Clue Me In: Semi-Supervised FGVC with Out-of-Distribution Data

Ruoyi Du, Dongliang Chang, Zhanyu Ma*, Yi-Zhe Song, and Jun Guo

Abstract

Despite great strides made on fine-grained visual classification (FGVC), current methods are still heavily reliant on fully-supervised paradigms where ample expert labels are called for. Semi-supervised learning (SSL) techniques, acquiring knowledge from unlabeled data, provide a considerable means forward and have shown great promise for coarse-grained problems. However, exiting SSL paradigms mostly assume in-distribution (i.e., category-aligned) unlabeled data, which hinders their effectiveness when re-proposed on FGVC. In this paper, we put forward a novel design specifically aimed at making out-of-distribution data work for semi-supervised FGVC, i.e., to “clue them in”. We work off an important assumption that all fine-grained categories naturally follow a hierarchical structure (e.g., the phylogenetic tree of “Aves” that covers all bird species). It follows that, instead of operating on individual samples, we can instead predict sample relations within this tree structure as the optimization goal of SSL. Beyond this, we further introduced two strategies uniquely brought by these tree structures to achieve inter-sample consistency regularization and reliable pseudo-relation. Our experimental results reveal that (i) the proposed method yields good robustness against out-of-distribution data, and (ii) it can be equipped with prior arts, boosting their performance thus yielding state-of-the-art results. Code is available at https://github.com/PRIS-CV/RelMatch.

1. Introduction

Progress on computer vision had been heavily reliant on having access to annotated data. This however represents a heavy constraint for the problem of fine-grained visual classification (FGVC), where labels can only come from experts, i.e., people who can tell the difference between two samples of the same category.

Figure 1. (a) Illustration of feature spaces for coarse-grained classification and fine-grained classification. The blue lines indicate the decision boundaries and the red circle represents the distribution of out-of-distribution category in the feature space. (b) The average Kullback-Leibler divergence of models’ prediction probabilities to the probability distribution centers.

“American Crow” and “Fish Crow”. This essentially means despite great strides made [9, 11–14, 27, 42, 44], our understanding of FGVC largely remains limited to fully supervised paradigms on categories where labels had been expertly curated for, e.g., Bird [41], Flower [32], Cars [21].

Elsewhere for coarse-grain problems, significant progress has already been made to address this “lack of label” phenomenon, notably through the means of semi-supervised learning (SSL) [15]. It can however be argued that as important as SSL is, the option of falling back to fully supervised remains conceivable for coarse-grain problems, i.e., more human (non-expert) labels can always be sourced with effort (and is arguably common practice in industry). The same nonetheless does not hold for FGVC, regardless of resource and effort – ultimately, there simply may never be enough experts available to label. Indeed, having sufficient labelled training data for FGVC might be a false argument to start with, thus making SSL an even more significant endeavour for FGVC.

Semi-supervised learning in the context of FGVC is however non-trivial. The difficulty is largely owing to the overwhelming presence of out-of-distribution samples in unlabeled data. This is easily observable in Figure 1(a), where when compared with coarse-grain, the feature space...
of the fine-grained classifier is a lot more compact making new out-of-distribution unlabelled data (in red circle) much easier to be confused. To further verify our point, we simulate two situations where we make certain categories either in-distribution or out-of-distribution to the model and measure how dispersed the probability distributions of model predictions are via Kullback-Leibler Divergence. From Figure 1(b), we can conclude that (i) the model generally yields more inconsistent predictions for an out-of-distribution category, and (ii) this phenomenon is considerably more significant for the fine-grained model. This essentially renders most of the existing SSL work that rely on pseudo-labeling [5, 23, 31] or consistency regularization [2, 22, 35, 39] significantly less effective when re-purposed for FGVC. This is because that they mostly work with the assumptions that (i) acquiring in-distribution unlabeled data is relatively easy, and (ii) out-of-distribution is not as salient a problem given the coarse decision space.

In this paper, we put forward a novel design that specifically aims at making out-of-distribution data work for the problem of FGVC — i.e., to “clue them in”. To reason around the inherently tight decision space of FGVC is however not straightforward — we need all the help we can get. For that, we take inspiration from [7] to utilize a label hierarchy of fine-grained categories (e.g., the phylogenetic tree of “Aves”). As shown in Figure 2, the phylogenetic tree provides an underlying structure that all bird species naturally obey. For example, the model can not tell the unlabeled bird is a “Collared Plover”, since it is totally out of the training label space. But the model can justifiably infer that it belongs to the same genus as “Semipalmated Plover” but not “Semipalmedated Plover”, i.e., owning the relation of “Same Genus, Different Species” with “Semipalmedated Plover” as defined in Figure 2. It follows that our main innovation lies in predicting sample relations within this hierarchy other than operating on individual samples.

More specifically, we utilize a simple multi-layer perceptron (MLP) that takes representations of sample pairs as input to work as a relation classifier. Predicting the relation of two birds can be regarded as a close-set classification problem because the defined phylogenetic tree includes all expert bird species. Therefore, by replacing the instance-level prediction goal in previous pseudo-labeling techniques with this relation-based prediction, we construct a novel relation-based pseudo-labeling strategy that importantly yields a common label space for both in-distribution and out-of-distribution data.

We further propose two novel strategies taking full advantage of the tree-structured hierarchy to better align in-distribution and out-of-distribution samples. First, we re-purpose the concept of rooted triplet [20], and introduce a triplet consistency mechanism to enable inter-sample regularization of samples (see Figure 2 for an example of rooted triplet). We conjecture that for any three unlabeled samples that form a rooted triplet, exactly one of them will form consistent relations with the other two (see Figure 3 (b) and Section 3.3.1 for proof). Second, in the scenario where two of three leaf samples are labeled, we can further infer consistency between unlabeled relations and ground-truths relations (see Figure 3 (d) and Section 3.3.3). Through this strategy we call label transfer, we significantly improve the quality of pseudo-relations learned which in turn helps with alignment.

We conduct experiments on semi-supervised FGVC benchmark datasets published by [37] under both their conventional setting and the realistic setting. Compared with state-of-the-art semi-supervised and self-supervised methods, we achieve better or comparable performance. In addition, we show that (i) the proposed framework can be easily integrated with prior arts in a plug-and-play manner, further boosting their performance, and (ii) our relation-based approach exhibits good robustness against out-of-distribution data and is able to achieve performance gains when trained from scratch. Ablative study further shows our method can also work with categories that are completely out of our phylogenetic tree (e.g., the whole of ImageNet [34]).

2. Related Work

2.1. Fine-Grained Visual Classification

Fine-grained visual classification (FGVC) tends to identify sub-categories that belong to a general class, e.g., distinguishing “American Crow” and “Fish Crow”. With a history of almost two decades [4], it has become a core problem of computer vision with many excellent works related (well summarized in a most recent survey [43]).

Early works mostly heavily relied on dense bounding-box annotations [3, 6] to perform part detection, due to the relatively large intra-category variations. After that, some weakly-supervised methods emerged that only trained with category labels [14, 42, 46]. However, due to the expertise needed for giving fine-grained labels, FGVC approaches are still limited to achieving further progress and also require higher labor costs in practice. Sparse data has become the bottleneck of FGVC in both academic research and industrial applications. Recently, researchers’ interests have shifted to saving expert effort during training, e.g., visual recognition with small-sample [25, 26, 47], web-supervised learning [38, 45], and leveraging layer persons annotations [10]. In this paper, we introduce a new lens to FGVC and propose a semi-supervised framework specifically aiming at FGVC tasks with out-of-distribution data.

2.2. Semi-Supervised Learning

Semi-supervised learning (SSL) is a popular topic for coarse-grained classification tasks and attract the attention
of many researchers. Most related arts to date follows two tracks: (i) Pseudo-labeling [5, 23, 31] that utilizes model predictions to generate synthetic labels for unlabeled data as training targets, which is also regarded as implicit entropy minimization [15], and (ii) Consistency Regularization [2, 22, 29, 35, 39] that applies stochastic disturbances on the input data or the model itself and encourages consistent model predictions can be given. Very recently, [36] and its variants [19, 24] broke the boundary between two trends and proposed a simple yet effective new paradigm by combining pseudo-labeling and consistency regularization.

Although prominent strides have been made, most of the SSL methods are evaluated and analyzed under an ideal setting that the categories of labeled data and the latent categories of unlabeled data are perfectly aligned. The work of [33] first put forward this very question and pointed out that when it comes to realistic settings where out-of-distribution unlabeled data is present, performances of state-of-the-art techniques show drastic degradation with almost no exception. The very recent work of [37] is most relevant to ours. Following [33], it evaluated state-of-the-art SSL methods in the FGVC task and obtained the same conclusion.

In this paper, as verified in Figure 1 (b), we argue that the devil is in the prediction probability space – it is hard for a model to make consistent predictions on samples with novel categories, which is obstructed to most existing SSL methods that operate on model predictions via whether pseudo-labeling, consistency regularization, or both. This is very intuitive since: (i) for pseudo-labeling methods, inconsistent predictions directly lead to meaningless artificial labels, and (ii) for consistency regularization methods, while model predictions sorely changing across samples, separately pursuing prediction consistency of individual sample also contribute less to discriminative knowledge mining.

3. Methodology
3.1. Overview

In this paper, focusing on tackling the semi-supervised fine-grained classification challenge with out-of-distribution data, we proposed RelMatch. It is composed by three main components: the feature exacter $F(\cdot)$, the category predictor $P_c(\cdot)$, and the relation predictor $P_r(\cdot, \cdot)$. Then we can use $P_c(F(\cdot))$ for category classification of individual samples and $P_r(P_c(F(\cdot)), P_c(F(\cdot)))$ for relation classification of sample pairs. Similar to existing methods, RelMatch simultaneously learns with labeled and unlabeled data. During the training phase, each batch both contains labeled samples $\{(x_i, y_i)\}_{i=1}^{N} \in X$ and unlabeled samples $\{u_j\}_{j=1}^{\mu N} \in U$, where $y_i$ are ground truth labels and $\mu$ is the ratio between the number of labeled samples and unlabeled samples. Generally, $\mu > 1$ for better benefiting from the huge quantity of unlabeled samples. At each iteration, RelMatch optimizes with three losses: (i) a supervised category classification loss $L_c$ to optimize $F(\cdot)$ and $P_c(\cdot)$, (ii) a supervised relation classification loss $L_r$ to optimize $P_r(\cdot, \cdot)$, and (iii) an unsupervised relation prediction loss $L_u$ to optimize $F(\cdot)$ and $P_r(\cdot)$. Then the overall loss function is:

$$L_{total} = L_c + L_r + L_u, \quad (1)$$

Specifically, $L_c$ is a standard cross-entropy loss:

$$L_c = -\frac{1}{N} \sum_{i=1}^{N} y_i \times \log(P_c(F(x_i))), \quad (2)$$
and the details of $L_c$ and $L_u$ are introduced in Section 3.2 and Section 3.3 respectively.

3.2. Relation-Based Pseudo-Labeling

3.2.1 Phylogenetic Tree

As mentioned above, due to the relative compact feature space of the fine-grained classifier, out-of-distribution data, whose latent labels totally out of the training label space, hinder the effectiveness of previous SSL techniques. Instead of directly predicting the categories of individual samples, we need a flexible tool to "clue" all these data in. Following the label hierarchy in [7], in this work, we leverage the phylogenetic tree of living beings as a bridge between in-distribution and out-of-distribution data. It can be regarded as a rooted tree where categories at different classification levels (e.g., Class, Order, Family, Genus, and Species) are nodes with different depths. After that, as illustrated in Figure 2, relations can be discretely defined within a close-set (e.g., "Same Family, Different Genus", "Same Genus, Different Species", etc.). These relations can be expressed in a more formulaic manner. Borrowing the concept of lowest common ancestor (LCA) in graph theory, for a pair of samples $x_i$ and $x_j$ with category labels $y_i$ and $y_j$, their LCA can be formulated by $LCA(y_i, y_j)$. Then the relation pair can be represented by the depth of their LCA $Dep(LCA(y_i, y_j))$. In fact, the LCA depth of a pair of samples is proportional to their distance along with the route on the tree and can be regarded as a measurement of their similarity. In the subsequent parts of this article, we let $S(y_i, y_j) = Dep(LCA(y_i, y_j))$ for simplifying.

3.2.2 Relation Prediction

Just to reemphasize, the purpose of RelMatch is conducting individual category prediction, and all designs for relation prediction should serve this ultimate goal. Hence, we let the relation predictor $P_r(\cdot, \cdot)$ take a pair of predictions from $P_c(\cdot)$ as sample representation for input, which enables both the feature extractor $F(\cdot)$ and the classifier $P_c(\cdot)$ can be optimized by the relation-based supervision. Specifically, let $p = P_c(F(x))$ be the category prediction probability, where $p \in \mathbb{R}^C$ and $C$ is the number of categories. Then, with $p_i$ and $p_j$ as the input probability pair, the relation predictor first use vector outer product to model their element-wise correlation as $m_{i,j} = p_i^T \times p_j$, $m_{i,j} \in \mathbb{R}^{C \times C}$. After this, $m_{i,j}$ is multiplied by a learnable transfer matrix $w_t \in \mathbb{R}^{C \times C \times R}$ along with the softmax function to obtain the final relation prediction probabilities. The predicted results $r_{i,j}$ can be expressed as:

$$r_{i,j} = P_r(P_c(F(x_i)), P_c(F(x_j))) = \text{Softmax}(p_i^T \times p_j \times w_t),$$

where $r_{i,j} \in \mathbb{R}^R$ and $R$ is the number of discrete relations that equals to the depth of the phylogenetic tree.

During the training phase, we simply combine the labeled samples $\{(x_i, y_i)\}_{i=1}^N \in X$ and its reverse version at the batch dimension to form a sample pair set $\{(x_i, y_i, x_{N-i}, y_{N-i})\}_{i=1}^N \in B_x$. With $s_{i,j} = S(y_i, y_j)$ obtained by species labels as relation ground-truth labels, the supervised relation classification loss can be formulated as:

$$L_r = -\frac{1}{N} \sum_{i=1}^N \log s_{i,N-i} \times s_{i,N-i} \times \log(r_{i,N-i}).$$

Note that, when we randomly sample pairs from the phylogenetic tree, the occurrence probabilities of various relations are significantly different, which leads to a long-tailed distribution. To alleviate this problem, we re-weight the losses of pairs according to their ground-truth relations. And $\{w_i\}_{i=1}^R \in W$ is generated according to the statistical probability of each relation level’s occurrence, where $\sum_{i=1}^R w_i = 1$. 

3.2.3 Naive Pseudo-Labeling via Relation Prediction

With predictions of pair relations, we can simply replace the instance-based prediction in previous SSL methods (e.g., pseudo-labeling [23]) with the relation-based prediction (shown in Figure 3 (a)). It seems to be a straight yet intuitive solution that not only enables the model to better learn from out-of-distribution data but also keeps the merits of previous arts. For the relation-based pseudo-labeling, the optimization goal can be formalized as:

$$L'_u = -\frac{1}{N} \sum_{j=1}^{\mu N} \sum_{k=1}^{\mu N} \tau_{j,k} \times w_{s_{j,k}} \times \hat{s}_{j,k} \times \log(r_{j,k}),$$

where $\hat{s}_{j,k} = \arg\max(r_{j,k})$ indicates the relation-level pseudo-labels, and $\tau_{j,k} = \hat{s}_{j,k} > t$ is used for sample selection with the confidence threshold $t$.

For semi-supervised learning, it is all about learning the underlying structure of a large amount of unlabeled data [30]. So far, we only rely on the phylogenetic tree to build a common label space shared by in-distribution and out-of-distribution data. In the next part, we will take a further step and introduce how the tree structure of fine-grained categories can help us achieve better alignment.

3.3. Triplet Consistency

3.3.1 Consistency of Rooted Triplet

A rooted triplet is a distinctly leaf-labeled, binary, rooted, unordered tree with three leaves [20] (illustrated in Fig-
where $\hat{s}_{j,k} = \text{argmax}(r_{j,k})$, and $H(\cdot)$ is an arbitrary function for distance measurement. And when $H(\cdot)$ is in a re-weighted cross entropy form, i.e., the consistency regularization is achieved by giving pseudo-labels, it can be regarded as an improved version of Equation 5 by introducing interactions between samples. In the following discussion, we will use this form by default.

3.3.3 Label Transfer via Triplet Consistency

For an unlabeled rooted triplet $\{u_j, u_k, u_l\}$ with $\hat{s}_{j,k} > \hat{s}_{j,l}$, $\hat{s}_{k,l}$ is the pseudo-label for $r_{j,l}$ due to $r_{j,l} = r_{k,l}$. We notice that, with the triplet consistency, quality of the pseudo-label is decided by qualities of both the relation prediction $\hat{s}_{k,l}$ and the relation comparison between $\hat{s}_{j,k}$ and $\hat{s}_{j,l}$, where still remains room for improvement.

To approach the optimal solution, i.e., the proposed RelMatch, we re-sample the rooted triplet with two labeled samples and one unlabeled sample as $\{x_j, x_k, u_l\}$. Then instead of unreliable prediction, the pseudo-label $\hat{s}_{k,l}$ is replaced by the ground-truth $s_{k,l}$ now. And once the relation comparison is correct, an exact pseudo-label can be given. In this way, the advantage brought by triplet consistency is quite remarkable. It is all about building a bridge between labeled data and unlabeled data – relation-based labels can be transferred from labeled pairs to unlabeled pairs with a simple relation comparison (shown in Figure 3 (d)). Ultimately, we reduce the difficulty of generating pseudo-labels from a multi-class classification problem to a binary classification problem. Our unsupervised optimization goal via label transfer is:

$$L_u'' = \frac{1}{\mu N^3} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \mu_N \left( \mathbb{1}(\hat{s}_{j,k} > \hat{s}_{j,l}) H(r_{j,l}, \hat{s}_{k,l}) \right), \quad (6)$$

$$L_u'' = \frac{1}{\mu N^3} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \mu_N \left( \mathbb{1}(\hat{s}_{j,k} > \hat{s}_{j,l}) H(r_{j,l}, \hat{s}_{k,l}) \right), \quad (7)$$

where $\tau_{k,l} = \mathbb{1}(E(s_{j,k}) - E(s_{j,l}) \geq 1)$ serves as the condition function for sample selection, $E(s_{j,k}) = \sum_{n=1}^{R} r_{j,k,n} \times 0$ indicates the mathematical expectation of $s_{j,k}$, and $E(s_{j,l})$ is the same.
To summarize, RelMatch can be regarded as a combination of pseudo-labeling and consistency regularization with (i) relation-based pseudo-labels \textit{instead of} instance-based pseudo-labels, (ii) inter-sample consistency regularization \textit{instead of} intra-sample consistency regularization, and (iii) label transfer for pseudo-label generation \textit{instead of} directly predicting them.

4. Experiments

4.1. Datasets

We adopt two SSL-FGVC benchmark datasets Semi-Aves and Semi-Fungi released by [37] for our comparison experiments and ablation studies. Each of them consists of three sub-sets: labeled samples \((x_i, y_i) \in X\), in-distribution unlabeled samples \(u_j \in U_{in}\) which shares the same label space with \(X\), and out-of-distribution unlabeled samples \(u_k \in U_{out}\) which include categories that are novel but still belong to the phylogenetic tree. We experiment with \(U_{in} + U_{out}\) as unlabeled data to simulate realistic applications where we cannot tell whether unlabeled data are in-distribution or not. Consistent with general SSL works, we also keep the setting that only considers \(U_{in}\) as unlabeled data for unsupervised optimization. Statistics of two datasets are displayed in Table 1.

| Dataset      | Categories | Images           |
|--------------|------------|------------------|
|              | \(X_{in}\) | \(U_{in}\) \(U_{out}\) | \(X_{in}\) \(U_{in}\) \(U_{out}\) |
| Semi-Aves    | 200 200 800 | 6K 27K 122K      |
| Semi-Fungi   | 200 200 1194 | 4K 13K 65K       |

Table 1. Statistics of Semi-Aves and Semi-Fungi dataset.

4.2. Baseline Methods

To demonstrate the superiority of RelMatch, we include following methods as baseline for comparison:

(1) \textbf{Vanilla Supervised Baseline}: The model is trained with only labeled data \(X\) and the Cross-Entropy loss.

(2) \textbf{Pseudo-Labeling} [23]: Following the implementation settings in [33], labeled data and unlabeled data are 1 : 1 sampled. The method selects unlabeled samples with maximum prediction probabilities greater than a pre-defined threshold, and then generates one-hot pseudo-labels for them by predictions from the model itself.

(3) \textbf{Curriculum Pseudo-Labeling (CPL)} [5]: Curriculum labeling multiplies the model training to several phases. At each training phase, the model is trained from scratch with only currently labeled set in a supervised manner. And after each phase, a certain proportion of unlabeled samples with highest predictions will be pseudo-labeled and added into the labeled set. The whole training process will keep iterating until all samples are added into the labeled set.

(4) \textbf{FixMatch} [36]: FixMatch combines the idea of pseudo-labeling and consistency regularization. For unlabeled data, pseudo-labels are given by their weakly augmented versions and are used to supervise their strong augmented versions.

(5) \textbf{Self-Training} [37]: “Self-Training” is a widely used term, and in [37] it refers specifically to a knowledge distillation [18] based procedure. Firstly, a vanilla supervised model is trained with only labeled data \(X\) to be the teacher model. And then, a student model is supervised by cross entropy loss on both labeled data and unlabeled data. Labels of unlabeled data are obtained by teacher model’s prediction results.

(6) \textbf{MoCo} [16]: As claimed in [8], self-supervised model can also be a good semi-supervised learner. Restricted by limited computational resource, we adopt MoCo to train the image encoder due to its independence of large batch size. And the whole training process consist of unsupervised pre-training on unlabeled data followed by supervised fine-tuning on labeled data.

Additionally, [37] shows that combination of various SSL methods leads to further performance boosting. In this paper, we also include (7) \textbf{Self-Training + MoCo} as a baseline model.

4.3. Implementation Details

For fair comparisons, we follow the experiment settings in [37]. We use ResNet50 [17] as the backbone network for all experiments. Input images are random-resize-cropped to \(224 \times 224\) during training and simply resized to \(224 \times 224\) during testing. We use SGD with a momentum of 0.9 and a cosine learning rate decay schedule [28] for optimization. To perform comprehensive comparisons with state-of-the-art methods, we train RelMatch from scratch, ImageNet [34] pre-trained model, and iNaturalist 2018 (iNat) [40] pre-trained model. Note that, iNat is a large-scale fine-grained dataset that contains 8142 species with no overlapping to Semi-Aves and Semi-Fungi. Our model is trained for 100k iterations and 50k iterations for training from scratch and pre-trained models, which is approximately aligned with the training setting of FixMatch in [37]. Besides, the learning rate and weight decay are set to be \(0.01, 0.001\) for training from scratch and \(0.001, 0.0001\) for training from pre-trained models. The batch size is set as 32 with \(\mu = 10\). For all experiments, we show the mean value and the standard deviation of 5 independent runs.

4.4. Comparison with SOTA Methods

We show comparison results on Semi-Aves and Semi-Fungi datasets in Table 2 and Table 3 respectively. When models are trained from scratch, MoCo [16] shows the great advantage of leveraging abundant data. It achieves impressive performance on \(U_{in}\) and obtains further boosting with the participation of \(U_{out}\). While other semi-supervised
methods are less effective. Instead of directly predicting pseudo-labels, with the label transfer strategy, the proposed RelMatch is able to generate more reliable pseudo-labels without pre-training. Therefore, we surpass MoCo with a large margin on $\mathcal{U}_{\text{in}}$ and obtain the best performance when coupling with MoCo under both two settings.

Under the setting of transfer learning, FixMatch [36] performs better than RelMatch [37] in most cases. However, it applies pseudo-labeling as the key component and is doomed to suffer performance degradation when faced with out-of-distribution unlabeled data. In comparison, Self-Training [37] is much more robust to novel categories, despite a slight degradation. On the contrary, the proposed RelMatch with shared label space can easily obtain state-of-the-art results on its own with out-of-distribution data. And RelMatch also significantly outperforms all comparison methods when being equipped with FixMatch. In general, the prominent merits of RelMatch are three-fold: (i) through building a shared label space and enabling inter-sample consistency, RelMatch can leverage out-of-distribution data to mine useful knowledge, (ii) through the label transfer strategy, RelMatch can generate more reliable pseudo-labels, and (iii) as a higher-level constraint, RelMatch can offer complementary supervising signals and boost previous SSL techniques.

4.5. Ablation Studies
4.5.1 Different Relation-based Variants
To discuss the effectiveness of each component upon the relation-based pseudo-labeling, we conduct ablation studies about its variants introduced above: (A) Relation-based Pseudo-Labeling as introduced in 3.2.3, (B) Triplet Consistency Regularization as introduced in 3.3.2, and (C) Label Transfer (RelMatch) as introduced in 3.3.3. As shown in Table 4, all experimental results are obtained by training the model from scratch. One important discovery is that all these relation-based techniques successfully leverage out-of-distribution unlabeled data and overcome the performance degradation. In addition, with the proposed triplet consistency regularization and label transfer strategy, RelMatch delivers greater performance gains when compared with the naive relation-based pseudo-labeling, which indicates the superiorities brought by the inter-sample consistency and the reliable pseudo-labels via label transfer that enable better alignment.

| Table 2. Comparison with other SOTA methods on Semi-Aves dataset. The best results are marked in red, and the second best results are marked in blue. |
| --- |
| Method | From Scratch (%) | From ImageNet (%) | From Nat (%) |
| --- | --- | --- | --- |
| Baseline | 56.0 ± 0.4 | 47.8 ± 0.7 | 54.0 ± 0.4 | 44.8 ± 0.5 |
| Pseudo-Label | 53.7 ± 0.2 | 38.8 ± 0.3 | 53.7 ± 0.2 | 38.8 ± 0.3 |
| FixMatch | 45.6 ± 0.5 | 37.8 ± 0.6 | 45.6 ± 0.5 | 37.8 ± 0.6 |
| Self-Training | 42.2 ± 0.4 | 34.1 ± 0.5 | 42.2 ± 0.4 | 34.1 ± 0.5 |
| MoCo | 45.0 ± 0.3 | 38.1 ± 0.4 | 45.0 ± 0.3 | 38.1 ± 0.4 |
| MoCo + Self-Training | 40.1 ± 0.2 | 32.1 ± 0.3 | 40.1 ± 0.2 | 32.1 ± 0.3 |
| RelMatch | 34.5 ± 0.2 | 27.0 ± 0.3 | 34.5 ± 0.2 | 27.0 ± 0.3 |
| RelMatch + MoCo | 37.9 ± 0.1 | 30.9 ± 0.1 | 37.9 ± 0.1 | 30.9 ± 0.1 |
| RelMatch + FixMatch | 36.7 ± 0.2 | 28.4 ± 0.2 | 36.7 ± 0.2 | 28.4 ± 0.2 |

Table 3. Comparison with other SOTA methods on Semi-Fungi dataset. The best results are marked in red, and the second best results are marked in blue.

| Table 3. Comparison with other SOTA methods on Semi-Fungi dataset. The best results are marked in red, and the second best results are marked in blue. |
| --- |
| Method | From Scratch (%) | From ImageNet (%) | From Nat (%) |
| --- | --- | --- | --- |
| Baseline | 31.0 ± 0.4 | 24.7 ± 0.8 | 31.0 ± 0.4 | 24.7 ± 0.8 |
| Pseudo-Label | 19.4 ± 0.4 | 14.3 ± 0.5 | 19.4 ± 0.4 | 14.3 ± 0.5 |
| FixMatch | 31.4 ± 0.6 | 25.0 ± 1.2 | 31.4 ± 0.6 | 25.0 ± 1.2 |
| Self-Training | 32.7 ± 0.2 | 26.9 ± 0.2 | 32.7 ± 0.2 | 26.9 ± 0.2 |
| MoCo | 30.6 ± 0.2 | 23.0 ± 0.3 | 30.6 ± 0.2 | 23.0 ± 0.3 |
| MoCo + Self-Training | 30.4 ± 0.3 | 24.4 ± 0.3 | 30.4 ± 0.3 | 24.4 ± 0.3 |
| RelMatch | 38.4 ± 0.4 | 31.5 ± 0.4 | 38.4 ± 0.4 | 31.5 ± 0.4 |
| RelMatch + MoCo | 38.8 ± 0.5 | 32.5 ± 0.5 | 38.8 ± 0.5 | 32.5 ± 0.5 |
| RelMatch + FixMatch | 39.4 ± 0.5 | 33.6 ± 0.5 | 39.4 ± 0.5 | 33.6 ± 0.5 |

Table 4. Ablation studies about different relation-based variants.

| Table 4. Ablation studies about different relation-based variants. |
| --- |
| Method | From Scratch (%) | From ImageNet (%) | From Nat (%) |
| --- | --- | --- | --- |
| Baseline | 27.9 ± 0.4 | 20.6 ± 0.4 | 27.9 ± 0.4 | 20.6 ± 0.4 |
| Pseudo-Label | 50.6 ± 0.4 | 30.0 ± 0.4 | 50.6 ± 0.4 | 30.0 ± 0.4 |
| FixMatch | 46.7 ± 0.5 | 32.4 ± 0.5 | 46.7 ± 0.5 | 32.4 ± 0.5 |
| Self-Training | 42.0 ± 0.3 | 30.4 ± 0.3 | 42.0 ± 0.3 | 30.4 ± 0.3 |
| MoCo | 39.3 ± 0.2 | 30.4 ± 0.3 | 39.3 ± 0.2 | 30.4 ± 0.3 |

Table 5. Ablation studies about the depths of phylogenetic tree that participate model training. For “Aves” and “Fungi”, the depths of their phylogenetic tree is 5 and 7, respectively.

4.5.2 Depth of The Phylogenetic Tree
The phylogenetic tree plays the most fundamental role in the proposed method. In this section, we conduct ablation studies about different tree depths. Specifically, we progressively reduce the depth of the tree from leaves to the root.
4.5.3 Affect of Training Iterations

To better understand the relation between model performances and numbers of training iterations, we conduct experiments with iteration numbers of \{1K, 5K, 10K, 15K, 20K\}. All models are trained from scratch in this section. As shown in Figure 4, we can conclude that model performances keep increasing with the increase of training iteration numbers and the margin of improvement will gradually decrease. Note that, as the number of iterations increases, the model will benefit more from the large amount of out-of-distribution data.

4.6. Discussion

How could RelMatch improve the quality of pseudo-labels?

To answer this question, we re-split the labeled set of Semi-Aves and Semi-Fungi to quantitatively evaluate pseudo-label quality. Specifically, we divide the 200 labeled categories in half, then we regard one half as in-distribution categories and the other half as out-of-distribution categories. With the divided categories, we can split the training set and the testing set accordingly. We train the model on the in-distribution training set and test on the in-distribution and out-of-distribution testing set separately. As shown in Table 6, when the model predicts relations of two unlabeled samples, the prediction accuracy only drops about 5% on out-of-distribution data, which solidly justifies the superiority of our relation-based prediction. When it comes to predicting relations between labeled and unlabeled data, the prediction accuracy naturally improves. Note that, this is not our contribution but just preparation for label transfer. And when we replace directly predicting relations with the proposed label transfer strategy via the triplet consistency, the prediction accuracy obtains significant improvement, and the gap between in-distribution and out-of-distribution accuracy continues to decrease, which demonstrates that the label transfer strategy itself does make pseudo-labels more reliable.

Can the learned tree structure generalize to other hierarchical data?

Till now we have only verified that our model can handle novel categories belonging to the same underlying tree. Here we devolve further and ask – can the learned tree structure generalize to fine-grained categories that are beyond the original phylogenetic tree of Aves/Fungi? To answer that, we experiment with iNat \[40\] and ImageNet \[34\] datasets as unlabeled data. More specifically, we use the variant of RelMatch without label transfer, \textit{i.e.}, the triplet consistency regularization introduced in Section 3.3.2. Results are shown in Table 7. It is interesting to see that these datasets do seem to contribute meaningfully and yield performance gains, despite with a smaller margin when compared with unlabeled data from the same latent tree. We attribute this to the proposed relation-based prediction still being able to be exploited as a higher-level constraint in other hierarchical datasets.

5. Conclusion

In this paper, we approached the fine-grained visual classification problem via a semi-supervised learning setting. Our key contribution is how best to train with out-of-distribution data. Our solution rests with leveraging the underlying tree structure of fine-grained categories to build a relation-based common label space. We further introduced a triplet consistency regularization to help with in-distribution and out-of-distribution alignment. We evaluated the proposed method on semi-supervised FGVC benchmark datasets and reported state-of-the-art results.
scale visual recognition challenge. *International journal of computer vision*, 2015. 2, 6, 8

[35] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In NeurIPS, 2016. 2, 3

[36] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In NeurIPS, 2020. 3, 6, 7

[37] Jong-Chyi Su, Zezhou Cheng, and Subhransu Maji. A realistic evaluation of semi-supervised learning for fine-grained classification. In CVPR, 2021. 2, 3, 6, 7

[38] Xiaoxiao Sun, Liyi Chen, and Jufeng Yang. Learning from web data using adversarial discriminative neural networks for fine-grained classification. In AAAI, 2019. 2

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In NeurIPS, 2017. 2, 3

[40] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In CVPR, 2018. 6, 8

[41] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report, California Institute of Technology*, 2011. 1

[42] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In CVPR, 2018. 1, 2

[43] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[44] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In CVPR, 2015. 1

[45] Chuanyi Zhang, Yazhou Yao, Huafeng Liu, Guo-Sen Xie, Xiangbo Shu, Tianfei Zhou, Zheng Zhang, Fumin Shen, and Zhenmin Tang. Web-supervised network with softly update-drop training for fine-grained visual classification. In AAAI, 2020. 2

[46] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In ICCV, 2017. 2

[47] Fangyi Zhu, Zhanyu Ma, Xiaoxu Li, Guang Chen, Jen-Tzung Chien, Jing-Hao Xue, and Jun Guo. Image-text dual neural network with decision strategy for small-sample image classification. *Neurocomputing*, 2019. 2