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

Supplementary Material

Here, we present the dataset details in Section 1, followed by the implementation details in Section 2. Finally, we provide additional experimental analysis in Section 3.

1. Dataset Details

In this section, we provide more details on the datasets we employ for experiments in this paper, including CIFAR-10, CIFAR-100, FOOD-101, Semi-Aves and STL-10.

CIFAR-10 [10] is an image classification dataset commonly used in various computer vision tasks and is also a classical benchmark for semi-supervised learning. The dataset consists of 60,000 images across 10 classes. Each class contains 6,000 images, which are split into 5000 images for training set and 1000 images for testing set. The image size is 32×32.

CIFAR-100 [10] is another popular SSL benchmark, which contains 100 fine-grained classes with 60,000 color images, where each class contains 600 images. The image size is also 32×32. The dataset is split in a fashion similar to CIFAR-10, with 50,000 images in the training set where each class contains 500 samples and 10,000 images in the testing set with 100 samples per class.

FOOD-101 [2] is an image classification dataset that is popular in various computer vision tasks, including SSL. The dataset comprises of 101,000 food images split across 101 classes. Each class contains 1,000 images, with 750 images in the training set, and 250 images in the testing set. As the images in this dataset are not fixed, we resize all images into 256×256 in our experiments

Semi-Aves [19] is a fine-grained image classification dataset containing 200 bird classes. The dataset includes a labeled training set comprising of 3959 images, a testing set with 8000 images, and finally an unlabeled training dataset with two parts. The first part (*i.e.*, in-distribution) has 26,640 unlabeled images are drawn from the same classes as the labeled training set. The second portion (*i.e.*, out-of-distribution) contains 122,208 samples from classes not present in the labeled training set.

STL-10 [3] is a 10-class image classification dataset extracted from ImageNet [4]. The dataset consists of 5000 labeled samples and 100,000 unlabeled samples for training set and 8000 samples for test set.

2. Implementation Details

Architecture. The general SSL architecture for vision tasks consists of an image encoder to obtain the image representation and a classifier to predict the class probability. In this work, we aim to introduce linguistic knowledge to guide the training. To create feature alignment between the image representation and the text embedding, we additionally add a projector head between the image encoder and the classifier, which maps the image representations into the same dimension space as the text embeddings. In particular, for ViT-VPT backbone, we adopt a fully connected layer as the projector head. For EfficientNet-B2 and ViT-FFT, we adopt a Multi-Layer Perceptron (MLP) layer, which consists of a fully connected layer with output dimensions as 1024, a ReLU activate layer, a Batch Normalization layer and a fully connected layer where the output dimensions relies on the text embedding dimension.

Training Details. Following BorLan [14], we utilize 80 handcraft prompts in CLIP [16] and use the frozen pre-trained Bert-Large [5] as the pre-trained language model to produce text embeddings for each class. Instead of using the [CLS] or [EOS] embedding, we directly utilize the output embedding associated with each class, following the approach of BorLan [14]. Moreover, we also calculate the covariance matrices for text embeddings with the diagonal form to calculate the feature alignment loss $\mathcal{L}s^{text}$ with the same manner in BorLan. 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 $15,000$. The momentum decay λ of EMA is set as 0.99 for all experiments. μ is set as 2. Following FINESSL [6], we train ViT-VPT with a learning rate of 0.03, where the parameters of CLIP are frozen. The learnable prompts has the length of 50 and we adopt a batch size of 32 and the weight decay is set as 5×10^{-4}. We fine-tune the model 30 epochs with 500 steps per epoch. Same with FINESSL, the threshold τ is set as 0.7. λ_u is set as 3 and λ_{ls} is set as 0.1. As for both EfficientNet-B2 and ViT-FFT backbones, we follow BorLan [14] to fine-tune the vision model with the learning rate of $1e{-}3$ and a $10\times$ larger value for classifier and projector head. The batch size of these two backbones are set as 64. The threshold τ is set as 0.7. λ_u and λ_{ls} are set as 1.

3. More Experimental Analysis

More Criteria. To comprehensively compare LADaS with the state-of-the-art methods in a classification setting, we further report the results of precision, recall, F1 score, and area under curve (AUC) on FOOD-101 with each class N2 setting, CIFAR-10 with N1 setting, and CIFAR-100 with

Datasets	FOOD-101-N2				CIFAR-10-N1				CIFAR-100-N4			
Criteria	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC
FixMatch [18]	72.37	73.30	69.09	98.41	53.11	67.64	57.86	94.17	76.79	76.91	74.98	99.11
DebiasPL [22]	85.15	86.11	85.20	99.48	77.30	77.20	77.24	85.76	79.45	79.69	79.40	99.24
FINESSL [6]	86.97	86.85	86.79	99.66	95.44	95.35	95.33	99.85	80.29	79.37	79.13	99.36
LADaS	**89.27**	**89.16**	**89.07**	**99.82**	**96.65**	**96.63**	**96.63**	**99.92**	**80.98**	**81.18**	**80.47**	**99.44**

Table 1. Precision, recall, F1 score, AUC results on FOOD-101, CIFAR-10 and CIFAR-100 datasets.

Dataset	CIFAR-10			CIFAR-100		
#Label	N1	N4	N25	N2	N4	N25
PL [11]	62.35 ± 3.1	11.79 ± 5.3	4.58± 0.4	36.66± 2.0	26.87 ± 0.9	15.72± 0.1
MT [20]	35.43 ± 4.9	12.85 ± 2.5	4.75 ± 0.5	40.50 ± 0.8	30.58 ± 0.9	17.09 ± 0.4
MixMatch [1]	34.96 ± 2.6	2.84 ± 0.9	2.05 ± 0.1	39.64 ± 1.3	27.74 ± 0.1	16.16 ± 0.3
VAT [15]	39.93 ± 6.3	6.67 ± 6.6	2.33 ± 0.2	34.11 ± 1.8	24.67 ± 0.4	16.58 ± 0.4
UDA [15]	21.24 ± 3.6	2.08 ± 0.2	2.04 ± 0.1	34.51 ± 1.6	24.15 ± 1.6	16.19 ± 0.2
FixMatch [18]	33.50 ± 15.1	2.56 ± 0.9	2.05 ± 0.1	34.71 ± 1.4	24.48 ± 0.1	16.02 ± 0.1
FlexMatch [26]	29.46 ± 9.6	2.22 ± 0.3	2.12 ± 0.2	36.24 ± 0.9	25.99 ± 0.5	16.28 ± 0.2
Dash [24]	25.64 ± 4.5	3.37 ± 2.0	2.10 ± 0.3	36.67 ± 0.4	25.46 ± 0.2	15.99 ± 0.2
AdaMatch [17]	14.85 ± 20.4	2.06 ± 0.1	2.08 ± 0.1	26.39 ± 0.1	21.41 ± 0.4	15.51 ± 0.1
InstanT [12]	12.68 ± 10.2	2.07 ± 0.1	1.92 ± 0.1	25.83 ± 0.3	21.20 ± 0.4	15.72 ± 0.5
FreeMatch [23]	–	–	–	21.07 ± 0.72	15.97 ± 0.24	–
SemiReward [13]	–	–	–	20.06 ± 0.41	15.62 ± 0.71	–
InterLUDE [9]	31.90 ± 4.1	1.78 ± 0.1	1.55 ± 0.1	35.66 ± 1.9	21.19 ± 0.2	13.39 ± 0.1
InterLUDE+ [9]	12.29 ± 7.3	1.55 ± 0.1	1.49 ± 0.1	23.60 ± 1.2	16.32 ± 0.3	12.93 ± 0.2
LADaS	**7.61 ± 5.34**	**1.17± 0.02**	**1.14 ± 0.01**	**18.31 ± 1.65**	**15.04 ± 0.09**	**12.31 ± 0.32**

Table 2. The complete error rate (%) with ViT-FINE backbone on CIFAR-10 and CIFAR-100 in table 4 in the main paper. The error rate and 95% confidence interval are reported based on three random seeds. The results of FreeMatch and SemiReward are copied from SemiReward, while other results are directly copied from InterLUDE [9]. The best results are highlighted in bold and the second-best underlined.

N4 setting. We adopt ViT-VPT as the backbone and select the top two strongest baselines, *i.e.*, DebiasPL and FINESSL, and the most popular SSL method FixMatch as comparison baselines. Performance results are shown in Tab. 1. As can been seen, our proposed LADaS achieves the best performance across all metrics on different datasets. This demonstrates the effectiveness of LADaS over the baseline methods.

Detailed Performance on ViT-FFT Backbone. We provide complete results on CIFAR-10 and CIFAR-100 datasets with ViT-FFT backbone in Tab. 2. As can be seen, our LADaS achieves the best performance compared to all the baselines, which demonstrates the superiority of our proposed method. Meanwhile, we notice that InterLUDE+ outperforms InterLUDE, which introduces the threshold-adjusting scheme of FreeMatch into InterLUDE. The reason behind this maybe that the threshold-adjusting scheme can play the role of implicit pseudo-label debiasing, as it assigns different thresholds to select pseudo-labels for each class. The same observation can be found by comparing FreeMatch and FixMatch. These observations verify the importance of pseudo-label debiasing in the foundation model-based semi-supervised learning. In contrast, our LADaS introduces explicit pseudo-label debiasing during the training process, leading to the best performance.

Convergence Speed. We visualize the change of top-1 accuracy on testing set during training process in Fig. 1, and compare LADaS with FixMatch, DebiasPL and FINESSL. From this figure, we can see that our LADaS converges very quickly, requiring fewer than 10 epochs to achieve near-optimal performance. Although FixMatch and DebiasPL also converge quickly, they tend to be constrained by sub-optimal performance. In contrast, our LADaS converges efficiently while achieving superior performance.

Impact of Different PLMs. To investigate the influence of different pre-trained language models, we conduct experiments on FOOD-101 dataset under the N2 setting with different PLMs, including SigLIP [25], BertL [5] and

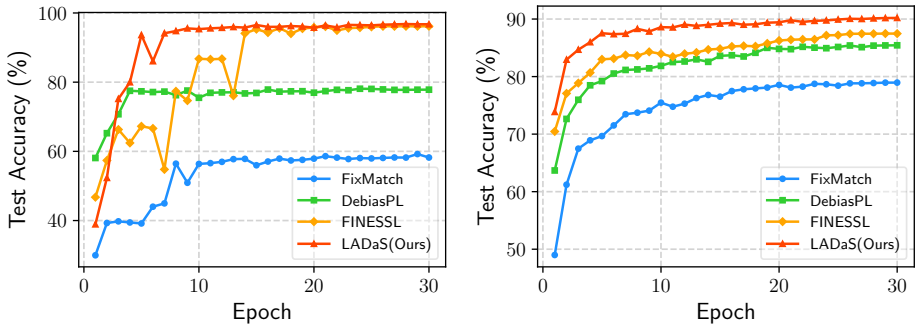

Figure 1. The top-1 accuracy on testing set for each epoch on (a) CIFAR-10 with one sample per class and (b) FOOD-101 with two samples per class.

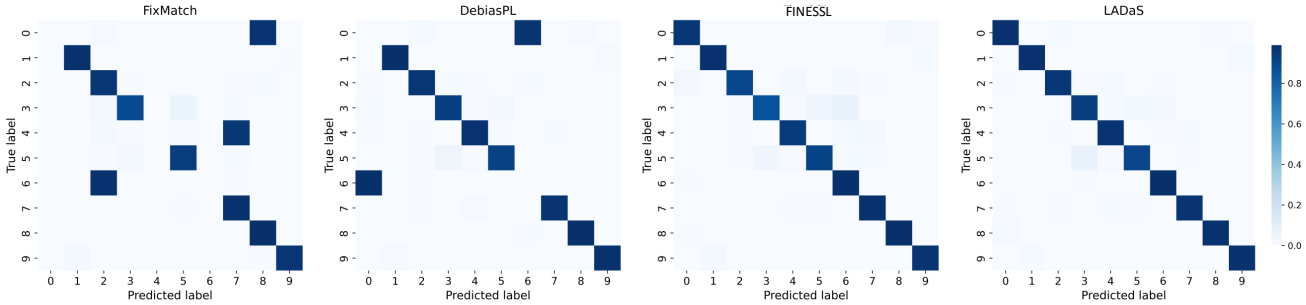

Figure 2. Confusion matrix on the testing set of CIFAR-10 with N1 setting.

CLIP [16], denoted as w/ SigLIP, w/ BertL and w/ CLIP, respectively. The performance is shown in Tab. 3, where w/o LAN denotes without linguistic knowledge. From this table, we notice that without linguistic knowledge, w/o LAN performs the worst, which indicates the importance of using linguistic knowledge for SSL task. Meanwhile, we find that w/ SigLIP performs worse than w/ BertL and w/ CLIP. The reason may be that the pre-trained PLMs inherently contains biases, which may influence the performance.

Method	w/o LAN	w/ SigLIP	w/ BertL	w/ CLIP
	13.29	11.64	10.85	10.67

Table 3. Error rate of different PLMs on the FOOD-101 dataset with the N2 setting.

Different Parameter-Efficient Fine-Tuning (PEFT). To get deeper insights in our proposed LADaS and verify its robustness on different PEFT strategies, we conduct experiments on CIFAR-100 and FOOD-100 with different PEFT strategies, including VPT, LoRA [8] and Adapter [7]. The performance is illustrated in Tab. 4. From this table, we notice that our proposed LADaS is robust to diverse PEFT strategies. Meanwhile, LoRA outperforms VPT for all settings across different datasets, which implies that LoRA is the more effective strategy than VPT for foundation model-based SSL task.

Settings	CIFAR-100		FOOD-101	
	N4	N25	N2	N4
VPT	18.75	15.49	10.85	10.13
LoRA	18.55	**15.11**	**10.33**	**9.77**
Adapter	**17.98**	15.44	11.34	9.91

Table 4. Error rate for different PEFT strategies with LADaS on CIFAR-100 and FOOD-101 datasets.

Confusion Matrix. We provide the confusion matrix which compares the true labels (y-axis) with predicted labels (x-axis) for FixMatch, DebiasPL, FINESSL and our LADaS on CIFAR-10 with N1 setting. The darker diagonal elements represent better classification performance. The off-diagonal elements indicate incorrect classifications. The results are shown in Fig. 2. From this table, we can observe that our LADaS achieves the best class-wise accuracy, as the majority of the dark elements are concentrated along the diagonal.

Impact of Hypermeter in Soft Pseudo-label. To study the influence of ϵ, we conduct experiments with ϵ values

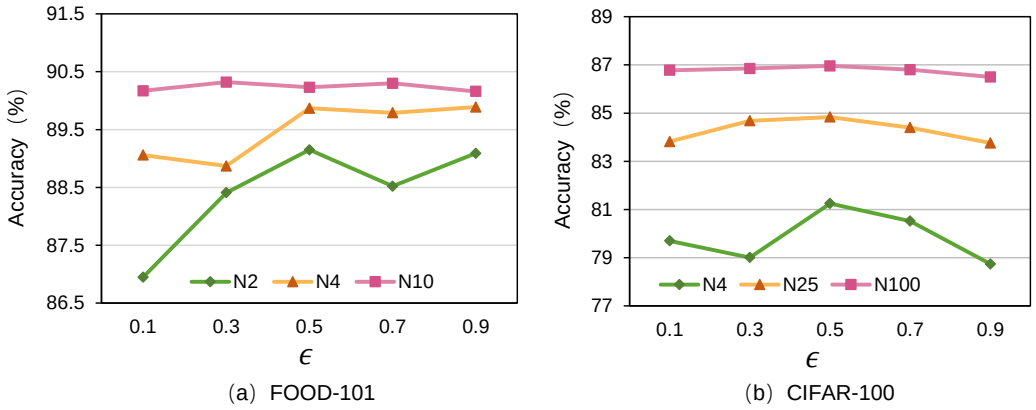

Figure 3. Impact of the hyperparameter ϵ in the soft pseudo-label on (a) FOOD-101 with N2, N4 and N10 settings and (b) CIFAR-100 with N4, N25 and N100 settings.

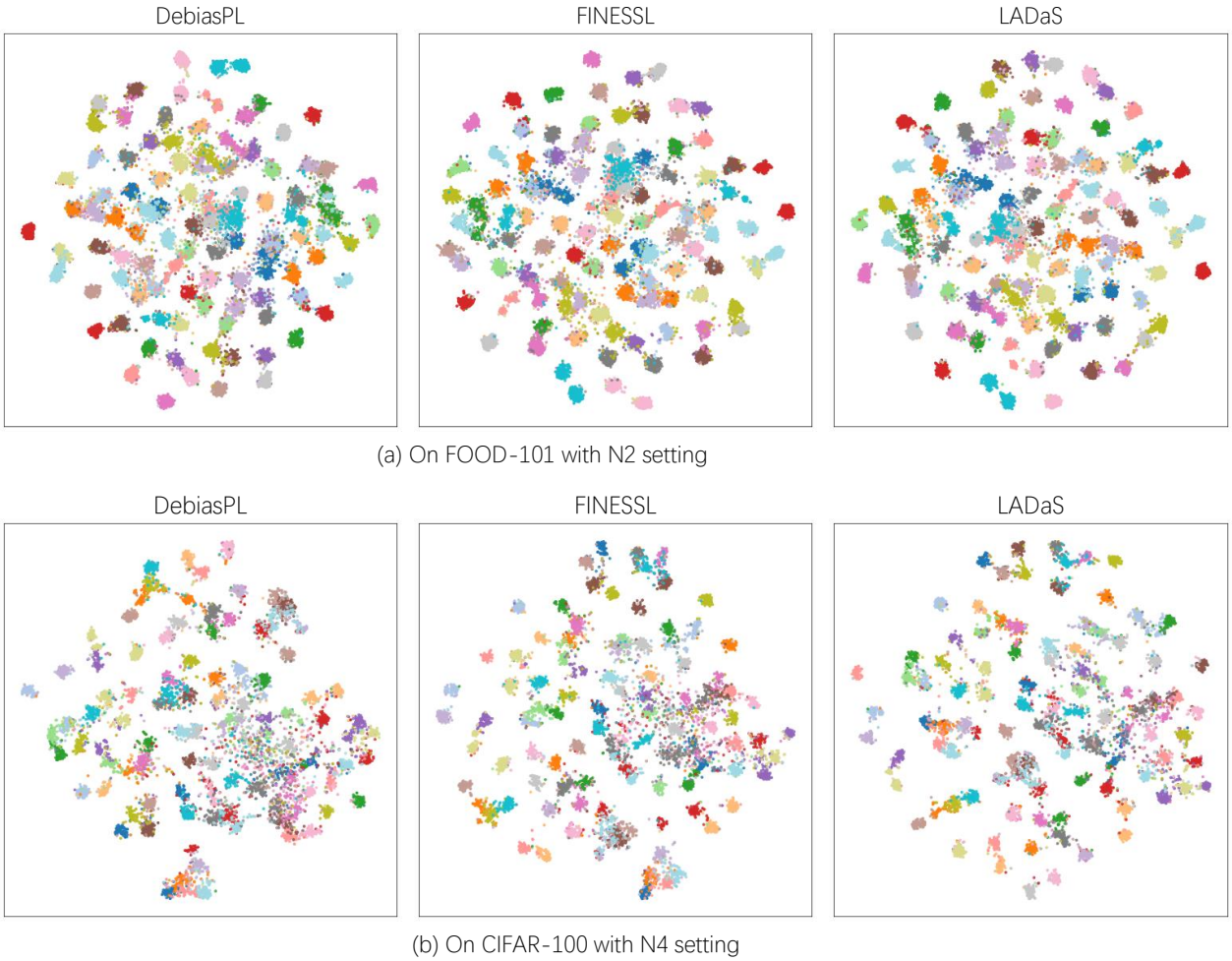

Figure 4. T-SNE visualization of features for test data on CIFAR-100 with 4 samples per class and FOOD-101 with 2 samples per class, respectively.

ranging from 0.1, 0.3, 0.5, 0.7 to 0.9. As shown in Fig. 3, the performance grows with ϵ increasing from 0.1 to 0.5. In a sense, a small value of ϵ corresponds to assigning relatively large probabilities to the predicted pseudo-labels in the soft pseudo-labels. However, since the pseudo-labels for low-confidence samples predicted by the vision model can be noisy, the resulting soft pseudo-labels may also inherit this noise. In contrast, the performance drops from 0.5 to 0.9, suggesting that there exist clean predicted pseudo-labels among the low-confidence samples. A large ϵ might lower the probabilities assigned to these clean pseudo-labels, ultimately degrading the performance.

Visualization on Feature Space. To gain deeper insights into our LADaS, we provide visualizations of learned representations on the testing data and compare it with DebiasPL and FINESSL, for both FOOD-101 with N2 setting and CIFAR-100 with N4 setting. The results are obtained with t-SNE [21] as shown in Fig. 4. Different colors represent different classes. As can be seen, our LADaS tends to separate samples of different classes into different groups across both FOOD-101 and CIFAR-100 datasets. On the contrary, the DebiasPL is incapable of discriminating samples of different classes well, especially on CIFAR-100 with N4 setting. Moreover, our LADaS show superiority over FINESSL across all datasets. The reason may be attributed to our proposed language-assisted pseudo-label debiasing module can help the model produce more accurate and balanced pseudo-labels for model training and hence promote the performance.

References

[1] 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.

[2] 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.

[3] 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.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[6] 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.

[7] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Jong-Chyi Su and Subhransu Maji. The semi-supervised inaturalist-aves challenge at fgvc7 workshop, 2021.

[20] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[21] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008.

[22] 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.

[23] 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.

[24] 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.

[25] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[26] 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.