Continual Local Replacement for Few-shot Image Recognition

Canyu Le, Zhonggui Chen, Xihan Wei, Biao Wang, Lei Zhang

1School of Informatics, Xiamen University
2DAMO Academy, Alibaba Group

lecanyu@gmail.com, chenzhonggui@xmu.edu.cn, 
{xihan.wxh, wb.wangbiao, lei.zhang.lz}@alibaba-inc.com

Abstract

The goal of few-shot learning is to learn a model that can recognize novel classes based on one or few training data. It is challenging mainly due to two aspects: (1) it lacks good feature representation of novel classes; (2) a few labeled data could not accurately represent the true data distribution. In this work, we use a sophisticated network architecture to learn better feature representation and focus on the second issue. A novel continual local replacement strategy is proposed to address the data deficiency problem. It takes advantage of the content in unlabeled images to continuously enhance labeled ones. Specifically, a pseudo labeling strategy is adopted to constantly select semantic similar images on the fly. Original labeled images will be locally replaced by the selected images for the next epoch training. In this way, the model can directly learn new semantic information from unlabeled images and the capacity of supervised signals in the embedding space can be significantly enlarged. This allows the model to improve generalization and learn a better decision boundary for classification. Extensive experiments demonstrate that our approach can achieve highly competitive results over existing methods on various few-shot image recognition benchmarks.

1 Introduction

While deep learning has achieved remarkable results in image recognition [Krizhevsky et al., 2012; He et al., 2016], it is highly data-hungry and requires massive labeled training data. In contrast, human-level intelligence can achieve fast learning after observing only one or few instances [Lake et al., 2011]. To relieve this gap, researchers have devoted efforts to few-shot learning problem, such as similarity metric [Koch et al., 2015; Snell et al., 2017], meta learning [Finn et al., 2017; Rusu et al., 2019], and augmentation [Hariharan and Girshick, 2017; Wang et al., 2018; Chen et al., 2019d].

However, the few-shot learning problem still remains challenging. We argue that the main challenge comes from two aspects: (1) feature deterioration. The trained feature representation from the training dataset may lose its discriminative property on novel classes. For example, as illustrated in Fig. 1 (a), the ProtoNet [Snell et al., 2017] features deteriorate from training classes to novel testing classes. The model which works well on training datasets may not have good performance on novel classes. (2) Data deficiency. A single or a few of labeled data could not represent the true data distribution. As a result, it’s difficult to learn a good decision boundary even with a decent feature representation. This concept is showed in Fig. 1 (b). Most of previous approaches treat their model as a black box and suffer from both issues. Recent work [Chen et al., 2019a] shows that existing approaches like ProtoNet and MAML [Finn et al., 2017] could be beaten by the standard transfer learning baseline in some experiment settings.

In this paper, we focus on the data deficiency issue and use a sophisticated network architecture (e.g. ResNet) to alleviate feature deterioration as suggested by Kornblith et al. [Kornblith et al., 2019]. We come up with our approach from an important observation. As illustrated in Fig. 1 (c), the only difference between two few-shot episodic testing (i.e. first row and second row) is the labeled images (i.e. support set), but the testing accuracy varies significantly. It seems that representative training data play a vital role in few-shot learning. Indeed, it is hard to precisely define what a representative sample is since we may not have prior knowledge about the novel classes in practice. However, it should be generally helpful if model has the chance to see more data.

Based on this observation, we propose a continual local replacement approach which leverages the content of unlabeled images to constantly alter the labeled images. In particular, some regions of the labeled images will be replaced by unlabeled images and its contents can be learned through a supervised way. Note that the labeled and unlabeled images may be semantically different (i.e. different classes). In this case, the replacement can be seen as artifacts or random erasing [Zhong et al., 2020] which brings limited benefits. To select informative unlabeled data, we borrow the idea of pseudo-labeling [Lee, 2013] and design an algorithm which makes the data selection on the fly before each training epoch. This brings two advantages. First, semantically
similar data will be selected to enhance labeled images, and thus the model can learn new semantic information. Second, different data will be continually selected which diversifies semantic information and the supervised signals in the embedding space could be enlarged.

Our main technical contribution is a continual local replacement strategy which effectively addresses the data deficiency issue and learns a better classifier decision boundary. Our algorithm is built upon standard transfer learning and can be seen as an extension of the baseline method in [Chen et al., 2019a]. Extensive experiments show that this simple yet effective approach can improve the baseline performance and achieve state-of-the-art results on various image recognition datasets. Source code has been made available in https://github.com/Lecanyu/ContinualLocalReplacement. We hope this approach can be a strong baseline for future research.

2 Related Work

Our method takes unlabeled images to alter labeled ones for few-shot image recognition. Such a strategy is closely related to few-shot image recognition, semi-supervised learning, and data augmentation. We briefly review the most relevant works below.

Few-shot image recognition. The goal of few-shot image recognition is to endow models with the ability to recognize novel classes and datasets where only a limited amount of labeled data is available. Many approaches have been proposed for this goal. The former work like [Fei-Fei et al., 2006; Salakhutdinov et al., 2012] presented the Bayesian model for few-shot image recognition. The recent metric learning approaches [Koch et al., 2015; Snell et al., 2017; Sung et al., 2018] learn metrics from pairwise comparisons between image instances. And the meta learning methods [Finn et al., 2017; Zhoua et al., 2018; Rusu et al., 2019] learn a good model initialization for the few-shot adaptation. Besides, the attention-based [Wang et al., 2017], graph-based [Garcia and Bruna, 2018; Liu et al., 2019], and memory-based [Santor et al., 2016; Cai et al., 2018] strategies have also been investigated.

Semi-supervised learning. Unlike standard supervised learning, it attempts to leverage unlabeled data to help learning, especially when labeled data are insufficient. The previous work [Grandvalet and Bengio, 2005] proposed the entropy regularization to encourage learning low-density separations between classes. The pseudo-label approach [Lee, 2013] iteratively trains and updates the model on unlabeled data with guessed labels. The perturbation and consistency regularizations strategies [Sajjadi et al., 2016; Miyato et al., 2018] perturb unlabeled data and force consistent output to exploit potential data distribution. A hybrid approach [Berthelot et al., 2019] integrates previous strategies and achieves the new state-of-the-art. These semi-supervised methods provide ways and ideas to leverage unlabeled data. But they are usually inapplicable on few-shot learning where the number of labeled data is extremely small. More recently, [Ren et al., 2018] tackles few-shot image recognition by proposing a variant of ProtoNet which takes advantage of unlabeled data to further adjust the learned prototypes. In this paper, we take the idea of pseudo-labeling with image local replacement to address the data deficiency issue in few-shot learning.

Data augmentation. Data augmentation is widely adopted in various machine learning problems. The typical and standard image augmentations are rotating, flipping, cropping, color jittering and etc. Such augmentations can endow models with better generalization [Krizhevsky et al., 2012]. More sophisticated augmentation methods have also been explored, such as image synthesis [Wang et al., 2018], random erasing [Zhong et al., 2020], feature hallucination and augmentation [Hariharan and Girshick, 2017; Chen et al., 2019d].

In these strategies, the most relevant works are [Chen et al., 2019b; Chen et al., 2019c]. They introduced various patch-level image augmentation methods like mixup and replacement. However, they ignored the diversity of augmented images and used a fixed augmented set even though a learning-based augmentation strategy is adopted in [Chen et al., 2019c]. By contrast, we apply image local replacement on the fly in each training epoch. The semantic information could be directly diversified through this way. Our strategy is simple and effective. It can be seamlessly integrated with the standard transfer learning procedure.

3 Continual Local Replacement

The few-shot image recognition problem can be described on training and testing datasets. The training dataset has abundant labeled classes \( C_{\text{train}} \). After training on samples with labels from \( C_{\text{train}} \), the goal is to produce a model
for recognizing a disjoint set of novel classes $C_{test}$ (i.e. $C_{train} \cap C_{test} = \emptyset$) for which only a single or few of labeled samples are available. Formally, let $(x, y)$ denote an image and its label respectively. The training dataset $D_{train} = \{(x_i, y_i)\}_{i=1}^{N_{train}}$, where $y_i \in C_{train}$. The testing dataset $D_{test} = \{(x_i, y_i)\}_{i=1}^{N_{test}}$, where $y_i \in C_{test}$. Most of recent works on few-shot learning follow the episodic paradigm [Vinyals et al., 2016]. In particular, the episodic testing consists of hundreds of independent testing episodes. For each episode, it will randomly sample an episodic testing dataset $d_{test} = \{(x_i, y_i)\}_{i=1}^{m} \subset D_{test}$ which contains $n$ random classes from $C_{test}$ and total $m = (k + t) \times n$ images where $k$ and $t$ are the number of labeled images and testing images per class respectively. In few-shot literature, the $n \times k$ labeled images and $n \times t$ testing images are also referred to as the support set $d_{support}$ and query set $d_{query}$, respectively. To mimic the episodic testing, most previous metric learning and meta learning methods like [Snell et al., 2017; Finn et al., 2017] develop episodic training. Our approach instead is built on the standard transfer learning which trains on $D_{train}$ and fine tunes on $D_{test}$. With continual local replacement, this transfer learning approach demonstrates highly competitive performance over existing methods.

### 3.1 Image Local Replacement

Our primary goal is to provide a chance for the model to see some replacement methods. For example, a random region of an image is replaced by another one as showed in Fig. 2(a). The original image can be divided into several blocks and some of them are either replaced like Fig. 2(b) or linearly mixed with other images like Fig. 2(c). All these replacement methods are applicable for our purpose. Note that the random local replacement is crucial for the robustness. When two images are semantically different, the replaced regions may be interpreted as partial occlusions. This can be seen as an augmentation to improve the generalization ability [Chen et al., 2019b; Zhong et al., 2020]. When two images are semantically similar, the model has a chance to learn new semantic information from replaced regions. Except the existing replacement methods, we may also design other fancy methods as long as it can satisfy these two objectives. But this is beyond the scope of this paper since our technical contribution mainly lies in the following continual replacement approach.

### 3.2 Training

In the training stage, we first apply local replacement on the original training set $D_{train}$ to synthesize an augmented version training set $D_{train}^s$. Specifically, for each image $x_i \in D_{train}$, we randomly select another image $x_j \in D_{train}$, $i \neq j$ with 0.5 probability that $y_i = y_j$ (i.e. two different images with the same label). Then, one or several local regions of $x_i$ will be replaced by $x_j$ to synthesize a new image $x_i^\ast$. The new image $x_i^\ast$ is still with the original label $y_i$. The reason why we introduce $D_{train}^s$ in training is that we want the model to understand such local replacement and make it robust to partial occlusions.

After building up $D_{train}^s$, both the original images and synthesized images are simultaneously fed into CNN for training. This procedure is illustrated in Fig. 3. Specifically, we optimize the loss function Eq. 1 during the training:

$$
\arg \min_{\theta,w} \sum_{D_{train},D_{train}^s} \ell(C_w(f_\theta(x_i)), y_i),
$$

where $\ell(\cdot, \cdot)$ is the standard cross-entropy loss, $f_\theta$ denotes the feature extractor backbone with trainable parameters $\theta$ and $C_w$ is a classifier (e.g. fully connected layer) with parameters $w$.

### 3.3 Fine Tuning

After training, we fine tune the model to recognize novel classes in $D_{test}$. For each testing episode, additional $u$ images per class will be randomly sampled as the unlabeled set $d_{unlabel} = \{x_i\}_{i=1}^{u}$. In other words, there will be total $m = (k + t + u) \times n$ images including support and query sets.

We follow the standard transfer learning procedure to fine tune the trained model on $D_{train}$ and $D_{train}^s$. Specifically,
suppose the pre-trained feature extractor is reasonably good. We fix the pre-trained feature extractor parameters $\theta$ and create a new classifier with random initialized parameters $w$. The image local replacement is constantly applied when tuning the classifier. At the first fine-tune epoch, the original support set is fed into the model and classifier parameters $w$ are updated by minimizing the cross-entropy loss. Before the following tuning epochs, the latest model makes predictions for all unlabeled images. The pseudo labels $\tilde{y}_i$ are assigned to unlabeled set $d_u = \{(x_i, \tilde{y}_i)\}$. Then, local regions of an image $x_i \in d_{\text{support}}$ will be replaced by an unlabeled image $x_j \in d_{\text{unlabel}}$ with $\tilde{y}_j = y_i$ to synthesize a new image $x_i'$. And an augmented support dataset $d^p_{\text{support}} = \{(x_i', \tilde{y}_i)\}$ can be obtained. The $d^p_{\text{support}}$ is fed into CNN to fine tune the classifier in the next epoch and Eq. 2 will be optimized:

$$\arg \min_w \sum_{d^p_{\text{support}}} \ell(C_w(f_\theta(x_i)), y_i).$$ (2)

The complete fine tuning algorithm is summarized in Alg. 1.

**Algorithm 1** The fine tuning algorithm

**Input:** The pretrained feature extractor $f_\theta$; support, query, unlabeled sets $d_{\text{support}}, d_{\text{query}}, d_{\text{unlabel}}$

**Output:** The linear classifier $C_w$

Randomly initialize a new linear classifier $C_w$. Initialize $d^p_{\text{support}} \leftarrow d_{\text{support}}$.

for epoch = 1, 2, ... do

for mini-batch in $d_{\text{support}}$ do

Optimize classifier $C_w$ using Eq. 2.

end for

Predict labels $\tilde{y}_i$ for each image $x_i \in d_{\text{unlabel}}$.

Select a subset $d_{\text{unlabel}} \subset d_{\text{unlabel}}$ which is semantically similar with $d_{\text{support}}$.

$d^p_{\text{support}} \leftarrow$ locally replace $d_{\text{support}}$ using $d_{\text{unlabel}}$

end for

This approach is robust to wrong predictions on unlabeled images (i.e. $y_i \neq \tilde{y}_i$) by controlling the area of local replacement. We set a threshold $\alpha$ so that the maximum $\alpha \%$ area of the original image is allowed to be locally replaced. When using a small $\alpha$, the wrong prediction doesn’t hurt performance because it can be interpreted as partial occlusions or artifacts [Chen et al., 2019b; Zhong et al., 2020]. However, small replaced regions introduce little semantic information. So we prefer to set a larger $\alpha$ so that the model can learn new semantic contents. Indeed, large replaced regions may lead to a risk of learning wrong contents and the fine-tune procedure may be trapped in poor local minima. We alleviate this issue by randomly determining the position and the size of replaced regions. Note that the fine-tune procedure can still converge quickly since $d^p_{\text{support}}$ always contains the content of the original support set. Empirically, this simple strategy works well and $\alpha = 65\%$ can roughly achieve the best trade-off across all experiment benchmarks.

**4 Experiments**

The experiments are conducted on four widely used few-shot learning benchmarks: MiniImageNet [Ravi and Larochelle, 2017], TieredImageNet [Ren et al., 2018], Caltech-256 [Griffin et al., 2007] and CUB-200 [Wah et al., 2011]. These datasets cover from general objects (e.g. MiniImageNet) to fine-grained bird species (e.g. CUB-200) and they can provide a comprehensive evaluation for our method. We implement our approach based on a recent testbed [Chen et al., 2019a] and follow its basic experimental settings, hyper-parameters and dataset split. We use ResNet-18 backbone with a single fully connected classifier (FC) layer on all datasets. The ResNet-18 backbone is only trained on the training set $D_{\text{train}}$ and the augmented version $D'_{\text{train}}$. Note that $D_{\text{train}}$ and $D'_{\text{train}}$ are essentially the same dataset and thus the total capacity of training information is not changed. During the episodic testing stage, the weights of backbone are fixed and a new FC layer will be fine-tuned. We adopt the image block augmentation [Chen et al., 2019b] as the local replacement strategy. Additional $u = 15$ unlabeled images are randomly sampled and maximum 4 and 6 random blocks are allowed to be replaced during the training and fine tuning stages, respectively.

**4.1 Evaluating Standard Few-shot Performance**

We follow the episodic testing protocol [Vinyals et al., 2016] for the standard few-shot performance evaluation.

**Baselines and competitors.** There are a lot of existing few-shot learning methods. We compare our approach with the most relevant ones. The competitors include popular baselines like MAML [Finn et al., 2017], semi-supervised approaches like S.S. ProtoNet [Ren et al., 2018], augmentation-based methods like BlockAug [Chen et al., 2019b] and state-of-the-art approaches like LEO [Rusu et al., 2019]. Since different backbone architectures could significantly affect the results, we report the used architecture for each method along with the 5-way 1-shot and 5-way 5-shot accuracy.

In addition to our Continual Local Replacement (CLR) approach, we also report the results of two variants: Vanilla
and CLR+Imprinting. The vanilla version is a simple transfer learning approach. The model is trained on $D_{train}$ and $D_{train}^r$, but it is fine-tuned on the original support set $d_{support}$ without any local replacement. So it can be seen as a baseline of our approach. Our algorithm has good scalability and can be easily combined with existing methods like the weight line of our approach. Our algorithm has good scalability and can be seen as a base-line. D-fer learning approach. The model is trained on

$$\text{linear} [Qi, et al. 2017]$$

When the features and weights are normalized, the embedded features and the weights of FC layer during the classifier. In particular, the imprinting method normalizes the imprinting technique to our framework since it learns similarity [Qi, et al. 2018]. This validates our method outperforms two most related methods BlockAug [Chen, et al. 2019c] with a clear improvement (3%-7%). This validates the effectiveness of our continual local replacement strategy. (3) The imprinting variant is more improved. This means that our CLR technique can effectively improve the baseline. (3) The imprinting variant is more effective on fine grained recognition tasks than general object recognition tasks. This variant demonstrates strong performance on CUB-200 dataset but is slightly worse than CLR on other three general object recognition datasets. Interestingly, some previous metric-based learning methods like ProtoNet also show strong results on CUB-200. Since Imprinting and ProtoNet explicitly maximize the cosine and Euclidean similarity respectively, we may think these metric-based learning strategies are suitable for fine grained classification tasks through reducing the intra-class variation [Qi, et al., 2018; Chen, et al., 2019a].

### 4.2 Ablation Study

We conduct ablation studies to help understand our approach better.

**The number of replaced blocks.** We adopt the image block augmentation as the local replacement method. The replaced area can be controlled by the number of replaced blocks. We evaluate our approach with different numbers of replaced blocks on both training and fine-tuning stages. As shown in Tab.2, the best results are achieved when the maximum numbers of replaced blocks on training/fine-tuning are roughly 3 and 6. This indicates that it is necessary to replace some blocks to build up $D_{train}^r$ for training to obtain the best results on the fine tune stage.

| Methods | Arch. | MiniImageNet (%) | CUB-200 (%) | Caltech-256 (%) | TieredImageNet (%) |
|---------|-------|------------------|-------------|-----------------|--------------------|
| S.K. ProtoNet [Ren et al., 2018] | CONV4 | 50.41 ± 0.31 | 64.39 ± 0.24 | / | / |
| TPN [Liu et al., 2019] | ResNet-18 | 54.72 ± 0.84 | 69.25 ± 0.67 | / | / | 52.39 ± 0.44 | 69.88 ± 0.20 |
| MAML [Finn et al., 2017] | ResNet-18 | 49.61 ± 0.92 | 65.72 ± 0.77 | 69.96 ± 1.01 | 82.70 ± 0.65 | 57.33 ± 1.00 | 75.77 ± 0.70 | / | / |
| MatchNet [Venkatesh et al., 2018] | ResNet-18 | 52.91 ± 0.88 | 68.88 ± 0.69 | 72.36 ± 0.90 | 83.64 ± 0.60 | 62.24 ± 0.89 | 77.92 ± 0.66 | / | / |
| ProtoNet [Snell et al., 2017] | ResNet-18 | 54.16 ± 0.12 | 70.68 ± 0.65 | 71.88 ± 0.91 | 80.72 ± 0.48 | 60.17 ± 0.90 | 80.58 ± 0.64 | / | / |
| RelationNet [Sung et al., 2018] | ResNet-18 | 52.48 ± 0.86 | 69.83 ± 0.68 | 67.59 ± 1.02 | 82.78 ± 0.58 | 55.72 ± 0.90 | 77.42 ± 0.68 | / | / |
| Baseline [Chen et al., 2019a] | ResNet-18 | 51.75 ± 0.80 | 74.27 ± 0.63 | 65.51 ± 0.87 | 82.85 ± 0.55 | 57.72 ± 0.88 | 79.06 ± 0.67 | / | / |
| Baseline++ [Chen et al., 2019a] | ResNet-18 | 51.87 ± 0.77 | 75.68 ± 0.63 | 67.02 ± 0.90 | 83.58 ± 0.54 | 56.72 ± 0.90 | 77.24 ± 0.67 | / | / |
| BlockAug [Chen et al., 2019b] | ResNet-18 | 58.80 ± 1.36 | 76.91 ± 0.72 | / | / | / | / | / | / |
| IDSMe-Net [Chen et al., 2019c] | ResNet-18 | 59.14 ± 0.76 | 76.63 ± 0.74 | / | / | / | / | / | / |
| DEiML [Zhou et al., 2018] | ResNet-50 | 58.49 ± 0.91 | 71.28 ± 0.69 | 66.95 ± 1.06 | 77.11 ± 0.78 | 62.25 ± 1.00 | 79.32 ± 0.63 | / | / |
| DualEnt [Chen et al., 2019d] | ResNet-18 | 58.12 ± 1.37 | 76.92 ± 0.69 | 69.61 ± 0.46 | 84.10 ± 0.35 | 63.77 ± 0.62 | 80.53 ± 0.46 | / | / |
| LEO [Rusu et al., 2019] | WRN28-10 | 61.76 ± 0.08 | 77.59 ± 0.12 | / | / | / | / | / | 86.35 ± 0.05 | 81.44 ± 0.09 |
| MetaOptNet [Lee et al., 2019] | ResNet-12 | 61.41 ± 0.61 | 77.88 ± 0.46 | / | / | / | / | / | 65.36 ± 0.71 | 81.34 ± 0.52 |

| Vanilla (ours) | ResNet-18 | 64.56 ± 0.81 | 78.18 ± 0.60 | 65.91 ± 0.88 | 83.43 ± 0.51 | 59.32 ± 0.88 | 79.95 ± 0.67 | 63.93 ± 0.89 | 84.66 ± 0.62 |
| CLR+Imprinting (ours) | ResNet-18 | 60.59 ± 0.86 | 78.26 ± 0.61 | 73.11 ± 0.92 | 87.46 ± 0.48 | 62.88 ± 0.97 | 81.16 ± 0.64 | 70.13 ± 0.98 | 85.04 ± 0.64 |
| CLR (ours) | ResNet-18 | 61.49 ± 0.83 | 81.00 ± 0.60 | 68.00 ± 0.92 | 84.26 ± 0.53 | 64.04 ± 0.92 | 83.05 ± 0.60 | 70.26 ± 1.00 | 86.78 ± 0.66 |

Table 1: 1-shot 5-way and 5-shot 5-way testing results on MiniImageNet, Caltech-256, CUB-200 datasets and TieredImageNet.

| Table 2: The performance of different number of replaced blocks on MiniImageNet. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Training | Fine tuning | 0 | 6 | 9 |
| 0 | 76.96 ± 0.60 | 78.11 ± 0.61 | 78.68 ± 0.61 | 79.16 ± 0.66 |
| 3 | 77.16 ± 0.64 | 79.23 ± 0.60 | 80.83 ± 0.60 | 80.14 ± 0.62 |
| 6 | 76.44 ± 0.61 | 79.26 ± 0.61 | 80.07 ± 0.62 | 79.23 ± 0.64 |
| 9 | 76.41 ± 0.60 | 78.93 ± 0.56 | 79.15 ± 0.60 | 78.8 ± 0.68 |

The effect of different components. To verify the effectiveness of our CLR strategy, some variants of CLR are evaluated and compared: Vanilla, One Time Local Replacement (OTLR), CLR without Pseudo Labeling (CLR w/o PL), CLR with Ground Truth Label (CLR w GT), Continual All Replacement (CAR). The OTLR selects unlabeled images to apply local replacement continually, whereas CLR w GT uses ground truth label to
select unlabeled images which can be seen as an oracle and the upper bound of our method. The CAR use the whole of unlabeled image to tune classifier instead of replacing local regions. This strategy can be seen as the original pseudo labeling method.

The evaluation results on MiniImageNet are illustrated in Fig. 5. The CLR achieves the best results. This verifies all components of our approach are important for the best performance. Besides, the CLR w/o PL performs comparably with Vanilla. It is better than Vanilla in 1-shot but slightly worse than Vanilla in 5-shot testing. This indicates the robustness of continual local replacement when wrong unlabeled images are selected to replace.

4.3 Why Continual Local Replacement Works

We conduct experiments to quantitatively and qualitatively explain why CLR works.

CLR introduces plentiful semantic information. On every fine-tune epoch, CLR makes the features of locally replaced images $d_{support}$ vary in their embedded space. This provides a larger semantic space for classifier tuning. Hence, a better classification decision boundary can be learned. As illustrated in Fig. 7, four fine-tune epochs are showed with T-SNE visualization [Maaten and Hinton, 2008]. The stars denote the features of $d_{support}$ and triangles represent $d_{support}$. We can see that the positions of triangles move dramatically inside the big red and green circles. Similar phenomenons can also be observed in the other three semantic clusters. In addition, the saliency maps are shown in Fig. 8. The local replacement can be either interpreted as new semantic information (e.g. top-left figure) or partial occlusions (e.g. top-right figure).

Robustness of CLR. Fig. 6 shows the typical accuracy change during the whole of fine-tune procedure. Since $d_{support}$ contains stable semantic information from original support set, the learning can converge quickly. Moreover, we also evaluate CLR on high way few-shot testing. The results are showed in Tab. 3. Our method achieves the best results. Note that the performance of CLR can be constantly improved with more replaced blocks. It still arrives the best trade-off when a maximum of 6 blocks replaced. This indicates that (1) our approach still works well even though the overall accuracy is relatively low on high way setting. (2) Since 6 replaced blocks can always achieve the best results, the number of replaced blocks hyper-parameter is insensitive to different datasets and experimental settings.

5 Conclusion

In this paper, we propose a continual local replacement method for few-shot image recognition. It leverages the content of unlabeled images to synthesize new images for training. To introduce more useful semantic information, a pseudo labeling strategy is applied during the fine-tune stage. It continually selects semantically similar images to locally replace labeled ones. Our strategy is simple yet effective on few-shot image recognition. Extensive experiments show that it can significantly enlarge the capacity of semantic information and achieve new state-of-the-art results.

References

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

[Cai et al., 2018] Qi Cai, Yingwei Pan, Ting Yao, Chengan Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4080–4088, 2018.

[Chen et al., 2019a] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In International Conference on Learning Representations, 2019.
Figure 7: The T-SNE visualization of CLR features on four consecutive 5-shot 5-way fine-tune epochs. The embedded features of labeled images $d_{support}$ (marked as stars) are fixed. But the features of local replaced images $d_{support}$ (marked as triangles) can change on every fine-tune epoch. It provides a larger semantic space for classifier tuning and thus a better classification decision boundary can be learned. Best viewed in color with zoom.

Figure 8: The saliency maps on labeled, unlabeled and synthesized images. The model can learn new semantic information from the regions of unlabeled image. Best viewed in color with zoom.

[Chen et al., 2019b] Zitian Chen, Yanwei Fu, Kaiyu Chen, and Yu-Gang Jiang. Image block augmentation for one-shot learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3379–3386, 2019.

[Chen et al., 2019c] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8680–8689, 2019.

[Chen et al., 2019d] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. IEEE Transactions on Image Processing, 2019.

[Fei-Fei et al., 2006] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. IEEE Transactions on pattern analysis and machine intelligence, 28(4):594–611, 2006.

[Finn et al., 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1126–1135. JMLR. org, 2017.

[Garcia and Bruna, 2018] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In International Conference on Learning Representations, 2018.

[Grandvalet and Bengio, 2005] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In Advances in neural information processing systems, pages 529–536, 2005.

[Griffin et al., 2007] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[Hariharan and Girshick, 2017] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In Proceedings of the IEEE International Conference on Computer Vision, pages 3018–3027, 2017.

[He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[Koch et al., 2015] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, volume 2, 2015.

[Kornblith et al., 2019] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2661–2671, 2019.

[Krizhevsky et al., 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[Lake et al., 2011] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In Proceedings of the annual meeting of the cognitive science society, volume 33, 2011.

[Lee et al., 2019] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10657–10665, 2019.

[Lee, 2013] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural
networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.

[Liu et al., 2019] Yanbin Liu, Juho Lee, Minseop Park, Sae-hoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. 2019.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Miyato et al., 2018] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[Qi et al., 2018] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5822–5830, 2018.

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

[Rusu et al., 2019] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.

[Saijadi et al., 2016] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171, 2016.

[Salakhutdinov et al., 2012] Ruslan Salakhutdinov, Joshua Tenenbaum, and Antonio Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 195–206, 2012.

[Santoro et al., 2016] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016.

[Snell et al., 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[Sung et al., 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[Vinyals et al., 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[Wah et al., 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wang et al., 2017] Peng Wang, Lingqiao Liu, Chunhua Shen, Zi Huang, Anton van den Hengel, and Heng Tao Shen. Multi-attention network for one shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2721–2729, 2017.

[Wang et al., 2018] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7278–7286, 2018.

[Zhong et al., 2020] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[Zhou et al., 2018] Fengwei Zhou, Bin Wu, and Zhenguo Li. Deep meta-learning: Learning to learn in the concept space. *arXiv preprint arXiv:1802.03596*, 2018.