Language-Assisted Debiasing and Smoothing for Foundation Model-Based Semi-Supervised Learning

Na Zheng[1], Xuemeng Song[2], Xue Dong[3], Aashish Nikhil Ghosh[1], Liqiang Nie[4], Roger Zimmermann[1]
[1]National University of Singapore [2] City University of Hong Kong
[3] Tsinghua University [4]Harbin Institute of Technology (Shenzhen)
{zhengnagrape, sxmustc, dongxue.sdu, nieliqiang}@gmail.com
aa.ghosh@nus.edu.sg rogerz@comp.nus.edu.sg

Abstract

Recent studies have focused on introducing pre-trained foundation models into semi-supervised learning (SSL) tasks. Nevertheless, these foundation models can exhibit biases toward different classes and tend to generate imbalanced pseudo-labels for SSL. Thus, efforts have been made to introduce the logit adjustment offset to reduce the inherent bias in foundation models for SSL tasks. Despite their success, existing foundation model-based SSL methods face challenges: 1) unreliability in the estimated logit adjustment offset, 2) overlooking the potential of linguistic knowledge in capturing model biases and 3) fail to fully exploit the unlabeled samples. To address these issues, we propose ***L****anguage-****A****ssisted* ***D****ebiasing* ***a****nd* ***S****moothing framework, namely LADaS, for foundation model-based SSL. It consists of two components: 1) Language-assisted Pseudo-Label Debiasing (LPLD) to reduce biases in foundation models, and 2) Language-aware Pseudo-Label Smoothing (LPLS) to fully exploit low-confidence samples to facilitate SSL training. In particular, LPLD introduces a reliability score to dynamically assess the reliability of the logit adjustment. Additionally, it incorporates a language-oriented preference to reduce model biases using linguistic knowledge derived from pre-trained language models. Finally, LPLS introduces language-aware soft pseudo-labels and devises language-aware pseudo-label smoothing loss to guide the learning of unlabeled samples with low-quality pseudo-labels. Extensive experiments demonstrate the superiority of our proposed LADaS.*

1. Introduction

Deep learning [23, 33, 41–43] has witnessed great success in various fields using large-scale labeled datasets. However, obtaining such large labeled datasets through manual annotation is both time-consuming and labor-intensive, limiting its application. In contrast, unlabeled data is typically much more abundant and easily accessible. Therefore, semi-supervised learning (SSL) [9, 22, 35], which trains models with a small amount of labeled data and massive unlabeled data has attracted significant attention.

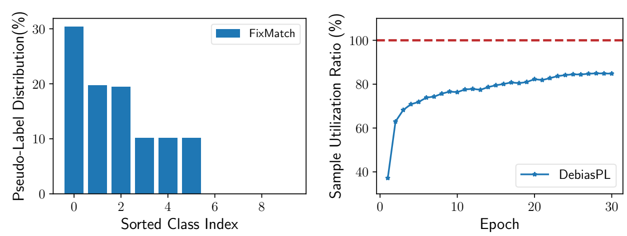


Figure 1. (a) The pseudo-label distribution and (b) sample utilization ratio, where pseudo-labels exceed a fixed threshold of 0.7, using DebiasPL on CIFAR-100 with 4 labeled samples per class.

Over the past decade, there has been extensive study of SSL regarding image classification [1, 2, 14, 15, 17, 31, 37]. Mainstream methods [22, 31] first apply weak and strong augmentations on unlabeled images, respectively, and pass them through a visual feature extractor and a classifier to predict their corresponding classes. Then they introduce a cross-entropy loss-based consistency regularization, which leverages the predicted classes of weakly-augmented images with prediction probabilities exceeding a certain confidence threshold as pseudo-labels to supervise the classification of strongly-augmented images. Early studies following the above paradigm train their visual feature extractors from scratch, and hence require a significant amount of training steps. Inspired by the great universality and generalization of pre-trained foundation models in vision tasks [20, 25], recent studies [9, 29, 30] replace the original vision extractor with the more powerful foundation models to enhance training efficiency and improve model performance.

Nevertheless, directly adapting foundation models to

SSL is non-trivial, as foundation models exhibit biases toward some classes [9]. Specifically, since there exist inherent imbalances in pre-training datasets for foundation models, they tend to produce more pseudo-labels for more frequently occurring classes. For illustration, directly using ViT as the image encoder in FixMatch [22] with visual prompting tuning (VPT) [11] would generate significantly imbalanced pseudo-labels for samples in CIFAR-10, as shown in Fig. 1 (a). Hence, several studies [9, 29] adopt the logit adjustment strategy [18] to encourage the model to generate more balanced pseudo-labels, where a logit adjustment offset is introduced into consistency regularization based on the class learning pace. The learning pace for a class is commonly measured by the number of unlabeled samples predicted as that class at each training step. Despite remarkable progress, existing foundation model-based methods still have the following limitations:

L1: Ignore the unreliability of the estimated logit adjustment offset. Existing studies [9, 29] directly incorporate the estimated logit adjustment offset during training while ignoring its reliability. In a sense, the logit adjustment offset is estimated from model predictions, which may be unreliable, especially when the model is unstable during the initial stage. Consequently, the unreliable offset may introduce noise and negatively impact the model training.

L2: Overlook the potential of linguistic knowledge in capturing the model bias. As aforementioned, previous efforts mainly estimate biases based on the class learning pace calculated by the predictions of the visual feature extractor, while ignoring the potential of linguistic knowledge encoded by pre-trained language models (PLMs) [39]. As PLMs are pretrained from large corpus, they can effectively encode the semantic information of each class into text embeddings. Intuitively, if a large number of unlabeled samples show high similarities to the text embedding of a given class, this implies the potential model bias toward that class.

L3: Fail to fully exploit the unlabeled samples. To avoid noise, previous studies usually discard the low-quality pseudo-labels, whose predicted probabilities fall below a confidence threshold. However, this may potentially lead to a poor utilization of unlabeled data. For instance, as depicted in Fig. 1 (b), when applying DebiasPL [29] to CIFAR-100 in our training setup, only 76% unlabeled samples are selected for optimization at the 10-th epoch.

To address above issues , we propose **L**anguage-**A**ssisted **D**ebiasing **a**nd **S**moothing framework, namely LADaS, for foundation model-based semi-supervised learning, as shown in Fig. 2. It consists of two components: 1) language-assisted pseudo-label debiasing, which adaptively mitigates biases in the foundation model, and 2) language-aware pseudo-label smoothing, which aims to fully exploit low-confidence unlabeled samples. In the first component, following previous studies, we integrate the logit adjustment into the consistency regularization for pseudo-label debiasing. In contrast, we design a linguistic knowledge-assisted logit adjustment, composed of self-adaptive logit adjustment and language-oriented preference. The former introduces a reliability score to dynamically assess the reliability of the learning pace, which starts at a relatively low value in the early training stage and gradually increases as training progresses. The latter estimates model bias with the linguistic knowledge, where the class centers are introduced to represent the semantic information for each class with a series of prompts and a PLM. Based on that, we estimate the language-oriented preference via the exponential moving average (EMA) of the language-guided class frequencies of unlabeled samples. For the second component, as linguistic knowledge provides supplementary semantic information, we leverage it to refine the original pseudo-labels predicted from the model to obtain language-aware soft pseudo-labels, which is then used to guide the learning of the strongly augmented samples. Extensive experiments on various backbones and datasets demonstrate the effectiveness of LADaS. Codes are available at `https://github.com/zhennaziyu/ladas`. We summarize our contributions as follows:

- We present a novel LADaS framework for foundation model-based SSL, which aims to fully exploit the unlabeled data with the guidance of linguistic knowledge.
- We design a language-assisted pseudo-label debiasing loss to mitigate model bias, where a reliability score and a language-oriented preference are introduced to enhance the estimation of logit adjustment offset and model bias.
- we propose a language-aware pseudo-label smoothing loss, where the language-aware soft pseudo-label with linguistic knowledge is introduced to guide the learning of samples with low-quality pseudo labels.

2. Related Work

This work is related to SSL and PLMs in vision tasks.

Semi-Supervised Learning. Mainstream methods [15, 22, 32] primarily adopt pseudo-labeling and consistency regularization strategies for unlabeled data, with significant research attention focused on selecting high-quality pseudo-labels. For example, FixMatch [22] selects the pseudo-labels with the predicted probability higher than a fixed threshold. Since this fixed threshold may result in low data utilization during the early training phases, later studies resort to designing adaptive thresholds that adjust as training progresses. For example, Dash [34] and AdaMatch [21] gradually increase a global threshold as training progresses. FlexMatch [37] argues that different classes should have distinct class thresholds and assigns class-specific thresholds based on the learning difficulty of each class. Despite great success, these classical SSL methods train the model from scratch, which may require significant training time. With the rapid development of foundation models, further research efforts [16, 17, 30] have been made to adapt them to SSL models. For example, Wang [30] adopted pre-trained transformers into SSL to reduce the training steps while improving the performance. Since foundation models tend to exhibit biases toward some classes, DebiasPL [29] and FINESSL [9] introduce the logit adjustment strategy into the consistency regularization to reduce the model biases. Despite significant performance, existing efforts overlook the use of linguistic knowledge for pseudo-label debiasing. Additionally, they mainly learn from the unlabeled data with high confidence while discarding unlabeled data with low confidence, which may cause data wastage.

Pre-trained Language Models for Vision Tasks. Recently, promoting the vision model with the help of linguistic knowledge has shown great success. In particular, the vision-language models (VLMs), *e.g.*, CLIP [20], which learn joint visual-linguistic representations based on large-scale image-text pairs, can be adapted for downstream vision tasks [27, 38, 40]. For example, CoOp [45] designs continual prompt learning strategy to adapt the pre-trained vision-language models to different recognition tasks. Subsequently, CoCoOp [44] and VT-CLIP [19] devise an intermediate network to derive image tokens as the conditional inputs to inject the language information with vision signals for enhancing the performance. However, these methods primarily rely on jointly pre-trained VLMs to derive a retrieval-based classifier, while failing to extend to individually pre-trained vision models. Therefore, some studies focus on decoupling pre-trained vision-language models and leveraging linguistic knowledge to improve vision models. For example, CaFo [39] generates human-like sentences with diverse contents with the large-scale pre-trained GPT-3 [5] to promote learning in the vision models. More recently, BorLan [17] and LSG [16] explicitly leverage the linguistic knowledge of PLMs to enhance vision model performance. In contrast, we focus on utilizing linguistic knowledge to enhance pseudo-label debiasing and facilitate the learning of samples with low-quality pseudo-labels.

3. Method

3.1. Preliminaries

Semi-Supervised Learning Pipeline. Formally, suppose we have a labeled dataset $\mathcal{D}^l = \{\boldsymbol{x}_i^l, y_i^l\}_{i=1}^N$ with N labeled samples and an unlabeled dataset $\mathcal{D}^u = \{\boldsymbol{u}_j\}_{j=1}^U$ with U unlabeled samples, where $N << U$. Then, $\boldsymbol{x}_i^l$ represents the i-th labeled sample, and $y_i^l \in \mathbb{R}^C$ is the corresponding ground truth label encoded as a one-hot vector, with C being the number of classes. The objective of SSL is to train a model $f(\boldsymbol{x})$ as a C-class classifier, parameterized by θ that utilizes both labeled and unlabeled datasets.

In general, recent SSL methods [9, 29] usually rely on the supervised loss $\mathcal{L}_s$ for the labeled data, and an unsupervised loss $\mathcal{L}_u$ for the unlabeled data as the optimization objective, defined as $\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u$, where λ_u is a scalar hyper-parameter. Specifically, given a mini-batch of B labeled data, the supervised loss usually utilizes the standard cross-entropy loss $\mathcal{H}$ as follows:

$$\mathcal{L}_s = \frac{1}{B}\sum_{i=1}^{B} \mathcal{H}(y_i^l, p(y|\omega(\boldsymbol{x}_i^l))), \tag{1}$$

where $\omega(\cdot)$ denotes the weak data augmentation and $p(y|\omega(\boldsymbol{x}_i^l)) = \mathrm{Softmax}(f(\omega(\boldsymbol{x}_i^l)))$ denotes the predicted class probability for weakly-augmented labeled sample $\boldsymbol{x}_i^l$.

For the unsupervised loss, the majority of existing studies [22, 31] depend on pseudo-labeling to generate pseudo-labels from the weakly-augmented unlabeled samples and followed by consistency regularization, which supervises the model predictions of the strongly-augmented samples based on the cross-entropy loss. To maintain the quality of the pseudo-labels, only samples whose predicted probability exceeds a confidence threshold τ are retained, while others are discarded. As foundation models tend to generate more pseudo-labels for their biased classes, recent studies [9, 29] integrate the logit adjustment strategy into the unsupervised loss to reduce the influence of model biases. To be specific, given a mini-batch of μB unlabeled data, the debiased unsupervised loss is formulated as follows,

$$\mathcal{L}_u = -\frac{1}{\mu B}\sum_{j=1}^{\mu B} \mathbb{I}(\max(\boldsymbol{q}_j) \geq \tau) \log \frac{\exp(\boldsymbol{z}_{\hat{y}_j} + \Delta_{\hat{y}_j}^t)}{\sum_{k=1}^{C} \exp(\boldsymbol{z}_k + \Delta_k^t)}, \tag{2}$$

where $\boldsymbol{z}_{\hat{y}} = f(\Omega(\boldsymbol{u}_j))$ is the model's prediction of the strongly augmented unlabeled sample $\boldsymbol{u}_j$. $\Omega(\cdot)$ is the strong augmentation and $\mathbb{I}(\cdot > \tau)$ denotes the indicator function to select reliable pseudo-labels which are higher than the threshold τ. μ refers to the hyper-parameter used to control the relative batch size ratio between labeled and unlabeled data. $\boldsymbol{q}_j = p(y|\omega(\boldsymbol{u}_j))$ is predicted class probability of the weakly augmented unlabeled sample $\boldsymbol{u}_j$, and $\hat{y}_j = \textbf{Onehot}\{\mathrm{argmax}(\boldsymbol{q}_j)\}$ is the pseudo-label of $\boldsymbol{u}_j$. Δ_k^t denotes the dynamic logit adjustment offset of the k-th class at the t-th training step, which indicates the model's bias to that class. In particular, existing studies [9] leverage the class learning pace (*e.g.*, the pseudo-label frequency) to determine the logit adjustment offsets, assigning larger offset values to biased classes.

3.2. Language-Assisted Pseudo-Label Debiasing

Existing efforts ignore the reliability of the learning pace as the training progresses. Furthermore, linguistic knowledge, which provides complementary information to images, has not been used to influence logit adjustment in SSL. Therefore, we design the linguistic knowledge-assisted logit adjustment, composed of the self-adaptive logit adjustment

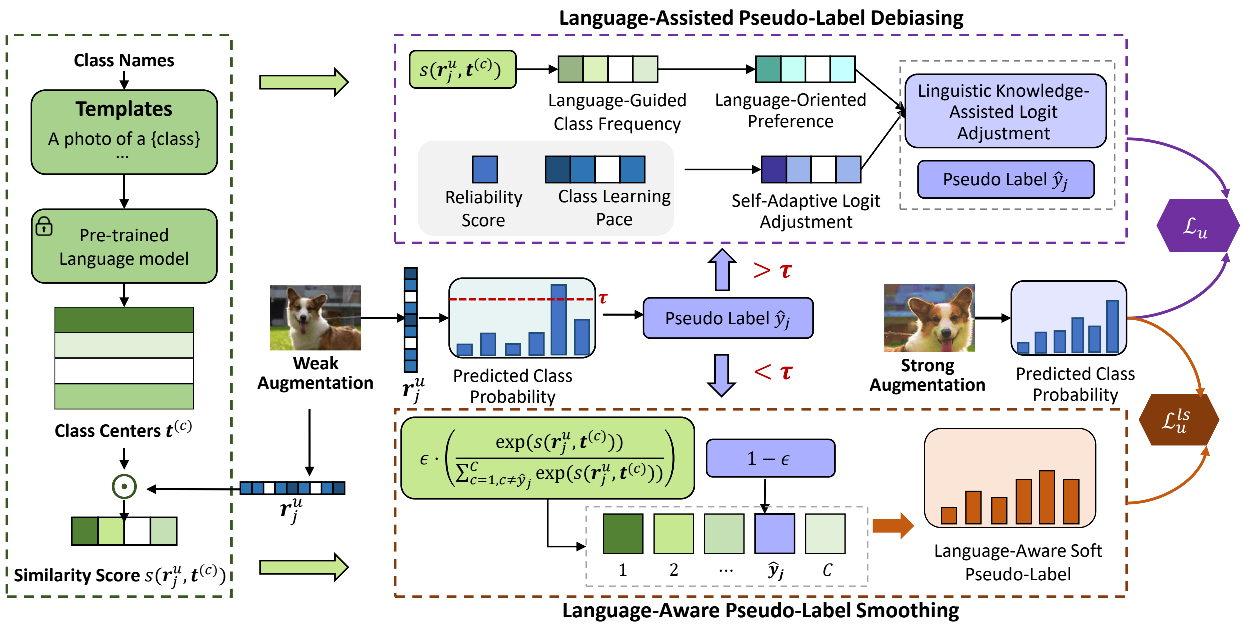


Figure 2. Illustration of LADaS, showcasing the pathway for unlabeled data, which consists of two key components: language assisted pseudo-labeling debiasing and language-aware pseudo-label smoothing.

and the language-oriented preference. The former works by dynamically injecting a reliability score which reflects the reliability of class learning pace. The latter focuses on estimating the model's preferences with linguistic knowledge.

Self-Adaptive Logit Adjustment. Similar to previous studies [29], given that the predicted class probability reflects the model's confidence in each class, we regard the average predicted probability for each class across all the unlabeled samples as the class learning pace. However, using all the unlabeled samples at each training step leads to significant time consumption. Hence, we follow [31] to estimate learning pace $l_t(c)$ using the EMA strategy. Specifically, we initialize $l_t(c)$ to $1/C$, since we assume all classes are balanced prior to training. Then we derive learning pace at the t-th training step for the c-th class as follows,

$$l_t(c) = \lambda l_{t-1}(c) + (1-\lambda)\frac{1}{\mu B}\sum_{j=1}^{\mu B} \boldsymbol{q}_j(c), \tag{3}$$

where $\boldsymbol{q}_j(c)$ denotes the predicted probability for class c in $\boldsymbol{q}_j$. $\lambda \in (0, 1)$ is the momentum decay of EMA.

Intuitively, during the early stages of training, the model is not stable and the estimated class learning pace can be unreliable and easily influenced by initialization. Directly incorporating this learning pace for pseudo-label debiasing could negatively affect model training. In light of this, we introduce the reliability score that is dynamically updated to reflect the reliability of the class learning pace throughout the training process. In essence, the reliability score should be related to model's learning status, where a relatively small value should be assigned at the initial training stage while gradually increasing to allow the model to concentrate on reducing its biases. Inspired by [31], we leverage the model's confidence in the unlabeled data which indicates the model's learning status to derive the reliability score. Similar to l_t, we initialize g_t to $1/C$ and update it with EMA at the t-th training step as follows,

$$g_t = \lambda g_{t-1} + (1-\lambda)\frac{1}{\mu B}\sum_{j=1}^{\mu B} \max(\boldsymbol{q}_j). \tag{4}$$

Based on class learning pace and the reliability score, we derive the self-adaptive logit adjustment offset $\Delta_t^v(c)$ for the c-th class at the t-th training step as follows,

$$\Delta_t^v(c) = \log(g_t \cdot \frac{l_t(c)}{\max\{l_t(c) : c \in [1, C]\}}), \tag{5}$$

where we apply maximum normalization to the class learning pace across all classes, following [9].

Language-Oriented Preference. We use a PLM to extract semantic information for each class. Specifically, given C class names $\{\mathcal{W}_c\}_{c=1}^{C}$, we adopt M predefined prompts, such as "a photo of a clean {class name}", to obtain complete M sentences for each class. We then obtain M text embeddings for each class by feeding the sentences into the frozen PLM, denoted by $\{\boldsymbol{t}_1^{(c)}, \boldsymbol{t}_2^{(c)}, \cdots, \boldsymbol{t}_M^{(c)}\}$, where $\boldsymbol{t}_m^{(c)} \in \mathbb{R}^{d_t}$, and d_t is the dimension of the text embedding. Finally, we define class centers as the mean of the L_2-normalized embeddings for each class, denoted by $\boldsymbol{t}^{(c)} = 1/M \sum_{m=1}^{M} \boldsymbol{t}_m^{(c)} / ||\boldsymbol{t}_m^{(c)}||_2$.

In essence, when most unlabeled samples tend to be closer to a certain class center, it may indicate an inherent bias of the model towards that class. Hence, we introduce the language-guided class frequency which represents the frequency of samples that are close to a given class center. In particular, inspired by CLIP [20], we first obtain the similarity scores between the unlabeled samples and each class center based on the cosine similarity. Specifically, given an unlabeled sample $\boldsymbol{u}_j$, we calculate its similarity score with the c-th class center as follows,

$$s(\boldsymbol{r}_j^u, \boldsymbol{t}^{(c)}) = \frac{\boldsymbol{r}_j^u \cdot \boldsymbol{t}^{(c)}}{||\boldsymbol{r}_j^u||_2 ||\boldsymbol{t}^{(c)}||_2}, \tag{6}$$

where $\boldsymbol{r}_j^u$ denotes the image representation derived from the vision model for the weakly-augmented unlabeled sample $\boldsymbol{u}_j$. Next, we determine the closest class center to $\boldsymbol{u}_j$ by selecting the class with highest similarity score as follows,

$$\hat{y}_j^l = \operatorname*{argmax}_{c \in C} s(\boldsymbol{r}_j^u, \boldsymbol{t}^{(c)}), \tag{7}$$

where $\hat{y}_j^l$ denotes the closest class center to the unlabeled sample $\boldsymbol{u}_j$. Thereafter, we obtain the language-guided class frequency $\beta_t^l(c)$ for each class as follows,

$$\beta_t^l(c) = \begin{cases} 1/C, & \text{if } t = 0, \\ \lambda \beta_{t-1}^l(c) + (1-\lambda)\dfrac{1}{\mu B}\displaystyle\sum_{j=1}^{\mu B} \mathbb{I}(\hat{y}_j^l = c), & \text{otherwise} \end{cases} \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Similar to $l_t(c)$, we employ the EMA strategy and initialize $\beta_t^l(c)$ as $1/C$.

Thereafter, we obtain the language-oriented preference $\Delta_t^l(c)$ for each class as follows,

$$\Delta_t^l(c) = \frac{\beta_t^l(c) - \min\{\beta_t^l(c) : c \in [1, C]\}}{\max\{\beta_t^l(c) : c \in [1, C]\} - \min\{\beta_t^l(c) : c \in [1, C]\}}, \tag{9}$$

where we apply min-max normalization to assign relatively larger preferences to classes with higher frequencies. Then we obtain the linguistic-knowledge injected logit adjustment offset by combining the self-adaptive logit adjustment offset and language-oriented preference, as follows,

$$\Lambda_c^t = \Delta_t^v(c) \cdot \Delta_t^l(c). \tag{10}$$

Finally, we define the linguistic knowledge adapted debiasing loss as follows,

$$\mathcal{L}_u = -\frac{1}{\mu B}\sum_{j=1}^{\mu B} \mathbb{I}(\max(\boldsymbol{q}_j) \geq \tau) \log \frac{\exp(\boldsymbol{z}_{\hat{y}} + \Lambda_{\hat{y}}^t)}{\sum_{c=1}^{C} \exp(\boldsymbol{z}_c + \Lambda_c^t)}. \tag{11}$$

3.3. Language-Aware Pseudo-Label Smoothing

To promote the utilization of unlabeled data, we focus on leveraging samples with probabilities below the confidence threshold, referred to as low-confidence samples, which are simply discarded in previous studies [9, 22]. Intuitively, directly using the pseudo-labels produced by the model for these samples may introduce noise. Since linguistic knowledge provides supplementary semantic information, it can be leveraged to refine the pseudo-labels. In light of this, we propose language-aware pseudo-label smoothing to exploit these low-confidence samples by refining pseudo-labels with the guidance of linguistic knowledge.

In particular, as label smoothing [46] has shown great success in mitigating the influence of noisy labels, we apply this strategy to exploit the low-confidence samples. A naive way is to directly utilize the vanilla method where the soft label is mixed with a uniform mixture over all possible labels. However, in our SSL context, as the obtained pseudo-labels may be noisy, we are more interested in increasing the probability of the potential ground-truth labels in the soft label. As the class centers provide useful semantic information related to each class, we leverage them to refine the pseudo-labels by increasing the probabilities of classes with higher similarity scores.

Specifically, given an unreliable sample $\boldsymbol{u}_j$, where $\max(\boldsymbol{q}_j) < \tau$, we first obtain its pseudo-label assigned by the vision model $\hat{y} = \operatorname{argmax}(\boldsymbol{q}_j)$. Then we calculate similarity scores between the samples and each class center based on the similarity by Eq. (6). Next, we define the language-aware soft pseudo-label $\boldsymbol{y}_j^s$ for the sample $\boldsymbol{u}_j$ as,

$$\boldsymbol{y}_j^s[c] = \begin{cases} 1-\epsilon, & \text{if } c = \hat{y}_j, \\ \epsilon \cdot \dfrac{\exp(s(\boldsymbol{r}_j^u, \boldsymbol{t}^{(c)}))}{\sum_{c=1, c\neq \hat{y}_j}^{C} \exp(s(\boldsymbol{r}_j^u, \boldsymbol{t}^{(c)}))}, & \text{otherwise.} \end{cases} \tag{12}$$

where $\boldsymbol{y}_j^s \in \mathbb{R}^C$ and $\boldsymbol{y}_j^s[c]$ is the c-th value in $\boldsymbol{y}_j^s$. We set the probability of the predicted pseudo-label as $1-\epsilon$, while the rest are derived based on the similarity scores.

Finally, we define the language-aware pseudo-label smoothing loss to regularize the consistency between the refined soft pseudo-label and the predicted class probability of each low-confidence sample under strong augmentation with the cosine similarity as follows,

$$\mathcal{L}_u^{ls} = -\frac{1}{\mu B}\sum_{j=1}^{\mu B} \mathbb{I}(\max(\boldsymbol{q}_j) < \tau)({\boldsymbol{y}_j^s}^\top p(y|\Omega(\boldsymbol{u}_j))). \tag{13}$$

3.4. Overall Loss Objective

Introducing linguistic knowledge requires ensuring feature alignment between the text embeddings and the image representations. Hence, we additionally incorporate the feature alignment loss $\mathcal{L}_s^{\text{text}}$ on labeled data, following BorLan [17]. Ultimately, the overall loss function of our proposed model is as follows,

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_s^{\text{text}} + \lambda_u \mathcal{L}_u + \lambda_{ls} \mathcal{L}_u^{ls}, \tag{14}$$

where λ_u and λ_{ls} are hyper-parameters.

Dataset	FOOD-101			CIFAR-10			CIFAR-100		
Settings	N2	N4	N10	N1	N2	N4	N4	N25	N100
PL [13]	41.78 ± 1.18	22.43 ± 0.93	15.38 ± 0.12	51.55 ± 2.99	25.67 ± 10.97	7.20 ± 0.38	33.83 ± 0.30	18.55± 0.30	15.35 ± 0.20
FixMatch [22]	25.08 ± 1.09	15.07 ± 0.70	11.28 ± 0.02	39.73 ± 1.89	12.69 ± 0.10	2.57 ± 0.14	22.90 ± 1.04	15.95 ± 0.16	13.83 ± 0.03
FlexMatch [37]	19.04 ± 1.68	11.49 ± 0.13	10.76 ± 0.11	38.29 ± 15.07	7.39 ± 7.31	2.59 ± 0.10	20.81 ± 0.23	16.23 ± 0.13	13.77 ± 0.25
FreeMatch [31]	25.69 ± 4.56	15.31 ± 1.11	11.03 ± 0.24	39.59 ± 1.78	12.96 ± 0.06	2.74 ± 0.14	23.81 ± 0.82	16.40± 0.16	13.66 ± 0.30
SoftMatch [6]	24.18 ± 5.24	14.95 ± 1.05	10.99 ± 0.08	41.26 ± 11.44	9.28 ± 5.28	2.65 ± 0.06	24.06 ± 0.46	16.15 ± 0.06	13.69 ± 0.07
DebiasPL [29]	13.49 ± 0.74	11.40 ± 0.34	10.68 ± 0.17	21.7 ± 16.85	3.17± 0.14	2.61 ± 0.06	20.43 ± 032	15.99 ± 0.07	13.84 ± 0.13
FINESSL [9]	12.96 ± 0.43	10.72 ± 0.14	10.31 ± 0.15	3.85 ± 0.13	3.13± 0.10	2.61 ± 0.05	19.56 ± 0.24	15.49 ± 0.01	13.34 ± 0.18
LADaS	**10.85 ± 0.79**	**10.13 ± 0.06**	**9.77 ± 0.05**	**3.15± 0.03**	**2.68 ± 0.01**	**2.53 ± 0.08**	**18.75 ± 0.14**	**15.16 ± 0.09**	**13.04 ± 0.11**

Table 1. Performance comparison with ViT-VPT backbone on FOOD-101, CIFAR-10 and CIFAR-100 datasets: means ± std of error rates (%) over 3 trials are reported. N# denotes the number of labeled samples for each class. Results of other methods are directly copied from FINESSL [9]. The best results are highlighted in **bold** and the second-best underlined.

Settings	$\mathcal{D}^u = \mathcal{D}^u_{in}$	$\mathcal{D}^u = \mathcal{D}^u_{in} \cup \mathcal{D}^u_{out}$
FixMatch [22]	34.48 ± 0.30	39.85 ± 0.16
FlexMatch [37]	35.40 ± 0.25	42.04 ± 0.11
FreeMatch [31]	35.43 ± 0.23	41.99 ± 0.38
SoftMatch [6]	35.25 ± 0.27	42.21 ± 0.52
DebiasPL [29]	33.18 ± 0.35	39.15 ± 0.37
FINESSL [9]	32.75 ± 0.11	**38.88 ± 0.07**
LADaS	**31.77± 0.16**	39.14 ± 0.21

Table 2. Error rate (%) with ViT-VPT backbone on Semi-Aves dataset. $\mathcal{D}^u$ denotes the unlabeled data sampled from the same classes as labeled data and $\mathcal{D}^u_{out}$ contains OOD samples.

Settings	N4	N25	N100
FixMatch [22]	57.97	29.99	21.69
ReMixMatch [3]	53.15	31.44	20.91
Co-Tuning [36]	57.58	30.94	22.22
FlexMatch [37]	56.89	30.13	22.70
Self-Tuning [28]	47.17	24.16	17.57
BorLan [17]	44.82	23.07	16.56
LADaS	**41.27**	**21.24**	**15.31**

Table 3. Error rate (%) with EfficientNet-B2 backbone on CIFAR-100 dataset with 4, 25 and 100 labeled samples for each class. Results of other methods are directly copied from BorLan [17].

4. Experiments

4.1. Experimental Settings

Datasets. Following previous studies [9, 10], we conduct extensive experiments on five publicly available datasets to evaluate our method. Specifically, we use CIFAR-10 [12], CIFAR-100 [12], STL-10 [7], FOOD-101 [4] and Semi-Aves [24]. Detailed descriptions for each dataset can be found in supplemental materials.

Implementation Details. We evaluate LADaS primarily based on VPT with ViT backbone in CLIP (ViT-VPT) [11] on CIFAR-10, CIFAR-100, FOOD-101 and Semi-Aves datasets. In addition, to comprehensively validate the effectiveness of LADaS, we further conduct experiments to fine-tune the pre-trained EfficientNet-B2 [26] on CIFAR-100 and pre-trained vision transformer (ViT-FFT) [14] on CIFAR-10, CIFAR-100 and STL-10 datasets. For all backbones, we adopt stochastic gradient descent (SGD) with a momentum of 0.9 as the optimizer and the total training steps are set to $15,000$. Following the previous study BorLan [17], we utilize 80 handcrafted prompts in CLIP [20] and use the frozen pre-trained Bert-Large [8] (BertL) as the pre-trained language model to produce text embeddings for each class. More implementation details can be found in supplemental materials.

Evaluation Metrics. For each dataset, we use different numbers of labeled samples to cover various SSL scenarios. For the ViT-VPT backbone, following previous studies [6, 9], we conduct experiments with $\{1, 2, 4\}$ labeled samples per class on CIFAR-10, $\{4, 25, 100\}$ labeled samples per class for CIFAR-100, and $\{2, 4, 10\}$ labeled samples per class for FOOD-101. As for Semi-Aves, we conduct experiments for two scenarios of unlabeled data: one with only in-distribution samples, and another with a mix of in-distribution and out-of-distribution (OOD) samples. As for the EfficientNet-B2 backbone, we follow BorLan [17] to select $\{4, 25, 100\}$ labeled samples per class for CIFAR-100. For the ViT-FFT backbone, we choose $\{1, 4, 25\}$ labeled samples per class for CIFAR-10, $\{2, 4, 25\}$ labeled samples per class for CIFAR-100 and $\{1, 4, 10\}$ labeled samples per class for STL-10. The classification performance is evaluated using the error rate on the test set.

4.2. Performance Comparison

We compare LADaS with existing state-of-the-art SSL methods, including PL [13], FixMatch [22], FlexMatch [37], FreeMatch [31], SoftMatch [6], DebiasPL [29], FINESSL [9], Co-Tuning [36], ReMixMatch [3], Self-

Dataset	CIFAR-10			CIFAR-100			STL-10		
Settings	N1	N4	N25	N2	N4	N25	N1	N4	N10
FixMatch [22]	33.50 $\pm$ 15.1	2.56 $\pm$ 0.9	2.05 $\pm$ 0.1	34.71 $\pm$ 1.4	24.48 $\pm$ 0.1	16.02 $\pm$ 0.1	59.87 $\pm$ 3.4	22.28 $\pm$ 4.4	11.59 $\pm$ 1.6
FlexMatch [37]	29.46 $\pm$ 9.6	2.22 $\pm$ 0.3	2.12 $\pm$ 0.2	36.24 $\pm$ 0.9	25.99 $\pm$ 0.5	16.28 $\pm$ 0.2	39.37 $\pm$ 12.9	21.83 $\pm$ 3.7	10.46 $\pm$ 1.3
InstanT [14]	12.68 $\pm$ 10.2	2.07 $\pm$ 0.1	1.92 $\pm$ 0.1	25.83 $\pm$ 0.3	21.20 $\pm$ 0.4	15.72 $\pm$ 0.5	30.61 $\pm$ 7.4	14.91 $\pm$ 2.8	10.65 $\pm$ 1.9
SemiReward [15]	–	–	–	<u>20.06 $\pm$ 0.41</u>	<u>15.62 $\pm$ 0.71</u>	–	–	9.72$\pm$ 0.62	7.10$\pm$ 0.39
InterLUDE [10]	31.90 $\pm$ 4.1	1.78 $\pm$ 0.1	1.55 $\pm$ 0.1	35.66 $\pm$ 1.9	21.19 $\pm$ 0.2	13.39 $\pm$ 0.1	27.49 $\pm$ 6.6	**3.14 $\pm$ 0.2**	**2.66 $\pm$ 0.1**
InterLUDE+ [10]	<u>12.29 $\pm$ 7.3</u>	<u>1.55 $\pm$ 0.1</u>	<u>1.49 $\pm$ 0.1</u>	23.60 $\pm$ 1.2	16.32 $\pm$ 0.3	<u>12.93 $\pm$ 0.2</u>	<u>25.83 $\pm$ 9.9</u>	4.56 $\pm$ 0.9	3.23 $\pm$ 0.3
LADaS	**7.61 $\pm$ 5.34**	**1.17$\pm$ 0.02**	**1.14 $\pm$ 0.01**	**18.31 $\pm$ 1.65**	**15.04 $\pm$ 0.09**	**12.31 $\pm$ 0.32**	**10.87 $\pm$ 1.81**	<u>3.91 $\pm$ 0.77</u>	<u>2.69 $\pm$ 0.36</u>

Table 4. Error rate (%) with ViT-FFT backbone. Note that for limited space, more results can be found in supplemental materials.

AM	Base	l_t	g_t	Δ_t^l	$\mathcal{L}_u^{ls}$	CIFAR-10			CIFAR-100			FOOD-101		
						N1	N2	N4	N4	N25	N100	N2	N4	N10
1	✓					31.16	12.72	2.63	23.91	16.07	13.79	23.11	14.90	10.71
2	✓	✓				21.97	3.69	2.72	21.92	15.94	13.58	15.20	12.91	10.01
3	✓	✓	✓			5.06	3.85	2.62	20.58	15.98	13.52	13.29	10.99	10.02
4	✓	✓	✓	✓		3.24	3.00	2.63	19.13	15.38	12.96	12.80	11.51	9.83
5	✓				✓	12.83	3.65	2.46	22.23	15.35	13.26	21.08	13.36	10.36
6	✓	✓	✓		✓	4.43	3.09	2.47	19.15	15.31	13.42	12.02	10.42	10.01
7	✓	✓	✓	✓	✓	3.15	2.68	2.53	18.75	15.16	13.04	10.85	10.13	9.77

Table 5. Ablation study with ViT-VPT backbone on CIFAR-10, CIFAR-100, and FOOD-101 datasets.

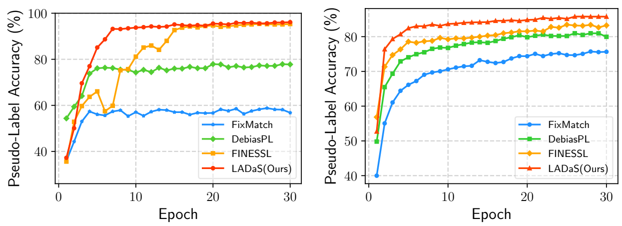
Tuning [28], BorLan [17]InstanT [14], SemiReward [15], InterLUDE [10] and InterLUDE+ [10]. The comparison results are shown in Tab. 1, Tab. 2 for ViT-VPT, Tab. 3 for EfficientNet-B2 and Tab. 4 for ViT-FFT, respectively.

From these tables, we have the following observations. 1) LADaS consistently outperforms existing methods for many scenarios across diverse backbones and datasets, while remaining competitive in others. This shows that the proposed method can be applied to various pure image pre-trained backbones and demonstrates the superiority of our proposed framework over the existing studies. 2) From Tab. 1 and Tab. 2, we notice that the methods that explicitly incorporate the pseudo-label debiasing strategies, including DebiasPL, FINESSL and our proposed LADaS, outperform the rest baselines. The reason behind this improvement may be that the pre-trained foundation models are biased to specific classes, and without debiasing strategies, the training process may accumulate confirmation biases and affect the final performance. 3) LADaS outperforms FINESSL and DebiasPL. The reason may be attributed to our introduction of linguistic knowledge to refine the logit adjustment offset. Moreover, we design the reliability score to determine the reliability of the logit adjustment derived by the vision model, which avoids the model to be disturbed. Meanwhile, from Tab. 3, we find that both BorLan and LADaS which inject linguistic knowledge for SSL, outperform the other baselines. These two findings indicate the importance of distilling linguistic knowledge for semi-supervised learning. 4) LADaS shows superiority over BorLan, the reason may be that we propose the linguistic knowledge adapted debiasing loss by introducing linguistic-knowledge injected logit adjustment offset into the consistency regularization, which can reduce the biases presented in a pre-trained foundation model.

4.3. Ablation Study

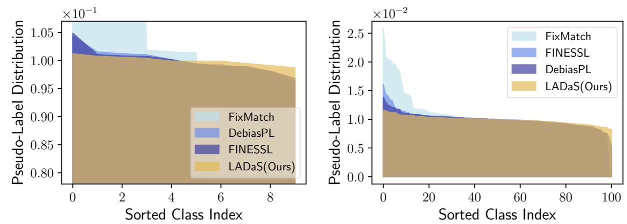
To verify the influence of each component in our model, we conduct the ablation study on CIFAR-10, CIFAR-100 and FOOD-101 datasets with ViT-VPT backbone, as shown in Tab. 5. In particular, the base model is implemented by removing the key components, *i.e.*, language-assisted pseudo-label debiasing and the language-aware pseudo-label smoothing from the whole framework. **AM** is the the abbreviation for Ablation Study. **AM7** denotes our full framework LADaS. Similarly, the performance is evaluated with error rate on test set for each ablation method.

From this table, we have the following observations. 1) Based on the error rate in **AM2** and **AM1**, we can see that introducing the logit adjustment based on the learning pace can boost the performance, demonstrating the importance of pseudo-label debiasing for adapting the foundation models to semi-supervised learning task. 2) **AM3** outperforms **AM2**, especially for the setting of N1 on CIFAR-10. This verifies the necessity of dynamically accessing the reliability of the class learning pace. 3) **AM4** outperforms **AM3**, and the same findings can be observed when



(a) (b)

Figure 3. (a) and (b): The pseudo-label accuracy on CIFAR-10 with N1 and FOOD-101 with N2 settings, respectively.



(a) (b)

Figure 4. (a) and (b): The pseudo-label distribution on CIFAR-10 with N1 and FOOD-101 with N2 settings, respectively. Note that FixMatch for CIFAR-10 exceeds beyond this figure to about 35%.

comparing **AM7** to **AM6**. These observations indicate that the designed language-oriented preference can facilitate the logit adjustment offset by leveraging the linguistic knowledge from the pre-trained language models. 4) Without the language-aware pseudo-label smoothing loss, **AM4** shows inferiority to our whole framework **AM7**. We also notice that adding $\mathcal{L}_u^{ls}$ to original base model, can significantly promote the performance. This demonstrates the effectiveness of incorporating low-confidence samples with the guidance of linguistic knowledge.

4.4. Qualitative Analysis

We conduct qualitative analyses of LADaS in comparison of other methods to get deeper insights. Additional visualization results are provided in the supplemental materials.

Quality of Pseudo-Labels. To get deeper insights, we visualize the pseudo-label accuracy for FixMatch, DebiasPL, FINESSL and LADaS, throughout training, which plays a vital role in boosting SSL performance. In particular, except for FixMatch, all other methods introduce debiased pseudo-labeling strategies. The results are shown in Fig. 3. From this figure, we notice the pseudo-labels generated by LADaS consistently show better accuracy than other methods during training, which verifies the effectiveness of our proposed language-assisted pseudo-label debiasing. Meanwhile, we observe that the methods equipped with debiased pseudo-labeling consistently predict better pseudo-labels than FixMatch. The reason behind may be that as the foundation models exhibit biases towards different classes, the pseudo-labels can be highly imbalanced and FixMatch accumulates such confirmation biases during training. Furthermore, we present the corresponding pseudo-label distributions in Fig. 4 (a) and (b). Similarly, we find that the pseudo-labels generated by our method are more balanced compared to FixMatch, DebiasPL and FINESSL, demonstrating the superiority of our proposed LADaS.

Different Strategies for Pseudo-Label Smoothing. We conduct comprehensive comparisons between the vanilla label smoothing (VLS) and language-aware pseudo-label smoothing (LPLS) on the base model to verify the effectiveness of the language-guided soft pseudo-label. As shown in Fig. 5, w/ LPLS consistently outperforms w/ VLS across all datasets. The reason may be that the language-guided soft pseudo-label can increase the probability of potential ground-truth labels for low-confidence samples, thus boosting the performance. Meanwhile, both w/ LPLS and w/ VLS outperform the base model, which demonstrates the importance of exploiting the low-confidence samples.

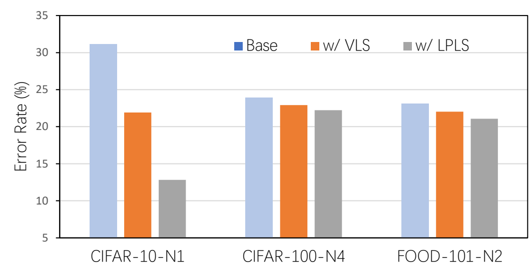


Figure 5. Comparison for different strategies of label smoothing with ViT-VPT. Dataset-N# denotes the dataset under N# setting.

5. Conclusion

In this paper, we propose the **L**anguage-**A**ssisted **D**ebiasing **a**nd **S**moothing (LADaS) method for the foundation model-based SSL task. Specifically, we design the language-assisted pseudo-label debiasing to mitigate biases inherent in foundation models. We dynamically inject a reliability score to determine the reliability of logit adjustment at each training step. Meanwhile, the language-oriented preference, which reflects the vision model's preferences with linguistic knowledge, is introduced to facilitate the pseudo-label debiasing. Furthermore, we design language-aware pseudo-label smoothing where we refine the pseudo-labels based on linguistic knowledge guidance to utilize samples with pseudo-labels below the confidence threshold. Extensive experiments verify the effectiveness of each key design in our LADaS. In particular, qualitative results show that our LADaS significantly promotes the pseudo-label quality. In the future, we plan to design more strategies leveraging linguistic knowledge to facilitate pseudo-label debiasing.

Acknowledgments: This research is supported by Singapore Ministry of Education Academic Research Fund Tier 2 under MOE's official grant number T2EP20221-0023.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International joint conference on neural networks*, pages 1–8. IEEE, 2020. 1

[2] 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*, 2019. 1

[3] 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*, 2020. 6

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 6

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and et al Askell. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 3

[6] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *International Conference on Learning Representations*, 2023. 6

[7] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 6

[8] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[9] Kai Gan and Tong Wei. Erasing the bias: Fine-tuning foundation models for semi-supervised learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 14453–14470. PMLR, 2024. 1, 2, 3, 4, 5, 6

[10] Zhe Huang, Xiaowei Yu, Dajiang Zhu, and Michael C Hughes. Interlude: Interactions between labeled and unlabeled data to enhance semi-supervised learning. *International Conference on Machine Learning*, 2024. 6, 7

[11] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 6

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. pages 32–33, 2009. 6

[13] 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*, page 896. Atlanta, 2013. 6

[14] Muyang Li, Runze Wu, Haoyu Liu, Jun Yu, Xun Yang, Bo Han, and Tongliang Liu. Instant: Semi-supervised learning with instance-dependent thresholds. In *Advances in Neural Information Processing Systems*, pages 2922–2938. Curran Associates, Inc., 2023. 1, 6, 7

[15] Siyuan Li, Weiyang Jin, Zedong Wang, Fang Wu, Zicheng Liu, Cheng Tan, and Stan Z. Li. Semireward: A general reward model for semi-supervised learning. In *The International Conference on Learning Representations*, 2024. 1, 2, 7

[16] Wenxuan Ma, Shuang Li, Jingxuan Kang, et al. Language semantic graph guided data-efficient learning. *Advances in Neural Information Processing Systems*, 36, 2023. 3

[17] Wenxuan Ma, Shuang Li, JinMing Zhang, Chi Harold Liu, Jingxuan Kang, Yulin Wang, and Gao Huang. Borrowing knowledge from pre-trained language model: A new data-efficient visual learning paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18786–18797. IEEE, 2023. 1, 3, 5, 6, 7

[18] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *International Conference on Learning Representations*, pages 1–13, 2021. 2

[19] Longtian Qiu, Renrui Zhang, Ziyu Guo, Ziyao Zeng, Zilu Guo, Yafeng Li, and Guangnan Zhang. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021. 3

[20] 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. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 5, 6

[21] Becca Roelofs, David Berthelot, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: a unified approach to semi-supervised learning and domain adaptation. In *International Conference on Learning Representations*, 2022. 2

[22] 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*, pages 596–608, 2020. 1, 2, 3, 5, 6, 7

[23] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. Neurostylist: Neural compatibility modeling for clothing matching. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 753–761. ACM, 2017. 1

[24] Jong-Chyi Su, Zezhou Cheng, and Subhransu Maji. A realistic evaluation of semi-supervised learning for fine-grained classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12966–12975. IEEE, 2021. 6

[25] Changchang Sun, Bin Duan, Hugo Latapie, Gaowen Liu, and Yan Yan. Dct: Divide-and-conquer transformer network with knowledge transfer for query-driven hoi detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1356–1364, 2024. 1

[26] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[27] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28578–28587. IEEE, 2024. 3

[28] Ximei Wang, Jinghan Gao, Mingsheng Long, and Jianmin Wang. Self-tuning for data-efficient deep learning. In *International Conference on Machine Learning*, pages 10738–10748. PMLR, 2021. 6, 7

[29] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657. IEEE, 2022. 1, 2, 3, 4, 6

[30] Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35:3938–3961, 2022. 1, 3

[31] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, , Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023. 1, 3, 4, 6

[32] Jianlong Wu, Haozhe Yang, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. Chmatch: Contrastive hierarchical matching and robust adaptive threshold boosted semi-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772. IEEE, 2023. 2

[33] Zhiliang Wu, Kang Zhang, Changchang Sun, Hanyu Xuan, and Yan Yan. Flow-guided deformable alignment network with self-supervision for video inpainting. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[34] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International conference on machine learning*, pages 11525–11536. PMLR, 2021. 2

[35] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022. 1

[36] Kaichao You, Zhi Kou, Mingsheng Long, and Jianmin Wang. Co-tuning for transfer learning. In *Advances in Neural Information Processing Systems*, pages 17236–17246. Curran Associates, Inc., 2020. 6

[37] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, pages 18408–18419, 2021. 1, 2, 6, 7

[38] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 3

[39] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222. IEEE, 2023. 2, 3

[40] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3479–3488. IEEE, 2023. 3

[41] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM International Conference on Multimedia*, page 266–274. ACM, 2019. 1

[42] Na Zheng, Xuemeng Song, Qingying Niu, Xue Dong, Yibing Zhan, and Liqiang Nie. Collocation and try-on network: Whether an outfit is compatible. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 309–317, 2021.

[43] Na Zheng, Xuemeng Song, Tianyu Su, Weifeng Liu, Yan Yan, and Liqiang Nie. Egocentric early action prediction via adversarial knowledge distillation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–21, 2023. 1

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825. IEEE, 2022. 3

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[46] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14890–14900. IEEE, 2023. 5