TRANSDUCTIVE CLIP WITH CLASS-CONDITIONAL CONTRASTIVE LEARNING

Junchu Huang[1,*], *Weijie Chen*[3,2], *Shicai Yang*[2], *Di Xie*[2], *Shiliang Pu*[2,✉], *Yueting Zhuang*[3,✉]

[1]South China University of Technology, Guangzhou, China
[2] Hikvision Research Institute, Hangzhou, China
[3] Zhejiang University, Hangzhou, China

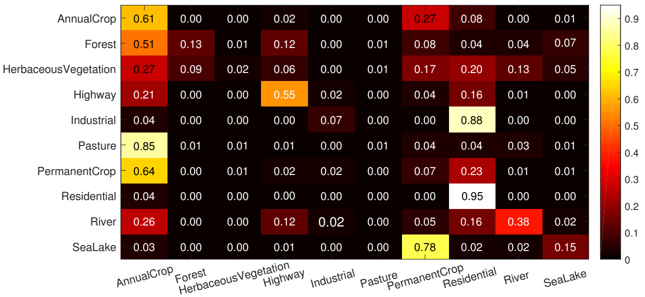
ABSTRACT

Inspired by the remarkable zero-shot generalization capacity of vision-language pre-trained model, we seek to leverage the supervision from CLIP model to alleviate the burden of data labeling. However, such supervision inevitably contains the label noise, which significantly degrades the discriminative power of the classification model. In this work, we propose Transductive CLIP, a novel framework for learning a classification network with noisy labels from scratch. Firstly, a *class-conditional contrastive learning* mechanism is proposed to mitigate the reliance on pseudo labels and boost the tolerance to noisy labels. Secondly, *ensemble labels* is adopted as a pseudo label updating strategy to stabilize the training of deep neural networks with noisy labels. This framework can reduce the impact of noisy labels from CLIP model effectively by combining both techniques. Experiments on multiple benchmark datasets demonstrate the substantial improvements over other state-of-the-art methods.

Index Terms— Vision-Language Pre-trained Model, Transductive Learning, Noisy Label Learning, Contrastive Learning, Unsupervised Model Optimization

1. INTRODUCTION

The revolutionized successes of deep neural networks in a variety of computer vision applications are conferred by large databases with accurate annotation [1, 2, 3]. In many real-world scenarios, data labeling is very costly in terms of resource and time consumption. Several efforts had been made for unsupervised model optimization in the past [4, 5, 6, 7, 8]. Recently, Contrastive Language-Image Pretraining (CLIP) [9] has emerged as a promising alternative for generalizing vision tasks. To alleviate the burden of data labeling, there is a strong motivation to leverage the unlabeled data supervised from the CLIP model in a transductive learning manner. However, it cannot achieve satisfactory performance by directly learning from the pseudo labels predicted by CLIP model since the classification model is prone to fit and memorize the label noise [10], leading to the performance degeneration.



	AnnualCrop	Forest	HerbaceousVegetation	Highway	Industrial	Pasture	PermanentCrop	Residential	River	SeaLake
AnnualCrop	0.61	0.00	0.00	0.02	0.00	0.00	0.27	0.08	0.00	0.01
Forest	0.51	0.13	0.01	0.12	0.00	0.01	0.08	0.04	0.04	0.07
HerbaceousVegetation	0.27	0.09	0.02	0.06	0.00	0.01	0.17	0.20	0.13	0.05
Highway	0.21	0.00	0.00	0.55	0.02	0.00	0.04	0.16	0.01	0.00
Industrial	0.04	0.00	0.00	0.00	0.07	0.00	0.00	0.88	0.00	0.00
Pasture	0.85	0.01	0.01	0.01	0.00	0.01	0.04	0.04	0.03	0.01
PermanentCrop	0.64	0.00	0.01	0.02	0.02	0.00	0.07	0.23	0.01	0.01
Residential	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00
River	0.26	0.00	0.00	0.12	0.02	0.00	0.05	0.16	0.38	0.02
SeaLake	0.03	0.00	0.00	0.01	0.00	0.00	0.78	0.02	0.02	0.15

Fig. 1. The realistic noise matrix in the EuroSAT dataset [11] based on the CLIP model. The label noise is significantly unbalanced among different categories, where the accuracy of 6-th category ("Pasture") is only 0.01 while the accuracy of 8-th category ("Residential") is up to 0.95.

To explore robust learning from noisy labels, a series of studies have been conducted, which can be roughly divided into three categories: 1) label correction methods, 2) loss correction methods, and 3) refined training strategies. Label correction methods focus on rectifying noisy labels with the help of complex noise models, i.e., directed graphical models [12] and conditional random fields [13]. However, in order to obtain the noise model, the support from extra clean data is indispensable. The idea of loss correction methods [14, 15] is to modify the objective functions for training deep neural networks robustly. It holds the belief that assigning importance weights to examples increases the robustness of the training objective. The methods of refined training strategies [16, 17, 18] promote the robustness of deep neural networks via modifying the training paradigms. For instance, Co-teaching [19] is proposed to maintain two peer networks simultaneously during training, in which one network is back-propagated by the selected confident samples from another network to alleviate the accumulated error.

To make noisy label learning more trackable, it is required to leverage the intrinsic information of the unlabeled data. For example, the consistency regularization from the semi-supervised learning holds the assumption that the prediction of an instance should not be too different from its perturbation one [21]. Under this inspiration, recent works explore to handle the label noise by integrating the wisdom of semi-supervised learning technology. In order to achieve

[*] This work is done when Junchu Huang was an intern in Hikvision Research Institute. ✉ Corresponding Authors

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

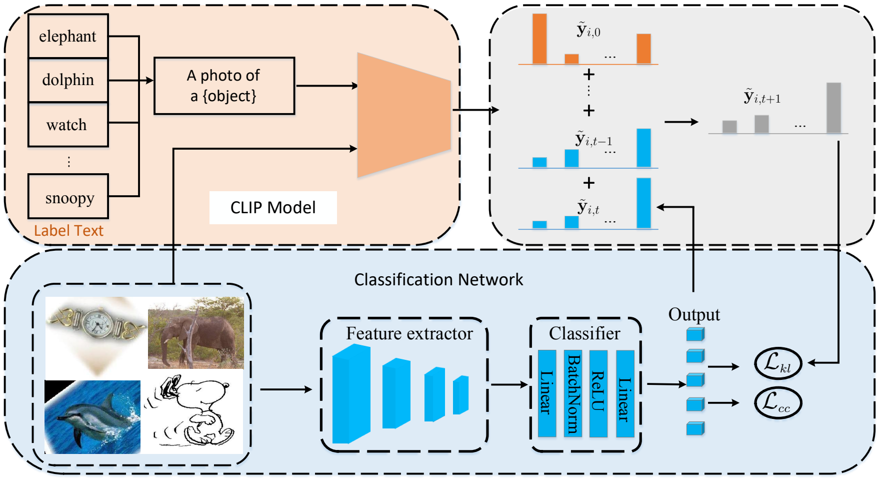


Fig. 2. The pipeline of the proposed method. Here CLIP model is merely required to obtain the label text of each category to generate text embedding so as to annotate the unlabeled images. Modules in blue is trained from scratch with the initial supervision from CLIP model and the training labels are updated in an iterative learning strategy to filter out the label noise. Specifically, top right is the *ensemble labels* module while bottom is the noisy label learning process regularized by *class-conditional contrastive learning* loss. In this paper, we use VIT-B/32 [20] as the backbone of CLIP model for zero-shot pseudo labeling and use ResNet-50 as the backbone of the classification network for transductive learning.

this goal, DivideMix [22] designs two diverged networks: one network uses the dataset division from another one alternately to separate the clean data (considered as labeled data) and the noisy data (considered as unlabeled data). Then the semi-supervised training are conducted with the improved MixMatch [23] to perform label co-refinement and label co-guessing on the labeled and unlabeled data, respectively.

Despite the remarkable empirical results achieved by the above mentioned methods, they hold the same assumption: the noisy label is simulated and balanced with known noise rate of each category. As shown in Figure 1, the noisy labels generated from CLIP model violate the above assumption. As a result, an obvious obstacle in the existing methods is the confirmation bias [24]: the performance is restricted when learning from severely inaccurate pseudo labels. To escape from the dilemma, we propose in this paper a novel design of Transductive CLIP. Specifically, this paper has the following contributions: 1) This paper proposes to leverage the unlabeled data that supervised from CLIP model in a transductive learning manner to alleviate the burden of data labeling. 2) To tackle confirmation bias problem, a *class-conditional contrastive learning* (C^3L) mechanism is proposed to mitigate the reliance on pseudo labels and boost the tolerance to noisy labels. 3) To stabilize the training of deep neural networks, an *ensemble labels* scheme is utilized to update the incorrect pseudo labels in an iterative learning strategy.

2. METHODOLOGY

We describe the framework in detail as follows. Given a dataset of unlabeled instances $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ with the text name of categories included in the dataset. The initial training labels $\tilde{\mathbf{Y}}_0 = \{\tilde{\mathbf{y}}_{i,0}\}_{i=1}^n$ are generated from CLIP model, where $\mathbf{x}_i \in \mathcal{R}^d$ (d is the dimension of each instance) and $\tilde{\mathbf{y}}_{i,0} \in \mathcal{R}^C$ (C is the number of categories). The goal of this paper is to build a classification network from scratch, which can be trained on the noisy labels robustly and generate higher-quality pseudo labels for the unlabeled data.

2.1. Class-Conditional Contrastive Learning

To exploit the knowledge from the initial training labels, it is natural to force the output of the network consistent with the noisy labels. We apply temperature scaling to recalibrate the training labels to reduce the uncertainty of training labels. The training label of i^{th} instance can be reformulated as

$$(\tilde{\mathbf{y}}_{i,t})_j = \frac{((\tilde{\mathbf{y}}_{i,t})_j)^\tau}{\sum_{j'=1}^{C}((\tilde{\mathbf{y}}_{i,t})_{j'})^\tau} \tag{1}$$

where τ is the rescaling factor and is set as 2 in all experiments (the value of τ is relatively stable over a large interval). The subscript j denotes the category index. t is the training epoch and the initialization $\tilde{\mathbf{y}}_{i,0}$ of training label $\tilde{\mathbf{y}}_{i,t}$ comes from the output of the CLIP model. The training label will be updated in each training epoch which will be introduced in Sec.2.2. The training objective of model f on the unlabeled data is obtained by minimizing the following Kullback-Leibler divergence loss (l_{kl}),

$$\mathcal{L}_{kl} = \sum_{i=1}^{m} \ell_{kl}\left(f_t(\mathbf{x}_i) \,||\, \tilde{\mathbf{y}}_{i,t}\right), \tag{2}$$

where $f_t(\mathbf{x}_i)$ is the output of classification model through a Softmax function. m is the batch size in the training phase.

$\mathcal{L}_{kl}$ tends to focus on optimizing the error between network output and training labels, which is effective and efficient in clean-annotated labels. However, the $\mathcal{L}_{kl}$ is not robust enough for the noisy labels. Training samples with noisy labels will be stuck into wrong category predictions in the early stage of training, and are difficult to be corrected later due to the memory effect of deep neural network, which is known as confirmation bias issue. When it comes to noisy label, the classification model trained by $\mathcal{L}_{kl}$ will be easily confused by false pseudo-labels since it focuses on learning a hyperplane for discriminating each class from the other classes.

To remedy the class discrimination, contrastive learning [4, 6] improves the quality of the learned representations by exploring the intrinsic structure of instances. However, the optimization of standard contrastive learning loss is independent of the training labels, leaving the useful discriminative information on the shelf. To address this problem, we propose to integrate the class discrimination and instance discrimination to cope with the training of noisy label. Consequently, the hyperplane is joint optimized. This strategy aims to uncover the underlying structure of network's output to reduce the overconfidence of the network on its predictions. Utilizing the discriminative information conveyed in the classifier predictions, the class-conditional contrastive learning (C^3L) loss is defined as:

$$\mathcal{L}_{cc} = -\sum_{i=1}^{m} \log \frac{\exp\left(\operatorname{sim}\left(f_t(\mathbf{x}_i), f_t(\tilde{\mathbf{x}}_i)\right)/T\right)}{\sum_{k=1}^{m} \exp\left(\operatorname{sim}\left(f_t(\mathbf{x}_i), f_t(\tilde{\mathbf{x}}_k)\right)/T\right)} \quad (3)$$

where the $\operatorname{sim}(,)$ is selected as the cosine function and T is the temperature parameter following the setting in contrastive learning. $f_t(\tilde{\mathbf{x}}_i)$ is the Softmax output of the model with the augmented input $\tilde{\mathbf{x}}_i$, which implicitly encodes the normalized distance between the instance feature and the learnable class prototypes. The C^3L loss $\mathcal{L}_{cc}$ maximizes the similarity of category prediction between differently augmented views of the same data point. According to the properties of the Softmax function adopted in equation (3), the similarity of category prediction between different data point is minimized, which could push category prediction away from noisy labels. The optimization of the predicted probability will lead to the optimization of the hyperplane. The clean label will win this prediction competition, since their prediction are easier to fit and achieve higher score, compared to that of false ones. Therefore, the model can not cause serious over-fitting in the training of pseudo-labels, which greatly alleviates the problem of confirmation bias. The overall training loss is

$$\mathcal{L}_{total} = \mathcal{L}_{kl} + \lambda \mathcal{L}_{cc} \quad (4)$$

λ is set to 1 by default in this paper.

2.2. Noisy Labels Rectification with Noise Filtering

The training labels with false predictions tend to fluctuate. For example, the predictions of one sample may be predicted as one object in one epoch and another object in another epoch, which usually contains higher label noise. To obtain reliable prediction on unlabeled examples and improve the quality of training labels, we adopt the *ensemble labels* scheme to update the training label in an iterative learning strategy. In particular, the training labels are updated by:

$$\tilde{\mathbf{y}}_{i,(t+1)} \leftarrow \frac{\tilde{\mathbf{y}}_{i,0} + \sum_{j=1}^{t} f_j(\mathbf{x}_i)}{t+1}, \forall \mathbf{x}_i \in \{\mathbf{x}_i\}_{i=1}^{n} \quad (5)$$

thus the training labels with higher label noise can be suppressed in this way. During the label updating progress, note that the predictions in different training epochs contribute equally which avoids negative updating.

3. EXPERIMENTS

3.1. Datasets

Caltech101 [25] consists of images from 101 object categories and a background class from Google search. Each category contains 31 to 800 images with medium resolution, which is around 300×300 pixels. **Describable Textures Dataset (DTD)** [26] is a texture database collected from Google and Flickr. It consists of 5640 images, which are organized according to a list of 47 categories inspired from human perception. **EuroSAT** [11] consists of 27,000 labeled images with 10 different land use and land cover classes. **NWPU-RESISC** [27] is a benchmark for Remote Sensing Image Scene Classification (RESISC), created by Northwestern Polytechnical University (NWPU). This dataset contains 31,500 images, covering 45 scene classes with 700 images in each class. **Flower102** [28] is an image classification dataset consisting of 102 flower categories.

3.2. Baseline Methods

We evaluate the effectiveness of our proposed framework by re-implementing other state-of-the-art methods on our proposed experimental settings:

- **FixMatch** [21] is a state-of-the-art semi-supervised learning method which retains the pseudo-label of weakly-augmented unlabeled images when the prediction of model is higher than the confidence-threshold (0.95 by default).
- **Re-weighting** is a popular loss correction method in noisy label learning to reduce the label noise and the instance weight is set as the maximum classification probability.
- **Dash** [29] is a state-of-the-art method which dynamically selects instance whose loss value is less than a dynamic threshold at each optimization step to train the models.
- **DivideMix** [22] is a state-of-the-art noisy label learning method. At each mini-batch, one network performs an improved Mixmatch method with the clean set and noisy set that are dynamically divided by another diverged network.

Table 1. Classification accuracies (%). VIT-B/32 (CLIP Backbone) → Resnet-50 (Target Backbone).

Methods	Caltech101	DTD	EuroSAT	NWPU-RESISC	Flower102	Average
CLIP Model	81.33	43.12	30.77	54.83	64.90	54.99
FixMatch	82.09	43.17	33.00	60.50	66.61	57.07
Re-weighting	83.56	44.46	33.53	60.42	67.36	57.87
Dash	82.91	44.07	34.46	60.83	68.56	58.17
DivideMix	80.51	45.51	37.38	70.77	70.72	60.98
Ours	86.29	50.83	45.76	72.00	74.25	65.83
Gain	+4.96	+7.71	+14.99	+17.17	+9.35	+10.84

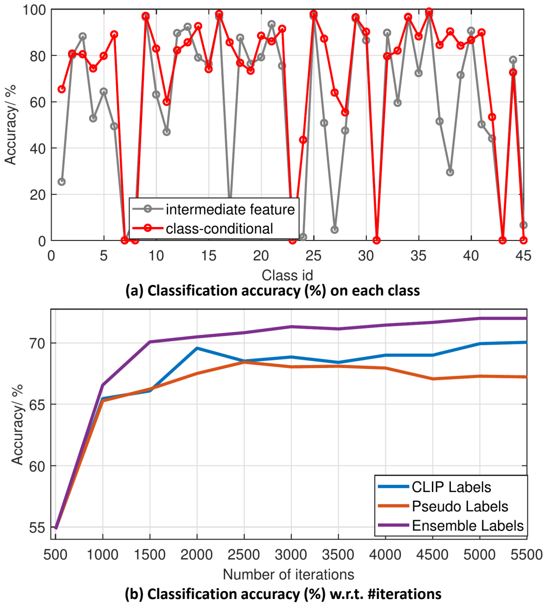
3.3. Implementation Details

Network architecture. In this paper, we employ the VIT-B/32 [20] as the backbone module of CLIP model and select the ResNet-50 model pre-trained on ImageNet as the backbone module for transductive optimization. Note that VIT-B/32 is substantially larger than ResNet-50. The classifier module is constructed by a two-layer MLP in conjunction with a batchnorm layer as shown in Figure 2.

Network hyper-parameters. Following the common setting in contrastive learning, the temperature T is adopted as 0.07. The learning rate of the classifier is set to 0.01 and the batch size is set to 128 in all experiments. The learning rate of the classifier is 10 times as the backbone following the common fine-tuning principle. Stochastic gradient descent (SGD) with a momentum of 0.9 is adopted to optimize the model.

3.4. Performance Analysis

Firstly, the classification accuracies of the proposed method and the four baseline methods on the five datasets are shown in Table 1. We observe that the proposed method achieves much better performance than the baseline methods with statistical significance. The classification accuracy of our method on NWPU-RESISC is 72.00% and the performance improvement is 17.17% compared to the CLIP model. Note that, this task is very difficult since most of the baseline methods can only achieve slight improvement and may perform very poorly on many of the datasets. Secondly, in Figure 3 (a), we conduct contrastive learning strategy on the intermediate feature [4] to show the superiority of the proposed class-conditional contrastive learning. Compared with standard contrastive learning loss, C^3L loss can achieve higher performance when label noise is high (for example, object with class id 38), which is inseparable from the contribution of model category information during contrastive regularization. Thirdly, we compare our proposed *ensemble labels* strategy with other training label updating strategies, and the result is shown in Figure 3 (b). Even if the training label is not updated (namely *CLIP labels*), the performance of the model can reach around 70 % on the NWPU-RESISC dataset. This shows that C^3L can handle noisy labels well. For *pseudo labels*, we have $\tilde{\mathbf{y}}_{i,(t+1)} \leftarrow f_t(\mathbf{x}_i), \forall \mathbf{x}_i \in \{\mathbf{x}_i\}_{i=1}^{n}$. It updates training labels according to the prediction in the previous epoch, which induces negative updating. Compared with other strategies, the proposed *ensemble labels* can achieve higher performance and a more stable training process.



(a) Classification accuracy (%) on each class

(b) Classification accuracy (%) w.r.t. #iterations

Fig. 3. Effectiveness verification.

4. CONCLUSION

To alleviate the burden of data labeling, in this paper we explore an interesting, realistic but challenging task where only the vision-language pre-trained model is provided to the unlabeled data as the supervision. To handle label noise, we propose a simple yet effective framework called Transductive CLIP. As a component of the proposed method, the *class-conditional contrastive learning* loss integrates the class discrimination and instance discrimination to mitigate the reliance on pseudo labels and boost the tolerance to noisy labels. Furthermore, we adopt the *ensemble labels* scheme to update the training label in an iterative learning strategy to rectify noisy labels. Extensive results on five image classification tasks verify the efficacy of our proposed method.

5. REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[2] Weijie Chen, Di Xie, Yuan Zhang, and Shiliang Pu, "All you need is a few shifts: Designing efficient convolutional neural networks for image classification," in *CVPR*, 2019.

[3] Luojun Lin, Lingyu Liang, Lianwen Jin, and Weijie Chen, "Attribute-aware convolutional neural networks for facial beauty prediction," in *IJCAI*, 2019.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[5] Weijie Chen, Shiliang Pu, Di Xie, Shicai Yang, Yilu Guo, and Luojun Lin, "Unsupervised image classification for deep representation learning," in *ECCVW*, 2020, pp. 430–446.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.

[7] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang, "A free lunch for unsupervised domain adaptive object detection without source data," in *AAAI*, 2021.

[8] Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu, Yueting Zhuang, and Wenqi Ren, "Self-supervised noisy label learning for source-free unsupervised domain adaptation," *CoRR*, vol. abs/2102.11614, 2021.

[9] 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 *arXiv preprint arXiv:2103.00020*, 2021.

[10] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," in *NeurIPS*, 2019, pp. 1917–1928.

[11] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019, vol. 12, pp. 2217–2226.

[12] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang, "Learning from massive noisy labeled data for image classification," in *CVPR*, 2015, pp. 2691–2699.

[13] Arash Vahdat, "Toward robustness against label noise in training deep discriminative neural networks," in *NeurIPS*, 2017.

[14] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *CVPR*, 2017, pp. 1944–1952.

[15] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama, "Masking: A new perspective of noisy supervision," in *NeurIPS*, 2018.

[16] Eran Malach and Shai Shalev-Shwartz, "Decoupling "when to update" from "how to update"," in *NeurIPS*, 2017.

[17] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *ICML*, 2018, pp. 2304–2313.

[18] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey, "Dimensionality-driven learning with noisy labels," in *ICML*, 2018, pp. 3355–3364.

[19] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *arXiv preprint arXiv:1804.06872*, 2018.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[21] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, 2020, pp. 596–608.

[22] Junnan Li, Richard Socher, and Steven CH Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *ICLR*, 2020.

[23] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel, "Mixmatch: A holistic approach to semi-supervised learning," 2019.

[24] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *IJCNN*, 2020, pp. 1–8.

[25] Li Fei-Fei, Rob Fergus, and Pietro Perona, "One-shot learning of object categories," in *TPAMI*, 2006, vol. 28, pp. 594–611.

[26] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, "Describing textures in the wild," in *CVPR*, 2014.

[27] Gong Cheng, Junwei Han, and Xiaoqiang Lu, "Remote sensing image scene classification: Benchmark and state of the art," in *Proceedings of the IEEE*, 2017, vol. 105, pp. 1865–1883.

[28] Maria-Elena Nilsback and Andrew Zisserman, "Automated flower classification over a large number of classes," in *ICCVGIP*, 2008, pp. 722–729.

[29] 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 *ICML*, 2021, pp. 11525–11536.