Machine Learning (2023) 112:1847–1869
https://doi.org/10.1007/s10994-022-06208-6

A viable framework for semi-supervised learning on realistic dataset

Hao Chang[1] · Guochen Xie[1] · Jun Yu[1] · Qiang Ling[1] · Fang Gao[2] · Ye Yu[3]

Received: 5 November 2021 / Revised: 19 April 2022 / Accepted: 30 May 2022 /
Published online: 21 June 2022

Abstract
Semi-supervised Fine-Grained Recognition is a challenging task due to the difficulty of data imbalance, high inter-class similarity and domain mismatch. Recently, this field has witnessed giant leap and many methods have gained great performance. We discover that these existing Semi-supervised Learning (SSL) methods achieve satisfactory performance owe to the exploration of unlabeled data. However, on the realistic large-scale datasets, due to the abovementioned challenges, the improvement of the quality of pseudo-labels requires further research. In this work, we propose Bilateral-Branch Self-Training Framework (BiSTF), a simple yet effective framework to improve existing semi-supervised learning methods on class-imbalanced and domain-shifted fine-grained data. By adjusting stochastic epoch update frequency, BiSTF iteratively retrains a baseline SSL model with a labeled set expanded by selectively adding pseudo-labeled samples from an unlabeled set, where the distribution of pseudo-labeled samples is the same as the labeled data. We show that BiSTF outperforms the existing state-of-the-art SSL algorithm on Semi-iNat dataset. Our code is available at https://github.com/HowieChangchn/BiSTF.

Editors: Bo Han, Tongliang Liu, Quanming Yao, Mingming Gong, Gang Niu, Ivor W. Tsang, Masashi Sugiyama.

✉ Jun Yu
harryjun@ustc.edu.cn

Hao Chang
changhaoustc@mail.ustc.edu.cn

Guochen Xie
xiegc@mail.ustc.edu.cn

Qiang Ling
qling@ustc.edu.cn

Fang Gao
fgao@gxu.edu.cn

Ye Yu
yuye@hfut.edu.cn

[1] University of Science and Technology of China, Jinzhai Road, Hefei 230026, Anhui, China

[2] Guangxi University, Daxue East Road, Nanning 530004, Guangxi, China

[3] Hefei University of Technology, Tunxi Road, Hefei 230009, Anhui, China

Keywords Semi-supervised learining · Fine-grained · Class imbalance · Domain mismatch

1 Introduction

With the emergence of research on deep Convolutional Neural Networks (CNNs), the property of image recognition has witnessed surprising progress. The performance benefit is mainly conferred by the use of the large datasets, which easily leads to a significant cost due to the human labor for labeling data. This motivates the advance of semi-supervised learning (SSL), which alleviates the requirement of labeled data by utilizing unlabeled data. During the construction of SSL benchmark datasets, maintaining balanced data distribution and same domain between labeled and unlabeled data is generally made implicitly as a common assumption. However, in many realistic scenarios, this assumption holds untrue. For example, the Semi-Supervised iNaturalist (Semi-iNat) dataset (Su & Maji, 2021), as a fine-grained image recognition dataset, has skewed distributions with a long tail. More challengingly, the Semi-iNat dataset does not make an evident distinction between in-class and out-of-class unlabeled data.

Class-rebalancing and SSL have been widely explored. A plethora of researches show that models trained on imbalanced data will bias towards majority classes and away from minority classes, which will seriously affect the performance of the model. Some popular class-rebalancing methods can be viewed as relying on labels to rebalance the biased model and thus mitigate bias, such as re-sampling (Buda et al., 2018; Byrd & Lipton 2019) and re-weighting (Cui et al., 2019; Cao et al., 2019). When training models on a large class-imbalanced semi-supervised dataset, the missing labels become the primary cause of poor performance of class-rebalancing methods. SSL provides a reasonable means of utilizing unlabeled data to satisfy the requirement for labeled data. Therefore, a large number of SSL methods (Lee, 2013; Tarvainen & Valpola, 2017; Berthelot et al., 2019; Sohn et al., 2020) with high performance gains at low cost have been proposed.

In contrast, semi-supervised fine-grained image recognition on realistic large-scale dataset with class imbalance and domain mismatch has been understudied. Actually, high inter-class similarity precludes the representative ability of representation learning, which exposes some limitations of existing class-rebalancing and SSL approaches. In SSL algorithms, pseudo-labels generated by the model trained on labeled data are utilized to retrain the model in following stages. However, pseudo-labels take the risks of bias towards majority classes and distortion due to out-of-class data if they are generated by an initial model trained on imbalanced and domain-shifted data. What is worse, in following stages biased and distorted pseudo-labels will intensify the bias and distortion, as well as deteriorate the model quality. Both class imbalance and domain mismatch have not been thoroughly evaluated in most existing SSL algorithms. Besides, these algorithms are focused on researching standard SSL image recognition benchmarks (Tarvainen & Valpola, 2017; Berthelot et al., 2019; Sohn et al., 2020; Xie et al., 2020) instead of fine-grained image recognition benchmarks.

In this work, we investigate the performance of SSL in the context of domain-shifted and imbalanced semi-supervised fine-grained classification, as illustrated in Fig. 1. Concretely, to figure out how SSL works, we select two similar species in the Semi-iNat dataset and

analyze the predictions produced by FixMatch (Sohn et al., 2020), a representative SSL algorithm with state-of-the-art performance on balanced SSL benchmarks. We observe that

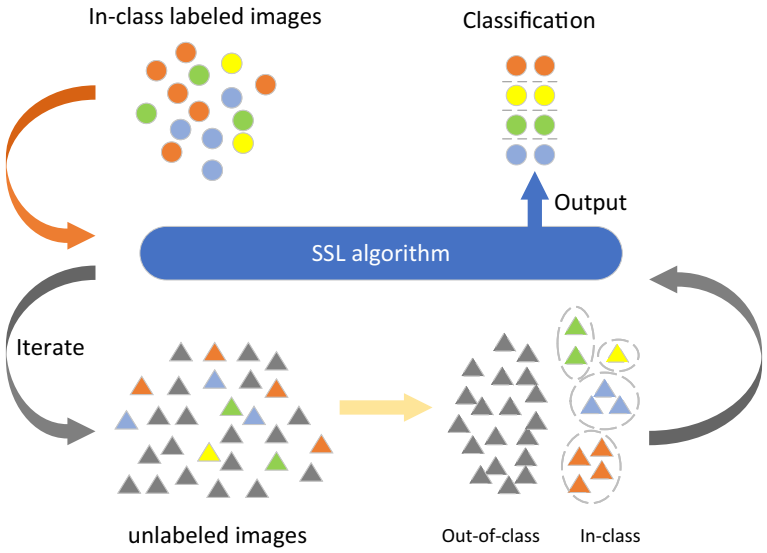


Fig. 1 Expected performance of SSL algorithm on Semi-iNat. We use circles to represent in-class (the target class of our recognition task) labeled images and label each category with a different color, while the number of each category reflects the class imbalance. We use triangles to represent unlabeled images, where colored triangles represent in-class data and gray triangles represent out-of-class data (classes other than the target class), and unlabeled data reflect domain mismatch. In addition, the out-of-class data and in-class data are highly similar, which indicates that our task is a fine-grained recognition task. Faced with the challenges of class imbalance and domain mismatch, the expected SSL algorithm should be competent to select more correct in-class samples (colored triangles) from the unlabeled set (triangles) and add them to the training. At the same time, the model can avoid the bias towards the majority classes (Color figure online)

the highly similar classes, class imbalance, and domain mismatch between the labeled and unlabeled data are causes of the undesired performance of FixMatch. The high similarity between classes urges the pseudo-labels to be misclassified with a high probability. Therefore, the addition of pseudo-labels, which may be misclassified or from out-of-class data, aggravates the degradation of the model's representation ability in the early stage of training. On the other side, the imbalanced recall obtained by biased model results in a low accuracy on balanced validation set in the mass.

Therefore, in this paper, for exhaustively improving the performance of semi-supervised fine-grained classification in realistic scenarios, we introduce a bilateral-branch self-training framework (BiSTF), which trains imbalanced data through a bilateral-branch structure, and samples pseudo-labeled data while maintaining the same data distribution through a stochastic epoch update strategy. In order to improve the fine-grained learning ability of the model, BiSTF utilizes a backbone with an attention mechanism. Rather than updating the labeled set in each iteration, we instead use a stochastic epoch update strategy to moderate the noise from out-of-class data in which the frequency of update increases as training progresses. In addition, to avoid the model being biased towards the majority classes, the proposed method samples the unlabeled data with the same distribution as the labeled data set and adds them into the labeled set for retraining in the next epoch.

In experiments, empirical results on two balanced semi-supervised benchmarks show that our model obviously outperforms existing state-of-the-art SSL methods. We also show in experiments that BiSTF improves over FixMatch (Sohn et al., 2020) by a large margin on imbalanced semi-supervised benchmarks. Furthermore, on Semi-iNat dataset (Su & Maji, 2021), our method outperforms FixMatch by as much as 10.25% in accuracy. Extensive ablation studies further demonstrate that our method particularly helps to improve ability to extract pseudo-labeled data from domain-shifted unlabeled data, making it a viable solution for class-imbalanced and domain-shifted semi-supervised fine-grained image recognition.

2 Related work

2.1 Class imbalance sampling

Recent years have observed many prominent researches on class-imbalanced data. Important class re-balancing methods can be mainly divided into re-sampling (Buda et al., 2018; Byrd & Lipton, 2019) and re-weighting (Cui et al., 2019; Cao et al., 2019). Re-sampling rebalances the data distribution by over-sampling the minority classes or under-sampling the majority classes, however this method may cause overfitting to the minority classes, while some valuable training samples will be lost, which deteriorates the model quality in large-scale datasets. Re-weighting adjusts the network training by rebalancing the contribution of each class in expectation closer to the test distributions. However, when applying re-weighting to large-scale realistic scenarios, it generally is difficult to optimize. In addition, there are some other methods to alleviate the impact of long-tailed data on model representation ability, such as FocalLoss (Lin et al., 2017). These methods assume that all labels of data fed into the model are available, and due to the missing label information in the SSL, the performance is to a great extent unknown.

2.2 Semi-supervised learning

Semi-supervised learning has gradually matured in recent years, with a lot of outstanding advancements (Lee, 2013; Miyato et al., 2018; Berthelot et al., 2019; Sohn et al., 2020; Xie et al., 2020, 2019; Laine & Aila, 2016). Due to the generality, self-training has been applied in many domains, such as image classification (Lee, 2013), object detection (Rosenberg et al., 2005). Pseudo-labeling (Lee, 2013; Sohn et al., 2020) as a special variant is one of the basic SSL techniques, which exploits the pseudo-labeled target predicted by the model itself to train a classifier. Consistency regularization (Miyato et al., 2018; Laine & Aila, 2016) improves the consistency of prediction between different random perturbations of unlabeled data through soft (Miyato et al., 2018; Berthelot et al., 2019; Laine and Aila 2016) and hard (Sohn et al., 2020) pseudo-labels to learn the classifier. In general, including Dropout (Srivastava et al., 2014), Mixup (Zhang et al., 2017), some data augmentations and stochastic regularization are leveraged to generate multiple disturbed views. In recent years, some holistic approaches have attempted to integrate pseudo-labeling and consistency regularization methods in a framework to obtain better performance, such as MixMatch (Berthelot et al., 2019) and FixMatch (Sohn et al., 2020). The quality of pseudo-labels is critical to the performance of recent SSL methods. In a class-imbalanced dataset,

particularly in a large-scale realistic dataset, the model bias may notably threaten the quality of pseudo-labels, however the abovementioned researches have not studied this field.

2.3 Class-imbalanced semi-supervised learning

Although SSL and class rebalancing have been extensively studied, it is still underexplored for large-scale realistic datasets with both tricky problems. Recently, Yang and Xu (2020) shown that by using SSL, self-supervised learning can be beneficial to class imbalanced learning. Hyun et al. (2020) proposed a method to suppress the loss of minority classes by suppressing the consistency loss. Although these works have done some researches on SSL under imbalanced data distribution, there is no more discussion for domain mismatch, neither for fine-grained recognition.

3 Domain-shifted and class-imbalanced SSL

In this section, we first set up the research problem and introduce the official baseline and baseline SSL algorithms. Next, we study the biased behavior of existing SSL algorithms on Semi-iNat.

3.1 Problem setup and baselines

First, we will construct a fine-grained recognition problem and partition the dataset into two sets C_{in} and C_{out} based on categories, as in-class and out-of-class data, respectively. Then, for each category in C_{in}, we will sample about β (e.g., $\beta = 10\%$) images for each class as labeled data L_{in}, and the rest as unlabeled data U_{in}. For the category in C_{out}, all images are contained in U_{out}. We then merge the two sets of unlabeled data $U = U_{in} \bigcup U_{out}$ and do not provide domain labels. As for the class imbalance, we use γ to measure the degree of imbalance, which illustrates the proportion of the most class to the least class in the dataset, e.g. $\gamma = \frac{N_{max}}{N_{min}}$. Our task is to efficiently use both the in-class labeled set and unlabeled set to maximize the accuracy of the model on the in-class validation and test set.

Many state-of-the-art SSL methods exploit unlabeled data by assigning a pseudo-label using the classifier's prediction. We choose FixMatch, one of the state-of-the-art SSL algorithms proposed for class-balanced SSL and as our semi-supervised comparison baseline. Before using FixMatch, we will outline its core idea. The ingenuity of FixMatch is the combination of pseudo-labeling and consistency regularization via a cross-entropy loss function. Specifically, the method performs consistency regularization by cross-entropy on the pseudo-label of weakly data augmented image and the probability distribution of strongly data augmented image of the same image. In addition, FixMatch makes predictions on unlabeled data at each iteration and uses only unlabeled data with high confidence to participate in training. CReST (Wei et al., 2021) is another semi-supervised baseline of our choice, who is a class-rebalancing self-training framework for imbalanced semi-supervised learning. CReST iteratively retrains labeled set by adding pseudo-labeled samples from unlabeled set, where pseudo-labeled samples from minority classes are selected more frequently based on the estimated class distribution.

In addition, the official of Semi-iNat presents a result of fully-supervised model on the labeled set using ResNet-50 (Krizhevsky & Hinton, 2009) models trained from ImageNet pretrained model. They built a general recognition network with basic training strategies.

3.2 How SSL performs on semi-iNat?

In this section, we attempt to investigate the misclassified and biased behavior of existing SSL algorithms on Semi-iNat. Many existing state-of-the-art SSL methods assign a pseudo-label with the classifier's prediction to leverage unlabeled data. Then both labeled and selected unlabeled samples with corresponding high-confidence pseudo-labels are utilized to optimize the classifier. These SSL algorithms have achieved excellent performance on standard class-balanced benchmarks, because the quality of the classifier improves during the training process, owing to the addition of pseudo-labeled data. Since the quality of online pseudo-labels is crucial to the performance of SSL algorithms, we

propose a conjecture that in realistic large-scale fine-grained dataset the online pseudo-labels will be biased along with the bias of the classifier due to high inter-class similarity, skewed class distribution and shifted domain, which may further aggravate the domain mismatch and class imbalance issue and result in terrible performance on validation set.

Instead of extending the protocol, which utilizes various class-imbalanced ratios to produce long-tailed versions of benchmark datasets, such as CIFAR (Krizhevsky & Hinton, 2009), and retains a fraction of training data as labeled and the rest as unlabeled, Semi-iNat has been split into train, validation, public test, and private test set. In order to justify our conjecture, we observe and analyze the processing of FixMatch, one of the state-of-the-art SSL algorithms proposed for class-balanced SSL, for pseudo-labels during the training process. Concretely, we artificially choose two species with a high inter-class similarity, and calculate the correctness of the pseudo-labels selected by FixMatch in different epochs.

As shown in Fig. 2, we have made confusion matrices for the species under the Lamiaceae family in the validation set. From the figure, we can find that the model confuses the species within the same genus more obviously, while the confusion between the genera is weaker. For example, the confusion among Salvia tesquicola, Salvia aethiopis, Salvia nemorosa and Salvia carduacea species under the Salvia genus is more obvious than other species.

To more vividly demonstrate weak learning ability of FixMatch for fine-grained features, we selected two highly similar species as examples. As shown in the Fig. 3, Salvia nemorosa speices and Salvia tesquicola speices are members of Salvia genus. The two species are highly similar and we selected two images belonging to Salvia nemorosa for testing on FixMatch. However, during training phase, an image belonging to Salvia tesquicola is classified incorrectly as Salvia nemorosa by FixMatch. When FixMatch inferred the unlabeled data at the end of each iteration, the images that satisfy the confidence threshold are given pseudo-labels and added to the training. At this point, because of FixMatch's obfuscation of fine-grained, pseudo-labels are unreliable to some extent, and incorrect pseudo-labels are added to training. For the two test species we selected, FixMatch adds a real-label image of Salvia tesquicola to the Salvia nemorosa species in the next iteration in the pseudo-label generation phase.

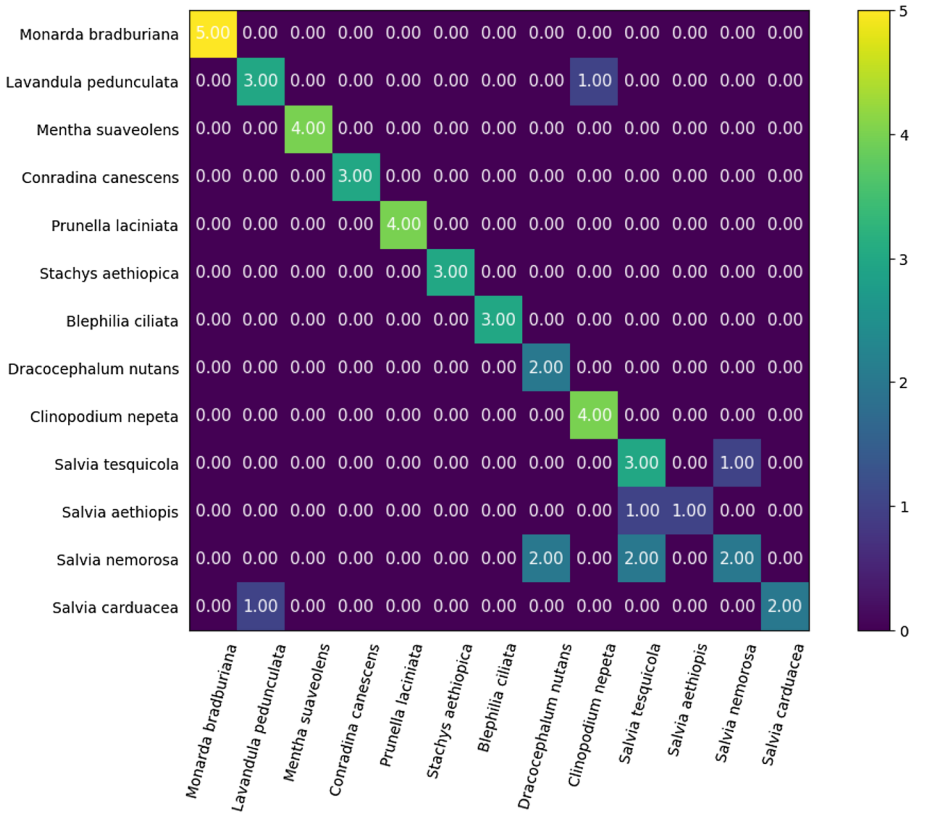


Fig. 2 Confusion matrices for the species under the Lamiaceae family. There are ten genuses under the Lamiaceae family, including four species under Salvia genus. FixMatch is more confused among species under the same genus

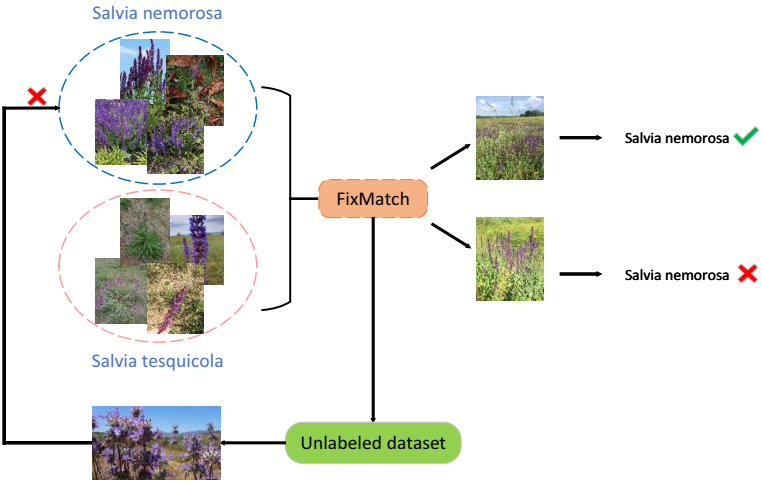
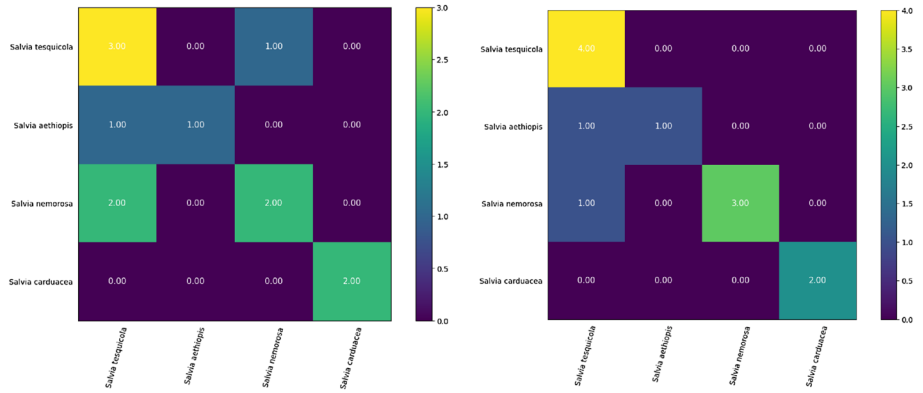


Fig. 3 Performance of FixMatch on Semi-iNat. An image of Salvia tesquicola is classified incorrectly due to highly similar classes. Besides, FixMatch introduces noise from out-of-class data during training phase



(a) Control group **(b)** Experimental group

Fig. 4 The effect of out-of-class data bias. The introduction of extraclassical data can confuse highly similar species even more

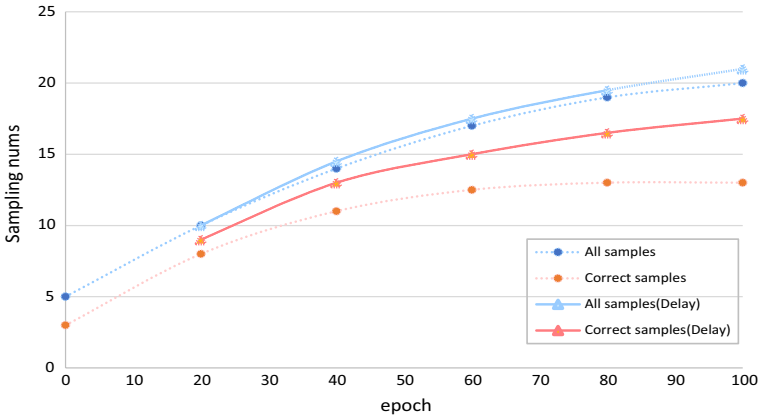


Fig. 5 Performance of FixMatch on Semi-iNat. By delaying the addition of pseudo-labels, the quality of pseudo-labels has been improved

Furthermore, we found that FixMatch also introduces noise from out-of-class data when selecting pseudo-labels from unlabeled data, so that in the next iteration, the model is heavily biased towards out-of-class data. To further explore the effect of out-of-class data bias, we use the original unlabeled dataset as a control group and remove the species under the Salvia genus other than the in-class data as the experimental group. To further explore the effect of out-of-class data bias, we remove species of the Salvia genus other than in-class data as the control unlabeled dataset, and we continue to use Salvia nemorosa speices and Salvia tesquicola speices as examples. We find that some of the images of Salvia genus belonging to the out-of-class data are incorrectly identified as Salvia nemorosa, and some of the images are incorrectly identified as Salvia tesquicola. The in-class data introduces a degree of noise from the out-of-class data,

biasing the model with the out-of-class data and making the highly similar species more confusing, as further evidenced by the confusion matrix in Fig. 4.

The statistics of the accuracy of pseudo-labels has shown in Fig. 5. We find that in the early stage of training, the pseudo-labeled data introduced by FixMatch is partly wrong, although they have more than 95% confidence, and with training, the proportion of errors will continue to increase, which is consistent with our conjecture. When we postpone the addition of pseudo-labels, the precision of pseudo-labels surprisingly improves. For example, the pseudo-labels of Salvia nemorosa achieves 89.7% accuracy on the epoch 40, while only achieving relatively low precision by FixMatch. This indicates that appropriate improvement to the pseudo-labels generation strategy will help the model to alleviate the bias towards out-of-class and noise.

In addition, the official of Semi-iNat presents a result of fully-supervised model on the labeled set using ResNet-50 (He et al., 2016) models trained from ImageNet pre-trained model. They built a general recognition network with basic training strategies. By calculating the verification recall rate and validation precision of the official baseline and FixMatch on Semi-iNat dataset, we find that both models achieve very high and poor recall, respectively, on majority classes and minority classes and FixMatch only has limited help for fine-grained image recognition on Semi-iNat, shown in Fig. 6. The quality of pseudo-labels is reduced due to class imbalance, also resulting in the poor performance on Semi-iNat. These empirical findings motivate us to improve the model's ability to learn fine-grained features and alleviate the impact of class imbalance and domain mismatch.

To achieve this goal, we introduce BiSTF, a bilateral-branch self-training framework for domain-shifted and class-imbalanced semi-supervised fine-grained recognition illustrated in Fig. 7.

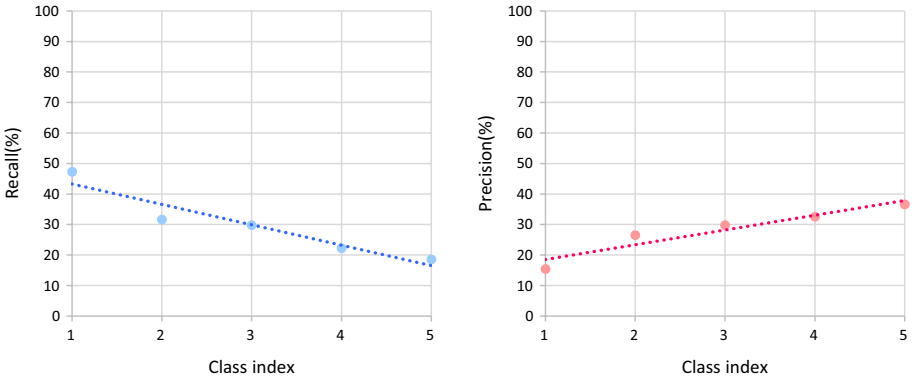


Fig. 6 Performance of FixMatch on Semi-iNat. The recall and precision show that FixMatch is biased towards the majority classes

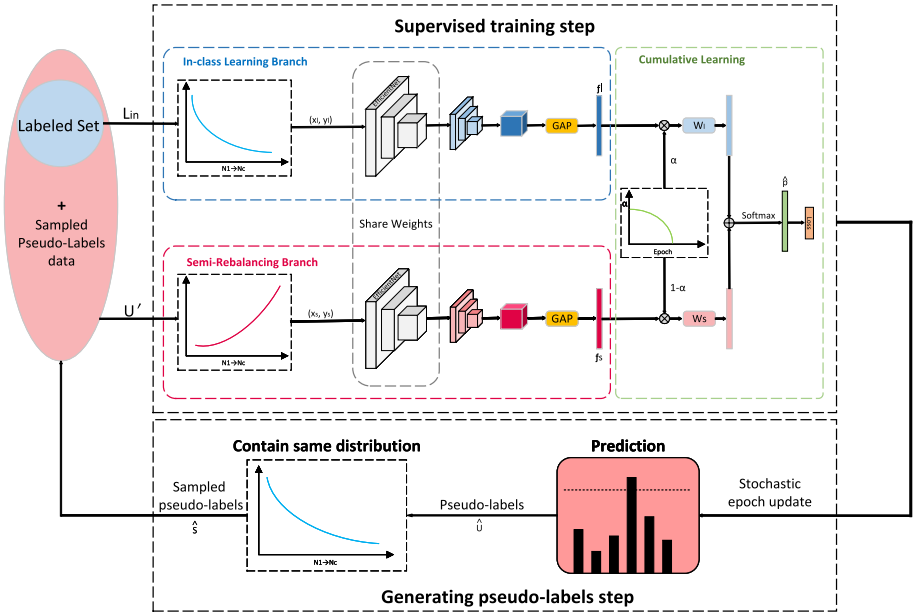


Fig. 7 Bilateral-Branch Self-Training Framework (BiSTF). By adjusting the update frequency through stochastic epoch update, BiSTF iteratively retrains a baseline SSL model with a labeled set expanded by adding pseudo-labeled samples from an unlabeled set, where pseudo-labeled samples contain same data distribution with the labeled dataset. See text for details

4 Approach

4.1 Overall framework

In SSL, self-training, a popular iterative method, trains the model for multiple epochs, in which each epoch involves two training steps, supervised training step and generating pseudo-labels step.

4.1.1 Supervised training step

As shown in Fig. 7, our BiSTF consists of abovementioned two main steps. In supervised training step, the model contains three main components. Concretely, we design two branches for in-class representation learning and semi-supervised rebalancing classifier learning, termed "in-class learning branch" and "semi-rebalancing branch", respectively.

Existing SSL algorithms usually ignore the subtle but discriminative features in fine-grained recognition, hence a network structure with an attention mechanism is necessary. In order to satisfy the needs of fine-grained recognition, both branches use the same EfficientNet (Tan & Le, 2019) network structure, where

SENet (Hu et al., 2018) can effectively capture the fine-grained features rather than residual network structure, and share all the weights except for the last MB-Conv block.

Considering the labeled set, let x denote a training sample with its corresponding label $y \in 1, 2, \ldots, C$ for a C-class recognition task. For the bilateral branches, we separately

apply uniform and reversed samplers proposed in BBN (Zhou et al., 2020) to each of them. After sampling, two samples (x_l, y_l) and (x_s, y_s) are obtained as the input data, where (x_l, y_l) is for the in-class learning branch and (x_s, y_s) is for the semi-rebalancing branch. Then, two samples are loaded into corresponding branch, and by global average pooling(GAP) the feature vectors $\boldsymbol{f}_l \in \mathbb{R}^D$ and $\boldsymbol{f}_s \in \mathbb{R}^D$ can be acquired.

Furthermore, inspired by BBN, we also introduce the specific cumulative learning strategy to shift the mode's learning "attention" in the supervised training step. The outputs will be integrated together by element-wise addition after feeding the weighted feature vectors $\alpha \boldsymbol{f}_l$ and $(1-\alpha)\boldsymbol{f}_s$ into the classifiers $\boldsymbol{W}_l \in \mathbb{R}^{D\times C}$ and $\boldsymbol{W}_s \in \mathbb{R}^{D\times C}$ respectively. The predicted output $z \in \mathbb{R}^C$ is illustrated as

$$z = \alpha \boldsymbol{W}_l^\top \boldsymbol{f}_l + (1-\alpha)\boldsymbol{W}_s^\top \boldsymbol{f}_s \tag{1}$$

Then softmax function will utilize each component in z, i.e., $[z_1, z_2, \ldots, z_C]^\top$, to calculate the probability for each category $i \in \{1, 2, \ldots, C\}$ by

$$\hat{\boldsymbol{p}}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_i}} \tag{2}$$

Generally, we denote the output probability distribution as $\hat{\boldsymbol{p}} = [\hat{p_1}, \hat{p_2}, \ldots, \hat{p_C}]^\top$, $E(\cdot,\cdot)$ as the cross-entropy loss function. Therefore, our in-class representation learning step generates a weighted cross-entropy recognition loss, which is formulated as

$$L = \alpha E(\hat{\boldsymbol{p}}, y_l) + (1-\alpha)E(\hat{\boldsymbol{p}}, y_s) \tag{3}$$

4.1.2 Generating pseudo-labels step

As observed in Sect. 3, out-of-class unlabeled data will easily be introduced in the early stage of training and cause interference to the model. Therefore, instead of iteration update strategy of FixMatch, we propose "stochastic epoch update" strategy when supervised training step has thoroughly trained. We define whether to update or not as a flag F_{update}.

When F_{update} is true, the algorithm really enters the generating pseudo-labels step. Pseudo-labeling exploits the idea of using the model obtained from the previous step to generate artificial labels for unlabeled data. Concretely, when the maximum class probability exceeds the predefined threshold, the hard label is retained as a pseudo label. Letting x_u is a sample in unlabeled set, and $q = p(y \mid x_u)$, pseudo-labeling uses the following function:

$$\hat{y} = f(max(q) \geq \tau) \tag{4}$$

where τ is the threshold and f is the classifier learned from the previous step. The generated pseudo-labeled set $\hat{U} = \{(x_u, \hat{y})\}$.

To accommodate the class-imbalance, this paper proposes "contain same distribution" strategy, that instead expands the labeled set with a selected subset $\hat{S} \subset \hat{U}$, i.e., $U' = L_{in} \cup \hat{S}$, rather than with all samples in $\hat{U}$. For the next epoch, the labeled set L_{in} and the union set U' will be fed into the "in-class learning branch" and "semi-rebalancing branch", respectively. The pseudo-code of our proposed method can be found in Algorithm 1.

Algorithm 1 BiSTF algorithm.

Input: labeled set $L_{in} = \{(x_l, y_l)\}$, unlabeled set $U = \{(x_u, y_u)\}$, adaptive trade-off parameter α, threshold τ

Require: $U' = L_{in} \cup \hat{S}$

$\hat{S}$ is None
for t in $[1, num\_epochs]$ **do**
  for each minibatch B **do**
    $\mathbf{f}_l \Leftarrow x_{l \in L_{in}}$ (Through the "in-class learning branch" and GAP)
    $\mathbf{f}_s \Leftarrow x_{s \in U'}$ (Through the "semi-rebalancing branch" and GAP)
    $z = \alpha \mathbf{W}_l^\top \mathbf{f}_l + (1-\alpha)\mathbf{W}_s^\top \mathbf{f}_s$
    $\hat{\mathbf{p}}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_i}}$
    $L = \alpha E(\hat{\mathbf{p}}, y_l) + (1-\alpha) E(\hat{\mathbf{p}}, y_s)$
  end for
  if F_{update} is True **then**
    $\hat{U} = \{(x_u, \hat{y})\} \Leftarrow \hat{y} = f(max(p(y \mid x_u)) \geq \tau)$
    $\hat{S} \subset \hat{U}$ (Sampling the pseudo-labels to contain same distribution)
    update U'
  end if
  update α
end for
return f

4.2 Details in supervised training step

Differences between bilateral branches The input data for the in-class learning branch is from labeled set L_{in} through a uniform sampler, which can reserve the original characteristics of data distribution. Thus, in-class learning branch benefits the representation learning. While the purpose of semi-rebalancing branch is alleviating the extreme class imbalance and leveraging the unlabeled data to improve the quality of classifier, so its input is from the U' through a reversed sampler, which is the union set of L_{in} and sampled pseudo-label set $\hat{S}$ generated by generating pseudo-labels step. Specifically, for one class, the sampling possibility in the reversed sampler is proportional to the reciprocal of the number of available samples.

α in cumulative learning strategy Different from BBN, the adaptive trade-off parameter α directly affects the model's ability to bias towards in-class data and re-balance the data by controlling the weights for f_l and f_s. The working mechanism is that the universal patterns in in-class labeled data is first learned and then the model pays more attention to selected pseudo-labels and minority classes. We follow the settings in BBN, the α decreases with the progress of epoch in the way of Parabolic decay.

4.3 Details in generating pseudo-labels step

Stochastic epoch update Specifically, at the beginning of training, to avoid introducing out-of-class noise to affect the model performance, we perform pseudo-labeling on the unlabeled dataset with a small probability and selectively add them to the training phase, and then the probability of epoch updating gradually increases.

Contain same distribution If pseudo-labels are generated by the model trained on imbalanced data, when they join the training, they will exacerbate the bias. Consequently, in the selection process, we follow the strategy of keeping the selected pseudo-labeled data distribution consistent with the in-class labeled data distribution, which avoids that the data distribution of U' is gradually biased towards majority classes.

5 Experiments

5.1 Datasets

LT-Cifar10 and LT-Cifar100 For fair comparisons, we use long-tailed CIFAR dataset introduces in Zhou et al. (2020), Sohn et al. (2020), containing CIFAR10 and CIFAR100 with 10 and 100 categories, respectively. Both two datasets split 50000 images for training set and 10000 images for validation set. As for the class imbalance, we use γ to measure the degree of imbalance, which illustrates the proportion of the most class to the least class in the dataset, e.g. $\gamma = \frac{N_{max}}{N_{min}}$. In our experiments, γ is set to 10, 50 and 100. After suffering the data, we select $\beta = 10\%$ and 30% of data from training set to build the labeled data, and the rest to build unlabeled data. To evaluate the efficacy of model, the validation set remains balanced. Besides, in order to actually evaluate the generalization performance of the model, we use all validation set for verification, which may be different from other SSL algorithms' settings.

Semi-iNat Different from standard SSL image recognition benchmarks, Semi-iNat dataset (Su & Maji, 2021) is full of challenges for semi-supervised recognition with fine-grained categories, a long-tailed distribution of classes, and domain mismatch between labeled and unlabeled data, as shown in Fig. 8. Semi-iNat contains images of species from three kingdoms in the natural taxonomy: Animal, Plants, and Fungi (Table 1).This dataset is at a larger scale for a total of $\approx$ 330k images. Specially, it is split into two sets C_{in} with 810 in-class species and C_{out} with 1629 out-of-class species. For each species in C_{in}, 5/10/10 images are selected for validation, public test, and private test set. Among the rest of the images, around 90% of the images are sampled as unlabeled data U_{in} and the rest as labeled data L_{in}. In addition, each class is guaranteed to have at least 5 labeled images. For species in C_{out}, all of them are included in U_{out}. The two sets of unlabeled data are then combined $U = U_{in} \bigcup U_{out}$, and more challengingly, no domain labels are provided but coarse taxonomic labels for the unlabeled data

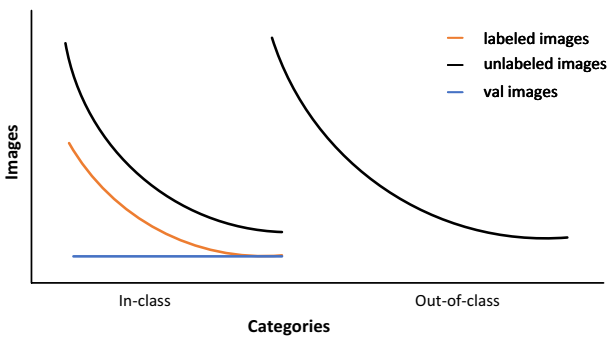


Fig. 8 Class imbalance and domain mismatch on Semi-iNat Dataset. Both labeled and unlabeled sets are class-imbalanced and domain-shifted, where the most majority class has 16× more samples than the most minority class. The validation and test set remains balanced

Table 1 The number of species in the taxonomy

Kingdom	Phylum	C_{in}	C_{out}
Animalia	Mollusca	11	24
	Chordata	113	228
	Arthropoda	301	605
	Echinodermata	4	8
Plantae	Tracheophyta	336	674
	Bryophyta	6	12
Fungi	Basidiomycota	29	58
	Ascomycota	10	20

For each phylum, around one-third of the species are selected for the in-class set C_{in} and the rest for the out-of-class set C_{out}

are provided, such as kingdom and phylum. In addition, a parameter that is critical to our experiments is the maximum imbalance rate with the value of $\gamma = 16$. In this paper, the official splits of train, validation, public test, and private test set are utilized for fair comparisons.

In order to more intuitively demonstrate the class imbalance data distribution and domain mismatch of Semi-iNat, we introduce the distribution graph of Semi-iNat dataset, in which the first 810 classes are in-class data and the last 2439 classes are out-of-class data, and all of them show long-tail distribution, as shown in Fig. 9.

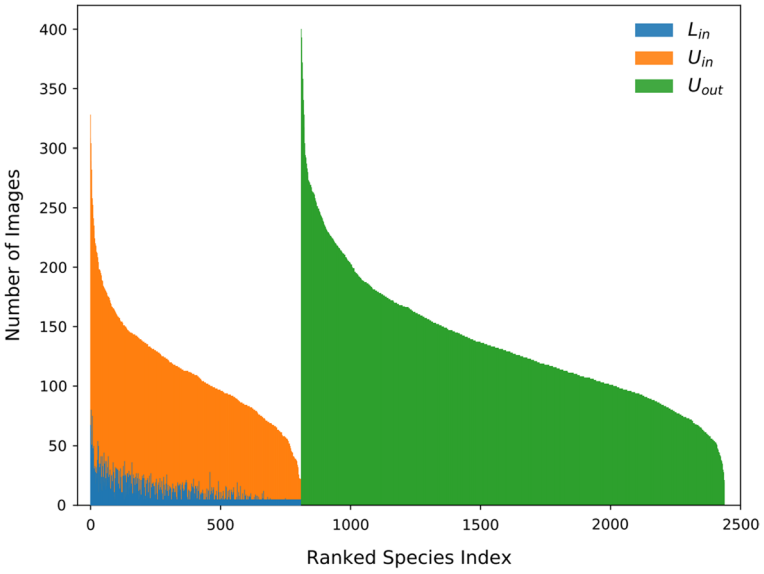


Fig. 9 Class distribution of Semi-iNat dataset. Su and Maji (2021) note that the distinction between U_{in} and U_{out} are only shown here for visualization. We did not provide the domain labels for unlabeled data

5.2 Implementation details

To be fair, we use the same network architecture and training protocol, including the data preprocessing, optimizer, learning rate schedule. In detail, we train the ResNet-50 (He et al., 2016) leveraged by the official as our backbone network by standard mini-batch stochastic gradient descent (SGD) with momentum of 0.9, weight decay of 1×10^{-4}. Our experiments follow the data augmentation strategies proposed in Sohn et al. (2020): resize image to 224×224, random horizontal flip with 50% probability, randomly resizecrop a 224×224 patch from the original image or its horizontal flip with scale from 0.2 to 1.0 and ratio from 0.75 to 1.333333333, as well as RandAugment (Cubuk et al., 2020) keeping the same settings in Sohn et al. (2020).We train all the models on a single NVIDIA A100 GPU with batch size of 64. The initial learning rate is set to 0.01 and the learning rate during subsequent training is decayed by multistep scheduler. Experiments are also conducted using MindSpore.

5.3 Main results

Comparisons on balanced semi-supervised dataset First, we compare BiSTF with various SSL methods on CIFAR10 and CIFAR100, and present the results in Table 2. Specially, we randomly select $\beta = 10\%$, 30% and 80% of training data as labeled data, the rest as unlabeled data, and all validation set for verification. Our results are state-of-the-art on all settings except for $\beta = 10\%$ on CIFAR10, where FixMatch performs a bit better. It shows that on the traditional SSL benchmarks, our model also achieves a performance gain to a certain extent.

Comparisons on imbalanced semi-supervised dataset We compare the performance of all abovementioned SSL algorithms along with BiSTF on LT-CIFAR10 and LT-CIFAR100 in Table 3. We omit results for Pseudo-Labeling, Mean Teacher, and MixMatch since the performances are poor on LT-CIFAR100. We find that the existing SSL algorithms perform severely on imbalanced datasets, and their accuracies decrease significantly with increasing imbalance ratio. In contrast, we also observe that BiSTF works particularly well on imbalanced data with high imbalance ratio.

For example, BiSTF achieves 67.01% accuracy on LT-CIFAR10 with $\beta = 30\%$ and $\gamma = 100$. As a point of reference, the highest accuracy achieved by FixMatch, used same experimental settings with BiSTF, is 59.45%.

Table 2 Accuracy for CIFAR10, CIFAR100

Method	CIFAR10			CIFAR100		
	$\beta = 10\%$	$\beta = 30\%$	$\beta = 80\%$	$\beta = 10\%$	$\beta = 30\%$	$\beta = 80\%$
Pseudo-labeling	68.15	75.94	77.98	19.77	44.73	61.21
Mean teacher	76.28	80.89	81.58	23.24	45.11	61.52
MixMatch	78.91	82.46	83.85	37.21	52.63	66.34
FixMatch (RA)	**81.15**	84.51	86.26	48.86	58.34	61.73
BiSTF	80.02	**85.16**	**88.47**	**50.37**	**60.58**	**67.58**

Best values are given in bold

All baseline models [Pseudo-Labeling (Lee, 2013), Mean Teacher (Tarvainen & Valpola, 2017), MixMatch (Berthelot et al., 2019), and FixMatch (Sohn et al., 2020)] are tested using the same settings

Table 3 Accuracy for LT-CIFAR-10, LT-CIFAR-100 of FixMatch, CReST (Wei et al., 2021) and BiSTF

Method	LT-CIFAR-10						LT-CIFAR-100					
	$\beta = 10\%$			$\beta = 30\%$			$\beta = 10\%$			$\beta = 30\%$		
	$\gamma = 10$	$\gamma = 50$	$\gamma = 100$	$\gamma = 10$	$\gamma = 50$	$\gamma = 100$	$\gamma = 10$	$\gamma = 50$	$\gamma = 100$	$\gamma = 10$	$\gamma = 50$	$\gamma = 100$
FixMatch (RA)	64.23	45.18	41.47	**78.91**	67.96	59.45	29.1	17.95	13.24	46.65	34.25	29.50
CReST	64.45	49.12	44.67	78.87	70.64	63.43	33.56	21.63	17.95	48.01	36.84	31.45
BiSTF	**66.61**	**57.02**	**47.62**	78.79	**72.73**	**67.01**	**35.06**	**24.41**	**21.40**	**48.20**	**37.11**	**32.74**

Best values are given in bold

It can be seen that our proposed BiSTF has higher recognition performance than FixMatch in most cases. However FixMatch and our model have similar accuracies in the partitioning of datasets with a high proportion of labeled datasets and a low proportion of imbalance, and we attribute this situation to the fact that the ten categories of the LT-CIFAR-10 dataset are inherently easy to distinguish, and the $\beta = 30\%$ labeled dataset and the $\gamma = 10$ imbalance rate are sufficient for both FixMatch and BiSTF to recognize effectively.

Comparisons on realistic large-scale dataset First, we compare our model with baseline reproduced according to the official and FixMatch, and present the results in Table 4. Due to the utilization of data augmentation, the accuracy of the baseline reproduced by us is improved by 6.21% over the result by the official that is for reference only. Although FixMatch performs reasonably well on public and private test set, its improvement is not as obvious as in the basic SSL benchmark. In contrast, BiSTF improves the accuracy of FixMatch and achieves as much as 1.79% absolute performance gain.

We also observe that our model works particularly well and achieves 1.75% and 1.79% accuracy gain on public test and private test data, respectively. We hypothesize the reason is that by stochastic epoch update strategy our model finds more correctly pseudo-labeled samples to augment the labeled set instead of iteration update strategy of FixMatch.

In order to confirm our hypothesis, we apply BiSTF to calculate the correctness of the pseudo-labels on Salvia nemorosa speices in different epochs, which is compared with the results illustrated in Sect. 3. We observe that the number of pseudo-labels sampled by BiSTF has increased, however the error rate of pseudo-labels has decreased, which proves that BiSTF facilitates the selection of pseudo-labels, shown in Fig. 10. In addition, by observing the performance on the validation set, it shows that the ability of BiSTF to learn fine-grained features has been improved, shown in Fig. 11.

We further report the performance of BiSTF with different backbones and image sizes in Table 5. After resizing images to 600×600, this paper first directly evaluates several common backbones, including Resnet101 (He et al., 2016), ResneXt101 (Xie et al., 2017), EfficientNet-b5-7 (Tan & Le, 2019). All the backbones are able to further boost the performance by another few points, resulting in 5.0–10.25% absolute accuracy improvement compared to BiSTF with backbone of Resnet50. Introducing the noisy student (Xie et al., 2020) to EfficientNet, the results of BiSTF can be further improved. Among these backbones, EfficientNetb7_ns achieves the best performance, so we take it as the final baseline. Finally, applying fivefold cross-validation after expanding the validation data to training data, our BiSTF model further gives accuracy gains, producing the best results.

Table 4 We compare BiSTF with baseline methods

	Val (%)	Public test	Private test
Official	31.00	–	–
Baseline	37.21	34.86%	35.60%
Fixmatch	37.53	36.15%	36.40%
BiSTF	**39.31**	**37.90%**	**38.19%**

Best values are given in bold

All models are trained with the same settings for fair comparison

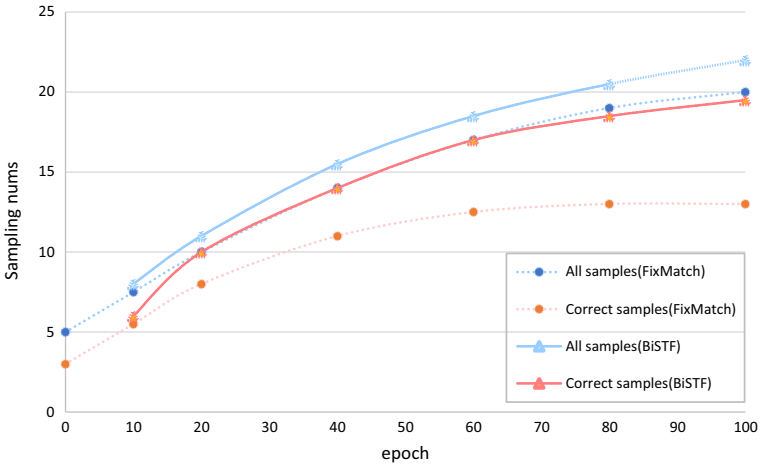


Fig. 10 Performance of BiSTF on Semi-iNat compared with FixMatch. BiSTF increases the sampling number of pseudo-labels, and also improves the accuracy of pseudo-labels

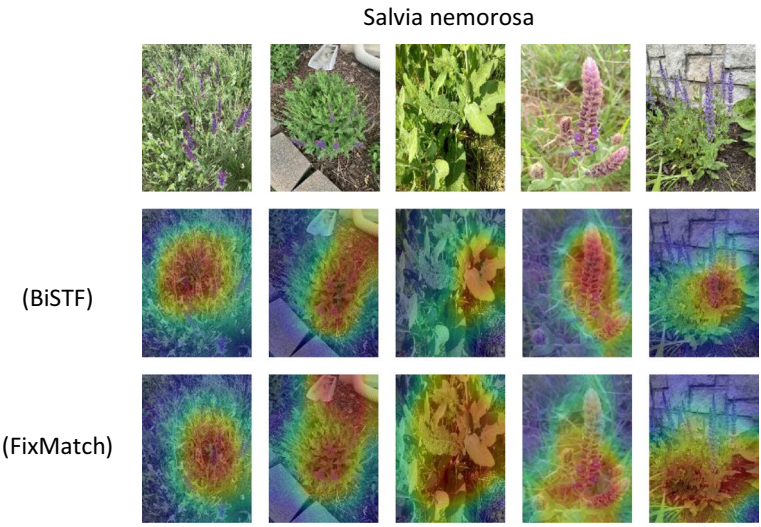


Fig. 11 Class activation mappings of BiSTF on Salvia nemorosa speices compared with FixMatch. By observing class activation mappings, BiSTF can focus more on pivotal fine-grained features and ignore minor elements than Fixmatch

5.4 Ablation studies

We perform an extensive ablation study to evaluate and understand the contribution of critical component in BiSTF. The experiments in this section are all performed with BiSTF on Semi-iNat.

Table 5 Top-1 Accuracy (%) of BiSTF with different backbones on Semi-iNat

	Val	Public test (%)	Private test (%)
ResNet50	44.33%	42.18	42.85
ResNet101	49.42%	47.03	47.52
ResNeXt101	61.86%	59.96	60.12
EfficientNet-b5	53.78%	52.33	53.10
EfficientNet-b7	64.77%	62.63	63.47
EfficientNet-b7_ns (expand-data)	–	**77.00**	**77.48**

Best values are given in bold

Effect of update probability. BiSTF introduces the "Stochastic epoch update strategy" that controls the update frequency. Figure 12 shows how update strategy influences performance over generations.

When P = 1 in the whole process of training, our method updates the dataset in each epoch. Besides, two other update strategies are also tested. Specifically, the update probability varies linearly and in separated stages with the epoch, respectively. To show the source of accuracy improvements, in Table 6 we present accuracy on the validation set of Semi-iNat. The results suggest that various stochastic update strategies affect the learning ability of the model to a certain extent by changing the pseudo-labeled data sampled for training, where the linear strategy produces the best result.

Effect of α. In the cumulative learning strategy, the adaptive trade-off parameter α affects the ability to bias in-class data and rebalance the data. We explore several different strategies to adjust α according to the number of training epochs, such as linear decay, cosine decay, etc., as shown in Table 7. It can be seen that the learning strategy of learning "in-class learning branch" first and then "semi-rebalancing branch" gives better results than the other strategies, among which the best strategy to generate α method is the proposed parabolic decay method.

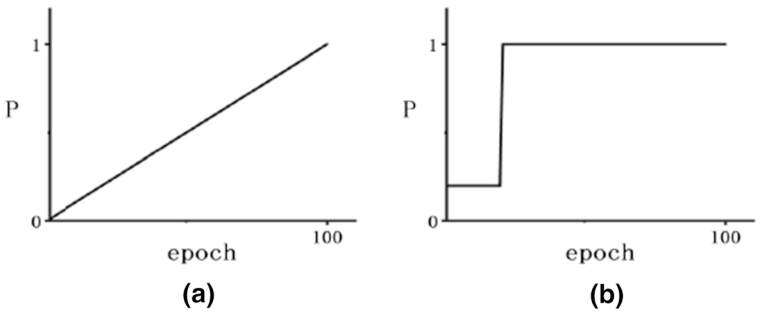


Fig. 12 Stochastic epoch update strategy

Table 6 Top-1 Accuracy (%) of BiSTF with different "Stochastic epoch update" strategies on Semi-iNat

Stochastic epoch update strategy	Val (%)
All	38.00
Seperated stage	38.59
Linear	**39.31**

Best values are given in bold

Update probability is **a** linear or **b** seperated stage with epoch

Table 7 Top-1 Accuracy (%) of different adaptor strategies of BiSTF on Semi-iNat

Adaptor	α	Val (%)
Equal weight	0.5	34.63
β-distribution	$Beta(0.2, 0.2)$	34.87
Parabolic increment	$(\frac{T}{T_{max}})^2$	35.65
Linear decay	$1 - \frac{T}{T_{max}}$	37.74
Cosine decay	$\cos(\frac{T}{T_{max}} \cdot \frac{\pi}{2})$	38.09
Parabolic decay (Ours)	$1 - (\frac{T}{T_{max}})^2$	**39.31**

Best value is given in bold

T is the current training epoch, T_{max} is the total number of training epochs

Effect of τ. On the one hand, high threshold can lead to higher quality pseudo-labels, but on the other hand, high threshold can also reduce the number of pseudo-labels and reduce the improvement of model performance. To better understand the role of threshold τ in BiSTF, we show the accuracy of the validation set under different threshold values in Table 8. In fact, our method is not particularly sensitive to the selection of τ values, and as we can see from the table, the effect of τ on the accuracy is within 0.1%. We believe this is mainly because our cumulative learning strategy and stochastic epoch update strategy make the pseudo-label generated by the model after it has already learned better representation ability does not have lower confidence in the prediction category.

Effect of stochastic epoch update. We apply iteration update and stochastic epoch update for BiSTF respectively. In the usual experimental setup, an epoch trains the entire labeled data, while an iteration selects only part of the labeled data for training. We count the accuracy of pseudo-labels at each epoch and obtain similar results as in Fig. 5, where the stochastic epoch update has a higher correct rate of pseudo-labels than the iteration update. Further, we visualize the class activation mappings for the 10th epoch, as shown in Fig. 13. And we find that iteration update lacking sufficient recognition of each class in the early stage, are more likely to introduce false pseudo-labels than stochastic epoch iterations, leading to poorer final performance of the model.

6 Conclusion

In this work, we present a bilateral-branch self-training framework, named BiSTF for domain-shifted and imbalanced semi-supervised fine-grained recognition. BiSTF is motivated by the observation that in addition to ignoring the subtle but discriminative features, existing SSL algorithms are vulnerable to class imbalance and domain mismatch. BiSTF

Table 8 Top-1 Accuracy (%) of different threshold of BiSTF on Semi-iNat

τ	Val (%)
0.75	39.30
0.9	39.31
0.95	39.31
0.99	39.28

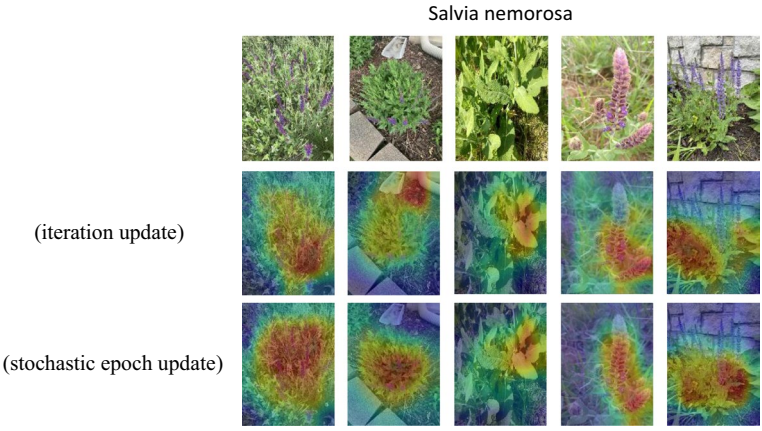


Fig. 13 Class activation mappings of BiSTF using iteration update or stochastic epoch update on the 10th epoch. Iteration update lacking sufficient recognition of each class in the early stage, are more likely to introduce false pseudo-labels than stochastic epoch iterations

iteratively refines a baseline SSL model with a labeled set expanded by adding pseudo-labeled samples from an unlabeled set, where pseudo-labeled samples contain same data distribution with the labeled dataset. Over epochs of self-training, the model becomes less biased towards majority classes and out-of-class data, focusing more on in-class data. Extensive experiments on Semi-iNat datasets demonstrate that the proposed BiSTF outperforms the existing state-of-the-art SSL algorithm.

The Semi-iNat dataset is labeled by naturalistic experts. There are still some limitations of our approach in the face of large-scale realistic datasets, such as the inability to have a large amount of noise in the labeled data. In fact, in the context of domain mismatch, when there is a lot of noise in the labeled data, the labels become less important and the semi-supervised framework will no longer be applicable.

Author contributions H.C. is first author. J.Y. is corresponding author. G.X. and Q.L. are co-authors.

Funding This work is sponsored by CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B), Natural Science Foundation of China (61906061), Anhui Province Key Research and Development Program (202104a05020007) and USTC Research Funds of the Double First-Class Initiative (YD2350002001).

Availability of data and materials The datasets used in our paper can be found at https://github.com/cvl-umass/semi-inat-2021.

Code availability Our code is available at https://github.com/HowieChangchn/BiSTF.

Declarations

Conflict of interest The authors declare that they have no conict ofinterest.

Consent to participate The authors agree to participate.

Consent for publication The authors agree to the publication of the data and images in this paper.

Ethics approval Not applicable.

References

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. (2019). Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106,* 249–259.

Byrd, J., & Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *International conference on machine learning* (pp. 872–881). PMLR.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. arXiv preprint arXiv:1906.07413

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 702–703).

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268–9277).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Hyun, M., Jeong, J., & Kwak, N. (2020). Class-imbalanced semi-supervised learning. arXiv preprint arXiv:2002.06815

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images

Laine, S., & Aila, T. (2016). Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning* (Vol. 3). ICML.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Miyato, T., Maeda, S.-I., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1979–1993.

Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-supervised self-training of object detection models

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., & Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Su, J.-C., & Maji, S. (2021). The semi-supervised inaturalist challenge at the fgvc8 workshop. arXiv preprint arXiv:2106.01364

Tan, M., & Le, Q. (2019). Efficientnet: rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780

Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F. (2021). Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10857–10866).

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., Le, Q. V. (2019). Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10687–10698)

Yang, Y., & Xu, Z. (2020). Rethinking the value of labels for improving class-imbalanced learning. arXiv preprint arXiv:2006.07529

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412

Zhou, B., Cui, Q., Wei, X.-S., & Chen, Z.-M. (2020). Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9719–9728).