



ProtoSimi: label correction for fine-grained visual categorization

Jialiang Shen¹ · Yu Yao¹ · Shaoli Huang¹ · Zhiyong Wang¹ · Jing Zhang¹ · Ruxing Wang² · Jun Yu³ · Tongliang Liu¹

Received: 14 August 2022 / Revised: 14 August 2022 / Accepted: 29 January 2023 /
Published online: 28 March 2023
© The Author(s) 2023

Abstract

Deep models trained by using clean data have achieved tremendous success in fine-grained image classification. Yet, they generally suffer from significant performance degradation when encountering noisy labels. Existing approaches to handle label noise, though proved to be effective for generic object recognition, usually fail on fine-grained data. The reason is that, on fine-grained data, the category difference is subtle and the training sample size is small. Then deep models could easily overfit the noisy labels. To improve the robustness of deep models on noisy data for fine-grained visual categorization, in this paper, we propose a novel learning framework named ProtoSimi. Our method employs an adaptive label correction strategy, ensuring effective learning on limited data. Specifically, our approach considers the criteria of exploring the effectiveness of both global class-prototype and part class-prototype similarities in identifying and correcting labels of samples. We evaluate our method on three standard benchmarks of fine-grained recognition. Experimental results show that our method outperforms the existing label noisy methods by a large margin. In ablation studies, we also verify that our method is non-sensitive to hyper-parameters selection and can be integrated with other FGVC methods to increase the generalization performance.

Keywords Label noise · Fine-grained visual categorization · Deep learning

Jialiang Shen, Yu Yao and Tongliang Liu have contributed equally to this work.

Editors: Yu-Feng Li, Prateek Jain.

✉ Jialiang Shen
jshe9143@uni.sydney.edu.au

✉ Yu Yao
yyao0814@uni.sydney.edu.au

✉ Tongliang Liu
tongliang.liu@sydney.edu.au

Extended author information available on the last page of the article

1 Introduction

Fine-grained visual categorization (FGVC) aims at recognizing subordinate categories such as species of animals (Wah et al., 2011) and models of aircraft (Maji et al., 2013). Since the category difference is subtle, data labeling usually requires expert-level knowledge, without which, labeling error easily appears. Although existing deep ConvNets have made great improvements of performance in FGVC, this success is limited to cases of training with clean data. When training data contain label noise, these models can have poor generalization ability because deep ConvNets try to fit all data points during training, including noisy labels. This issue is also discussed by Arpit (Arpit et al., 2017) for deep models.

To against label noise, existing mainstream approaches use model predictions to extract confident examples (whose labels are likely to be correct) (Han et al., 2018; Yu et al., 2019; Li et al., 2020) by exploiting the memorization effect (Arpit et al., 2017; Wei et al., 2020). These methods are demonstrated effective in coarse-grained object recognition, but it is hard to employ them for fine-grained recognition. It is because that modern deep neural networks used for fine-grained tasks are usually trained via fine-tuning. The parameters of the final classification layer are trained on the target dataset from scratch, and they are much more sensitive to label noise, while their former counterparts are quite robust. (Bai et al., 2021). As illustrated in Fig. 1, we visualize representations before and after the final linear layer by t-SNE (Van der Maaten & Hinton, 2008). It shows that, by comparing with the clusters of the representations after the final linear layers shown in Fig. 1a, c, the clusters of the representations after the final linear layers shown in Fig. 1b, d are more noisy.

Furthermore, existing methods merely apply global-level information in images to deal with label noise. The part-level features which characterize the critical discriminative information in recognizing fine-grained objects are ignored. Last but not least, fine-grained datasets are usually small in scale. Existing approaches to extracting confident examples reduce the number of training samples, further exacerbating the data scarcity and overfitting during model training. As a result, existing methods to extract confident examples are therefore unreliable.

To tackle the label noise problem on the FGVC task, in this paper, we propose a novel class-prototype similarity-based learning framework. Specifically, during the training phase, training data is divided into two categories: confident examples and unconfident examples (whose labels are likely to be incorrect). To select confident examples and to correct unconfident examples, we exploit the presentations extracted before the final linear layer (feature-level information) rather than model predictions. Specifically, *global-class*

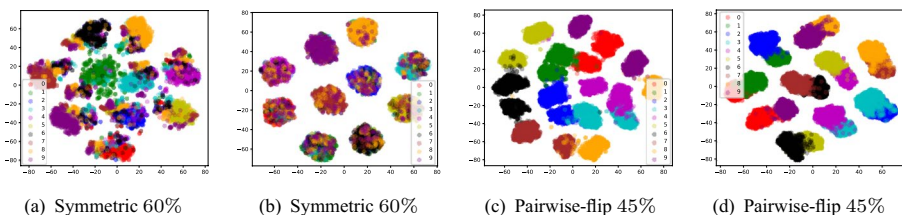


Fig. 1 Figures are generated by using t-SNE visualisation. ResNet-18 pretrained on ImageNet is fine-tuned on noisy FashionMNIST dataset. The dataset is split to training set and validation set. The validation set is used for model selection. The different colors represent different clean labels. **a, c** Visualise the representations before the final linear layer, respectively. **b, d** Visualise the representations after the final linear layer, respectively

and *part-class prototypes* calculated by averaging the global and part features of training examples from the same class are employed as references. The confident examples are identified with high similarities to their corresponding class prototypes; while unconfident examples are corrected with new possible clean labels according to their similarity to class prototypes, which address the data scarcity issue.

We extensively evaluate the proposed method on fine-grained datasets under different settings of label noise, i.e., symmetric and pairwise-flip label noise with various levels of noise rates. The experimental results show that our approach outperforms all the baselines, making significant improvements in generalization performance. We also show that our selection process of confident examples is non-sensitive to hyper-parameters. These results validate the robustness and the reliability of the proposed method. Moreover, our learning framework is flexible and can be regarded as a general component in the FGVC to tackle the label-noise issue. Our contribution is summarized as follows:

- By exploiting feature-level information to identify confident examples in the noisy fine-grained dataset, we propose a novel prototype similarity-based approach named Proto-Simi. By considering characteristics of fine-tuned models, the method largely increases the reliability of identifying confident examples and alleviates the impacts of memorization effects in ConvNets.
- As a response to data scarcity in fine-grained problems, we explore the feasibility of revising incorrect labels by jointly referencing global and part class-prototypes similarities, performing in a semi-supervised learning style.
- We establish the first benchmark of fine-grained classification with close-set synthesised label noise on three popular datasets. Our approach obviously improves the generalization performance by gradually increasing the clean-label ratio of training datasets via the proposed label-correction strategy.

The rest of this paper is organized as follows. In Sect. 2, we briefly review the current popular label-noise algorithms and the FGVC algorithms. Section 3 introduces the proposed approach in detail, which is then experimentally compared with other baselines as discussed in Sect. 4. Section 5 presents the ablation study on the CUB-200-2011 dataset to comprehensively analyse the principle of our proposed method and the hyper-parameters settings. Section 6 concludes the current work.

2 Related work

In this section, we briefly review the related literature with fine-grained visual categorization and label noise.

2.1 Fine-grained visual categorization

Fine-grained recognition has been an active research topic in recent years. This task is difficult as the critical information to distinguish categories usually hides in subtle part regions. To meet this challenge, the early literature includes a series of part-based pipelines that localize the object parts based on strongly supervised learning (Huang et al., 2016).

Considering the expensive cost of acquiring part annotations, some techniques (Zhang et al., 2016; Zheng et al., 2017; Sun et al., 2018; Zheng et al., 2019) attempt to learn part

features in a weakly-supervised manner. For example, Zhang et al. (2016) first picked distinctive filters which were then used to learn part detectors through an iteratively alternating strategy. MA-CNN (Zheng et al., 2017) obtained part regions by clustering feature maps of intermediate convolutional layers. Yet, the performance of these methods remains limited because of the insufficient training data for this task. Recent works (Chen et al., 2019; Zhuang et al., 2020) try to ease this limitation by consolidating the data enhancement mechanisms into their learning frameworks. For instance, Zhang et al. (2019) progressively cropped out the discriminative regions to generate diversified data sets for training network experts. Chen et al. (2019) destructed the images into part regions and designed a region alignment network to restore the original spatial layout of the part regions. Huang et al. (2022) designed a few-shot based algorithm by employing the intra-class variance and inter-class fluctuations.

Although these methods have achieved promising performance on fine-grained tasks, the label-noise issue existing in real applications can easily degrade the model's effectiveness. In this work, we propose a new learning framework that optimizes the learning of deep modes on noisy fine-grained data.

2.2 Label noise

The research on label noise in various computer vision tasks has been promoted by deep learning-based approaches. However, these approaches generally suffer from the overfitting issue on noisy examples due to the memorization effects of ConvNets. This issue could be relieved by dropouts, data-augmentation (Zhang et al., 2017), regularization paradigms (Liu et al., 2020; Xia et al., 2021), specially designed loss functions Huang et al. (2020), novel optimizer (Foret et al., 2020), loss correction (Patrini et al., 2017; Song et al., 2019), etc.

The sample selection strategy is developed to discard all unconfident examples and only train a model on the remained confident examples (Malach & Shalev-Shwartz, 2017; Jiang et al., 2018; Han et al., 2018; Yu et al., 2019; Liu et al., 2021). These methods reduce the number of training samples unfortunately aggravates the data scarcity problem for FGVC with label noise. Zhang et al. (2020) and Sun et al. (2021) use a sample selection strategy which discards unconfident examples on web fine-grained datasets and treats the remaining as confident examples to help generalization ability of a learning model. The label correction methods (Li et al., 2020) are developed to efficiently mine the information from abandoned some unconfident examples and then give pseudo labels to them, which is a complement strategy of sample selection. Gong et al. (2023) tackles the noisy labels in a class-wise way to ease the entire noise correction task. However, all mentioned methods are generally based on the small-loss criterion, the effectiveness directly depends on the reliability of the model predictions, whereas the model predictions are not reliable in the FGVC task. Han (Han et al., 2019) firstly adopted a class-prototype concept to estimate pseudo labels. However, the method is not specifically designed for FGVC, i.e., it uses multiple examples as prototypes in each class to capture geometry information. CleanNet (Lee et al., 2018) also adopted a class-prototype concept, but it requires an extra clean sample to produce prototypes, which increases labeling costs.

Statistic consistency can also be exploited to alleviate the label noise issue, which aims to infer the clean class posterior distribution by using the noisy class posterior and a label noise transition matrix (Liu & Tao, 2015; Yao et al., 2020; Li et al., 2021). However, this

group of methods suffers from the magnified estimation error of the transition matrix, especially when the number of classes increases.

Several works are designed to relieve label noise problem by leveraging self-learning techniques such as contrastive learning that could adaptively prevent the noisy information from dominating the network optimization. (Yang et al., 2022b, 2021, 2022a)

However, all the existing methods do not consider part-level features which usually characterize the critical discriminative information in recognizing fine-grained objects and are important for improving generalization performance.

3 Our approach

In this section, we propose a novel prototype similarity-based label-correction (ProtoSim) framework which employs both global-level and part-level features for FGVC. Our algorithm can be divided into two stage. At the first stage, the global-level and part-level features are extracted by exploiting feature maps trained on a target fine-grained dataset. Then a self-adaptive sample selection and label-correction strategy is proposed to increase the quality of training data. An illustration of the learning framework is given in Fig. 2, and the algorithm description is given in Algorithm 1.

3.1 Global-level and part-level feature extraction

To generate global-level features, feature maps learned from a target fine-grained dataset are employed. Specifically, let $D = \{(x_i, y_i)\}_{i=1}^n$ be the target fine-grained dataset. Let \mathcal{G}_Φ be an extractor for feature maps fine-tuned on a target noisy dataset. Let $\mathcal{G}_\Phi(x) \in \mathbb{R}^{k \times h \times w}$ be the feature maps of an instance x , which are outputs of the last convolutional block, where k, h, w are the number of channels, height and width of feature maps, respectively. By applying global average pooling (GAP) operation on feature maps, a feature vector Φ_x is obtained, which we called a global-level feature of the instance x .

To further improve the robustness of our algorithm, the part-level features are also important and can not be ignored. Specifically, because the intra-class variation is subtle for fine-grained datasets, the labels are usually determined by small regions of images. Thereby, to increase the diversity of features and their discrimination, we introduce part-level features which are extracted from discriminative regions of the input image. The

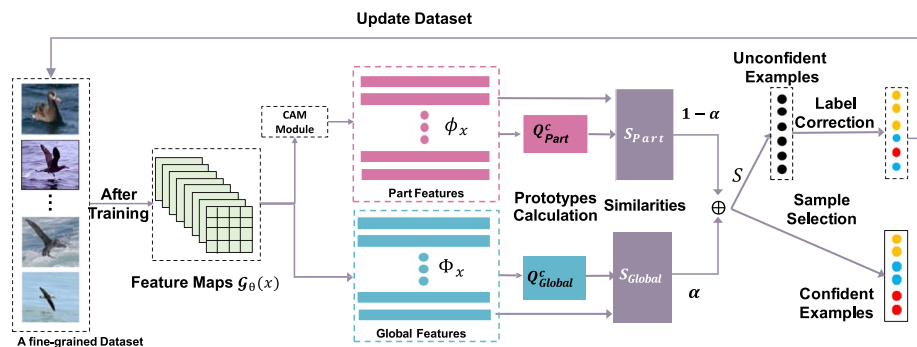


Fig. 2 The overview of ProtoSim training paradigm

part-level features can be consider as auxiliary semantic information for the global-level features $\mathcal{G}_\Phi(\mathbf{x})$.

To generate part-level features, class activation mapping (CAM) method (Zhou et al., 2016) is employed. The workflow is illustrated in Fig. 3. Specifically, let $\mathcal{G}_\Phi(\mathbf{x})^j$ be the j -th feature map, and $\mathbf{w}_y \in \mathbb{R}^d$ is the classifier weight vector corresponding to a label y . Then we can obtain the class activation map $CAM(\mathbf{x})$ as:

$$CAM(\mathbf{x}) = \sum_{j=0}^k \mathbf{w}_y^j \mathcal{G}_\Phi(\mathbf{x})^j.$$

Here, we ignore the bias term which could be contained in the weight vector and by including additional feature maps with all ones to $\mathcal{G}_\Phi(\mathbf{x})$. After normalizing $CAM(\mathbf{x})$, the result can be seen as an attention map that describes which position is focused by the model with respect to a particular class. We then apply element-wise multiplication between the normalized $CAM(\mathbf{x})$ and each feature map to extract the classification-aware discriminative regional feature maps, which is followed by a GAP operation. In this way, the final part-level features are dominant to the class characteristics, and therefore can be used for validating whether the label is correct.

3.2 Self-adaptive sample selection and label-correction strategy

We propose a class-prototype similarity-based method which exploits the feature-level information in sample selection. In a nutshell, after fine-tuning on the dataset D , we measure the similarities of all examples to their class prototypes. The details of our sample selection method are explained in the following paragraph.

To obtain c -th global class-prototype Q_{global}^c , we average all the global-level features in c -th class, i.e.,

$$Q_{global}^c = \frac{1}{n} \sum_{x,y=c} \Phi_x,$$

where n is the number of examples in c -th class. To obtain global similarity $S_{global}(\mathbf{x}, y)$ between a example (\mathbf{x}, y) and its global class-prototype Q_{global}^y , cosine similarity is employed, i.e.,

$$S_{global}(\mathbf{x}, y) = \frac{\Phi_x \cdot Q_{global}^y}{\|\Phi_x\|_2 \cdot \|Q_{global}^y\|_2}.$$

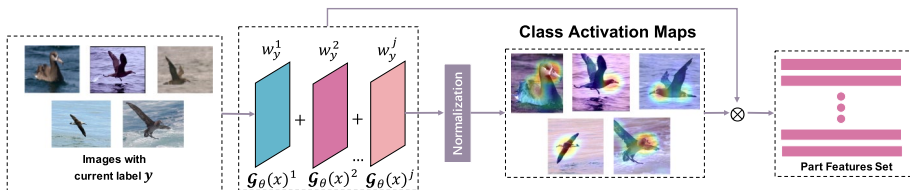


Fig. 3 The process of generating part-level feature by CAM

The part class prototype Q_{part}^c and the part similarity $S_{part}(x, y)$ are calculated in the same fashion as the global class prototype and the global similarity, respectively.

The $(1 - \epsilon)$ examples in each class with the highest similarities are selected as confident examples, where ϵ is the noise rate of the target dataset and can be estimated from noisy data (Liu & Tao, 2015; Yao et al., 2020). To take both part-level features and global-level features into consideration for sample selection, we propose a score $S(x, y)$ that linearly combines part similarity S_{part}^y and global similarity S_{global}^y , i.e.,

$$\begin{aligned} S(x, y) &= (1 - \alpha)S_{part}(x, y) + \alpha S_{global}(x, y) \\ &= (1 - \alpha) \frac{\phi_x \cdot Q_{part}^y}{\|\phi_x\|_2 \cdot \|Q_{part}^y\|_2} + \alpha \frac{\Phi_x \cdot Q_{global}^y}{\|\Phi_x\|_2 \cdot \|Q_{global}^y\|_2}, \end{aligned} \quad (1)$$

where $\alpha \in [0, 1]$ is a hyper-parameter to balance the contributions between part similarity and global similarity.

In fine-grained datasets, the number of examples in each class is usually limited. During the phase of sample selection, a large number of unconfident examples could be discarded, which further reduces the number of training examples. This will significantly reduce the generalization performance of the data-hungry Modern ConvNets. Therefore, the problem that how to sufficiently use unconfident examples to improve the generalization performance is essential in the FGVC task. To solve this problem, we propose a self-adaptive label-correction strategy that performs in a semi-supervised manner.

Specifically, after the global and part feature-based sample selection, unconfident examples can be obtained, which are regarded as unlabeled data. Then we approximate their clean labels based on the score S , i.e., given an instance x , the class with the highest score is the approximated clean label y of the instance. Formally,

$$\hat{y} = \arg \max_{i \in \{1, \dots, C\}} S(x, i), \quad (2)$$

where C is the total number of classes. In such a way, a new dataset is generated by combining all the confident and unconfident examples and is used to train a new model. The new model then is employed to select new confident examples and correct the label of new unconfident examples.

$$h_\theta^* = \arg \min_{h_\theta} \frac{1}{\|D_l\|} \sum_{(x,y) \in D_l} \ell_{ce}(h_\theta, x, y) + \frac{1}{\|D_u\|} \sum_{(x,\hat{y}) \in D_u} \ell_{ce}(h_\theta, x, \hat{y}), \quad (3)$$

where ℓ_{ce} is the cross-entropy loss function, and h_θ is a learning model parameterized by a learnable parameter θ . For the set of confident examples D_l , the original labels are employed to train the model; For the set of unconfident examples D_u , the approximated clean labels are employed to train the model.

With the help of sample selection and label correction strategy, the trained model will receive correct label information from refined training data, thus the latter learnt feature representation may further increase the precision of sample selection and label correction strategy in next round. Therefore, we repeat this strategy every interval epochs.

Furthermore, the extra computational cost for calculating feature similarities is not expensive, which is linear. Specifically, for c classes and n examples, the time complexity to get the both local and global class-prototype is $O(n)$. The time complexity of

calculating the local and global similarity between n samples and class-prototypes is $O(2c * n)$. The total time complexity is $O((2c + 1)n)$.

Algorithm 1 ProtoSimi

Input: Training dataset $D = \{(x_i, y_i)\}_{i=1}^n$, noise rate ϵ , class label set C , λ stride of label correction.

Output: Optimized classifier f .

```

1: while Current epoch  $T$  is smaller than total epochs do
2:   Update parameters  $\{\theta, \psi\}$  for the classifier  $f = v_\psi GAP(\mathcal{G}_{\bar{\Phi}})$  by dataset  $D$ .
3:   if  $T \bmod \lambda \neq 0$  and  $T \neq 1$  then
4:     Extract global features and part features.
5:     Calculate global class prototype  $Q_{global}^c$  and part class prototype  $Q_{part}^c$ .
6:     Get scores  $S(x, y)$  of all examples in  $D$  by Eq. 1.
7:     Pick first  $1 - \epsilon$  portion of examples from descent sorted scores list of each
      class as confident examples. The remains are unconfident examples.
8:     Update labels of unconfident examples by referencing score  $S(x, c)$ 
      (Eq. 1) and Eq. 2 for each class  $c$ .
9:     Update Dataset  $D$  by union confident examples with relabeled unconfi-
      dent examples via (3).
10:   end if
11: end while

```

4 Experiments

In this section, we illustrate the generalization ability of our methods on different datasets with different types of noise.

4.1 Experimental settings

We verify the effectiveness of our approach on three benchmark datasets, i.e., fine-grained benchmarks including CUB-200-2011 (Wah et al., 2011), Stanford-Cars (Krause et al., 2013), and FGVC-aircraft (Maji et al., 2013). We conduct experiments on the widely used noise types. Specifically, we use symmetric noise (Han et al., 2018) with noise rates 20%, 40%, and 60%; pairwise-flip noise with noise rates 20% and 40%. Since conventional label noise methods are applied without bounding boxes or part annotations, this external auxiliary information is excluded in our experiments. We compare our method with the fine-grained approaches that only need class labels. For fairness of comparison, we re-implement all the competitors by fine tuning. The training phase employs regular data augmentation in the fine-grained settings. The input images are first resized into 512×512 and then randomly cropped to 448×448 , followed by random horizontal flipping. All experiments are repeated 3 times.

4.2 Implementation details

ResNet-50 (He et al., 2016) is used as the backbone network, and the GAP output of the last convolutional block is a 2048-dimensional feature vector. The learning rate is initialized as

0.01 for the classification layer and 0.001 for all other layers, and then it decayed by 0.5 in every 30 epochs. The SGD optimizer with momentum 0.9 is utilized to optimize the model. The model is trained for 50 epochs with a fixed batch size of 16, and then tested on clean data to compute the top-1 classification accuracy. The ϵ is the noise rate which can be estimated from the noisy data (Liu & Tao, 2015). The weight α for balancing the global and part similarity scores is set to 0.7 in all datasets. The stride λ is set to 15 for the CAR dataset with 60% symmetric noise and 5 for all other experiments. The way of perturbing clean labels remains the same as Han et al. (2018).

4.3 Baselines

To make a sufficient comparison, we adopt 12 different methods as the competitors including both label noise and fine-grained approaches. Specially, Cross-Entropy (He et al., 2016) follows basic ResNet-50 training process. Mixup (Zhang et al., 2017), Early Learning Regularization (ELR) (Liu et al., 2020), Self-Adaptive Training with Cross-Entropy (SAT-CE) (Huang et al., 2020) and Sharpness-Aware Minimization (SAM) (Foret et al., 2020) are proposed to alleviate overfitting effects. Co-teaching (Han et al., 2018) and Co-teaching⁺ (Yu et al., 2019) are two classical sample selection methods with small-loss criterion. Especially, recent work Advanced-Softly-Update-Drop (ASUD) (Liu et al., 2021), they propose a model prediction based sample selection method for fine-grained classification against label noise. Re-weighting (Liu & Tao, 2015) is an important statistically consistent method. DCL (Chen et al., 2019), MGE-CNN (Zhang et al., 2019) and API-Net (Zhuang et al., 2020) are representative SOTA part-based weakly-supervised methods in FGVC.

4.4 Classification accuracy

The results are illustrated in Table 1. ProtoSimi outperforms all the baselines for most of the experiments and achieves a large margin improvement in generalization performance. Specifically, the baseline methods Mixup, ELR, SAT-CE, and SAM have limited improvements in the fine-grained task. The performance of ASUD, Co-teaching and Co-teaching⁺ also degenerate significantly in the case of high noise rate. Compared with the part-based fine-grained methods DCL, MGE-CNN and API-Net, we have a clear advantage when the noise rate is large.

5 Ablation study on CUB dataset

5.1 Small-loss and feature based sample-selection criteria

The sample-selection criterion based on sample selection is a popular way to select confident examples (Li et al., 2021), it considers that the examples having small losses are more likely to be clean examples in model training, and therefore, they should be selected as confident examples. By contrast, our method employs both global-level and part-level features to select confident examples. In Fig. 4, we compare the clean ratio obtained by utilizing small-loss and feature-based sample-selection methods for different types of label

Table 1 Average test accuracy over the last three epochs on CUB-200-2011, Stanford-CAR and Stanford-Aircraft dataset

Datasets	Methods	Symmetric noise			Pairwise-flip noise		
		40%			20%		
		20%	60%	40%	20%	60%	40%
CUB	Cross-entropy	74.81 ± 0.98	61.07 ± 1.17	40.50 ± 2.08	73.75 ± 0.53	52.86 ± 0.62	
	Mixup (Zhang et al., 2017)	77.84 ± 0.53	62.93 ± 1.06	41.87 ± 1.34	75.69 ± 0.97	54.87 ± 0.07	
	ELR (Liu et al., 2020)	73.71 ± 0.40	58.38 ± 1.49	38.20 ± 1.18	72.09 ± 0.38	51.44 ± 0.26	
	SAT-CE (Huang et al., 2020)	77.07 ± 0.21	66.20 ± 0.67	51.47 ± 0.57	74.74 ± 0.31	53.22 ± 1.30	
	SAM (Foret et al., 2020)	76.48 ± 0.43	62.47 ± 0.30	41.62 ± 0.81	74.33 ± 0.40	52.59 ± 1.28	
	Co-teaching (Han et al., 2018)	76.18 ± 1.24	59.51 ± 0.40	32.22 ± 2.05	69.49 ± 1.36	45.44 ± 1.66	
	Co-teaching* (Yu et al., 2019)	62.21 ± 0.79	51.24 ± 0.98	29.73 ± 1.03	54.51 ± 0.06	39.88 ± 0.01	
	Re-weighting (Liu & Tao, 2015)	72.00 ± 1.55	54.32 ± 3.98	33.94 ± 3.13	72.20 ± 2.60	56.00 ± 1.65	
	DCL (Chen et al., 2019)	78.15 ± 0.92	60.38 ± 2.09	39.99 ± 1.59	76.92 ± 1.80	54.21 ± 1.29	
	MGE-CNN (Zhang et al., 2019)	79.60 ± 3.59	66.07 ± 4.78	43.36 ± 1.63	78.20 ± 0.84	56.16 ± 1.43	
	API-Net Zhuang et al. (2020)	61.01 ± 7.14	47.00 ± 2.92	27.55 ± 2.05	66.50 ± 1.81	46.24 ± 1.70	
	ASUD (Liu et al., 2021)	71.11 ± 2.03	57.49 ± 1.35	37.33 ± 0.93	68.60 ± 1.57	50.55 ± 0.19	
	ProtoSimi	81.95 ± 0.10*	76.56 ± 0.06*	63.46 ± 0.74*	79.97 ± 0.32*	64.75 ± 2.21*	
	<i>p</i> value	<i>p</i> = 0.007	<i>p</i> = 0.003	<i>p</i> = 0.001	<i>p</i> = 0.001	<i>p</i> = 0.0	
	CAR	Cross-entropy	80.35 ± 0.28	62.40 ± 0.72	36.54 ± 2.29	78.61 ± 0.67	54.00 ± 0.10
Mixup (Zhang et al., 2017)		80.19 ± 0.97	65.38 ± 1.08	39.97 ± 2.07	80.47 ± 0.71	56.89 ± 0.68	
ELR (Liu et al., 2020)		81.02 ± 0.16	63.05 ± 1.01	35.44 ± 1.67	79.83 ± 0/06	55.21 ± 0.89	
SAT-CE (Huang et al., 2020)		82.51 ± 0.13	69.66 ± 0.10	44.70 ± 5.90	80.60 ± 0.12	58.87 ± 1.85	
SAM (Foret et al., 2020)		81.72 ± 0.18	66.59 ± 0.76	7.11 ± 3.41	79.82 ± 0.28	57.55 ± 1.25	
Co-teaching (Han et al., 2018)		78.77 ± 2.75	62.60 ± 0.57	32.78 ± 0.77	71.23 ± 3.21	48.00 ± 2.17	
Co-teaching* (Yu et al., 2019)		30.14 ± 0.74	40.52 ± 0.75	28.08 ± 0.67	32.33 ± 0.64	32.21 ± 0.37	
Re-weighting (Liu & Tao, 2015)		71.60 ± 3.00	51.88 ± 2.56	28.45 ± 0.67	73.51 ± 1.68	49.35 ± 3.17	
DCL (Chen et al., 2019)	82.89 ± 0.84	64.44 ± 1.56	36.76 ± 1.64	80.47 ± 0.19	56.21 ± 0.38		
MGE-CNN (Zhang et al., 2019)	83.77 ± 1.49	68.63 ± 0.93	40.79 ± 6.59	81.16 ± 1.13	58.40 ± 2.09		
API-Net (Zhuang et al., 2020)	66.62 ± 9.53	48.94 ± 12.93	25.65 ± 13.82	75.16 ± 2.31	52.19 ± 2.29		

Table 1 (continued)

Datasets	Methods	Symmetric noise			Pairwise-flip noise		
		40%			20%		
		20%	60%	40%	20%	60%	40%
AIR	ASUD (Liu et al., 2021)	79.66 ± 0.69	63.98 ± 1.24	77.19 ± 1.14	56.53 ± 0.75		
	ProtoSimi	83.87 ± 0.17	69.10 ± 2.10	84.03 ± 0.30*	66.00 ± 1.65*		
	<i>p</i> value	<i>p</i> = 0.08	<i>p</i> = 0.29	<i>p</i> = 0.001	<i>p</i> = 0.0		
	Cross-entropy	75.65 ± 0.61	59.23 ± 0.56	73.25 ± 0.64	53.83 ± 1.52		
	Mixup(Zhang et al., 2017)	76.80 ± 0.80	62.29 ± 0.46	75.93 ± 0.59	55.04 ± 1.17		
	ELR (Liu et al., 2020)	76.97 ± 0.37	59.68 ± 1.01	75.89 ± 0.11	53.41 ± 1.31		
	SAT-CE (Huang et al., 2020)	78.33 ± 0.48	68.04 ± 1.29	76.01 ± 0.89	56.44 ± 0.83		
	SAM (Foret et al., 2020)	79.64 ± 0.35	69.44 ± 0.57	78.29 ± 1.06	57.65 ± 0.52		
	Co-teaching (Han et al., 2018)	79.51 ± 0.87	63.94 ± 1.81	70.81 ± 0.76	48.00 ± 2.17		
	Co-teaching* (Yu et al., 2019)	60.27 ± 4.94	53.11 ± 1.17	55.91 ± 3.05	37.95 ± 1.01		
	Re-weighting (Liu & Tao, 2015)	70.15 ± 0.24	53.97 ± 2.86	72.40 ± 0.39	53.25 ± 0.27		
	DCL (Chen et al., 2019)	77.38 ± 0.57	59.02 ± 0.74	76.11 ± 0.46	55.44 ± 0.41		
	MGE-CNN (Zhang et al., 2019)	80.89 ± 1.57	65.85 ± 1.38	78.97 ± 0.75	57.11 ± 1.49		
	API-Net (Zhuang et al., 2020)	67.41 ± 3.21	48.91 ± 2.26	68.71 ± 2.17	50.89 ± 0.45		
	ASUD (Liu et al., 2021)	75.81 ± 0.77	62.83 ± 0.35	74.97 ± 0.97	55.82 ± 0.74		
ProtoSimi	81.17 ± 0.53	73.07 ± 0.68*	79.20 ± 1.09 *	66.14 ± 1.10*			
<i>p</i> value	<i>p</i> = 0.104	<i>p</i> = 0.001	<i>p</i> = 0.002	<i>p</i> = 0.0			

All the best experimental results are marked in bold

We employ the one-sided Wilcoxon signed rank test by comparing the accuracy of ProtoSimi and the accuracy of the best baseline method (marked in italics in each column) in each setting. The proposed method provides significantly more accurate than most of baseline with the 0.05 significant level

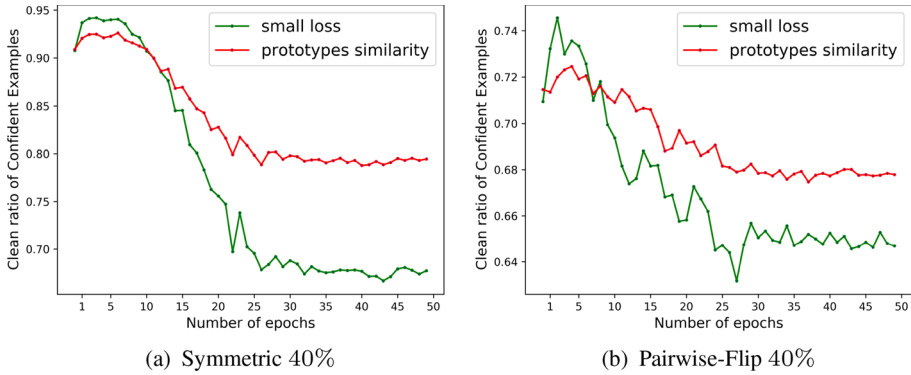


Fig. 4 ResNet-50 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) is fine-tuned on synthetic-noise CUB-200-2011 dataset (Wah et al., 2011). After each training epoch, we rank the similarity scores in descending order and the loss values in ascending order. We choose the first 60% examples in each class, and then calculate the clean ratio of all the selected examples

noise. It shows that, by utilizing the feature-level information, the clean ratio of selected examples is higher than the sample-selection method based on sample selection.

5.2 Investigation on class-prototype similarity

We investigate that how global and part similarities contribute to the model performance. Specifically, we compare three different settings, including only part similarity, only global similarity, and the combined class-prototype similarity with balancing weight $\alpha = 0.7$. Table 2 shows that the combined class-prototype similarity performs the best among three settings. Furthermore, both only global similarity and only part similarity achieve better performance than training with naive

5.3 Impacts on label correction stride λ

We compare the performance of different λ values, as illustrated in Fig. 5. The experiments are conducted in the case of symmetric noise with 40% noise rate. Figure 5a, b show the clean ratio of training examples and test accuracy for each epoch, respectively. The best test accuracy can be observed when $\lambda = 5$. The highest clean ratio in the training set is achieved when $\lambda = 3$, while $\lambda = 5$ has a similar clean ratio. Both metrics degenerate severely when $\lambda = 10$, and the model tends to overfit noisy labels.

Table 2 Generalization performance comparison between different class-prototype similarity modes

Methods	Symmetric noise			Pairwise-flip noise	
	20%	40%	60%	20%	40%
Only global prototype	81.28	74.99	60.65	79.29	61.34
Only part prototype	80.53	74.87	59.53	79.02	62.65
Combined	81.96	76.49	63.07	80.33	63.05

All the best experimental results are marked in bold

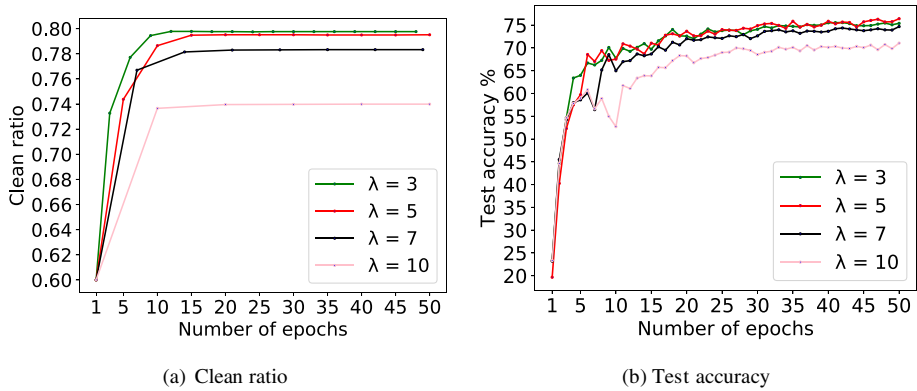


Fig. 5 The comparison of different λ , **a** is clean ratio of confident examples, and **b** is the test accuracy of each epoch

5.4 Impacts on biased noise rate ϵ

In previous experiments, we assume the noise rate ϵ is known. However, the actual noise rate is always biased in real applications, and the estimation error may be noticeable. In this section, we investigate that how a biased noise rate affects the model performance. The actual symmetric noise rate is 40%. We manually add bias to the actual noise rate, which yields a set of biased noise rates {20%, 30%, 35%, 45%, 50%, 60%}. Figure 6 shows the influence of biased noise rate on the generalization ability of our method is not large. It indicates that our approach is not sensitive to the noise rate estimation, and therefore, the class-prototype similarity criteria is robust to label noise.

5.5 Significant testing on weight α

In Table 3, we use the one-sided Wilcoxon signed rank test to check whether using a combination of part-level and global-level features has significant performance gain. The opponent is using global-level features along, where results have been shown in Table 2. The smaller p value indicates the result is more significant. We underline the p values which are

Fig. 6 The comparison of noise rate ϵ with estimation error

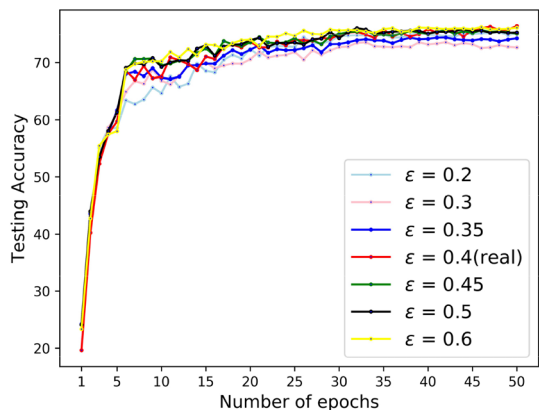


Table 3 One-sided Wilcoxon signed significant tests on CUB dataset, which shows that our method is robust with different α values

noise type	0.0	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Sym-40%	74.87%	76.29%	76.01%	76.19%	76.16%	76.49%	75.92%	75.74%
	$p = 1.0$	$p = 0.21$	$p = 0.21$	$p = 0.03$	$p = 0.03$	$p = 0.03$	$p = 0.21$	$p = 0.82$
Sym-60%	59.53%	61.45%	61.87%	63.19%	63.51%	63.07%	63.66%	63.38%
	$p = 1.0$	$p = 0.82$	$p = 0.59$	$p = 0.00$	$p = 0.00$	$p = 0.01$	$p = 0.00$	$p = 0.00$
Pair-40%	62.65%	61.65%	62.09%	62.57%	62.67%	63.05%	62.05%	61.53%
	$p = 0.52$	$p = 0.92$	$p = 0.73$	$p = 0.04$	$p = 0.02$	$p = 0.01$	$p = 0.02$	$p = 0.45$

Bold indicates our results are significant in one-sided Wilcoxon signed ranktest

smaller than the 0.05. It shows that when $\alpha = 0.7$, the performance gain of combing both part-level and global-level features is significant compared with using global-level features along.

5.6 Integration with other methods

We examine whether the proposed learning framework could improve the robustness of existing methods. We employ MGE-CNN and API-Net as base methods. Specifically, since all these approaches use ResNet-50 as the backbone, we extract the global feature representation before feeding it to the dedicated network. Considering that the last convolutional features of those models have different semantic meanings, we omit the computation of the part class-prototype similarity. To be explicit, we only use the global class-prototype similarity to improve the above FGVC models.

The results are listed in Table 4, the integrated methods produce noticeable performance improvement on symmetric label noise, and a small improvement on pairwise-flip noise. It is noted that in the case of 40% noise rate, our strategy yields a large improvement. In the case of API-Net on 60% symmetric noise, the strategy fails because the basic architecture is too vulnerable to provide discriminative features. In summary, our learning framework can be used as a general component to increase the robustness of learning models.

Table 4 Performance comparison with original model and combined model in CUB dataset, symbol ‘-’ means the label correction strategy is not work

Methods	Symmetric noise			Pairwise-flip noise	
	20%	40%	60%	20%	40%
ProtoSimi	81.83	76.00	62.68	78.31	62.49
MGE-CNN	82.01	68.19	44.62	78.83	56.79
MGE-CNN + ProtoSimi	83.22	74.71	56.46	80.87	58.08
API-Net	66.57	44.34	26.33	67.66	44.41
API-Net + ProtoSimi	69.92	54.37	–	67.73	47.46

All the best experimental results are marked in bold

Significant testing is not necessary for this experiment, because the performance hugely depends on the training of basic fine-grained model. If the model is too vulnerable to label noise, ProtoSimi not always has obviously positive effects on these methods

Table 5 Test accuracy on CUB dataset with extreme noise rate

	ProtoSimi	Deep self-learning	SAT-CE
Sym-80%	25.60 \pm 0.20	16.10 \pm 0.73	23.56 \pm 0.05
Pair-45%	54.31 \pm 0.07	48.67 \pm 0.76	47.36 \pm 0.30

All the best experimental results are marked in bold

Table 6 Test accuracy on Web-Bird dataset with reported 0.35 noise rate (Sun et al., 2021)

	Cross-entropy	Only global prototype	Combined
Web-Bird	65.53 \pm 0.18	71.40 \pm 0.53	73.48 \pm 0.04

All the best experimental results are marked in bold

5.7 Investigation on extreme noisy data

We have compared the classification accuracy of our method with Deep Self-learning and SAT-CE under the extreme noisy data settings, the noise rate is set to 80% for symmetric noise and 45% for pairwise-flip noise, respectively. Deep Self-learning has the similar philosophy to generate pseudo labels by referencing class-prototypes and SAT-CE has the best performance among all baselines, therefore we only compare these two algorithms with ours to investigate the model performance.

In Table 5, our algorithm outperforms a large margin in both Sym-80% and Pair-45% situations, this result strongly shows that the power of global and part combined prototype-based label correction strategy also has competitive advantages in tackling extreme noisy labels. Deep Self-Learning also exploits the prototypes information for the robust training, however, the prototypes are poorly estimated when the training sample contains too much noise, thus the generated pseudo labels lead to a worse generalization ability.

5.8 Investigation on real-world web dataset

We further investigate the performance of ProtoSimi on recently released real-world dataset “Web-Bird” (Sun et al., 2021). The experiments settings are same as the settings on CUB dataset. In Table 6, we illustrate the accuracies obtained by employing naive cross-entropy, only global class-prototype, and combination of both part and global class-prototypes. It shows that combing both part and global class-prototypes produces the best results.

6 Conclusion

In this paper, we propose a novel class-prototype similarity-based learning framework to tackle the label noise problem on the FGVC task. This approach takes advantage of the global and part similarities between sample features and class prototypes. The learning process can improve the clean ratio of the training examples by identifying confident examples and correcting labels on unconfident examples. As a result, we can alleviate the data scarcity problem in the FGVC task and reduce overfitting during learning. Extensive experiments validate the superiority of our model compared with the current state-of-the-art methods. Ablation studies indicate that features are much more robust on label noise than model predictions, and the proposed model is insensitive to hyper-parameter settings.

Acknowledgements Jun Yu is sponsored by Natural Science Foundation of China (62276242), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B, CAAIXSJLJJ-2022-001A), Anhui Province Key Research and Development Program (202104a05020007), USTC-IAT Application Sci. & Tech. Achievement Cultivation Program (JL06521001Y). Ruxin Wang is partially supported by Yunnan Foundation Research Plans 202201AU070034, 202201AT070173, and NSFC 62101480.

Author contributions JS and YY are both first authors, equal contributions. JS mainly contributes to idea verification, experiments designing, and experimental part paper writing. YY contributes theory research, detailed algorithms designing, and theoretical part paper writing. SH and ZW are both experts in fine-grained visual categorization. They give us very strong supports and novel ideas in the computer vision techniques. Shaoli Huang proposes to use Class Activation Mapping (CAM) method to extract discriminative part features, and suggest to use both global and part class-prototypes to enhance the reliability of generated class-prototypes. ZW proposes the self-adaptive label correction strategy to iteratively raise the clean ratio of training examples, and he also helps us in choosing the compared baselines in FGVC methods. RW, JY and JZ conceive and design the empirical settings and analyses. RW provides strong supports in ablation studies, i.e., the studies in hyper-parameters setting and their influences to algorithm. Moreover, he contributes the significant testing of experimental results. JY contributes the idea of viewing our algorithm as a plugin component in general FGVC methods to gain the robustness in label noise issue, he also helps us to design the experiments for verification. JZ contributes the idea of conducting experiments on extreme noisy data. TL is the leader of this project, expert in both label noise and robustness of machine learning systems. He proposes the background and motivations of FGVC problem with label noise. Novel ideas in label noise areas, i.e., the comparison experiments of utilizing features and small loss trick in selecting confident examples in corrupted fine-grained dataset. Most importantly, he contributes the fundamental idea that the deeper layers in DNNs are much prone to overfitting on label noise, this paper is mainly conducted on this assumption.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. No funds, Grants, or other support was received.

Data availability All datasets generated or analysed during this study are included in main article. And all these datasets are public datasets.

Code availability Our codes will be available after reviews.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arpit, D., Jastrzebski, S. L., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S. et al. (2017). A closer look at memorization in deep networks. In *ICML* (pp. 233–242).
- Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., & Liu, T. (2021). Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, *34*, 24392–24403.
- Chen, Y., Bai, Y., Zhang, W., & Mei, T. (2019). Destruction and construction learning for fine-grained image recognition. In *CVPR* (pp. 5157–5166).

- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR* (pp. 248–255).
- Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2020). Sharpness-aware minimization for efficiently improving generalization. arXiv preprint [arXiv:2010.01412](https://arxiv.org/abs/2010.01412)
- Gong, C., Ding, Y., Han, B., Niu, G., Yang, J., You, J., Tao, D., Sugiyama, M. (2023). Class-wise denoising for robust learning under label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 2835–2848. <https://doi.org/10.1109/TPAMI.2022.3178690>
- Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS* (pp. 8527–8537).
- Han, J., Luo, P., & Wang, X. (2019). Deep self-learning from noisy labels. In *ICCV* (pp. 5138–5147).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).
- Huang, H., Zhang, J., Yu, L., Zhang, J., Wu, Q., & Xu, C. (2022). TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2), 853–866. <https://doi.org/10.1109/TCSVT.2021.3065693>
- Huang, L., Zhang, C., & Zhang, H. (2020). Self-adaptive training: Beyond empirical risk minimization. In *NeurIPS* (Vol. 33).
- Huang, S., Xu, Z., Tao, D., & Zhang, Y. (2016). Part-stacked CNN for fine-grained visual categorization. In *CVPR* (pp. 1173–1182).
- Jiang, L., Zhou, Z., Leung, T., Li, L.-J., & Fei-Fei, L. (2018). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML* (pp. 2304–2313).
- Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3D object representations for fine-grained categorization. In *ICCVW* (pp. 554–561).
- Lee, K.-H., He, X., Zhang, L., & Yang, L. (2018). Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*.
- Li, J., Socher, R., & Hoi, S. C. (2020). Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*.
- Li, X., Liu, T., Han, B., Niu, G., & Sugiyama, M. (2021). Provably end-to-end label-noise learning without anchor points. In *International conference on machine learning* (pp. 6403–6413).
- Liu, H., Zhang, C., Yao, Y., Wei, X., Shen, F., Zhang, J., & Tang, Z. (2021). Exploiting web images for fine-grained visual recognition by eliminating noisy samples and utilizing hard ones.
- Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C. (2020). Early-learning regularization prevents memorization of noisy labels. In *NeurIPS* (Vol. 33).
- Liu, T., & Tao, D. (2015). Classification with noisy labels by importance reweighting. *TPAMI*, 38(34), 47–461.
- Maji, S., Kannala, J., Rahtu, E., Blaschko, M., & Vedaldi, A. (2013). Fine-grained visual classification of aircraft (Technical Report).
- Malach, E. & Shalev-Shwartz, S. (2017). Decoupling “when to update” from “how to update”. arXiv preprint [arXiv:1706.02613](https://arxiv.org/abs/1706.02613)
- Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *CVPR* (pp. 1944–1952).
- Song, H., Kim, M., & Lee, J.-G. (2019). Selfie: Refurbishing unclean samples for robust deep learning. In *ICML* (pp. 5907–5915).
- Sun, M., Yuan, Y., Zhou, F., & Ding, E. (2018). Multi-attention multi-class constraint for fine-grained image recognition. In *European Conference on Computer Vision*.
- Sun, Z., Yao, Y., Wei, X.-S., Zhang, Y., Shen, F., Wu, J., & Shen, H. T. (2021). Webly supervised fine-grained recognition: benchmark datasets and an approach. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10602–10611).
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).
- Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset (Technical Report No CNS-TR-2011-00). California Institute of Technology.
- Wei, H., Feng, L., Chen, X., & An, B. (2020). Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13726–13735).
- Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., & Chang, Y. (2021). Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*.
- Yang, M., Huang, Z., Hu, P., Li, T., Lv, J., & Peng, X. (2022). Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14308–14317).

- Yang, M., Li, Y., Hu, P., Bai, J., Lv, J. C., & Peng, X. (2022). Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2022.3155499>
- Yang, M., Li, Y., Huang, Z., Liu, Z., Hu, P., & Peng, X. (2021). Partially view-aligned representation learning with noise-robust contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 1134–1143).
- Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., & Sugiyama, M. (2020). Dual t: Reducing estimation error for transition matrix in label-noise learning.
- Yu, X., Han, B., Yao, J., Niu, G., Tsang, I. W., & Sugiyama, M. (2019). How does disagreement help generalization against label corruption? arXiv preprint [arXiv:1901.04215](https://arxiv.org/abs/1901.04215)
- Zhang, C., Yao, Y., Shu, X., Li, Z., Tang, Z., & Wu, Q. (2020). Data-driven meta-set based fine-grained visual classification. arXiv preprint [arXiv:2008.02438](https://arxiv.org/abs/2008.02438)
- Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint [arXiv:1710.09412](https://arxiv.org/abs/1710.09412)
- Zhang, L., Huang, S., Liu, W., & Tao, D. (2019). Learning a mixture of granularity-specific experts for fine-grained categorization. In *ICCV* (pp. 8331–8340).
- Zhang, X., Xiong, H., Zhou, W., Lin, W., & Tian, Q. (2016). Picking deep filter responses for fine-grained image recognition. In *CVPR* (pp. 1134–1142).
- Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV* (pp. 5209–5217).
- Zheng, H., Fu, J., Zha, Z.-J., & Luo, J. (2019). Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5012–5021).
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *CVPR* (pp. 2921–2929).
- Zhuang, P., Wang, Y., & Qiao, Y. (2020). Learning attentive pairwise interaction for fine-grained classification. In *AAAI* (pp. 13130–13137).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Jialiang Shen¹  · Yu Yao¹ · Shaoli Huang¹ · Zhiyong Wang¹ · Jing Zhang¹ · Ruxing Wang² · Jun Yu³ · Tongliang Liu¹

Shaoli Huang
shaoli.huang@sydney.edu.au

Zhiyong Wang
zhiyong.wang@sydney.edu.au

Jing Zhang
jing.zhang1@sydney.edu.au

Ruxing Wang
rosinwang@gmail.com

Jun Yu
harryjun@ustc.edu.cn

¹ School of Computer Science, The University of Sydney, 1 Cleveland Street, Sydney, NSW 2050, Australia

² School of Software, Yunnan University, No. 2 North CuiHu Road, Kunming 650000, Yunnan, People's Republic of China

³ Department of Automation, University of Science and Technology of China, No. 96, JinZhai Road, Hefei 230026, Anhui, People's Republic of China