red numbers on the bottom row are copied from the main paper (corresponding to the MTL setting:  $SS \Theta$ , meta-batch) and shown here for the convenience of comparison.

To get the first row, we train 4CONV net by large-scale data (same to the pre-training of ResNet-12) and get inferior results, as we declared in the main paper. Results on the second and third rows show the performance drop when changing the single FC layer  $\theta$  to multiple layers, *e.g.* 2 FC layers and 3 FC layers. Results on the fourth row show the performance drop when updating both  $\Theta$  and  $\theta$  for the base-learning. The reason is that  $\Theta$  has too many parameters to update with too little data.

## F. Interpretation of meta-learned $SS$

In Figure S3, we show the statistic histograms of learned  $SS$  parameters, taking miniImageNet 1-shot as an example setting. Scaling parameters  $\Phi_{S_1}$  are initialized as 1 and shifting parameters  $\Phi_{S_1}$  as 0. After meta-train, we observe that these statistics are close to Gaussian distributions respectively with (0.9962, 0.0084) and (0.0003, 0.0002) as (mean, variance) values, which shows that the uniform initialization has been changed to Gaussian distribution through few-shot learning. Possible interpretations are in three-fold: 1) majority patterns trained by a large number of few-shot tasks are close to the ones trained by large-scale data; 2) tail patterns with clear scale and shift values are the ones really contributing to adapting the model to few-shot tasks; 3) tail patterns are of small quantity, enabling the fast learning convergence.



Meta-learning	Base-learning	FC dim of $\theta$	Feature extractor	miniImageNet		1-shot	FC100	
				1-shot	5-shot		5-shot	10-shot
$\Phi_{S_1}, \Phi_{S_2}$	$\theta$	5	4 CONV (pre)	$45.6 \pm 1.8$	$61.2 \pm 0.9$	$38.0 \pm 1.6$	$46.4 \pm 0.9$	$56.5 \pm 0.8$
$\Phi_{S_1}, \Phi_{S_2}$	$\theta$ (2-layer)	512, 5	ResNet-12 (pre)	$59.1 \pm 1.9$	$70.7 \pm 0.9$	$40.3 \pm 1.9$	$53.3 \pm 0.9$	$54.1 \pm 0.8$
$\Phi_{S_1}, \Phi_{S_2}$	$\theta$ (3-layer)	1024, 512, 5	ResNet-12 (pre)	$56.2 \pm 1.8$	$68.7 \pm 0.9$	$40.0 \pm 1.8$	$52.3 \pm 1.0$	$53.8 \pm 0.8$
$\Phi_{S_1}, \Phi_{S_2}$	$\Theta, \theta$	5	ResNet-12 (pre)	$59.6 \pm 1.8$	$71.6 \pm 0.9$	$43.3 \pm 1.9$	$54.6 \pm 1.0$	$60.7 \pm 0.8$
$\Phi_{S_1}, \Phi_{S_2}$	$\theta$	5	ResNet-12 (pre)	$60.2 \pm 1.8$	$74.3 \pm 0.9$	$43.6 \pm 1.8$	$55.4 \pm 0.9$	$62.4 \pm 0.8$

Table S1. Additional ablative study. On the last row, we show the red numbers which are reported in our main paper (corresponding to the MTL setting:  $SS [\Theta; \theta]$ , meta-batch).

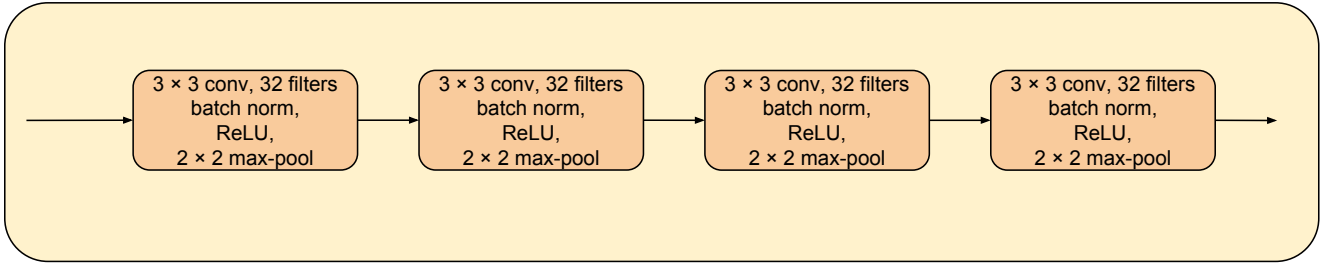
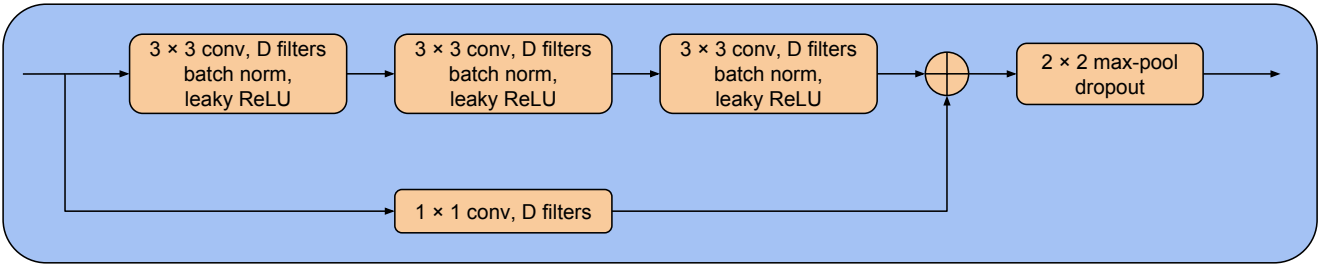


Figure S1. Network architecture of 4CONV

(a) Residual block, D filters



(b) Feature extractor

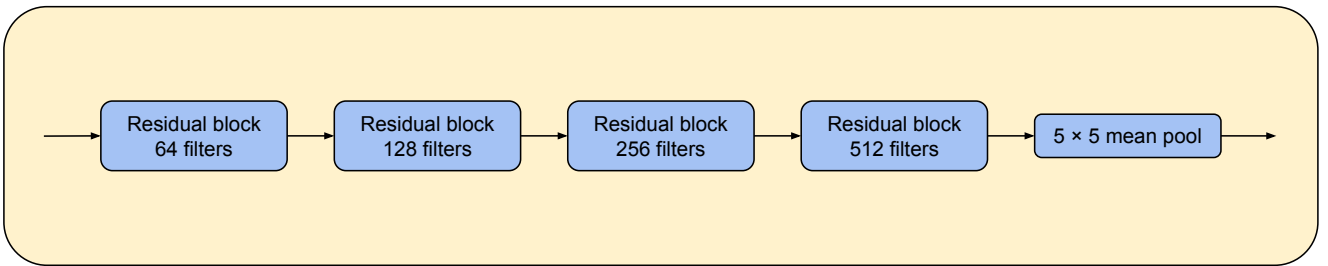


Figure S2. Network architecture of ResNet-12

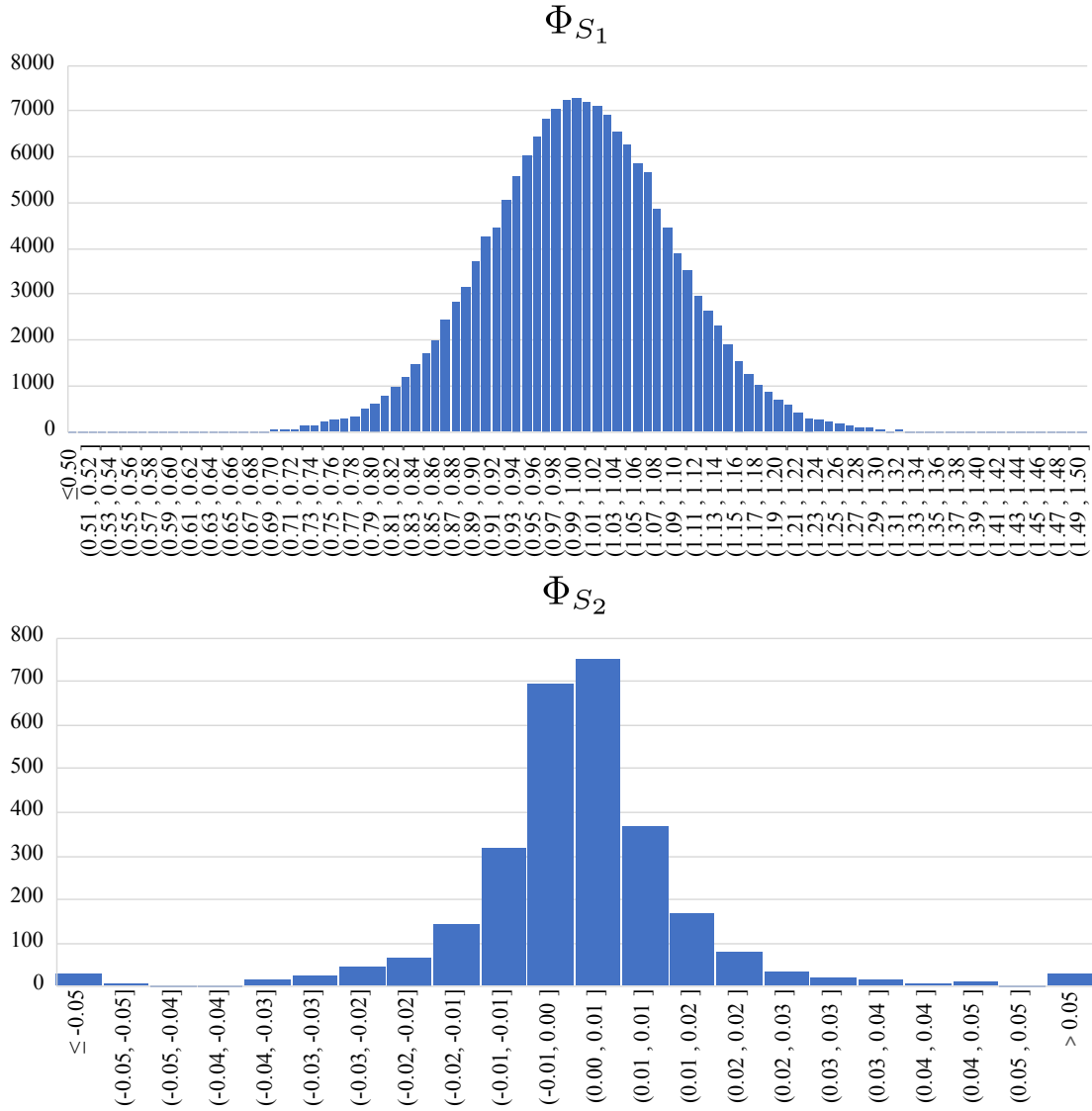


Figure S3. The statistic histograms of learned SS parameters, taking miniImageNet 1-shot as an example setting.