Debiased Learning from Naturally Imbalanced Pseudo-Labels

Xudong Wang[1] Zhirong Wu[2] Long Lian[1] Stella X. Yu[1]
[1]UC Berkeley / ICSI [2]Microsoft Research

Abstract

Pseudo-labels are confident predictions made on unlabeled target data by a classifier trained on labeled source data. They are widely used for adapting a model to unlabeled data, e.g., in a semi-supervised learning setting.

Our key insight is that pseudo-labels are naturally imbalanced due to intrinsic data similarity, even when a model is trained on balanced source data and evaluated on balanced target data. If we address this previously unknown imbalanced classification problem arising from pseudo-labels instead of ground-truth training labels, we could remove model biases towards false majorities created by pseudo-labels.

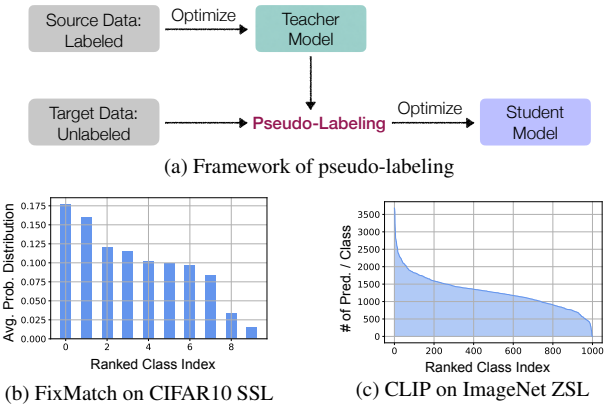
We propose a novel and effective debiased learning method with pseudo-labels, based on counterfactual reasoning and adaptive margins: The former removes the classifier response bias, whereas the latter adjusts the margin of each class according to the imbalance of pseudo-labels. Validated by extensive experimentation, our simple debiased learning delivers significant accuracy gains over the state-of-the-art on ImageNet-1K: 26% for semi-supervised learning with 0.2% annotations and 9% for zero-shot learning. Our code is available at: https://github.com/frank-xwang/debiased-pseudo-labeling.

1. Introduction

Real-world observations, as well as non-curated datasets, are naturally long-tail distributed [19, 61]. Imbalanced classification [10, 25, 64] tackles such data biases to prevent models from being dominated by head-class instances. Developing visual recognition systems capable of counteracting biases also has significant social impacts [37].

While existing methods focus on debiasing from imbalanced ground-truth labels collected by human annotators, we discover that pseudo-labels produced by machine learning models are naturally imbalanced, creating another source for widespread biased learning!

Pseudo-labels are highly confident predictions made by an existing (teacher) model on unlabeled data, which then become part of the training data for supervising the (student) model adaptation to unlabeled data (Fig. 1a). When the student model is the teacher model itself, the learning process is also known as *self-training* [4, 5, 30, 57, 70]. Pseudo-labeling is widely used in semi-supervised learning (SSL) [33, 57], domain adaptation [26, 40], and transfer learning [1].



(a) Framework of pseudo-labeling

(b) FixMatch on CIFAR10 SSL

(c) CLIP on ImageNet ZSL

Figure 1. We study the pseudo-labeling-based Semi-Supervised Learning (SSL) and transductive Zero-Shot Learning (ZSL), where both tasks require transferring semantic information learned from labeled source data to unlabeled target data via pseudo-labeling. Surprisingly, we find that pseudo-labels of target data produced by typical SSL and ZSL methods (i.e., FixMatch [57] and CLIP [49]) are highly biased, even when both source and target data are class-balanced or even sampled from the same domain.

We examine pseudo-label distributions in two common tasks. **1)** In zero-shot transfer learning (ZSL) where the source and target domains are different, a pretrained CLIP model [49] produces highly imbalanced predictions on the curated and balanced ImageNet-1K dataset, although the training set of CLIP is approximately balanced (Fig. 1c). More than 3500 instances are predicted as class 0, 3 times the actual number of samples in class 0. **2)** In semi-supervised learning where the source and target domains are the same, FixMatch [57] trained on labeled CIFAR10 images generates highly biased pseudo-labels on unlabeled images, although both the labeled and unlabeled sets are balanced (Fig. 1b).

That is, pseudo-labels created by machines are naturally imbalanced, just like ground-truth labels created by humans. If we address this previously unknown imbalanced classification problem arising from pseudo-labels instead of ground-truth training labels, we could improve model learning based

on pseudo-labels and remove the model bias towards false majorities created by pseudo-labels.

We propose a novel and effective debiased learning method with pseudo-labels, without any knowledge about the distribution of actual classification margins that are readily available to debiased learning with ground-truth labels [23, 34, 62]. It consists of an adaptive debiasing module and an adaptive marginal loss. The former dynamically removes the classifier response bias through counterfactual reasoning, whereas the latter dynamically adjusts the margin of each class according to the imbalance of pseudo-labels.

Validated by our extensive experiments, our simple debiased learning not only improves the state-of-the-art on ImageNet-1K by 26% for SSL with 0.2% annotations and 9% for ZSL, but is also a universal add-on to various pseudo-labeling methods with more robustness to domain shift. The imbalanced pseudo-labeling issue is even more severe when the unlabeled raw data is naturally imbalanced, and the model tends to mislabel tail-class samples as head-class. By applying debiased learning, we improve SSL performance under long-tailed settings by a large margin.

Our work makes four major contributions. **1)** We systematically investigate and discover that pseudo-labels are naturally imbalanced and create biased learning. **2)** We propose a simple debiased learning method with pseudo-labeled instances, requiring no knowledge of their actual classification margins. **3)** We improve the ZSL/SSL state-of-the-art by a large margin and demonstrate that our debiasing is a universal add-on to various pseudo-labeling models. **4)** We establish a new effective ZSL/SSL pipeline for applying vision-and-language pre-trained models such as CLIP.

2. Related Work

Semi-Supervised Learning integrates unlabeled data into training a model given limited labeled data. There are four lines of approaches. **1)** Consistency-based regularization methods impose classification invariance loss on unlabeled data upon perturbations [39, 55, 60, 69]. **2)** Pseudo-labeling expands model training data from labeled data to additional unlabeled but confidently pseudo-labeled data [4, 5, 30, 32, 57, 70]. **3)** Transfer learning trains the model first on large unlabeled data through self-supervised representation learning, e.g., contrastive learning, and then on small labeled data through supervised classifier learning [2, 13]. **4)** Data-centric SSL assumes that labeled data are not given but can be optimally selected among unlabeled data for labeling [65]. Focusing on this practical issue of labeled data selection turns out to bring substantial gains for SSL.

CReST [67] improves existing SSL methods on class-imbalanced data by leveraging a class-rebalanced sampler, which samples more frequently for the minority class according to the labeled data distribution. CReST does not work when the labeled data is balanced. In contrast, our approach does not assume any prior distribution for the labeled set.

Although previous literature has achieved tremendous success in SSL, the implicitly biased pseudo-labeling issue in SSL is previously unknown and has not been thoroughly analyzed, which, however, has a great impact on the learning efficiency. The focus of this work is on proposing a simple yet effective debiasing module to eliminate this critical issue.

Zero-shot Classification refers to the problem setting where a zero-shot model classifies images from novel classes into correct categories that the model has not seen during training [48, 51, 63]. Several strategies have been considered from various sets of viewpoints: **1)** hand-engineered attributes [16, 28]; **2)** pretrained embeddings that incorporate prior knowledge in form of semantic descriptions of classes [17, 56]; **3)** modeling relations between seen and unseen classes with knowledge graphs [24, 41]; **4)** learning generic visual concepts with vision-language models, allowing zero-shot transfer of the model to a variety of downstream classification tasks [8, 49].

Long-Tailed Recognition (LTR) aims to learn accurate "few-shot" models for classes with a few instances, without sacrificing the performance on "many-shot" classes, for which many instances are available. **1)** re-balancing/re-weighting method τ-norm [25] tackles LTR problem by giving more importance to tail classes; **2)** margin-based method LDAM [10] proposes a label-distribution-aware margin loss to improve the generalization of minority classes by encouraging larger margins for tail classes; **3)** post-hoc adjustment approach modifies a trained model's predictions according to the prior knowledge of class distribution, such as LA [38], or pursues the direct causal effect by removing the paradoxical effects of the momentum, such as Causal Norm [59]; **4)** ensemble-based approach RIDE [64] optimizes multiple diversified experts and a dynamic expert routing module to reduce model bias and variance on long-tailed data.

In stark contrast to previous works on LTR which either requires the prior knowledge of class distribution or are applied post-hoc to a trained model, the proposed debias module does not require any prior knowledge and focuses on the biased pseudo-labels issue which is previously unknown.

3. Pseudo-Labels are Naturally Imbalanced

In contrast to previous work that concentrated on biases caused by trained on *imbalanced* data, our focus is on pseudo-label biases, even when trained on *balanced* data. In this section, we provide an analysis of this previously unknown issue hidden behind the tremendous success of FixMatch [57] on SSL and CLIP [49] on ZSL, both of which require the use of "pseudo-labeling" to transfer knowledge learned in source data to target data.

We first describe the backgrounds for pseudo-labeling approaches and then analyze their bias issue. We attribute the cause of bias to the inter-class correlation problem.

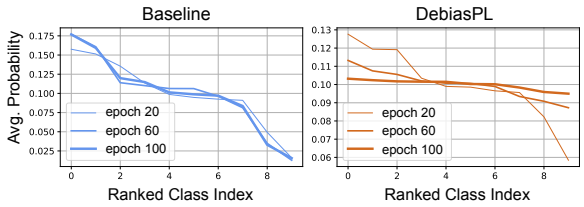


Figure 2. FixMatch's pseudo-labels are highly imbalanced across different training stages, even though the unlabeled and labeled data it trains on is class-balanced. In contrast, DebiasPL produces nearly balanced pseudo-labels at late stages. The probability distributions of FixMatch and DebiasPL are averaged over all unlabeled data. The class indices are sorted by average probability. We conduct experiments on CIFAR10 with 4 labeled instances per class.

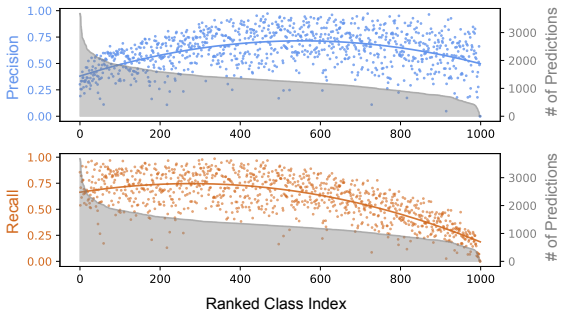


Figure 3. Per-class precision and recall of pseudo-label predictions on 1.3M ImageNet instances with a pre-trained CLIP. The majority classes with high recall often have less precise pseudo-labels.

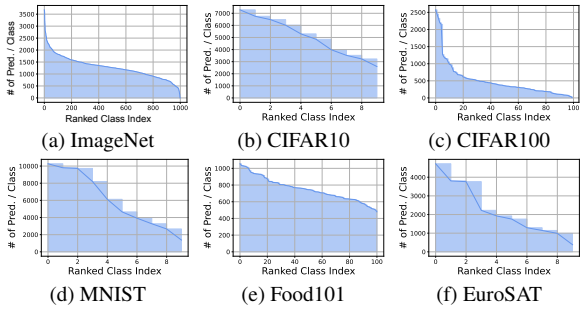


Figure 4. CLIP's zero-shot predictions are highly biased for various datasets and benchmarks.

3.1. Background

FixMatch for semi-supervised learning. The core technique of FixMatch [57] is pseudo-labeling [30]. It selects unlabeled samples with high confidence as training targets.

Suppose we have a labeled dataset $X_L = \{(x_i, y_i)\}_{i=1}^{L}$ with L labeled instances, and an unlabeled dataset $X_U = \{(x_i)\}_{i=L+1}^{L+U}$ with U instances. x_i is the input instance and $y_i = [y_i^1, ..., y_i^C] \subseteq \{0,1\}^C$ is a discrete annotated target with C classes. X_U and X_L share the same semantic labels. The optimization objective consists of two terms: $\mathcal{L} = \mathcal{L}_\text{s} + \lambda_u \mathcal{L}_\text{u}$, i.e., the supervised loss $\mathcal{L}_s$ applied to labeled data and an unsupervised loss $\mathcal{L}_u$ applied to unlabeled data, and λ_u is a scalar hyperparameter.

The supervised loss $\mathcal{L}_s$ is the cross-entropy between the model predictions and the ground truth: $\mathcal{L}_\text{s} = \frac{1}{B}\sum_{i=1}^{B} \text{H}(y_i, p(\alpha(x_i)))$, where α is the weak augmentation, and B is the batch size. The pseudo-labels $\hat{y}_i$ for unlabeled instances are generated from the weakly-augmented unlabeled samples, which are used to supervise the model prediction of the strongly-augmented samples. Instances whose largest probability fall under a confidence threshold τ are regarded as unreliable samples and discarded. Formally, the unsupervised loss $\mathcal{L}_u$ can be formulated as:

$$\mathcal{L}_\text{u} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \mathbb{1}[\max(p(\alpha(x_i))) \geq \tau] \cdot \text{H}(\hat{y}_i, p(\beta(x_i))) \quad (1)$$

where β is a strong augmentation [15], and μ determines the ratio of labeled and unlabeled samples in the minibatch.

CLIP for zero-shot learning. CLIP [49] is an efficient and scalable way to learn image representations from scratch on a dataset of 400M image-text pairs, which is manually curated to be approximately query-balanced. At pre-training time, an image encoder and a text encoder are optimized by maximizing (minimizing) the similarity between paired (unpaired) captions and visual images.

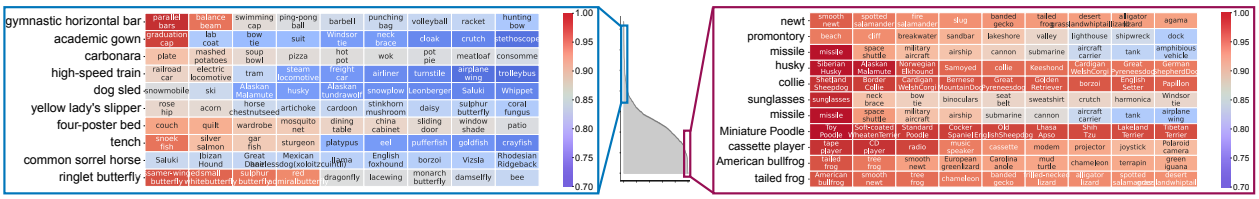
For producing pseudo-labels of unlabeled data, natural language prompting is used to enable zero-shot transfer to target datasets: CLIP uses the names or descriptions of the target dataset's classes as the set of potential text pairings (e.g. "a photo of a dog") and predicts the most probable class according to the cosine similarity of image-text pairs. Specifically, the feature embedding of the image and the feature embedding of the set of possible texts are first computed by their respective encoders. The cosine similarity of these embeddings is then evaluated, and normalized into a probability distribution via a softmax function.

3.2. Biases in Semi-supervised Learning

Fig. 2 visualizes the FixMatch probability distributions averaged on all unlabeled data at various training epochs. Surprisingly, even when labeled and unlabeled data are both curated (class-balanced), the pseudo-labels are still highly class-imbalanced, most notably at the early training stage. As the training progresses, this situation persists.

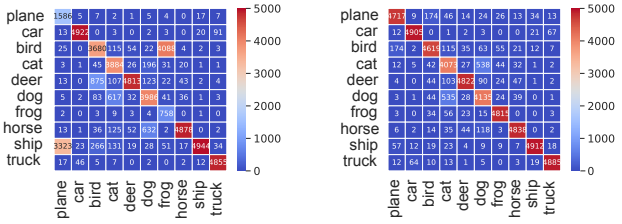
A student model will inherit the implicitly imbalanced pseudo-labels and, in turn, reinforces the teacher model's biases. Once confusing samples are wrongly pseudo-labeled, the mistake is almost impossible to be self-corrected. On the contrary, it may even mislead the model and further amplify existing bias to produce more wrong predictions. Without intervention, the model will get trapped in irreparable biases.

On the contrary, as in Fig. 2, although DebiasPL is also



a) top-10 classes b) least-10 classes

Figure 5. The low-frequency classes of ImageNet, with the least-10 number of CLIP predictions per class, usually have strong inter-class correlations, while the high-frequency classes are the opposite. We compare the cosine similarity between each class's image embedding centroid and embedding centroids of its nine closest "negative" classes. (better view zoomed in)



(a) FixMatch (b) DebiasPL

Figure 6. The cause for pseudo-label biases can be partially attributed to inter-class confounding. For example, FixMatch often misclassifies "ship" as "plane". The confusion matrix of FixMatch's and our DebiasPL's pseudo-labels are visualized.

troubled by the imbalanced pseudo-labels at the beginning, this situation can be significantly alleviated, and, eventually, we can obtain an almost balanced distribution through dynamically debiasing the model.

3.3. Biases in Zero-Shot Learning

CLIP actually generates highly biased predictions on ImageNet, which is hidden behind CLIP's tremendous success in terms of overall zero-shot prediction accuracy.

Except for the imbalance problem, the precision and recall of many high-frequency classes are much lower than many medium-/few-shot classes, as illustrated in Fig. 3. Thresholding the CLIP predictions based on the confidence score may help. However, simply setting a higher confidence score threshold could lead to even more imbalanced distributions (more details in appendix). There is a trade-off between imbalance ratio and precision/recall.

Highly biased zero-shot predictions are not unique to ImageNet. They are widely present on many benchmarks, such as EuroSAT [21], MNIST [29], CIFAR10 [27], CIFAR100 [27], and Food101 [7], as shown in Fig. 4.

3.4. Inter-Class Correlations

To delve into the causes of biased pseudo-labels, we provide an analysis of inter-class correlations. For CLIP, we first compute one image centroid per class by taking the mean of the normalized image features, extracted by the image encoder of a pre-trained CLIP model, that belong to this class. The cosine similarity between the image centroid of classes with top-10/least-10 prediction frequency and their closest "confusing" classes are visualized. The prediction confusions indicate image similarities at the class level. Fig. 5 shows that the low-frequency classes of ImageNet, with the least-10 number of CLIP predictions per class, usually have strong inter-class confusions.

Fig. 6a shows the confusion matrix of FixMatch's pseudo-labels. It is observed that many instances in some categories tend to be misclassified into one or two specific negative classes; for instance, "ship" is often misclassified as "plane".

Based on our analysis of the inter-class correlations, we believe that the blame for the pseudo-label bias can be largely attributed to inter-class confounding, which the proposed DebiasPL can successfully address as in Fig. 6b. DebiasPL will be introduced in the next section.

4. Debiased Pseudo-Labeling

This section introduces Debiased Pseudo-Labeling (DebiasPL) and methods to integrate it into ZSL and SSL tasks. It is worth noting that the proposed simple yet effective approach is universally applicable to various networks and benchmarks, not limited to the ones introduced here.

4.1. Adaptive Debiasing

Our DebiasPL approach aims at dynamically alleviating biased pseudo labels' influence on a student model without leveraging any prior knowledge on marginal class distribution, even when exposed to source and target data that follow different distributions. An adaptive debiasing module with counterfactual reasoning and an adaptive marginal loss is proposed to fulfill this goal, described next.

Adaptive Debias w/ Counterfactual Reasoning. Causal Inference is the undertaking of deriving counterfactual conclusions using only factual premises, in which causal graphical models represent the interventions among the variables [18, 44, 46, 52, 53]. It has been widely studied and applied in various tasks to remove selection bias which is pervasive in almost all empirical studies [3], eliminating the confounding effect using causal intervention [72], disentangling the desired direct effects with counterfactual reasoning [6], etc.

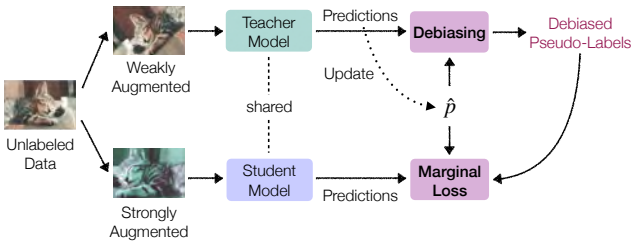


Figure 7. Diagram of the proposed Adaptive Debiasing module and Adaptive Marginal Loss, added to the top of FixMatch.

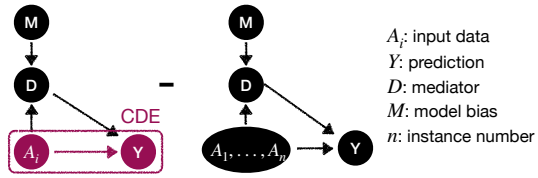


Figure 8. Causal graph of debiasing with counterfactual reasoning.

Motivated by this, to dynamically mitigate impacts of unwanted bias (*counterfactual*), we incorporate causality of producing debiased predictions through counterfactual reasoning [22, 44–47].

Given the proposed causal graph in Fig. 8, we can delineate our goal for generating debiased predictions: the pursuit of the direct causal effect along $A_i \rightarrow Y$, defined as Controlled Direct Effect (CDE) [18, 46, 47, 50, 59]:

$$\text{CDE}(Y_i) = [Y_i|do(A_i), do(D)] - [Y_i|do(\hat{A}), do(D)] \quad (2)$$

i.e. the contrast between the counterfactual outcome if the individual were exposed at $A = A_i$ (with $do(A_i)$ notation) and the counterfactual outcome if the same individual were exposed at $A = \hat{A} = \{A_1, ..., A_n\}$, with the mediator set to a fixed level D. CDE [18, 46] disentangles the model bias in a counterfactual world, where the model bias is considered as the Y's indirect effect when $A = \hat{A}$ but D retains the value when $A = A_i$.

However, measuring the counterfactual outcome via visiting all training samples is significantly computational expensive. We use Approximated Controlled Direct Effect (ACDE) instead. ACDE assumes that the model bias is not drastically changed, therefore, the momentum-updated counterfactual outcomes (Eqn. 4) can be served as an approximation to the actual $[Y_i|do(\hat{A}), do(D)]$. The debiased logits with counterfactual reasoning, which is later used to perform pseudo-labeling (i.e., replace $p(\alpha(x_i))$ in Eqn. 1), can be formulated:

$$\tilde{f}_i = f(\alpha(x_i)) - \lambda \log \hat{p} \quad (3)$$

$$\hat{p} \leftarrow m\hat{p} + (1-m)\frac{1}{\mu B}\sum_{k=1}^{\mu B} p_k \quad (4)$$

$m \in [0, 1)$ is a momentum coefficient, $f(\alpha(\cdot))$ refers to logits of weakly-augmented unlabeled instance, p_k is the probability distribution for instance $\alpha(x_k)$ obtained via a softmax function. λ denotes the debias factor, which controls the strength of the indirect effect. If the debias factor is too strong, it is hard for a model to fit on the data, while too small a factor can barely eliminate the biases and, ultimately, impairs the generalization ability. Since the scale of logits is unstable, most notably at the early training stage, we use the probability distribution p_k rather than directly using the logit vector in the second term of Eqn. 3. A log function is applied to rescale $\hat{p}$ to match the magnitude of logit.

Eqn. 3 can be associated with re-weighting and logits adjustment methods in long-tailed recognition, whereas ours is dynamically adaptive.

Adaptive Marginal Loss. As aforementioned in Sec. 3, the biases in pseudo-labels may be partially caused by inter-class confusion. Motivated by this, we apply adaptive margin loss to demand a larger margin between hardly biased and highly biased classes, so that scores for dominant classes, towards which the model highly biased, do not overwhelm the other categories. In addition, by enforcing a dynamic class-specific margin, inter-class confusion can be greatly counteracted, which is further empirically evidenced in Fig. 6. $\mathcal{L}_{\text{AML}}$ can be formulated as:

$$\mathcal{L}_{\text{AML}} = -\log \frac{e^{(z_{\hat{y}_i} - \Delta_{\hat{y}_i})}}{e^{(z_{\hat{y}_i} - \Delta_{\hat{y}_i})} + \sum_{k \neq \hat{y}_i}^{C} e^{(z_k - \Delta_k)}} \quad (5)$$

where $\Delta_j = \lambda \log(\frac{1}{\hat{p}_j})$ for $j \in \{1, ..., C\}$, $z = f(\beta(x_i))$. We use $\mathcal{L}_{\text{AML}}$ to replaced $\text{H}(\hat{y}_i, f(\beta(x_i)))$ in Eqn. 1. We then get the final unsupervised loss by updating Eqn. 1 with Eqn. 3 and Eqn. 5.

(Optional) All unlabeled instances with low probabilities do not contribute to the final loss. We find it beneficial to apply cross-level instance-group discrimination loss CLD [66] to unlabeled instances to leverage their information fully.

4.2. Distinctions and Connections with Alternatives

Please refer to Sec. 2 for an introduction to LA, LDAM, and Causal Norm. Another often adopted method in SSL distribution alignment (DA) [4] is also compared. It aims to encourage the actual marginal distribution of the model's predictions to match the *actual* marginal class distribution.

Please refer to Tab. 1 to check the distinctions and connections with these alternatives handling distribution mismatch and long-tailed recognition in key properties, and Tab. 2 and Tab. 3 to compare experimental results.

The use of a momentum updated $\hat{p}$ for debiasing pseudo-labels with counterfactual reasoning and applying adaptive marginal loss is crucial to the success of DebiasPL, which also enables our training objective does not necessarily need to use the true marginal class distribution as prior knowledge. Furthermore, since more training samples per class

Desired Properties	LA or LDAM	Causal Norm	DA	Ours
Improve representation learning at training time	✓	✗	✓	✓
No prior knowledge on true marginal class distribution	✗	✓	✗	✓
Adaptive as the training progresses	✗	✗	✓	✓
Applicable to both imbalanced and balanced data	✗	✗	✓	✓
Source and target data can come from varying distributions	✗	✗	✗	✓

Table 1. Our method is the only one with all these desired properties. Comparisons with previous works concentrating on resolving training data distribution issues, including LA [38], LDAM [10], DA [4], Causal Norm [59] and our DebiasPL, in key properties. Desired (undesired) properties are in green (red).

do not necessarily lead to a higher model bias against it, dynamically adjusting the margin rather than measuring margins based on the number of samples per class as in LA and LDAM could better respect the degree of bias against each class. The number of samples alone can not determine the degree of bias. Also, unlike previous works, e.g., LA/LDAM and Causal Norm, that use fixed margins or adjustments, we argue that the degree of bias of each class should never be a fixed value, but is in a process of dynamic change. The cause of bias cannot be attributed to the data alone, but the cause of the interaction between model and data.

For DA, the biggest issue is that it is limited to scenarios where either *true* marginal class distribution is available, or source and target data are collected from the same distribution, which is too ideal in the real world.

Experiments on several benchmarks are made to show the validity and feasibility of DebiasPL. *For imbalanced data*, Tab. 1 shows that integrating LA [38] into FixMatch lags far behind FixMatch w/ DebiasPL. *For balanced data*, since the adjustment or re-weighting vector is calculated based on the true class distribution, most existing long-tailed methods that rely on true marginal class distribution are no longer applicable without major changes (balanced class distribution leads to identical treatment for all classes).

4.3. DebiasPL for T-ZSL and SSL

For semi-supervised learning, the proposed DebiasPL can be integrated into FixMatch, as in Fig. 7, by adopting the adaptive debiasing module and adaptive marginal loss. To further boost the performance of SSL and exploit the power of the vision-language pre-trained model, during the training time, we can also integrate CLIP into FixMatch/DebiasPL by pseudo-labeling the discarded unlabeled instances with CLIP. Because the instances CLIP are not confident on may be noisy, only these unlabeled instances with a CLIP confidence score greater than τ_{clip} are pseudo-labeled by CLIP. We could get CLIP's predictions on all training data and store it in a dictionary without re-predicting per iteration. Therefore, the computational overheads introduced by using the CLIP model are negligible. We only leverage CLIP in large-scale datasets since using CLIP on low-resolution datasets like CIFAR10 can only observe marginal gains, partly due to the lack of scale-based data augmentation in CLIP [49].

For transductive zero-shot learning, to better exploit knowledge learned from the vision-language pre-trained model and alleviate the domain shift problem when transferring the knowledge to downstream ZSL tasks, a new framework to conduct transductive zero-shot learning (T-ZSL) based on FixMatch and CLIP is developed.

Specifically, we again make use of the *pseudo-labeling* idea by leveraging the one-hot labels (i.e., the $\arg\max$ of the model's output) and retaining pseudo labels whose largest class probability fall above a confidence threshold τ_{clip} ($= 0.95$ by default). These instances, along with their pseudo labels, are considered "labeled data" in SSL.

After this, we could follow the original FixMatch pipeline to optimize "labeled" and "unlabeled" data jointly. To make a fair comparison with previous works and simplify the overall system, all other training recipes and settings are consistent with the original FixMatch+EMAN settings, including the model initialization part. The diagram is in the appendix.

Because CLIP is highly biased, the vanilla FixMatch + CLIP framework under-performs the original CLIP zero-shot learning, confirming our earlier hypothesis that learning from a biased model may further amplify existing bias and produce more wrong predictions. Therefore, we update the unsupervised loss $\mathcal{L}_u$ with our Adaptive Marginal Loss for alleviating the inter-class confusion and Adaptive Debias for producing debiased pseudo-labels as in Sec. 4.1.

5. Experiment

In this section, we conduct empirical experiments to show that DebiasPL: 1) delivers state-of-the-art results on both semi-supervised and zero-shot learning benchmarks; 2) works as a universal add-on and brings consistent performance gains to various methods; 3) exhibits stronger robustness to domain shifts; 4) is capable of improving performance on long-tailed, balanced and even hybrid data.

5.1. Semi-supervised Learning

Dataset. We perform comprehensive evaluations of DebiasPL on multiple SSL benchmarks, including CIFAR10 [27], long-tailed CIFAR10 (CIFAR10-LT) [27], and ImageNet-1K [54], with varying amounts of labeled data. For the balanced benchmarks, the performance almost saturates when using more than 2% labeled data. We put our focus on the extremely low-shot settings, i.e., 0.08%/0.16%/2% on CIFAR10 and 1%/0.2% on ImageNet-1K. For imbalanced

Method	CIFAR10-LT: # of labels (percentage)				CIFAR10: # of labels (percentage)		
	γ=100		γ=200				
	1244 (10%)	3726 (30%)	1125 (10%)	3365 (30%)	40 (0.08%)	80 (0.16%)	250 (2%)
UDA [68] §	-	-	-	-	71.0 ±6.0	-	91.2 ±1.1
MixMatch [5] §	60.4 ±2.2	-	54.5 ±1.9	-	51.9 ±11.8	80.8 ±1.3	89.0 ±0.9
CReST w/ DA [67]	75.9 ±0.6	77.6 ±0.9	64.1 ±0.22	67.7 ±0.8	-	-	-
CReST+ w/ DA [67]	78.1 ±0.8	79.2 ±0.2	67.7 ±1.4	70.5 ±0.6	-	-	-
CoMatch w/ SimCLR [12, 32]	-	-	-	-	92.6 ±1.0	94.0 ±0.3	95.1 ±0.3
FixMatch [57] §	67.3 ±1.2	73.1 ±0.6	59.7 ±0.6	67.7 ±0.8	86.1 ±3.5	92.1 ±0.9	94.9 ±0.7
FixMatch w/ DA w/ LA [4, 38, 57, 67] §	70.4 ±2.9	-	62.4 ±1.2	-	-	-	-
FixMatch w/ DA w/ SimCLR [4, 12, 57] §	-	-	-	-	89.7 ±4.6	93.3 ±0.5	94.9 ±0.7
DebiasPL (w/ FixMatch)	**79.2 ±1.0**	**80.6 ±0.5**	**71.4 ±2.0**	**74.1 ±0.6**	**94.6 ±1.3**	**95.2 ±0.1**	**95.4 ±0.1**
gains over the best FixMatch variant	**+8.8**	**+7.5**	**+9.0**	**+6.4**	**+4.9**	**+1.9**	**+0.5**

Table 2. Without any prior knowledge of the marginal class distribution of unlabeled/labeled data, the performance of DebiasPL on *both* **CIFAR *and* CIFAR-LT SSL benchmarks** surpasses previous SOTAs, which are *either designed for balanced data or meticulously tuned for long-tailed data*. DibasMatch is experimented with the same set of hyper-parameters across all benchmarks. § states the best-reported results of counterpart methods, copied from [32], [57] or [67]. γ: imbalance ratio. We report results averaged on 5 different folds.

Method	B.S.	#epochs	Pre-train	1%		0.2%	
				top-1	top-5	top-1	top-5
FixMatch w/ DA [4, 57]	4096	400	✗	53.4	74.4	-	-
FixMatch w/ DA [4, 57]	4096	400	✓	59.9	79.8	-	-
FixMatch w/ EMAN [9, 57]	384	50	✓	60.9	82.5	43.6*	64.6*
DebiasPL w/ FixMatch	384	50	✓	**63.1** (+2.2)	**83.6** (+1.1)	**47.9** (+3.7)	**69.6** (+5.0)
DebiasPL (multi-views)	768	50	✓	**65.3** (+4.4)	**85.2** (+2.7)	**51.6** (+8.0)	**73.3** (+8.7)
DebiasPL (multi-views)	768	200	✓	**66.5** (+5.6)	**85.6** (+3.1)	**52.3** (+8.7)	**73.5** (+8.9)
DebiasPL (multi-views)	1536	300	✓	**67.1** (+6.2)	**85.8** (+3.3)	-	-
DebiasPL w/ CLIP [49]	384	50	✓	**69.1** (+8.2)	**89.1** (+6.6)	**68.2** (+24.6)	**88.2** (+23.6)
DebiasPL w/ CLIP (multi-views) [49]	768	50	✓	**70.9** (+10.0)	**89.3** (+6.8)	**69.6** (+26.0)	**88.4** (+23.8)
CLIP (few-shot) [49, 73]	256	50	✓	53.4	-	40.0	-
SwAV [11]	4096	50	✓	53.9	78.5	-	-
SimCLRv2 (+ Self-distillation) [13]	4096	400	✓	60.0	79.8	-	-
PAWS (multi-crops) † [2]	4096	50	✓	66.5	-	-	-
CoMatch (multi-views) [32]	1440	400	✓	67.1	87.1	-	-

Table 3. DebiasPL delivers state-of-the-arts results on **ImageNet-1K semi-supervised learning** with various fractions of labeling samples, especially for extremely low-shot settings. All results are produced with a backbone of ResNet-50. †: unsupervised pre-trained for 800 epochs, except for PAWS [2], which is pre-trained for 300 epochs with pseudo-labels generated non-parametrically. *: reproduced.

benchmarks, we follow the settings in [67] and test DebiasPL on CIFAR10-LT under various pre-defined imbalance ratios γ, where $\gamma \in [100, 200]$, and percentage of labeled data, including 10% and 30%. More details about datasets are included in the appendix.

Setup. For all experiments on *both* long-tailed CIFAR10 and CIFAR10 datasets, we follow previous works [57, 67] to use the network architecture WRN-28-2 [20, 71]. We also follow the same set of hyper-parameters in FixMatch, except we reduce the total optimization iterations by half.

For experiments on ImageNet-1K, we use ResNet50 as the backbone network and follow the training recipes introduced in FixMatch w/ EMAN [9], which is also the default baseline of all experiments on ImageNet-1K. The model is initialized with MoCo v2 + EMAN as in [9]. For the setting with multiple views, we perform two strong augmentations and two weak augmentations on each unlabeled sample. Each strongly-augmented instance is paired with one weakly-augmented instance, and we jointly optimize the two pairs via pseudo-labeling as in the original setting of Fig. 7. Multi-views could increase the convergence speed and stabilize the training process.

DebiasPL is simple yet effective. Tab. 2 and Tab. 3 show that DebiasPL delivers state-of-the-art performance on all experimented benchmarks, outperforming current approaches by a large margin. Without using CLIP, DebiasPL can outperform CoMatch on CIFAR, and is comparable to CoMatch on ImageNet-1K. DebiasPL wins on its merit of simplicity. Leveraging the power of CLIP could significantly improve the performance of DebiasPL, surpassing CoMatch by about 4% on ImageNet-1K SSL.

DebiasPL is agnostic to source/target data distribution. Tab. 2 shows that, for both CIFAR and long-tailed CIFAR SSL benchmarks, using a unified framework and the same set of hyper-parameters, DebiasPL can surpass previous state-of-the-art methods, which are either designed for balanced

Method	Labeled: **LT**; 10% labeled, $\gamma = 200$	
	Unlabeled: **LT**	Unlabeled: **Balanced**
FixMatch [57]	62.3 ±1.6	72.1 ±2.3
DebiasPL	**71.4** ±**2.0** (+9.1)	**83.5** ±**2.4** (+11.4)

Table 4. DebiasPL consistently improves the performance of SSL when the unlabeled data is either the sames as labeled data, i.e., long-tailed distributed, or different with labeled data, i.e., balanced distributed across semantics. We report results averaged on 5 folds.

	FixMatch	MixMatch	UDA
Baseline	89.7 ± 4.6	47.5 ± 11.5	29.1 ± 5.9
+ DebiasPL	**94.6** ± **1.3**	**61.7** ± **6.1**	**43.2** ± **5.2**

Table 5. **DebiasPL is a universal add-on.** Top-1 accuracies of various SSL methods on CIFAR10, averaged on 5 folds, are compared. 4 instances per class are labeled.

Method	#param	Accuracy (%) top-1	top-5
ConSE [43]	-	1.3	3.8
DGP [24]	-	3.0	9.3
ZSL-KG [41]	-	3.0	9.9
Visual N-Grams [31]	-	11.5	-
CLIP (prompt ensemble) [49]	26M	59.6	-
(ours) CLIP + FixMatch	26M	55.7	80.6
(ours) CLIP + DebiasPL	26M	**68.3** (+8.7)	**88.9** (+8.3)
CLIP (few-shot) [49, 73] †	26M	53.4	-
CLIP + CoOp (few-shot) [73] †	26M	60.9	-
CLIP (ViT-B/32) [49]	398M	63.2	-
CLIP (ResNet50x4) [49]	375M	65.8	-

Table 6. DebiasPL delivers state-of-the-art results of **zero-shot learning on ImageNet-1K**, outperforming CLIP with bigger models or fine-tuned with labels. †: CoOp and CLIP (few-shot) are fine-tuned with about 1.5% annotated data.

data or meticulously tuned for long-tailed data. Furthermore, Tab. 4 shows that when tested in scenarios where labeled and unlabeled data follow different distributions, DebiasPL produces an even greater gain (11.4%) to the baseline.

The fewer labeled data, the more significant gains can be observed in Tab. 2 and Tab. 3, almost eliminating the gap between fully-supervised and semi-supervised learning.

DebiasPL is also a universal add-on as illustrated in Tab. 5. Incorporating DebiasPL into various SSL methods can achieve consistent performance improvements.

5.2. Tranductive Zero-Shot Learning

Dataset. We evaluate the efficiency of DebiasPL in T-ZSL on ImageNet-1K [54]. EuroSAT [21], MNIST [29], CIFAR10 [27], CIFAR100 [27], and Food101 [7] are also used as evaluation datasets to show the robustness to domain shift.

Setup. T-ZSL assumes that the list of possible class candidates is known for the target data. Following this setting, we do not use any semantic labels for target data. We apply DebiasPL on CLIP in a similar way as we apply DebiasPL on FixMatch, except that the labeled data is "labeled" by CLIP rather than a human annotator. Specifically, all unlabeled instances whose CLIP confidence score greater than τ_{clip} are pseudo-labeled by CLIP and considered as "labeled" data. A backbone of ResNet50 and a threshold τ_{clip} of 0.95 are used for all datasets. The same default hyper-parameters and training recipes as in FixMatch + EMAN are utilized for fair comparisons. More details are in the appendix.

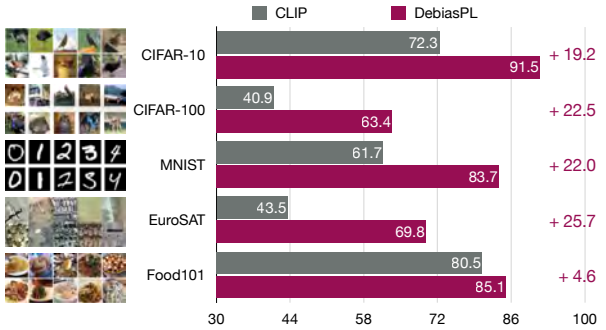


Figure 9. DebiasPL exhibits stronger robustness to domain shift when conducting **zero-shot learning on various datasets**. We experiment with ResNet-50 as a backbone network. CLIP results are reproduced with official codes.

DebiasPL delivers SOTA results on zero-shot learning, even surpassing CLIP [49] and CoOP [73] that are fine-tuned on partial human-labeled data. Moreover, DebiasPL with a backbone of ResNet50 can significantly outperform CLIP with 15× larger backbones, as shown in Tab. 6. The time cost of zero-shot training DebiasPL w/ CLIP (without using any human annotations) for 100 epochs is less than 0.01% of CLIP's overall training time.

DebiasPL exhibits stronger robustness to domain shift than zero-shot CLIP without accessing any semantic labels, as depicted in Fig. 9. Also, DebiasPL can observe greater gains (more than 20%) on datasets with larger domain shifts, e.g., an astonishing 25.7% gains can be obtained on the satellite image dataset EuroSAT [21].

6. Summary

In this paper, we conduct research on the previously unknown biased pseudo-labeling issue. A simple yet effective method DebiasPL is proposed to dynamically alleviate biased pseudo-labels' influence on a student model, without leveraging any prior knowledge of true data distribution. As a universal add-on, DebiasPL delivers significantly better performance than previous state-of-the-arts on both semi-supervised learning and transductive zero-shot learning tasks and exhibits stronger robustness to domain shifts.

Acknowledgements. This work was supported, in part, by US Government fund through Etegent Technologies on Low-Shot Detection and Semi-supervised Detection.

Algorithm 1: PyTorch-style pseudocode for semi-supervised learning with DebiasPL

```
# initialize p_hat with 1/C, C is the number of classes
p_hat = torch.ones([1, C]) / C
# load a batch with unlabeled and labeled samples
# x:  labeled samples ; target:  labels for x ; u:  unlabeled samples
for (x, target), u in loader:
    # augment x with weak augmentation and get two versions of u with strong and weak augmentations
    x, u_s, u_w = weak(x), strong(u), weak(u)
    # model forward
    l_x, l_us, l_uw = model(x, u_s, u_w)

    # get debiased pseudo-labels
    p_uw = F.softmax(l_uw - tau * torch.log(p_hat), dim=1)
    max_probs, pseudo_label = torch.max(p_uw, dim=-1)

    # get mask for filtering instances with low confidence score
    mask = max_probs.ge(thresh).float()
    # update p_hat
    p_hat = momentum * p_hat + (1 - momentum) * p_uw.detach().mean(dim=0)

    # calculate loss_x for labeled instances
    loss_x = F.cross_entropy(l_x, target)
    # calculate marginal loss loss_u for unlabeled instances
    l_us = l_us + lambda * torch.log(p_hat)
    loss_u = (F.cross_entropy(l_us, pseudo_label, reduction='none') * mask).mean()
    # total loss
    loss = loss_x + lambda_u * loss_u

    # optimization step
    loss.backward()
    optimizer.step()
# update the ema model
model.momentum_update_ema()
```

7. Appendix

7.1. Details on Datasets and Implementations

The PyTorch-style pseudocode for semi-supervised learning with DebiasPL is available at Algo. 1.

We conduct experiments on several benchmarks to prove the effectiveness and universality of DebiasPL. Here we provide more details on datasets and implementations for each benchmark:

CIFAR10 [27]: The original version of CIFAR10 contains 50,000 images on the training set and 10,000 images on the validation set with 10 categories for CIFAR10. For semi-supervised learning on CIFAR10, we conduct the experiments with a varying number of labeled examples from 40 to 250, following standard practice in previous works [4, 5, 32, 57]. The reported results of each previous method in the paper are directly copied from the best-reported results in MixMatch [5], ReMixMatch [4], FixMatch [57], CoMatch [32], etc.

We keep all hyper-parameters the same as FixMatch, except for the number of training steps. We use WideResNet-28-2 [20, 71] with 1.5M parameters as a backbone network for CIFAR10. The SGD optimizer with a Nesterov momentum of 0.9 is used for optimization. The learning rate is initialized as 0.03 and decayed with a cosine learning rate scheduler [36], which sets the learning rate at training step k as $\cos(\frac{7\pi k}{16K})$ times the initial learning rate, where $K = 2^{19}$ is the total number of training steps, i.e., about 512 epochs, and is 2 times fewer than the original number of FixMatch training steps. The model is trained with a mini-batch size of 512, which contains 64 labeled samples and 448 unlabeled samples, on one V100 GPU. As in previous works, an exponential moving average of model parameters is used to produce the final performance. The weight decay is set as 0.0005 for CIFAR10. Unless otherwise stated, the only independent hyperparameter of DebiasPL λ is fixed and set to 0.5 in all experiments. Each method is tested under 5 different folds, and we report the mean and the standard deviation of accuracy on the test set.

CIFAR10-LT [27, 35, 67]: The long-tailed version of CIFAR10 follows an exponential decay in sample sizes across different categories. CIFAR10-LT is constructed by sampling a subset of CIFAR10 following the Pareto distribution with the power value $\gamma \in [100, 200]$. Then, we select 10% or 30% of all CIFAR10-LT instances to construct the SSL benchmark labeled dataset, and the others are regarded as the unlabeled datasets. Each algorithm is tested under 5 different folds of labeled data, and we report the mean and the standard deviation of accuracy on the test set. As in previous works, an exponential moving average of model parameters

is used to produce the final performance.

To demonstrate the universality of the proposed method DebiasPL and its insensitivity to data distribution, we follow the same hyperparameters and training formulas in CIFAR10. We do not specifically adjust any hyperparameters when conducting experiments in the long-tail SSL benchmarks.

ImageNet-1K [54]: ImageNet-1K is a curated dataset with approximately class-balanced data distribution, containing about 1.3M images for training and 50K images for validation.

For semi-supervised learning, ImageNet-1K with varying amounts of labeled data is experimented with, i.e., 0.2% and 1%. The FixMatch model is trained with a batch size of 64 (320) for labeled (unlabeled) images with an initial learning rate of 0.03. Following [9], we replace batch normalization (BN) layers with exponential moving average normalization (EMAN) layers in the teacher model. EMAN updates its statistics by exponential moving average from the BN statistics of the student model. ResNet-50 is used as the default network and the default hyperparameters in the corresponding papers [9, 57] are applied. The model is initialized with MoCo v2 + EMAN pre-trained model as in [9]. To make fair comparisons, we report results of FixMatch with EMAN as the baseline model, and all hyper-parameters of FixMatch with EMAN are untouched unless noted otherwise.

For zero-shot learning, no manual annotation is leveraged in the training process. We train CLIP + DebiasPL and CLIP + FixMatch following the same hyperparameters and training recipes as FixMatch with EMAN, except that the labeled data is "labeled" by CLIP rather than a human annotator. Specifically, all unlabeled instances whose CLIP confidence score greater than τ_{clip} are pseudo-labeled by CLIP (with a backbone of ResNet50) and considered as "labeled" data. A backbone of ResNet50 and a threshold τ_{clip} of 0.95 are used. The same default hyper-parameters and training recipes as in FixMatch + EMAN are utilized for fair comparisons. The framework of transductive zero-shot learning with DebiasPL is illustrated in Fig. 10.

For experiments on other benchmarks of ZSL, including EuroSAT [21], MNIST [29], DTD [14], GTSRB [58] and Flowers102 [42], we follow the training recipe of ImageNet-1K.

7.2. Ablation Study

In this section, we conduct additional ablation studies on the influence of the two components of DebiasPL (Table. 7) for SSL, DebiasPL's unique hyperparameter λ (Table. 8) for SSL, and CLIP's confidence score threshold τ_{clip} (Table. 9) for T-ZSL.

As shown in Table. 7, the two components of DebiasPL lead to significant improvements to *both* CIFAR10 and CIFAR10-LT SSL benchmarks. Compared with the balanced benchmark, the performance improvement obtained by introducing the marginal loss is relatively smaller than the unbalanced benchmark.

Debiasing	Magirnal Loss	CIFAR10	CIFAR10-LT
		86.1	73.5
✓		93.3	79.6
✓	✓	94.6	80.6

Table 7. **Ablation study** on the **contribution of each component** of DebiasPL. Experimented on CIFAR10 and CIFAR10-LT ($\gamma = 100$) SSL, in which 4 out of 5,000 samples are labeled per class for CIFAR10 and 30% instances are labeled for CIFAR10-LT. Results averaged over 5 different folds are reported.

λ	0.0	0.25	0.5	0.75	1.0	2.0
DebiasPL	73.5	79.5	80.6	80.5	80.5	77.7

Table 8. **Ablation study** on CIFAR10-LT ($\gamma = 100$) semi-supervised learning with DebiasPL under various **weight** λ of debiasing module and marginal loss. 30% samples are labeled. The model is identical to FixMatch when $\lambda = 0$. Results averaged over 5 different folds are reported.

Table. 8 illustrates the influence of debias factor λ. When the value of λ is set to 0, DebiasPL is identical to FixMatch. Adding a debiasing module and marginal loss can improve the performance on CIFAR10-LT by more than 7% when selecting the optimal choice of λ 0.5, which is marginally better than the default value of 1.0. However, there is a trade-off. Suppose the debias factor λ is too strong. In that case, it is hard for a model to fit on the data, while a too-small factor can barely eliminate the biases, ultimately impairs the generalization ability.

τ_{clip}	0.2	0.4	0.6	0.8	0.9	0.95
DebiasPL + CLIP	55.9	63.2	66.2	67.1	67.7	67.7

Table 9. **Ablation study** on ImageNet-1K zero-shot Learning with DebiasPL + CLIP [49] under various **threshold** τ_{clip}.

As illustrated in the main paper, the CLIP predictions are class-imbalanced. Therefore, the natural question is whether we can obtain a more balanced prediction by filtering instances with a threshold τ_{clip}? Unfortunately, no, on the contrary, when filtering predictions with a larger threshold, a higher imbalance rate is observed, as in Fig. 11. Furthermore, when filtering instances with a threshold of 0.95, more than 60 categories get zero predictions.

The dilemma is that using a smaller threshold τ_{clip} can obtain a smaller imbalanced ratio, which is the desired property. However, it also leads to a lower precision, introducing many outliers and misclassified samples. Therefore, a module to eliminate biases captured by the CLIP model when CLIP is pre-trained on source data is needed to yield a good performance on target data.

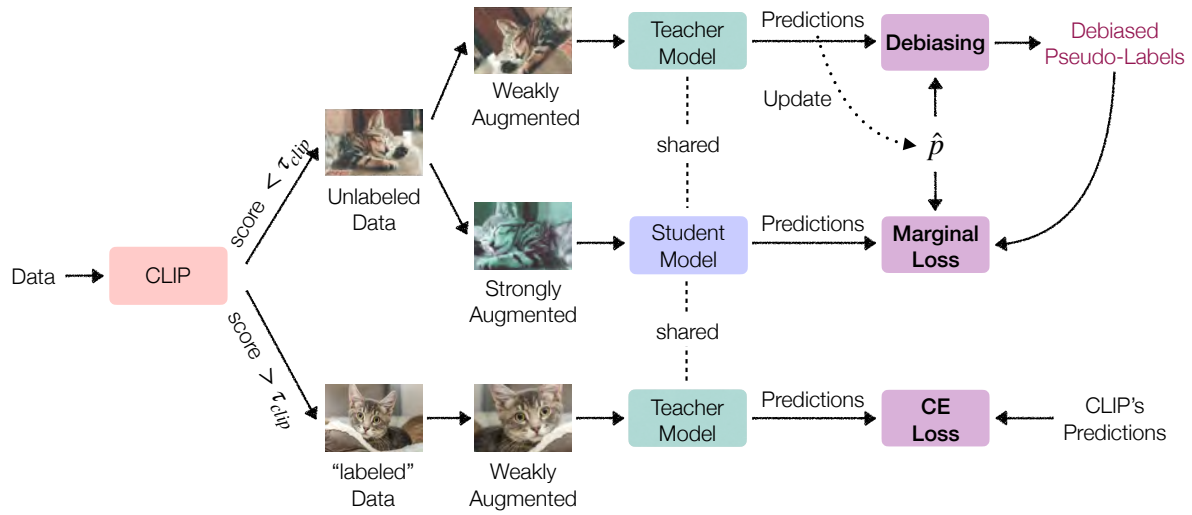
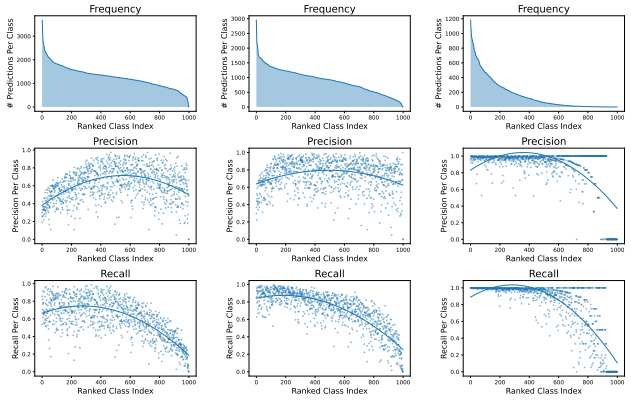


Figure 10. The overall framework of transductive zero-shot learning with CLIP + DebiasPL. CLIP + FixMatch can be realized by removing the debiasing module and replacing the marginal loss with cross-entropy loss.



(a) $\tau \geq 0.0$ (b) $\tau \geq 0.4$ (c) $\tau \geq 0.95$

Figure 11. A higher imbalanced ratio is obtained when filtering CLIP's zero-shot predictions with a larger threshold, analyzed on CLIP's zero-shot predictions on 1.3M almost class-balanced ImageNet training samples. Per class number of predictions (row 1), precision (row 2), and recall (row 3) of samples passing various confidence score thresholds τ are visualized. Zero-shot predictions are produced with an ensemble of 80 prompts and a backbone of ResNet50, using official codes.

Table. 9 shows that using a threshold of 0.95 can get the optimal performance on the ImageNet zero-shot learning task, which indicates that the high precision of the labeled data, realized by using a high threshold, is essential for better performance on target data. At the same time, our proposed DebiasPL can greatly alleviate the trouble of a higher imbalance ratio caused by using a larger threshold, eventually obtaining more than 10% performance gains.

References

[1] Andrew Arnold, Ramesh Nallapati, and William W Cohen. A comparative study of methods for transductive transfer learning. In *Seventh IEEE international conference on data mining workshops (ICDMW 2007)*, pages 77–82. IEEE, 2007. 1

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. *arXiv preprint arXiv:2104.13963*, 2021. 2, 7

[3] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2012. 4

[4] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2019. 1, 2, 5, 6, 7, 9

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32:5049–5059, 2019. 1, 2, 7, 9

[6] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Learning Representations*, 2019. 4

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 4, 8

[8] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakan-

tan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[9] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 194–203, 2021. 7, 10

[10] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019. 1, 2, 6

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 7

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 7

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 2, 7

[14] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 10

[15] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 3

[16] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009. 2

[17] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013. 2

[18] Sander Greenland, James M Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Statistical science*, pages 29–46, 1999. 4, 5

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 9

[21] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 4, 8, 10

[22] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. 5

[23] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008. 2

[24] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11487–11496, 2019. 2, 8

[25] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Learning imbalanced datasets with label-distribution-aware margin loss. In *International Conference on Learning Representations*, pages 1567–1578, 2020. 1, 2

[26] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 1

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4, 6, 8, 9

[28] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013. 2

[29] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 4, 8, 10

[30] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013. 1, 2, 3

[31] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017. 8

[32] Junnan Li, Caiming Xiong, and Steven Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. *arXiv preprint arXiv:2011.11183*, 2021. 2, 7, 9

[33] Bin Liu, Zhirong Wu, Han Hu, and Stephen Lin. Deep metric transfer for label propagation with limited annotated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[34] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016. 2

[35] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 9

[36] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 9

[37] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 1

[38] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. 2, 6, 7

[39] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 2

[40] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1103, 2021. 1

[41] Nihal V Nayak and Stephen H Bach. Zero-shot learning with common sense knowledge graphs. *arXiv preprint arXiv:2006.10713*, 2020. 2, 8

[42] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 10

[43] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 8

[44] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009. 4, 5

[45] Judea Pearl. *Causality*. Cambridge university press, 2009. 5

[46] Judea Pearl. Direct and indirect effects. *arXiv preprint arXiv:1301.2300*, 2013. 4, 5

[47] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018. 5

[48] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 6, 7, 8, 10

[50] Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5):1511–1519, 2013. 5

[51] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR, 2015. 2

[52] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 4

[53] Donald B Rubin. Essential concepts of causal inference: a remarkable history and an intriguing future. *Biostatistics & Epidemiology*, 3(1):140–155, 2019. 4

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6, 8, 10

[55] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *arXiv preprint arXiv:1606.04586*, 2016. 2

[56] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013. 2

[57] 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. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 3, 7, 8, 9, 10

[58] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 10

[59] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 5, 6

[60] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 2

[61] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 1

[62] Feng Wang, Weiyang Liu, Haijun Liu, and Jian Cheng. Additive margin softmax for face verification. *arXiv preprint arXiv:1801.05599*, 2018. 2

[63] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019. 2

[64] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. 1, 2

[65] Xudong Wang, Long Lian, and Stella X Yu. Data-centric semi-supervised learning. *arXiv preprint arXiv:2110.03006*, 2021. 2

[66] Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level instance-group discrimination.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12586–12595, 2021. 5

[67] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. 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*, pages 10857–10866, 2021. 2, 7, 9

[68] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 7

[69] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[70] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 1, 2

[71] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 7, 9

[72] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33, 2020. 4

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 7, 8