Improved Visual Fine-tuning with Natural Language Supervision

Junyang Wang[1*] Yuanhong Xu[2] Juhua Hu[3] Ming Yan[2] Jitao Sang[1,4] Qi Qian[5†]
[1] School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
[2] DAMO Academy, Alibaba Group, Hangzhou, China
[3] School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA
[4] Peng Cheng Lab, Shenzhen, China
[5] DAMO Academy, Alibaba Group, Bellevue, WA 98004, USA
{junyangwang, jtsang}@bjtu.edu.cn, {yuanhong.xuyh, ym119608, qi.qian}@alibaba-inc.com, juhuah@uw.edu

Abstract

Fine-tuning a visual pre-trained model can leverage the semantic information from large-scale pre-training data and mitigate the over-fitting problem on downstream vision tasks with limited training examples. While the problem of catastrophic forgetting in pre-trained backbone has been extensively studied for fine-tuning, its potential bias from the corresponding pre-training task and data, attracts less attention. In this work, we investigate this problem by demonstrating that the obtained classifier after fine-tuning will be close to that induced by the pre-trained model. To reduce the bias in the classifier effectively, we introduce a reference distribution obtained from a fixed text classifier, which can help regularize the learned vision classifier. The proposed method, Text Supervised fine-tuning (TeS), is evaluated with diverse pre-trained vision models including ResNet and ViT, and text encoders including BERT and CLIP, on 11 downstream tasks. The consistent improvement with a clear margin over distinct scenarios confirms the effectiveness of our proposal. Code is available at `https://github.com/idstcv/TeS`.

1. Introduction

Fine-tuning a pre-trained visual deep model has become a prevalent paradigm for vision categorization [21]. By initializing the model with parameters pre-trained on a large-scale data set, fine-tuning can effectively transfer the semantic information from the pre-training data to diverse downstream tasks, which is essential to mitigate the over-fitting problem on data sets with limited training examples [19].

While supervised pre-training methods [21] require full label information for the whole data set, the recent progress on self-supervised learning shows that an applicable pre-trained model can be obtained from unlabeled data [18]. Consequently, a lot of research efforts have been devoted to exploring unsupervised pre-training, which can eliminate the labeling cost for a vast amount of examples. Many effective self-supervised methods have been developed, e.g., instance discrimination [7, 18], cluster discrimination [5, 32], and masked image modeling [17]. Compared to the supervised counterparts, fine-tuning pre-trained models from these unsupervised methods can achieve comparable or even better performance on downstream tasks.

With the success of pre-training, fine-tuning has thus attracted much attention to leverage the pre-trained model appropriately for downstream tasks. One main challenge for fine-tuning is the catastrophic forgetting in pre-trained backbone [9], that is, after sufficient learning with a large learning rate, the fine-tuned model becomes far away from

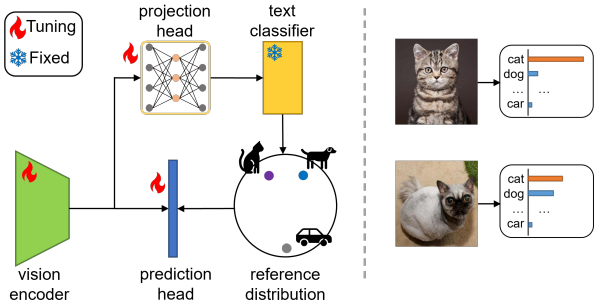

Figure 1. Illustration of the proposed method, TeS. During fine-tuning, the vision classifier is regularized by reference distributions defined with a fixed text classifier, which is obtained from class names. In TeS, diverse reference distributions can be obtained for different examples even from the same class (e.g., cats on the right side).

*Work done during internship at DAMO Academy, Alibaba Group.
†Corresponding author

the pre-trained model, incurring the over-fitting problem. The challenge has been extensively studied and can be tackled by simply setting a small learning rate for backbone or constraining the distance from the fine-tuned model to the pre-trained model explicitly [24, 25].

The other challenge for fine-tuning is the potential bias existing in a pre-trained model. It should be noted that a pre-trained model is optimized by a specific pretext task using only the pre-training data. Therefore, if the data distribution of a target task is different from that of the pre-training data, the pre-trained model is biased to the pre-training data distribution [30]. This bias problem becomes more challenging in the scenario together with catastrophic forgetting, where the bias cannot be eliminated sufficiently with a constraint making the fine-tuned model and pre-trained model close. A recent work [26] proposes to add a subset of pre-training data that is related to the target task for fine-tuning. The optimization with partial pre-training data can help preserve effective information in the pre-trained model to avoid over-fitting while reducing the bias for the target task. However, it requires the access of the pre-training data, which is infeasible in many real applications.

Recently, it has been found that side information from other related modalities can help visual pre-training. For example, CLIP [33] pre-trains dual encoders by optimizing image-text pairs that align images with their corresponding text descriptions. The obtained model shows a strong zero-shot transfer performance, which can classify images using proxies consisting of text representations extracted from class names. The observation implies that text information is capable of guiding visual representation learning. However, the superior zero-shot performance highly depends on the paired vision and text encoders in pre-training, and thus is not applicable for an arbitrary vision encoder.

Inspired by vision-language pre-training [33], we introduce natural language supervision to fine-tuning, so as to mitigate the contradiction problem between the bias in pre-trained vision models and catastrophic forgetting in fine-tuning them. First, we show that without any side information, the classifier induced by the pre-trained vision model will be largely preserved after fine-tuning on the target data, which demonstrates the potential bias exiting in the conventional fine-tuning pipeline. To reduce the bias in the learned classifier, we propose to include text supervision as the reference information. Concretely, with a fixed pre-trained text encoder, a text classifier for the target task can be obtained by extracting proxies with class names. Given the text classifier, both class-level and instance-level reference distributions can be obtained for the target vision task. Then, the pre-trained vision models can be fine-tuned with the appropriate reference distributions for the vision classifier. Since the reference regularization is independent from the backbone, our method can reduce the bias without catastrophic forgetting.

The proposed **Te**xt **S**upervised fine-tuning (TeS) method is illustrated in Fig. 1. It should be noted that reference distributions are extracted from a fixed text encoder, and thus the overhead for the text supervision is negligible. The main contributions of this work can be summarized as follows.

- This work proposes to leverage text supervision from a fixed text encoder for fine-tuning an arbitrary pre-trained vision model. The text encoder can be pre-trained on a large corpus with rich context information, making it a good complement to vision tasks.
- With the classifier consisting of proxies from text encoder, we investigate the class-level and instance-level distributions for regularizing the fine-tuning of the vision classifier. By minimizing the derived cross entropy loss, the vision model can exploit the side information from the text supervision to reduce the bias from the pre-trained models.
- Experiments on extensive downstream tasks demonstrate the effectiveness of text supervision for visual fine-tuning. In addition, the CLIP text encoder outperforms the BERT encoder [11] and it implies that the text encoder pre-trained with vision-language pretext tasks is more appropriate for supervising visual tasks.

2. Related Work

In this section, we briefly review visual pre-training and visual fine-tuning.

2.1. Visual Pre-training

A generic visual pre-training paradigm includes both supervised and self-supervised approaches. Supervised pre-training requires a large number of labeled data and can learn rich semantic information for fine-tuning [30, 34, 42].

To eliminate the cost of labeling, self-supervised learning is developed to obtain pre-trained models from unlabeled data. Many pretext tasks were proposed for effective learning, e.g., instance discrimination that considers each instance as an individual class and optimizes random augmentations from the same instance [6, 18], cluster discrimination that explores the relationship between different instances [5, 32] and masked image modeling that leverages information within each image [16]. Moreover, [43] demonstrates that self-supervised pre-training improves supervised pre-training with strong data augmentations.

2.2. Visual Fine-tuning

After pre-training, fine-tuning aims to facilitate the downstream tasks by transferring semantic information in the pre-trained models [21]. One main challenge for fine-tuning is from catastrophic forgetting that the fine-tuned

model over-fits target data and becomes far away from the pre-trained model [9, 24, 25]. To mitigate the problem, [24] constrains the distance between the fine-tuned model and its pre-trained counterpart. [25] proposes a data-dependent strategy that limits the distance between representations of target data before and after fine-tuning.

Besides catastrophic forgetting, another challenge is from the bias in pre-trained models. Due to the gap between learning tasks and data for pre-training and fine-tuning, the bias in the pre-trained model will degenerate the performance on the target task [30]. Without any side information, the bias is hard to be eliminated. Recent work [26] shows that the bias can be reduced by fine-tuning the target data with a related subset selected from the pre-training data for the target task. In this work, we investigate another form of side information from text supervision. Compared with the information from pre-training data, the text supervision is easier to be accessed with a pre-trained text encoder.

3. Visual Fine-tuning with Text Supervision

Given an image data set of $\{x_i, y_i\}_{i=1}^n$, where x_i denotes an image and y_i is the corresponding label, a classification model can be learned by minimizing the empirical risk as

$$\min_{\theta, W} \frac{1}{n} \sum_i^n \ell(\mathbf{x}_i, y_i) \tag{1}$$

where $\mathbf{x}_i = f(x_i) \in \mathbb{R}^d$ is the representation extracted from a neural network $f(\cdot)$ and θ is the parameter of $f(\cdot)$. ℓ can be an appropriate loss function and cross entropy loss is prevalent for classification, which is also adopted in this work. The C-class classification loss can be written as

$$\ell(\mathbf{x}_i, y_i) = -\log \frac{\exp(\mathbf{x}_i^\top \mathbf{w}_{y_i})}{\sum_j^C \exp(\mathbf{x}_i^\top \mathbf{w}_j)}$$

where $W = [\mathbf{w}_1, \ldots, \mathbf{w}_C] \in \mathbb{R}^{d \times C}$ is a linear classifier consisting of a single proxy from each class [31].

Due to the over-parameterization property of deep models [1], i.e., the number of parameters in θ can be larger than that of training examples, directly optimizing the problem in Eqn. 1 may incur the over-fitting problem, especially on data sets with a limited number of examples. To mitigate the problem, the model can be initialized by parameters pre-trained on a large-scale data set. Fine-tuning a pre-trained model with a small learning rate becomes prevalent for visual categorization [21].

Compared with a randomly initialized model, the pre-trained model contains sufficient semantic information from the pre-training data that can be transferred to downstream tasks. However, the bias from the specific pre-training task and data can result in a sub-optimal performance when data distribution shifts. In this work, we propose to include the text supervision from the class names in the target task to help reduce the bias from the pre-trained model.

3.1. Biased Classifier in Conventional Fine-tuning

First, we demonstrate that the bias in a pre-trained model will be preserved in the classifier W after conventional fine-tuning. Moreover, even with a large learning rate solely for the classifier during fine-tuning, the obtained W is still close to that implied by the pre-trained model.

Given fixed representations from a pre-trained model, the sub-problem of optimizing the classifier with cross entropy loss is convex and a global optimum can be obtained. Let $W^0 = [\mathbf{w}_1^0, \ldots, \mathbf{w}_C^0]$ be the optimal solution as

$$W^0 = \arg\min_W \mathcal{L}(\theta^0, W) = \frac{1}{n} \sum_i \ell(\mathbf{x}_i^0, y_i) \tag{2}$$

where $\mathbf{x}_i^0 = f_0(x_i)$ is extracted from the pre-trained model $f_0(\cdot)$, θ_0 is the parameter of $f_0(\cdot)$, and $\mathbf{w}_j^0$ denotes the proxy for the j-th class obtained from the pre-trained model.

After fine-tuning T iterations with stochastic gradient descent (SGD), we can observe (θ^T, W^T) for the problem

$$\min_{\theta, W: \|\theta - \theta^0\|_F \leq \epsilon} \mathcal{L}(\theta, W) = \frac{1}{n} \sum_i \ell(\mathbf{x}_i, y_i) \tag{3}$$

where $W^T = [\mathbf{w}_1^T, \ldots, \mathbf{w}_C^T)]$ and ϵ denotes the distance to the pre-trained backbone. A small ϵ is essential to avoid catastrophic forgetting in fine-tuning. We find that the proxies in W^T are close to W^0 (i.e., the solution implied by the pre-trained model) as follows. All detailed proof of this work can be found in the supplementary.

Theorem 1. *Let W^0 and W^T be the optimal solution for Eqn. 2 and that by SGD for Eqn. 3. Assuming that $\mathcal{L}(\theta^0, W)$ is m-strongly convex in W and $\mathcal{L}(\theta, W^T)$ is $L/2$-Lipschitz continuous in θ, we have*

$$\|W^T - W^0\|_F^2 \leq \frac{L}{m} \epsilon$$

Remark 1 Theorem 1 indicates that the changes in the classifier after fine-tuning heavily depends on the distance between backbones before and after fine-tuning, i.e., ϵ. With a small learning rate for the backbone during the fine-tuning, the bias in the pre-trained model will be preserved, and thus can degenerate the performance on downstream tasks.

This can be further demonstrated by characterizing the gap ϵ between the pre-trained and fine-tuned models following the common practice in fine-tuning. During the fine-tuning, the backbone will be refined by an initial learning rate of η_0 with a cosine learning rate decay [21]. With SGD for optimization, ϵ can be bounded as follows.

Proposition 1. *Let θ^0 and θ^T denote the parameters of the pre-trained backbone and those after fine-tuning. Assuming the gradient in SGD is bounded by $\|\nabla_\theta \mathcal{L}\|_F \leq \delta$, we have*

$$\|\theta^0 - \theta^T\|_F \leq 0.5\eta_0\pi\delta$$

Remark 2 Proposition 1 shows that with the cosine decay, the gap between the pre-trained and fine-tuned backbones mainly depends on the initial learning rate for the backbone. If adopting a small learning rate, the refined model will be close to the initial one, which can preserve the bias. However, increasing the learning rate is ineffective since it will lose the knowledge from the pre-trained model, and thus incur catastrophic forgetting.

Incorporating the result in Proposition 1, the difference between proxies can be depicted as

Corollary 1. *Let W^0 and W^T be the optimal solution for Eqn. 2 and that by SGD for Eqn. 3. With assumptions in Theorem 1 and Proposition 1, we have*

$$\|W^T - W^0\|_F^2 \leq \frac{\eta_0\pi\delta L}{2m}$$

Remark 3 Corollary 1 shows that the distance between classifiers depends on the initial learning rate for the backbone. Therefore, even with a different and larger learning rate for the classifier, it will still be close to that suggested by the pre-trained model.

Since it is hard to balance the trade-off for fine-tuning by the learning rate, we propose to include reference distributions from text supervision to mitigate the contradiction issue between bias and catastrophic forgetting.

3.2. Reference Distribution from Text Supervision

To reduce the potential bias existing in the pre-trained model, we introduce a reference distribution for the target task. Concretely, a distance constraint is added during the fine-tuning

$$\min_{\theta,W} \mathcal{L}(\theta, W) \quad s.t. \quad D(W, W_r) \leq \alpha$$

where W_r is a reference distribution of W and $D(\cdot,\cdot)$ is a distance function, in which a squared Euclidean distance can be adopted as $D(W, W_r) = \|W - W_r\|_F^2$.

By constraining the distance to a reference distribution, the bias in the pre-trained model can be reduced effectively. In addition, the regularization is defined with proxies independent from the backbone, where a small learning rate is applicable for the backbone to avoid catastrophic forgetting.

However, obtaining an appropriate reference without any side information is challenging. Inspired by the recent progress in visual representation learning with natural language supervision [33], we consider leveraging the class name as text information to generate the reference distribution. Compared with images, tokens in text is organized discretely and text encoder can be pre-trained on a sufficiently large corpus with rich context, which is ideal to complement discrimination tasks such as classification.

Let $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_C] \in \mathbb{R}^{d_z \times C}$ denote the proxies extracted from a text encoder for classes, the problem with text supervision can be written as

$$\min_{\theta,W} \mathcal{L}(\theta, W) \quad s.t. \quad \|W - Z\|_F^2 \leq \alpha$$

which is equivalent to

$$\min_{\theta,W} \mathcal{L}(\theta, W) + \lambda\|W - Z\|_F^2$$

The Euclidean distance requires that the text feature has the same dimension as the vision feature, i.e., $d_z = d$, while these dimensions can vary with different encoders. By applying a projection function $h(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^{d_z}$, proxies of classes from different modalities can be manually aligned and the optimization problem becomes

$$\min_{\theta,W} \mathcal{L}(\theta, W) + \lambda\|h(W) - Z\|_F^2 \tag{4}$$

However, due to the inherent differences between modalities, matching representations directly will introduce text-specific noise to visual tasks, which can degenerate the performance on the target visual task. In addition, compared to matching proxies of different modalities, the relationship between classes from the text encoder can be more informative for the target task. Therefore, we consider transferring the pairwise similarity between proxies from the text reference distribution to the vision task.

Concretely, given an anchor class j, the distribution over all classes can be computed

$$P_{j,k} = \frac{\exp(\tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_k/\tau)}{\sum_k^C \exp(\tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_k/\tau)}; P'_{j,k} = \frac{\exp(\tilde{\mathbf{z}}_j^\top \tilde{\mathbf{z}}_k/\tau')}{\sum_k^C \exp(\tilde{\mathbf{z}}_j^\top \tilde{\mathbf{z}}_k/\tau')} \tag{5}$$

where P_j and P'_j denote the distribution defined by proxies from the vision encoder and text encoder, respectively. $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{z}}$ have the unit norm for $\mathbf{w}$ and $\mathbf{z}$, while τ and τ' are the temperature parameters for vision and text distributions, respectively. With the pairwise relations, we can constrain the KL-divergence between distributions as

$$\min_{\theta,W} \mathcal{L}(\theta, W) + \lambda \sum_j^C D_{KL}(P'_j || P_j)$$

To improve the efficiency of fine-tuning, the text encoder is fixed without fine-tuning. Therefore, representations of

proxies Z are only extracted once before fine-tuning the vision encoder. By eliminating the constant term from P'_j in KL-divergence, the problem can be simplified with the cross entropy constraint as

$$\min_{\theta,W} \mathcal{L}(\theta, W) - \lambda \sum_{j}^{C} \sum_{k}^{C} P'_{j,k} \log(P_{j,k}) \tag{6}$$

Compared with the problem in Eqn. 4, KL-divergence focuses on the similarity between classes, and thus can reduce the noise from the modality-specific information.

3.3. Instance-level Reference Distribution

The class-level regularization in Eqn. 6 can be further improved by including the target task examples as a data-dependent regularization. First, if replacing the anchor class j by an anchor example x_i, the revised instance-level distribution over vision classes becomes

$$P_{i,k} = \frac{\exp(\mathbf{x}_i^\top \mathbf{w}_k)}{\sum_k^C \exp(\mathbf{x}_i^\top \mathbf{w}_k)}$$

We can show that the class-level distribution is an approximation for the instance-level distribution and optimizing the instance-level distribution can help capture the variance in real data better.

Theorem 2. *Assuming that the norm of proxies is bounded by γ as $\forall k, \|\mathbf{w}_k\|_2 \leq \gamma$, the distribution defined by the anchor class is an approximation for the distribution defined by the anchor example as*

$$\forall k, \quad \frac{1}{c^2} P_{y_i,k} \leq P_{i,k} \leq c^2 P_{y_i,k}$$

where $c = \exp(\gamma \|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2)$.

When the intra-class distribution is compact as $\|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2 \to 0$, the approximation becomes tight.

Unlike the class-level distribution, the instance-level distribution over text proxies is hard to compute due to the lack of corresponding text features for x_i. To mimic the data distribution in text space, we propose to optimize the cross entropy loss with the fixed text classifier for representation learning as

$$\ell_T(\mathbf{x}_i, y_i) = -\log \frac{\exp(h'(\mathbf{x}_i)^\top \tilde{\mathbf{z}}_{y_i}/\tau')}{\sum_j \exp(h'(\mathbf{x}_i)^\top \tilde{\mathbf{z}}_j/\tau')}$$

where $h'(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^{d_z}$ is the projection head with the unit normalization to project vision representations to the text space. Compared with Eqn. 4, the text-specific information will be learned for projected examples, which will not introduce noisy patterns to vision proxies.

With the approximated text representations $\mathbf{x}'_i = h'(\mathbf{x}_i)$, the distribution over fixed text proxies can be computed as

$$P'_{i,k} = \frac{\exp(\mathbf{x}_i'^\top \tilde{\mathbf{z}}_k/\tau')}{\sum_k^C \exp(\mathbf{x}_i'^\top \tilde{\mathbf{z}}_k/\tau')}$$

By optimizing the vision encoder and text projection simultaneously, the final objective of our proposed **Te**xt **S**upervised fine-tuning (TeS) can be cast as

$$\min_{\theta,h',W} \underbrace{(1-\lambda_V)\mathcal{L}(\theta, W) - \frac{\lambda_V}{n} \sum_{i}^{n} \sum_{k}^{C} P'_{i,k} \log(P_{i,k})}_{\text{vision fine-tuning}} + \underbrace{\frac{\lambda_T}{n} \sum_i \ell_T(\mathbf{x}_i, y_i)}_{\text{text projection}} \tag{7}$$

where the former two terms fine-tune the network for the target vision task and the latter one is for the projection head to obtain approximated distribution from the text space. During the inference, the projection head will be discarded and only the vision network is adopted for evaluation.

Connection to label smoothing While conventional label smoothing assigns the uniform weight for unrelated classes [39], our method can be considered as smoothing labels for each example with the reference distribution from the text classifier. The instance-level label smoothing is more flexible to model the diversity in target task data.

4. Experiments

To evaluate the proposed method, we conduct experiments on 11 downstream tasks including Caltech-101 [15], CIFAR-10 [23], CIFAR-100 [23], Caltech-UCSD Birds 200 (CUB) [37], Describable Textures Dataset (DTD) [10], FGVC Aircraft (Aircraft) [27], Food-101 [2], Oxford-IIIT Pet (Pets) [29], Oxford 102 Flower (Flowers) [28], Stanford Cars (Cars) [22], and SUN397 [38]. Following the common practice [13, 33, 40, 41], we report mean per-class accuracy on Aircraft, Caltech, Flowers, and Pets, while Top-1 accuracy is reported for others.

Considering that label smoothing is closely related, two variants of label smoothing strategy are included in the comparison. First, the one-hot label in cross entropy loss can be smoothed by assigning uniform weights to unrelated classes [35], which is referred as "CE+LS". Besides, the uniform weight can be replaced by the distribution implied by a text encoder pre-trained by CLIP [33] as in Eqn. 5, which is denoted as "CE+TLS". Finally, fine-tuning the original cross entropy loss is included as "CE".

Method	Aircraft	Caltech	Cars	C10	C100	CUB	DTD	Flower	Food	Pet	SUN	Avg.
CE	87.70	91.39	89.98	97.75	84.50	76.63	73.19	96.39	87.70	90.13	59.03	84.95
CE + LS	87.10	91.41	90.95	97.44	84.64	76.18	73.67	96.00	87.18	90.95	59.56	85.01
CE + TLS	87.19	91.42	89.09	97.80	84.66	76.22	72.98	96.49	87.16	90.29	59.54	84.80
TeS + BERT	87.84	93.19	90.92	97.71	85.53	77.62	73.94	97.26	87.63	90.55	60.31	85.68
TeS + CLIP	87.87	92.86	91.84	98.00	86.95	78.41	75.00	97.15	87.90	91.39	61.86	86.29
	(+0.17)	(+1.44)	(+0.89)	(+0.20)	(+2.29)	(+1.78)	(+1.33)	(+0.66)	(+0.20)	(+0.44)	(+2.30)	(+1.28)

Table 1. Comparison with ResNet-50 pre-trained by MoCo-v2. "LS" and "TLS" denote conventional label smoothing and text guided label smoothing in Eqn. 5, respectively. The last row shows the accuracy improvement over the best baseline.

Method	Aircraft	Caltech	Cars	C10	C100	CUB	DTD	Flower	Food	Pet	SUN	Avg.
CE	79.26	92.36	86.52	96.71	83.80	78.06	71.54	96.28	83.10	91.37	59.84	83.53
CE+LS	79.11	92.34	86.69	96.23	83.42	78.86	71.65	96.25	82.79	91.76	59.59	83.52
CE+TLS	79.41	91.73	86.84	96.53	82.93	78.17	71.60	95.89	82.44	91.83	59.67	83.37
TeS + CLIP	79.42	92.57	88.66	96.94	84.09	79.27	73.09	96.62	83.68	91.86	60.34	84.23
	(+0.01)	(+0.21)	(+1.82)	(+0.23)	(+0.29)	(+0.41)	(+1.44)	(+0.34)	(+0.58)	(+0.03)	(+0.50)	(+0.70)

Table 2. Comparison with supervised pre-trained ResNet-18. The last row shows the accuracy improvement over the best baseline.

4.1. Implementation Details

For an extensive comparison, ResNet [20] and ViT [14], two prevalent but diverse visual backbones are included in experiments. For ResNet, ResNet-50 pre-trained by MoCo-v2 [8] and a supervised pre-trained ResNet-18 are adopted, while ViT-B/32 pre-trained by MAE [16] and CLIP [33] are applied for fine-tuning. Each model is fine-tuned with SGD [3] for 100 epochs. The batch size is 256 and the momentum is 0.9. The learning rate is searched in a range of 7 logarithmically-spaced values between 10^{-4} and 10^{-1} on a validation set. Weight decay is optional and if it is applied, the value is searched with the same setting between 10^{-6} and 10^{-3}. The standard augmentations, i.e., random crop and random horizontal flipping, are applied as in standard fine-tuning pipelines. The parameter λ_V in our method is fixed as 0.1 and shared by baseline methods with label smoothing, while tuning it may further improve our performance. The temperature $\tau' = 0.03$ and λ_T is searched in $[0.1, 1.5]$ with the validation set.

To obtain text supervision, two different language models are leveraged as the text encoder. First, we adopt the text encoder from CLIP [33] that is pre-trained with a visual encoder. Besides, we also include BERT [12] to explore the differences between the pure language model and the multi-modal language model. To extract features for each class name, the same prompt in [33], i.e., "a photo of a A" or "a photo of a A, a type of B" for fine-grained data, is applied as input for text encoder. The projection function h' consists of a 2-layer MLP suggested by the ablation experiment.

4.2. Comparison on ResNet

First, we compare different methods by fine-tuning a pre-trained ResNet. The comparison is mainly conducted on ResNet-50 as in Table 1, while the results on ResNet-18 are summarized in Table 2.

From Table 1, we can observe that conventional label smoothing methods cannot improve the performance over the baseline with a distinct margin. It is because that these methods are developed to avoid over-fitting in pre-training, while fine-tuning can mitigate over-fitting by leveraging the pre-trained model. Consequently, the bias in the pre-trained model is more critical for fine-tuning, which cannot be addressed by existing label smoothing techniques. With the instance-level reference distribution from the text encoder, our method TeS can effectively improve the performance over all data sets. By applying the text encoder pre-trained by CLIP, TeS outperforms CE by 1.34% in average with ResNet-50. Since the average performance with ResNet-50 already achieves about 85%, the improvement is relatively large. Finally, both text encoders can help fine-tune from ResNet-50 while the one from CLIP is 0.61% better than BERT. It shows that the text encoder pre-trained with vision data is more appropriate for supervising the fine-tuning of vision tasks. Therefore, the text encoder from CLIP will be adopted in the following experiments.

By further examining Table 2, we find that both supervised and unsupervised pre-trained models can benefit from TeS, which demonstrates that the proposed method is applicable for models pre-trained with different pretext tasks.

4.3. Comparison on ViT

Then, we compare the proposed method with ViT as the pre-trained model. The provided projection head in CLIP vision encoder is kept as pre-trained parameters for projection. Table 3 and 4 show the results with different pre-trained ViTs.

Method	Aircraft	Caltech	Cars	C10	C100	CUB	DTD	Flower	Food	Pet	SUN	Avg.
CE	79.69	93.13	88.65	98.23	87.68	78.08	73.08	93.63	89.76	92.72	65.34	85.45
CE + LS	78.82	93.65	88.67	98.15	87.50	78.98	74.26	93.02	89.96	92.98	66.76	85.71
CE + TLS	78.36	93.99	88.64	98.20	87.79	80.31	75.05	93.48	90.01	93.05	66.72	85.96
TeS + CLIP	81.24	94.10	91.12	98.45	88.92	81.60	74.84	94.52	90.50	93.27	67.62	86.93
	(+1.55)	(+0.11)	(+2.45)	(+0.22)	(+1.13)	(+1.29)	(-0.21)	(+0.89)	(+0.49)	(+0.22)	(+0.86)	(+0.97)

Table 3. Comparison with ViT pre-trained by MAE. The last row shows the accuracy improvement over the best baseline.

Method	Aircraft	Caltech	Cars	C10	C100	CUB	DTD	Flower	Food	Pet	SUN	Avg.
ZS	18.30	78.13	57.31	88.02	57.07	53.21	41.54	66.50	83.24	85.20	60.02	62.60
CE	76.93	94.30	89.99	97.84	87.31	80.79	76.11	96.41	89.08	91.47	67.04	86.12
CE + LS	77.28	94.78	89.25	97.98	88.79	78.82	76.33	96.32	88.42	91.57	70.28	86.35
CE + TLS	77.99	93.62	89.37	97.91	87.91	78.27	76.49	95.73	87.89	92.11	70.05	86.12
TeS + CLIP	78.10	94.91	90.14	98.08	88.62	80.40	77.13	96.70	88.57	92.23	70.96	86.90
	(+0.11)	(+0.13)	(+0.15)	(+0.10)	(-0.17)	(-0.39)	(+0.64)	(+0.29)	(-0.51)	(+0.12)	(+0.68)	(+0.55)

Table 4. Comparison with ViT pre-trained by CLIP. The last row shows the accuracy improvement over the best baseline.

Method	Aircraft	Caltech	Cars	C10	C100	CUB	DTD	Flower	Food	Pet	SUN	Avg.
CE	28.47	56.13	56.40	94.01	68.50	53.47	47.13	96.39	69.37	68.32	38.78	61.54
CE+LS	28.80	57.73	58.75	93.75	68.63	53.78	48.38	96.48	68.22	63.17	39.55	61.57
CE+TLS	29.92	56.62	58.85	94.31	67.09	53.24	47.21	96.10	65.63	64.92	39.52	61.22
TeS + CLIP	37.56	63.08	67.88	94.39	72.78	57.47	51.65	97.15	72.06	66.63	44.62	65.93
	(+7.64)	(+5.35)	(+9.03)	(+0.08)	(+4.15)	(+3.69)	(+3.27)	(+0.67)	(+2.69)	(-1.69)	(+5.07)	(+4.36)

Table 5. Comparison with 10% randomly sampled training examples and ResNet-50 pre-trained by MoCo-v2. The last row shows the accuracy improvement over the best baseline.

Evidently, a similar phenomenon as that with ResNet can be observed. First, the proposed TeS can improve the average performance for diverse downstream tasks, while conventional methods are ineffective for fine-tuning. Second, TeS surpasses CE by 1.48% when fine-tuning the ViT pre-trained by MAE, which shows the effectiveness of our method for different architectures. Note that the vision encoder pre-trained by CLIP already incorporates the text side information in contrastive pre-training, which makes the gain from our method less than that by MAE. Nevertheless, TeS can still improve the average performance by 0.78% over CE and the repeated experiments in supplementary confirm that TeS outperforms the best baseline significantly.

Third, fine-tuning vision encoders pre-trained by MAE and CLIP shows the similar average performance, while that for individual data sets varies. It demonstrates that different pre-training methods can encode diverse patterns, while the proposed method can obtain consistent improvement over various pre-trained models. Finally, compared with zero-shot transfer denoted as "ZS", fine-tuning outperforms ZS by a large margin of 24.3% and it suggests that fine-tuning is preferred when training examples are sufficient.

4.4. Comparison on Few-shot Learning

In addition to the above comparison with fine-tuning the whole data set, the proposed method is also evaluated in a challenging scenario, where each data set only contains 10% randomly sampled training examples for few-shot learning. The minimal number of examples will be set to 10 for each class. Pre-trained ResNet-50 is applied as the vision encoder and Table 5 demonstrates the comparison.

With only 10% training examples, it is challenging for fine-tuning to reduce the bias in the pre-trained model when distribution shifts. With text supervision, our method can explicitly constrain the learned distribution close to the target task and achieve 4.36% improvement on average. While our method demonstrates the superior performance on fine-tuning the whole data set, the experiment on few-shot learning implies that it can be more helpful when the number of target training examples is limited.

4.5. Comparison on Long-tailed Data

Finally, we investigate the proposed method for the data set with long-tailed distribution. Following [4], CIFAR-10 and CIFAR-100 with the imbalance ratio of 10 and 100 are included in the comparison. The performance of fine-tuning a pre-trained ResNet-50 is shown in Table 6.

Dataset	CIFAR-10		CIFAR-100	
Imba Ratio	100	10	100	10
CE	88.71	95.71	54.95	75.55
TeS	**89.80**	**96.00**	**57.76**	**77.69**

Table 6. Comparison on long-tailed CIFAR-10 and CIFAR-100 by fine-tuning ResNet-50.

With the reference distribution from text, the distribution of proxies from minor classes will not be overwhelmed by that from major classes. Thus, the performance of TeS surpasses the baseline CE with a clear margin. The improvement becomes larger when the imbalance ratio increases to 100 and it shows that our method is not sensitive for long-tailed data.

To further demonstrate the proposed method for long-tailed data, we evaluate TeS on a challenging data set iNaturalist-2017 [36] that contains 5,089 categories in Table 7. Evidently, our method outperforms cross entropy baseline with a large margin of 6.19% and it confirms the potential of text supervision for handling long-tailed data.

Method	Acc
CE	42.37
TeS	**48.56**

Table 7. Comparison of accuracy (%) on iNaturalist-2017 with ResNet-50.

Since the proposed method is not optimized for class imbalance learning, combining it with state-of-the-art methods for this specific task [4] is left as future work.

4.6. Ablation Study

In this subsection, we demonstrate the effect of different components in TeS by ablation study. Experiments are conducted with ResNet-50 on CIFAR-100.

#Layer in Projection h' h' contains multi-layer MLP to align vision representations with text classifier. We vary the number of layers and summarize the results in Table 8.

#Layer	1	2	3	4
Acc%	86.59	**86.95**	86.82	86.77

Table 8. Comparison of number of layers in projection function h'.

When h' has 1-layer MLP, it degenerates to a linear projection. If increasing the number of layers to 2, the non-linear mapping helps improve the performance by 0.36%. However, a more complicated head will not further improve the performance, which is consistent with the observation in [7]. Compared to ResNet-50, the computational overhead introduced by the 2-layer MLP is negligible, which keeps the efficiency of the proposed method.

Weight of Text Regularization λ_T λ_T in Eqn. 7 weights the loss for projecting the vision representation to the text space spanned by the text classifier. We vary it in $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5\}$ and Table 9 summarizes the result.

λ_T	0.1	0.3	0.5	0.7	0.9	1.1	1.3	1.5
Acc%	85.94	86.63	86.79	**86.95**	86.84	86.80	86.67	86.66

Table 9. Comparison of λ_T in Eqn. 7.

When λ_T is small, the projection head cannot be learned sufficiently, which results in an inaccurate estimation for the reference distribution. By increasing λ_T, the projection head can approximate the text space effectively and a satisfied performance can be obtained.

Temperature for Text Classifier τ' When approximating the representation in text space, a normalized Softmax operator is applied in cross entropy loss with a temperature parameter τ'. We vary it in $\{0.01, 0.03, 0.05, 0.07, 0.1, 0.2, 0.3\}$ and Table 10 summarizes the result. It is obvious that a small temperature is preferred to obtain a sharp distribution for supervision. Note that we have the standard Softmax for vision encoder and there is no temperature for vision classifier.

τ'	0.01	0.03	0.05	0.07	0.1	0.2	0.3
Acc%	86.56	**86.95**	86.83	86.28	85.99	85.70	85.27

Table 10. Comparison of the temperatures τ' in the text classifier.

Prompt for Text Supervision Despite that a single prompt for each class demonstrates the superior performance with TeS, an ensemble of prompts shows better performance for zero-shot transfer in CLIP [33]. Therefore, we also include the same ensemble strategy in TeS and the results are summarized in Table 11. While the ensemble is effective for zero-shot classification, the improvement for fine-tuning becomes marginal. Unlike the fixed pre-trained model in zero-shot learning, fine-tuning can leverage the labeled data to refine the pre-trained model. Hence, a single prompt is sufficient to generate appropriate text supervision for fine-tuning.

Variants of TeS Besides the instance-level distribution for regularization, the text supervision can be leveraged by directly mapping as in Eqn. 4 and class-level distribution as in Eqn. 6. We denote the two variants as TeS-M and TeS-C respectively and compare them to TeS in Table 12. In addition, the prediction from the text classifier is also evaluated.

Compared with the baseline CE, all of these methods can improve the fine-tuning performance on CIFAR-100. Moreover, TeS-C outperforms TeS-M and it shows that pairwise

Method	Aircraft	Caltech	Cars	C10	C100	CUB	DTD	Flower	Food	Pet	SUN	Avg.
CE	87.70	91.39	89.98	97.75	84.50	76.63	73.19	96.39	87.70	90.13	59.03	84.95
TeS+CLIP-S	87.87	92.86	**91.84**	98.00	**86.95**	**78.41**	75.00	97.15	**87.90**	**91.39**	61.86	86.29
TeS+CLIP-En	**88.23**	**93.18**	91.54	**98.08**	86.87	78.20	**75.32**	**97.18**	87.87	90.92	**61.99**	**86.31**

Table 11. Comparison of different prompt strategies on ResNet-50. "CLIP-S" and "CLIP-En" denote a single prompt and an ensemble of prompts for each class as in [33], respectively.

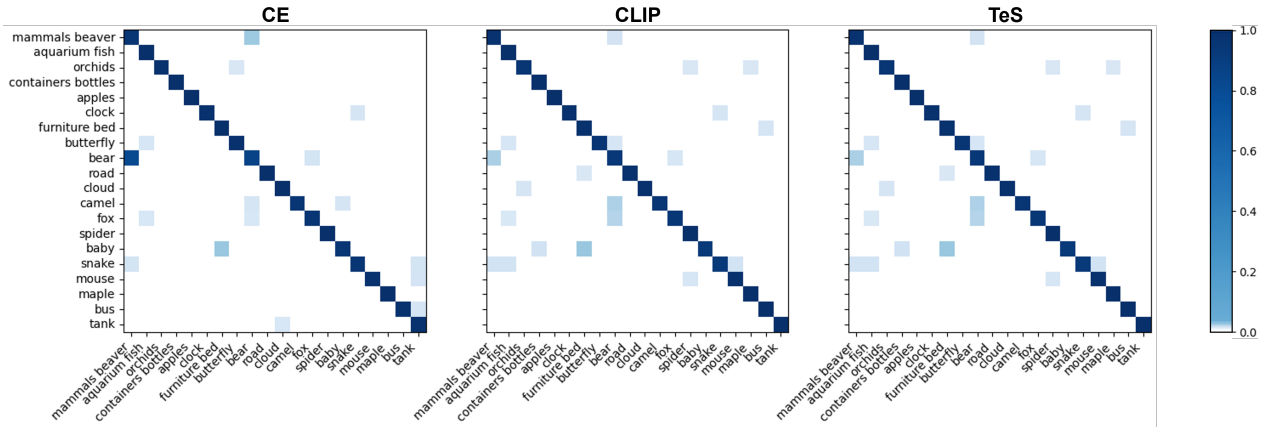

Figure 2. Illustration of confusion matrix from CE, CLIP, and TeS. Prediction on 20 classes from CIFAR-100 is shown for the comparison.

	CE	TeS-M	TeS-C	Text Head of TeS	TeS
Acc%	84.50	84.78	84.94	86.20	**86.95**

Table 12. Comparison of variants for TeS. TeS-M denotes optimizing the objective in Eqn. 4 while TeS-C minimizes the loss in Eqn. 6.

alignment is more effective for regularization than the direct mapping. With instance-level supervision as in TeS, the accuracy can be further improved by more than 2%. Finally, the prediction from the text head is 0.75% worse than that from the vision encoder, which demonstrates the modality gap between vision and language.

Illustration of Reference Distribution To illustrate the effect of reference distribution from text, we show the obtained confusion matrix from different methods in Fig. 2. Note that there are 20 superclasses in CIFAR-100, 20 target classes from different superclasses are randomly sampled for comparison.

First, it is obvious that CE and CLIP have a different off-diagonal distribution, which shows the difference between pre-trained vision encoder and text encoder. Since text encoder can be pre-trained with a much larger corpus containing context information, its distribution can be leveraged as the reference for fine-tuning small data sets. Comparing TeS to CLIP, the off-diagonal patterns, e.g., the relation between "baby" and "containers bottles" and that between "mouse" and "spider", can be well preserved. On the contrary, many biased patterns in CE, e.g., "orchids" and "butterfly", "bus" and "tank" are removed in TeS, which demonstrates our proposal.

5. Conclusion

While leveraging pre-trained models for fine-tuning becomes prevalent for downstream tasks, reducing the bias in pre-trained models has been less investigated. In this work, we propose to apply class names as natural language supervision and obtain reference distribution to help adapt the fine-tuned model to the target data distribution. Experiments on diverse downstream classification tasks demonstrate the efficacy of text supervision. After the success on classification, applying our method for different vision tasks can be our future work.

Limit To obtain the reference distribution, the exact name of each class is required as the input for the text encoder while it may be inaccessible due to the privacy policy. Therefore, the scenario without accurate class names can be challenging for the current method, which inspires a future research direction.

Acknowledgments This work is supported by the Beijing Natural Science Foundation (No. JQ20023).

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 2019.

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

[3] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20, 2007.

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1565–1576, 2019.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[9] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1906–1916, 2019.

[10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Kun Ding, Ying Wang, Pengzhang Liu, Qiang Yu, Haojian Zhang, Shiming Xiang, and Chunhong Pan. Prompt tuning with soft context sharing for vision-language models. *arXiv preprint arXiv:2208.13474*, 2022.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020.

[19] Kaiming He, Ross B. Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4917–4926. IEEE, 2019.

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

[21] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 2661–2671, 2019.

[22] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[24] Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2830–2839. PMLR, 2018.

[25] Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[26] Ziquan Liu, Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, Xiangyang Ji, Antoni B Chan, and Rong Jin. Improved fine-tuning by better leveraging pre-training data. In *Advances in Neural Information Processing Systems*, 2022.

[27] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

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

[29] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[30] Qi Qian, Juhua Hu, and Hao Li. Hierarchically robust representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7334–7342. Computer Vision Foundation / IEEE, 2020.

[31] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6449–6457. IEEE, 2019.

[32] Qi Qian, Yuanhong Xu, Juhua Hu, Hao Li, and Rong Jin. Unsupervised visual representation learning by online constrained k-means. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16619–16628. IEEE, 2022.

[33] 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.

[34] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[36] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[37] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[38] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[39] Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, and Rong Jin. Towards understanding label smoothing. *CoRR*, abs/2006.11653, 2020.

[40] 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, 2022.

[41] 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.

[42] Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. Debiased fine-tuning for vision-language models by prompt regularization. *arXiv preprint arXiv:2301.12429*, 2023.

[43] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020.