Unleashing the Power of Contrastive Self-Supervised Visual Models via Contrast-Regularized Fine-Tuning

Yifan Zhang[1]* **Bryan Hooi**[1]   **Dapeng Hu**[1]   **Jian Liang**[2]   **Jiashi Feng**[3]
[1]National University of Singapore   [2]Chinese Academy of Sciences   [3]SEA AI Lab

Abstract

Contrastive self-supervised learning (CSL) has attracted increasing attention for model pre-training via unlabeled data. The resulted CSL models provide instance-discriminative visual features that are uniformly scattered in the feature space. During deployment, the common practice is to directly fine-tune CSL models with cross-entropy, which however may not be the best strategy in practice. Although cross-entropy tends to separate inter-class features, the resulting models still have limited capability for reducing intra-class feature scattering that exists in CSL models. In this paper, we investigate whether applying contrastive learning to fine-tuning would bring further benefits, and analytically find that optimizing the contrastive loss benefits both discriminative representation learning and model optimization during fine-tuning. Inspired by these findings, we propose Contrast-regularized tuning (Core-tuning), a new approach for fine-tuning CSL models. Instead of simply adding the contrastive loss to the objective of fine-tuning, Core-tuning further applies a novel hard pair mining strategy for more effective contrastive fine-tuning, as well as smoothing the decision boundary to better exploit the learned discriminative feature space. Extensive experiments on image classification and semantic segmentation verify the effectiveness of Core-tuning.

1 Introduction

Pre-training a deep neural network on a large database and then fine-tuning it on downstream tasks has been a popular training scheme. Recently, contrastive self-supervised learning (CSL) has attracted increasing attention on model pre-training, since it does not rely on any hand-crafted annotations but even achieves more promising performance than supervised pre-training on downstream tasks [6, 7, 21, 23, 49]. Specifically, CSL leverages unlabeled data to train visual models via contrastive learning, which maximizes the feature similarity for two augmentations of the same instance and minimizes the feature similarity of two instances [62]. The learned models provide instance-discriminative visual representations that are uniformly scattered in the feature space [57].

Although there have been substantial CSL studies on model pre-training [24, 49], few have explored the fine-tuning process. The common practice is to directly fine-tune CSL models with the cross-entropy loss [6, 13, 21]. However, we empirically (cf. Table 1) find that different fine-tuning methods significantly influence the model performance on downstream tasks, and fine-tuning with only cross-entropy is not the optimal strategy. Intuitively, although cross-entropy tends to learn separable features among classes, the resulting model is still limited in its capability for reducing intra-class feature scattering [39, 59] that exists in CSL models. Meanwhile, most existing fine-tuning methods [34, 36] are devised for supervised pre-trained models and tend to enforce regularizers to prevent the fine-tuned models changing too much from the pre-trained ones. However, they suffer from the issue of negative transfer [9], since downstream tasks are often different from the pre-training contrastive task. In this sense, how to fine-tune CSL models remains an important yet under-explored question.

---

*Corresponding to: Yifan Zhang <yifan.zhang@u.nus.edu>

| + Positive Class | − Negative Class | ✚ Anchor Data | ⊞ Mixed Data |

(a) Sample features (b) Hard positive generation (c) Hard negative generation (d) Sharp classifier (e) Smooth classifier learning
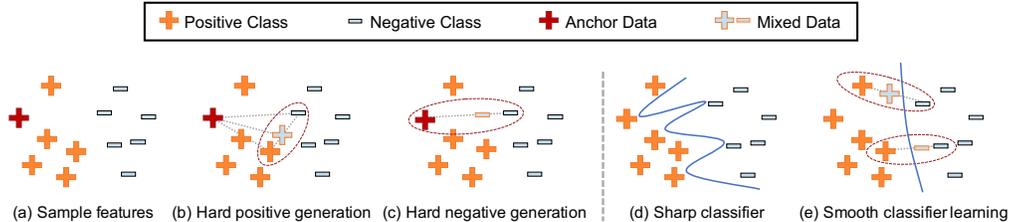
Figure 1: Illustration of two challenges in contrastive fine-tuning. (1) How to mine hard sample pairs for more effective contrastive fine-tuning. As shown in (a), the majority of sample pairs are easy-to-contrast, which may induce negligible contrastive loss gradients that contribute little to learning discriminative representations. (2) How to improve the generalizability of the model. As shown in (d), the classifier simply trained with cross-entropy is often sharp and near training data, leading to limited generalization performance.

Considering that optimizing the unsupervised contrastive loss during pre-training yields models with instance-level discriminative power, we investigate whether applying contrastive learning to fine-tuning would bring further benefits. To answer this, we analyze the contrastive loss during fine-tuning (cf. Section 3) and find that it offers two benefits. First, integrating the contrastive loss into cross-entropy can provide an additional regularization effect, as compared to cross-entropy based fine-tuning, for discriminative representation learning. Such an effect encourages the model to learn a low-entropy feature cluster for each class (*i.e.,* high intra-class compactness) and a high-entropy feature space (*i.e.,* large inter-class separation degree). Second, optimizing the contrastive loss will minimize the infimum of the cross-entropy loss over training data, which can provide an additional optimization effect for model fine-tuning. Based on the optimization effectiveness as well as the regularization effectiveness on representations, we argue that optimizing the contrastive loss during fine-tuning can further improve the performance of CSL models on downstream tasks.

Considering the above benefits, a natural idea is to directly add the contrastive loss to the objective for fine-tuning, *e.g.,* one recent study [19] simply uses contrastive learning to fine-tune language models. However, such a method cannot take full advantage of contrastive learning, since it ignores an important challenge in contrastive fine-tuning. That is, contrastive learning highly relies on positive/negative sample pairs, but the majority of sample features are easy-to-contrast (cf. Figure 1 (a)) [20, 60] and may produce negligible contrastive loss gradients. Ignoring this makes the method [19] fail to learn more discriminative features via contrastive learning and thus cannot fine-tune CSL models well.

In this paper, to better fine-tune CSL models and enhance their performance on downstream tasks, we propose a contrast-regularized tuning approach (termed Core-tuning), based on a novel hard pair mining strategy. Specifically, Core-tuning generates both hard positive and hard negative pairs for each anchor data via a new hardness-directed mixup strategy (cf. Figure 1 (b-c)). Here, hard positives indicate the positive pairs far away from the anchor, while hard negatives are the negative pairs close to the anchor. Meanwhile, since hard pairs are more informative for contrastive learning [20], Core-tuning further assigns higher importance weights to hard positive pairs based on a new focal contrastive loss. In this way, the resulting model is able to learn a more discriminative feature space by contrastive fine-tuning. Following that, we also explore how to better exploit the learned discriminative feature space in Core-tuning. Previous work has found that the decision boundary simply trained with cross-entropy is often sharp and close to training data [54], which may make the classifier fail to exploit the high inter-class separation degree in the discriminative feature space (cf. Figure 1 (d)), and also suffer from limited generalization performance. To address this, Core-tuning further uses the mixed features to train the classifier, so that the learned decision boundaries can be more smooth and far away from the original training data (cf. Figure 1 (e)).

The key contributions are threefold. 1) To our knowledge, we are among the first to look into the fine-tuning stage of CSL models, which is an important yet under-explored question. To address this, we propose a novel Core-tuning method. 2) We theoretically analyze the benefits of the supervised contrastive loss on representation learning and model optimization, revealing that it is beneficial to model fine-tuning. 3) Promising results on image classification and semantic segmentation verify the effectiveness of Core-tuning for improving the fine-tuning performance of CSL models. We also empirically find that Core-tuning benefits CSL models in terms of domain generalization and adversarial robustness on downstream tasks. Considering the theoretical guarantee and empirical effectiveness of Core-tuning, we recommend using it as a standard baseline to fine-tune CSL models.

2 Related Work

Contrastive self-supervised learning (CSL). Self-supervised learning is a kind of unsupervised learning method based on self-supervised proxy tasks, *e.g.,* rotation prediction [16], colorization prediction [31] and clustering [64]. Recently, CSL has become the most popular self-supervised paradigm, which treats each instance as a category to learn instance-discriminative representations. State-of-the-art CSL methods include InsDis [62], MoCo [21], SimCLR [6, 7] and InfoMin [49]. Most CSL studies are devoted to network pre-training, but few have explored the fine-tuning process.

As an effective data augmentation method, mixup [70] has recently been applied to instance augmentation for CSL [24, 27, 32, 47]. Among these methods, the work [24] uses mixup to generate hard negative pairs for better instance discrimination. However, all these methods focus on unsupervised pre-training and cannot accurately generate hard pairs regarding classes. Comparatively, Core-tuning focuses on the fine-tuning of CSL models and can generate accurate hard positive/negative pairs for each class. Note that the hardness-directed mixup strategy in Core-tuning is different from manifold mixup [54] that cannot be directly used to generate hard sample pairs.

Pre-training and Fine-tuning. In deep learning, it is a popular scheme to first pre-train a deep neural network on a large database (*e.g.,* ImageNet) and then fine-tune it on downstream tasks [36, 35]. Supervised learning is the mainstream method for pre-training [28], whereas self-supervised learning is attracting increasing attention since it does not rely on rich annotations [6, 7]. Most existing methods for fine-tuning, like L2-SP [36] and DELTA [34], are devised for supervised pre-trained models and tend to enforce some regularizer to prevent the fine-tuned models changing too much from the pre-trained ones. However, they may be unsuitable for contrastive self-supervised models, since downstream tasks are often different from the contrastive pre-training task, leading to negative transfer [9]. Very recently, one work [19] explored contrastive learning to fine-tune language models. However, it simply add the contrastive loss to the objective of fine-tuning and cannot theoretically explain why it boosts fine-tuning. More critically, it ignores the challenge of hard pair mining in contrastive fine-tuning and thus cannot fine-tune CSL models well.

3 Effects of Contrastive Loss for Model Fine-tuning

We start by analyzing the benefits of the contrastive loss during fine-tuning, which will motivate our new method. Before that, we first define the problem and notations.

Problem Definition and Notation. This paper studies the fine-tuning of contrastive self-supervised visual models that are pre-trained on a large-scale unlabeled database. During fine-tuning, let $\{(x_i, y_i)\}_{i=1}^n$ denote the target task dataset with n samples, where x_i is an instance with one-hot label $y_i \in \mathbb{R}^K$ and K denotes the number of classes. The neural network model is denoted by G, which consists of a pre-trained feature encoder G_e and a new predictor G_y specific to the target task. Based on the network, we extract visual representations by $z_i = G_e(x_i)$ and make a prediction by $\hat{y}_i = G_y(z_i)$. Such a contrastive self-supervised model is generally fine-tuned with the cross-entropy loss [13, 21].

Following [1], we define the random variables of samples and labels as X and Y, and those of embeddings and predictions as $Z|X \sim G_e(X)$ and $\hat{Y}|Z \sim G_y(Z)$, respectively. Moreover, let p_Y be the distribution of Y, $p_{(Y,Z)}$ be the joint distribution of Y and Z, and $p_{Y|Z}$ be the conditional distribution of Y given Z. We define the entropy of Y as $\mathcal{H}(Y) := \mathbb{E}_{p_Y}[-\log p_Y(Y)]$ and the conditional entropy of Y given Z as $\mathcal{H}(Y|Z) := \mathbb{E}_{p_{(Y,Z)}}[-\log p_{Y|Z}(Y|Z)]$. Besides, we define the cross-entropy (CE) between Y and $\hat{Y}$ by $\mathcal{H}(Y; \hat{Y}) := \mathbb{E}_{p_Y}[-\log p_{\hat{Y}}(Y)]$ and the conditional CE given Z by $\mathcal{H}(Y; \hat{Y}|Z) := \mathbb{E}_{p_{(Y,Z)}}[-\log p_{\hat{Y}|Z}(Y|Z)]$. Before our analysis, we first revisit contrastive loss.

Contrastive loss. We use the supervised contrastive loss [26] for fine-tuning, which is a variant of InfoNCE [45]. Specifically, given a sample feature z_i as anchor, the contrastive loss takes the features from the same class to the anchor as positive pairs and those from different classes as negative pairs. Assuming features are ℓ_2-normalized, the contrastive loss is computed by:

$$\mathcal{L}_{con} = -\frac{1}{n} \sum_{i=1}^n \frac{1}{|P_i|} \sum_{z_j \in P_i} \log \frac{e^{(z_i^\top z_j / \tau)}}{\sum_{z_k \in A_i} e^{(z_i^\top z_k / \tau)}}, \tag{1}$$

where τ is a temperature factor, while P_i and A_i denote the positive pair set and the full pair set of the anchor z_i, respectively. We next analyze the contrastive loss and find it has two beneficial effects.
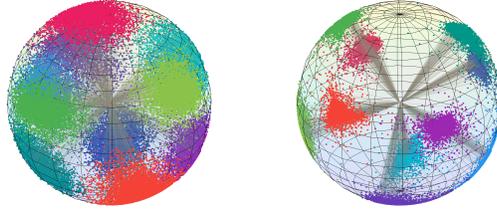
3.1 Regularization Effect of Contrastive Loss

We first show the contrastive loss has regularization effectiveness on representation learning based on the following theorem.

Theorem 1 *Assuming the features are ℓ_2-normalized and the classes are balanced with equal data number, minimizing the contrastive loss is equivalent to minimizing the class-conditional entropy $\mathcal{H}(Z|Y)$ and maximizing the feature entropy $\mathcal{H}(Z)$:*

$$\mathcal{L}_{con} \ \propto \ \mathcal{H}(Z|Y) \ - \ \mathcal{H}(Z)$$

Please see Appendix A for the proof. This theorem shows that $\mathcal{L}_{con}$ explicitly regularizes representation learning. On one hand, minimizing $\mathcal{L}_{con}$ will minimize $\mathcal{H}(Z|Y)$, which encourages learning a low-entropy cluster for each class (*i.e.,* high intra-class compactness). On the other hand, minimizing $\mathcal{L}_{con}$ will maximize $\mathcal{H}(Z)$ and tends to learn a high-entropy feature space (*i.e.,* large inter-class separation degree). This provides an additional regularization effect on the feature space, which can be observed by the feature visualization in Figure 2. As for the two assumptions, ℓ_2-normalized features can be satisfied by a non-linear projection in practice (cf. Section 4.1), while

(a) Training with $\mathcal{L}_{ce}$ (b) Training with $\mathcal{L}_{ce}+\mathcal{L}_{con}$

Figure 2: Visualizations of features learned by ResNet-18 on the CIFAR10 validation set. Compared to training with only cross-entropy $\mathcal{L}_{ce}$, the contrastive loss $\mathcal{L}_{con}$ helps to regularize the feature space and make it more discriminative. Best viewed in color.

contrastive fine-tuning also empirically performs well on class-imbalanced datasets (cf. Table 6). Note that this analysis is different from the analysis in unsupervised contrastive learning [57], which is specific to the (unlabeled) instance level rather than the class level.

3.2 Optimization Effect of Contrastive Loss

We next show that the contrastive loss improves the optimization effectiveness during model training via Theorem 2.

Theorem 2 *Assuming the features are ℓ_2-normalized and the classes are balanced, the contrastive loss is positive proportional to the infimum of conditional cross-entropy $\mathcal{H}(Y;\hat{Y}|Z)$, where the infimum is taken over classifiers:*

$$\mathcal{L}_{con} \ \propto \ \inf \underbrace{\mathcal{H}(Y;\hat{Y}|Z)}_{Conditional\ CE} \ - \ \mathcal{H}(Y)$$

Please see Appendix A for proofs. This theorem shows $\mathcal{L}_{con}$ boosts model optimization. Concretely, the label Y is given by datasets, so its entropy $\mathcal{H}(Y)$ is a constant and can be ignored. Hence, minimizing $\mathcal{L}_{con}$ will minimize the infimum of conditional cross-entropy $\mathcal{H}(Y;\hat{Y}|Z)$, which provides an additional optimization effect as compared to fine-tuning with only cross-entropy. More intuitively, pulling positive pairs together and pushing negative pairs further apart make the predicted label distribution closer to the ground-truth distribution, which further minimizes the cross-entropy loss.

4 Contrast-Regularized Tuning

Based on the above theoretical analysis, we are motivated to introduce contrastive learning to fine-tune contrastive self-supervised visual models on downstream tasks. Nevertheless, we empirically find that simply adding the contrastive loss to the fine-tuning objective is insufficient to obtain promising performance (cf. Table 2). One key cause is that contrastive learning highly relies on positive/negative sample pairs, but the majority of samples are easy-to-contrast pairs [20, 60] that may produce negligible contrastive loss gradients. This makes contrastive learning fail to learn more discriminative representations and thus suffer from unsatisfactory performance. To address this issue and better fine-tune contrastive self-supervised models, we propose a new contrast-regularized tuning (Core-tuning) method based on a novel hard sample pair mining strategy as follows.

4.1 Hard Sample Pair Mining for Contrastive Fine-Tuning

For more effective contrastive fine-tuning, Core-tuning generates both hard positive and hard negative pairs via a new hardness-directed mixup strategy, and meanwhile assigns higher importance weights to hard positive pairs via a new focal contrastive loss.

Hard positive pair generation. As shown in Figure 1 (b), for a given feature anchor z_i, we first find its hardest positive data (z_i^{hp}, y_i^{hp}) and hardest negative data (z_i^{hn}, y_i^{hn}) based on cosine similarity. That is, z_i^{hp} is the positive data (from the same class) with the lowest cosine similarity to the anchor, and z_i^{hn} is the negative data (from different classes) most similar to the anchor. We then generate a hard positive pair as a convex combination of the two hardest pairs:

$$z_i^+ = \lambda z_i^{hp} + (1-\lambda)z_i^{hn}; \quad y_i^+ = \lambda y_i^{hp} + (1-\lambda)y_i^{hn},$$

where $\lambda \sim \text{Beta}(\alpha, \alpha) \in [0, 1]$ [68], in which $\alpha \in (0, \infty)$ is a hyper-parameter to decide the Beta distribution. The generated positive pairs are located between positives and negatives and thus are harder to contrast. Note that the generated positive pairs do not have to be the hardest. In fact, as long as we can generate relatively hard pairs, the performance of contrastive fine-tuning could be improved.

Hard negative pair generation. As shown in Figure 1 (c), for a given feature anchor z_i, we randomly select a negative sample (z_i^n, y_i^n) to synthesize a semi-hard negative pair as follows:

$$z_i^- = (1-\lambda)z_i + \lambda z_i^n; \quad y_i^- = (1-\lambda)y_i + \lambda y_i^n,$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$. The reason why we select a random negative sample instead of the hardest negative is that generating too hard negatives may result in false negatives and degrade performance. Note that semi-hard negatives may even yield better performance in metric learning [61].

Hard pair reweighting. After generating hard sample pairs, we use an additional two-layer MLP head G_c to obtain ℓ_2-normalized contrastive features $v_i = G_c(z_i)/\|G_c(z_i)\|_2$, since a nonlinear projection improves contrastive learning [7, 8]. Based on these features, one may directly use $\mathcal{L}_{con}$ in Eq. (1) for fine-tuning. However, since hard pairs are more informative for contrastive learning, we propose to assign higher importance weights to hard positive pairs. Inspired by focal loss [37], we find that hard positive pairs generally lead to a low prediction probability $p_{ij} = \frac{\exp(v_i^\top v_j/\tau)}{\sum_{v_k \in A_i} \exp(v_i^\top v_k/\tau)}$.

Thus, we reweight $\mathcal{L}_{con}$ with $(1-p_{ij})$ and develop a focal contrastive loss:

$$\mathcal{L}_{con}^f = -\frac{1}{n}\sum_{i=1}^n \frac{1}{|P_i|}\sum_{v_j \in P_i}(1-p_{ij})\log\frac{e^{(v_i^\top v_j/\tau)}}{\sum_{v_k \in A_i} e^{(v_i^\top v_k/\tau)}},$$

where P_i, A_i denote the anchor's positive and full pair sets, which contain the generated hard pairs. Via the hard pair mining strategy, Core-tuning is able to learn a more discriminative feature space.

4.2 Overall Training Scheme and Smooth Classifier Learning

In fine-tuning, both the feature extractor and classifier need to be trained, so the final training scheme of Core-tuning[2] is to minimize the following objective:

$$\min \quad \underbrace{\mathcal{L}_{ce}^m}_{\text{cross-entropy loss}} + \underbrace{\eta\mathcal{L}_{con}^f}_{\text{focal contrastive loss}},$$

where η is a trade-off factor. Since hard sample mining has helped to learn a discriminative feature space, the remaining question is how to train the classifier for better exploiting such a feature space.

Smooth classifier learning. Previous work [54] has found that the classifier simply trained with cross-entropy is often sharp and close to data. This may make the classifier fail to exploit the high inter-class separation degree in the discriminative feature space due to closeness to training data, as well as suffer from limited generalization performance since the classifier near the training data may lead to incorrect yet confident predictions when evaluated on slightly different test samples. To address this, inspired by the effectiveness of mixup for helping learn a smoother decision boundary [42, 54], we further use the mixed data from the generated hard sample pair set (denoted by $\mathcal{B}$) for classifier training: $\mathcal{L}_{ce}^m = -\frac{1}{n}\sum_{i=1}^n y_i \log(\hat{y}_i) - \frac{1}{|\mathcal{B}|}\sum_{(z_j, y_j) \in \mathcal{B}} y_j \log(G_y(z_j))$. In this way, Core-tuning is able to learn a smoother classifier that is far away from the training data, and thus can better exploit the learned discriminative feature space and improve the model generalizability.

---

[2]The pseudo code is provided in the supplementary.

Table 1: Comparisons of various fine-tuning methods for the MoCo-v2 pre-trained ResNet-50 model on image classification in terms of top-1 accuracy. SL-CE-tuning denotes supervised pre-training on ImageNet and then fine-tuning with cross-entropy.

Method	ImageNet20	CIFAR10	CIFAR100	Caltech101	DTD	Aircraft	Cars	Pets	Flowers	Avg.
SL-CE-tuning	91.01	94.23	83.40	93.39	74.40	87.03	89.77	92.17	98.78	89.35
CE-tuning	88.28	94.70	80.27	91.87	71.68	86.87	88.61	89.05	98.49	87.76
L2SP [36]	88.49	95.14	81.43	91.98	72.18	86.55	89.00	89.43	98.66	88.10
M&M [66]	88.53	95.02	80.58	92.91	72.43	87.45	88.90	89.60	98.57	88.22
DELTA [34]	88.35	94.76	80.39	92.19	72.23	87.05	88.73	89.54	98.65	87.99
BSS [9]	88.34	94.84	80.40	91.95	72.22	87.18	88.50	89.50	98.57	87.94
RIFLE [35]	89.06	94.71	80.36	91.94	72.45	87.60	89.72	90.05	98.70	88.29
SCL [19]	89.29	95.33	81.49	92.84	72.73	87.44	89.37	89.71	98.65	88.54
Bi-tuning [78]	89.06	95.12	81.42	92.83	73.53	87.39	89.41	89.90	98.57	88.58
Core-tuning (ours)	**92.73**	**97.31**	**84.13**	**93.46**	**75.37**	**89.48**	**90.17**	**92.36**	**99.18**	**90.47**

5 Experiments

We first test the effectiveness of Core-tuning on image classification and then apply it to semantic segmentation. Next, since Core-tuning potentially improves model generalizability, we further study how it affects model generalization to new domains and model robustness to adversarial samples.

5.1 Results on Image Classification

Settings. As there is no fine-tuning method devoted to contrastive self-supervised models, we compare Core-tuning with advanced fine-tuning methods for general models (*e.g.,* supervised pre-trained models): L2SP [36], M&M [66], DELTA [34], BSS [9], RIFLE [35], SCL [19] and Bi-tuning [78]. We denote the fine-tuning with cross-entropy by CE-tuning.

Following [28], we test on 9 natural image datasets, including ImageNet20 (a subset of ImageNet with 20 classes), CIFAR10, CIFAR100 [30], Caltech-101 [15], DTD [10], FGVC Aircraft [41], Standard Cars [29], Oxford-IIIT Pets [46] and Oxford 102 Flowers [44]. Specifically, ImageNet20 is an ImageNet subset with 20 classes, by combining the ImageNette and ImageWoof datasets [22]. Here, we do not directly test on ImageNet [11], since all CSL models are pre-trained on the ImageNet dataset. These datasets cover a wide range of fine/coarse-grained object recognition tasks.

We implement Core-tuning in PyTorch[3]. Following [13], we use ResNet-50 ($1\times$), pre-trained by various CSL methods on ImageNet, as the network backbone. All checkpoints of pre-trained models are provided by authors or by the PyContrast repository[4]. Following [6], we perform parameter tuning for η and α from $\{0.1, 1, 10\}$ on each dataset. Moreover, we set the temperature $\tau = 0.07$. To make the generated negative pairs closer to negatives, we clip $\lambda \sim \text{Beta}(\alpha, \alpha)$ by $\lambda \geq \lambda_n$ when generating hard negative pairs, where λ_n is a threshold and we set it to 0.8. All results are averaged over 3 runs in terms of the top-1 accuracy. More dataset details, more implementation details and the parameter analysis are put in Appendices C and E.

Comparisons with previous methods. We report the fine-tuning performance of the MoCo-v2 pre-trained model in Table 1. When using the standard CE-tuning, the MoCo-v2 pre-trained model performs worse than the supervised pre-trained model on most datasets. This is because the self-supervised pre-trained model is less class-discriminative than the supervised pre-trained model due to the lack of annotations during pre-training. Moreover, the classic fine-tuning methods designed for supervised pre-trained models (*e.g.,* L2SP and DELTA) cannot fine-tune the contrastive self-supervised model very well. One reason is that the contrastive pre-training task is essentially different from the downstream classification task, so strictly regularizing the difference between the contrastive self-supervised model and the fine-tuned model may lead to negative/poor transfer. In addition, M&M, SCL and Bi-tuning use the triplet loss or the contrastive loss during fine-tuning. However, they ignore the two challenges in contrastive fine-tuning as mentioned in Figure 1, leading to limited model performance on downstream tasks. In contrast, Core-tuning handles those challenges well and improves the fine-tuning performance of CSL models a lot. This result demonstrates the superiority of Core-tuning. More results like the standard error are put in Appendix D.

---

[3]The source code of Core-tuning is available at: `https://github.com/Vanint/Core-tuning`.
[4]`https://github.com/HobbitLong/PyContrast`.

Table 2: Ablation studies of Core-tuning (Row 5) for fine-tuning MoCo-v2 pre-trained ResNet-50 in terms of top-1 accuracy, where cross-entropy is used in all baselines. Here, $\mathcal{L}_{con}$ is the original contrastive loss, while $\mathcal{L}_{con}^f$ is our focal contrastive loss. Moreover, "mix" denotes the manifold mixup, while "mix-H" indicates the proposed hardness-directed mixup strategy in our method.

$\mathcal{L}_{con}$	$\mathcal{L}_{con}^f$	mix	mix-H	ImageNet20	CIFAR10	CIFAR100	Caltech101	DTD	Aircraft	Cars	Pets	Flowers	Avg.
				88.28	94.70	80.27	91.87	71.68	86.87	88.61	89.05	98.49	87.76
√				89.29	95.33	81.49	92.84	72.73	87.44	89.37	89.71	98.65	88.54
		√		90.67	95.43	81.03	92.68	73.31	88.37	89.06	91.37	98.74	88.96
√			√	92.20	97.01	83.89	93.22	74.78	88.88	89.79	91.95	98.94	90.07
	√		√	**92.73**	**97.31**	**84.13**	**93.46**	**75.37**	**89.48**	**90.17**	**92.36**	**99.18**	**90.47**

Table 3: Fine-tuning results of ResNet-50, pre-trained by various methods. "Cont." indicates contrastive self-supervised pre-training; CE indicates cross-entropy.

Pre-training	Types	Caltech101		DTD		Pets	
		CE	ours	CE	ours	CE	ours
InsDis [62]		82.30	**88.60**	69.81	**70.94**	87.57	**89.59**
PIRL [43]	Cont.	84.23	**89.29**	68.95	**71.72**	86.87	**89.52**
MoCo-v1 [21]		85.74	**89.16**	69.91	**71.90**	88.16	**90.11**
InfoMin [49]		92.73	**94.01**	72.59	**74.89**	90.00	**92.34**
DeepCluster[2]		89.99	**92.34**	72.77	**75.21**	90.53	**93.17**
SwAV [3]	Non-Cont.	87.71	**91.34**	75.29	**77.41**	92.48	**93.29**
BYOL [17]		91.19	**93.25**	74.94	**76.56**	92.39	**93.74**
CE	Supervised	93.65	**94.20**	74.40	**77.27**	92.17	**93.82**

Table 4: Fine-tuning performance of various architectures. Here, ResNet (R) and ResNeXt (RX) are pre-trained by InfoMin; DeiT-S [50] is pre-trained by DINO [4].

Archs.	Caltech101		DTD		Pets	
	CE	ours	CE	ours	CE	ours
R-50	92.73	**94.01**	72.59	**74.89**	90.00	**92.34**
R-101	93.06	**94.33**	73.38	**75.09**	90.84	**92.91**
R-152	93.39	**94.66**	73.74	**75.42**	91.08	**92.97**
RX-101	93.71	**95.12**	74.43	**75.97**	91.97	**94.04**
RX-152	93.92	**95.19**	74.76	**76.22**	92.70	**94.49**
DeiT-S/16	91.24	**92.31**	71.35	**72.83**	92.43	**93.72**

Ablation studies of Core-tuning. We conduct ablation studies for Core-tuning regarding the focal contrastive loss and the hardness-directed mixup strategy. As shown in Table 2, each component improves the fine-tuning performance in Core-tuning. Note that the mixup in Row 3 is the manifold mixup [54], which is essentially designed for classification and is expected to outperform our hardness-directed mixup strategy regarding classification performance. However, our proposed Core-tuning (Row 5) still shows obvious improvement on all datasets, which strongly verifies the value of contrastive fine-tuning. More ablation results for verifying the effectiveness of hard pair generation and smooth classifier learning are put in Appendix E.

Results on different pre-training methods. In previous experiments, we fine-tune the MoCo-v2 pre-trained ResNet-50, but it is unclear whether Core-tuning can be applied to fine-tune models with other pre-training methods. Hence, we further use Core-tuning to fine-tune ResNet-50, pre-trained by other CSL methods (*i.e.,* InsDis [62], PIRL [43], MoCo-v1 [21] and InfoMin [49]), non-contrastive self-supervised methods (*i.e.,* DeepCluster-v2 [2], SwAV [3] and BYOL [17]), and supervised learning. As shown in Table 3, Core-tuning fine-tunes all pre-trained models consistently better than CE-tuning on 3 image classification datasets. Such results verify the generalizability of the proposed Core-tuning. More results on different pre-trained models are put in Appendix D.

Results on different network architectures. Previous experiments are based on ResNet-50, while it is unclear whether Core-tuning can be applied to other network architectures. Hence, we further use Core-tuning to fine-tune various residual network architectures (*i.e.,* ResNet-101 and 152; ResNeXt-101 and 152 [63]) pre-trained by InfoMin [49], and vision transformer (*i.e.,* DeiT-S/16 [50]) pre-trained by DINO [4]. As shown in Table 4, Core-tuning fine-tunes all network architectures well on all three datasets, showing strong universality.

Results on different data sizes. The labeled data may be scarce in downstream tasks. Hence, we further evaluate Core-tuning on ImageNet20 with different sampling rates of data. We report the results in Table 5, while the results on the full ImageNet20 have been listed in Table 1. Specifically, Core-tuning outperforms baselines in all cases. Note that when the data is very scarce (*e.g.,* 10%), the fine-tuning performance of CE-tuning degrades and fluctuates significantly, in which case Core-tuning obtains more significant improvement and achieves more stable performance.

Results on large-scale and class-imbalanced dataset. The real-world datasets may be large-scale and class-imbalanced [72, 73, 77], so we also evaluate Core-tuning on a long-tailed iNaturalist18 dataset [52], consisting 437,513 images from 8,142 classes. As shown in Table 6, Core-tuning also performs well on the large-scale and class-imbalanced dataset for fine-tuning contrastive pre-trained models. Note that in our theoretical analysis, we assume that the classes are balanced with the same data number to facilitate analysis. Nevertheless, this assumption does not mean that contrastive fine-tuning cannot handle class-imbalanced datasets. Here, the promising results on iNaturalist-18 verify the effectiveness of Core-tuning on highly class-imbalanced scenarios.

Table 5: Fine-tuning performance of the MoCo-v2 pre-trained ResNet-50 with various numbers of labeled data.

Method	Sampling Rates on ImageNet20			
	10%	25%	50%	75%
CE-tuning	52.97+/-3.96	63.17+/-3.94	81.78+/-1.37	85.85+/-0.11
Bi-tuning	60.50+/-1.11	75.86+/-0.74	83.18+/-0.27	87.19+/-0.19
Core-tuning	**78.64+/-0.58**	**84.48+/-0.34**	**89.09+/-0.40**	**90.93+/-0.24**

Table 6: Fine-tuning performance of the MoCo-v2 pre-trained ResNet-50 on large-scale and class-imbalanced iNaturalist18 in terms of top-1 accuracy.

Fine-tuning method	iNaturalist18
CE-tuning	61.72+/-0.18
CE-Contrastive-tuning	62.75+/-0.22
Core-tuning (ours)	**63.57+/-0.09**

Table 7: Fine-tuning performance on PAS-CAL VOC semantic segmentation based on DeepLab-V3 with ResNet-50, pre-trained by various CSL methods.

Pre-training	Fine-tuning	MPA	FWIoU	MIoU
Supervised	CE	87.10	89.12	76.52
InsDis	CE	83.64	88.23	74.14
	ours	**84.53**	**88.67**	**74.81**
PIRL	CE	83.16	88.22	73.99
	ours	**85.30**	**88.95**	**75.49**
MoCo-v1	CE	84.71	88.75	74.94
	ours	**85.70**	**89.19**	**75.94**
MoCo-v2	CE	87.31	90.26	78.42
	ours	**88.76**	**90.75**	**79.62**
InfoMin	CE	87.17	89.84	77.84
	ours	**88.92**	**90.65**	**79.48**

5.2 Results on Semantic Segmentation

We next apply Core-tuning to fine-tune contrastive self-supervised models on semantic segmentation.

Implementation details. We adopt the DeepLab-V3 framework [5] for PASCAL VOC semantic segmentation and use CSL pre-trained ResNet-50 models as the backbone. In Core-tuning, we enforce the contrastive regularizer after the penultimate layer of ResNet-50 via an additional global average pooling. Following [58], the model is fine-tuned on VOC train_aug2012 set for 30k steps via SGD based on two GPUs and evaluated on val2012 set. The image is rescaled to 513×513 with random crop and flips for training and with center crop for evaluation. The batch size and output stride are 16. Besides, we set the initial learning rate to 0.1 and adjust it via the poly decay schedule. Other parameters are the same as image classification. We use three metrics: Mean Pixel Accuracy (MPA), Frequency Weighted Intersection over Union (FWIoU) and Mean Intersection over Union (MIoU).

Results. As shown in Table 7, Core-tuning contributes to the fine-tuning performance of all CSL models in terms of MPA, FWIoU and MIoU. The promising results demonstrate the effectiveness of Core-tuning on semantic segmentation. Interestingly, we find that with standard fine-tuning, the models pre-trained by MoCo-v2 and InfoMin have already outperformed the supervised pre-trained model. One explanation is that self-supervised pre-training may keep more visual information, compared to supervised pre-training that mainly extracts information specific to classification [76]. In other words, unsupervised contrastive learning may extract more beneficial information for dense prediction, which inspires us to explore unsupervised contrastive regularizers in the future.

5.3 Effectiveness on Cross-Domain Generalization

The generalizability of deep networks to unseen domains is important for their application to real-world scenarios [12]. We thus wonder whether Core-tuning also benefits model generalization on downstream tasks, so we apply Core-tuning to the task of domain generalization (DG).

Implementation details. DG aims to train a model on multiple source domains and expect it to generalize well to an unseen target domain. Specifically, we use MoCo-v2 pre-trained ResNet-50 as the backbone, and evaluate Core-tuning on 3 benchmark datasets, *i.e.,* PACS [33], VLCS [14] and Office-Home [53]. For training, we use Adam optimizer with batch size 32. The learning rate is set to 5×10^{-5} and the training step is 20,000. More implementation details are put in Appendix C.

Results. We report the results on PACS and VLCS in Table 8 and the results on OfficeHome in Appendix D, from which we draw observations as follows. First, when fine-tuning with cross-entropy, the contrastive self-supervised model performs worse than the supervised pre-trained model. This results from the relatively worse discriminative abilities of the contrastive self-supervised model, which can also be found in Table 1. Second, enforcing the contrastive regularizer during fine-tuning improves DG performance, since the contrastive regularizer helps to learn more discriminative features (cf. Theorem 1) and also helps to alleviate distribution shifts among domains [25]. Last, Core-tuning further improves the generalization performance of models. This is because hard pair generation further boosts contrastive learning, while smooth classifier learning benefits model generalizability. We thus conclude that Core-tuning is beneficial to model generalization on downstream tasks.

Table 8: Domain generalization accuracies of various fine-tuning methods for MoCo-v2 pre-trained ResNet-50. CE means cross-entropy; CE-Con enhances CE with the contrastive loss. Moreover, A/C/P/S and C/L/V/S are different domains in PACS and VLCS datasets, respectively.

Pre-training	Fine-tuning	PACS					VLCS				
		A	C	P	S	Avg.	C	L	V	S	Avg.
Supervised	CE	83.65	79.21	96.11	81.46	85.11	98.41	63.81	68.55	75.45	76.56
MoCo-v2	CE	78.71	76.92	90.87	75.67	80.54	94.96	66.87	68.96	64.98	73.94
	CE-Con	85.11	81.77	95.58	80.12	85.65	95.94	67.76	69.31	73.57	77.67
	ours	**87.31**	**84.06**	**97.53**	**83.43**	**88.08**	**98.50**	**68.19**	**73.15**	**81.53**	**80.34**

Table 9: Adversarial training performance of MoCo-v2 pre-trained ResNet-50 on CIFAR10 under the attack of PGD-10 in terms of robust and clean accuracies. AT-CE indicates adversarial training (AT) with CE; AT-CE-Con enhances AT-CE with the contrastive loss; AT-ours uses Core-tuning for AT.

Method	ℓ_2-attack						ℓ_∞-attack					
	ϵ=0.5		ϵ=1.5		ϵ=2.5		ϵ= 2/255		ϵ= 4/255		ϵ= 8/255	
	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean
CE	50.25	94.70	48.29	94.70	46.82	94.70	25.13	94.70	12.28	94.70	4.57	94.70
AT-CE	86.59	92.00	89.60	94.28	89.16	94.15	83.20	93.05	75.82	91.99	69.27	92.79
AT-CE-Con	90.74	94.71	90.29	94.80	89.70	94.27	85.07	94.56	79.75	93.79	70.70	93.38
AT-ours	**92.97**	**96.82**	**92.32**	**96.90**	**92.05**	**96.87**	**86.92**	**96.29**	**82.01**	**95.95**	**74.83**	**95.90**

5.4 Robustness to Adversarial Samples

As is known, deep networks are fragile to adversarial attack [48]. We next study whether Core-tuning also benefits model robustness to adversarial samples in the setting of adversarial training (AT).

Implementation details. We use MoCo-v2 pre-trained ResNet-50 as the network backbone, and use the Projected Gradient Descent (PGD) [40] to generate adversarial samples with ℓ_2 attack (strength σ=0.5) and ℓ_∞ attack (strength σ=4/255). During AT, we use both original samples and adversarial samples for fine-tuning. Moreover, we use the clean accuracy on original samples and the robust accuracy on adversarial samples as metrics. More implementation details are put in Appendix C.

Results. We report the results on CIFAR10 in Table 9 and the results on Caltech-101, DTD and Pets in Appendix D. First, despite good clean accuracy, fine-tuning with cross-entropy cannot defend against adversarial attack, leading to poor robust accuracy. Second, AT with cross-entropy improves the robust accuracy significantly, but it inevitably degrades the clean accuracy due to the well-known accuracy-robustness trade-off [51]. In contrast, the contrastive regularizer improves both robust and clean accuracies. This is because contrastive learning helps to improve robustness generalization (*i.e.,* alleviating the distribution shifts between clean and adversarial samples). Last, Core-tuning further boosts AT and, surprisingly, even achieves better clean accuracy than the standard fine-tuning with cross-entropy. To our knowledge, this is quite promising since even the most advanced AT methods [65, 69] find it difficult to overcome the accuracy-robustness trade-off [67]. The improvement is because both contrastive learning and smooth classifier learning boost robustness generalization. We thus conclude that Core-tuning improves model robustness on downstream tasks.

6 Conclusions

This paper studies how to fine-tune contrastive self-supervised visual models. We theoretically show that optimizing the contrastive loss during fine-tuning has regularization effectiveness on representation learning as well as optimization effectiveness on classifier training, both of which benefit model fine-tuning. We thus propose a novel contrast-regularized tuning (Core-tuning) method to fine-tune CSL visual models. Promising results on image classification and semantic segmentation verify the effectiveness of Core-tuning. Also, we empirically find that Core-tuning is beneficial to model generalization and robustness on downstream tasks. We thus recommend using Core-tuning as a standard baseline to fine-tune CSL visual models, and also call for more attention to the fine-tuning of CSL visual models on understanding its underlying theories and better approaches in the future.

Limitation discussion. One potential limitation of Core-tuning is that it is specifically designed for and also focuses on the fine-tuning of CSL visual models. Considering the universality of Core-tuning (cf. Table 3), we will explore the extension of Core-tuning to better fine-tune supervised pre-trained and other self-supervised visual models and even language models on more tasks.

Acknowledgments

References

[1] Malik Boudiaf, Jérôme Rony, et al. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European Conference on Computer Vision*, 2020.

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv*, 2021.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, 2017.

[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*, 2020.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, 2020.

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

[9] Xinyang Chen, Sinan Wang, et al. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *Advances in Neural Information Processing Systems*, 2019.

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

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, 2019.

[13] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *IEEE Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

[14] Chen Fang, Ye Xu, et al. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *International Conference on Computer Vision*, 2013.

[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 *Computer Vision and Pattern Recognition Workshop*, 2004.

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Guo, Mohammad Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Neural Information Processing Systems*, 2020.

[18] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[19] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*, 2021.

[20] Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *International Conference on Computer Vision*, pages 2821–2829, 2017.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Computer Vision and Pattern Recognition*, 2020.

[22] Jeremy Howard. The imagenette and imagewoof datasets, 2020.

[23] Dapeng Hu, Qizhengqiu Lu, Lanqing Hong, Hailin Hu, Yifan Zhang, Zhenguo Li, Alfred Shen, and Jiashi Feng. How well self-supervised pre-training performs with streaming data? *ArXiv*, 2021.

[24] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

[25] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition*, 2019.

[26] Prannay Khosla, Piotr Teterwak, et al. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

[27] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. *ArXiv*, 2020.

[28] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Computer Vision and Pattern Recognition*, pages 2661–2671, 2019.

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

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

[31] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593, 2016.

[32] Kibok Lee, Yian Zhu, Kihyuk Sohn, et al. i-mix: A strategy for regularizing contrastive representation learning. In *International Conference on Learning Representations*, 2021.

[33] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *International Conference on Computer Vision*, pages 5542–5550, 2017.

[34] Xingjian Li, Haoyi Xiong, et al. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *International Conference on Learning Representations*, 2019.

[35] Xingjian Li, Haoyi Xiong, et al. Rifle: Backpropagation in depth for deep transfer learning through re-initializing the fully-connected layer. In *International Conference on Machine Learning*, 2020.

[36] Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, 2018.

[37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision*, pages 2980–2988, 2017.

[38] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Computer Vision and Pattern Recognition*, 2017.

[39] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, pages 507–516, 2016.

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[41] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, 2013.

[42] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Winter Conference on Applications of Computer Vision*, pages 2218–2227, 2020.

[43] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

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

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, 2018.

[46] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Computer Vision and Pattern Recognition*, 2012.

[47] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, and Trevor Darrell. Rethinking image mixture for unsupervised visual representation learning. *ArXiv*, 2020.

[48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[49] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021.

[51] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[52] 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 *Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[53] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition*, 2017.

[54] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447, 2019.

[55] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. *Arxiv*, 2020.

[56] Meihong Wang and Fei Sha. Information theoretical clustering via semidefinite programming. In *International Conference on Artificial Intelligence and Statistics*, pages 761–769, 2011.

[57] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 2020.

[58] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[59] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.

[60] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *International Conference on Computer Vision*, pages 2840–2848, 2017.

[61] Mike Wu, Milan Mosse, Chengxu Zhuang, Daniel Yamins, and Noah Goodman. Conditional negative sampling for contrastive learning of visual representations. In *International Conference on Learning Representations*, 2021.

[62] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Computer Vision and Pattern Recognition*, 2018.

[63] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition*, 2017.

[64] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Computer Vision and Pattern Recognition*, 2020.

[65] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *Advances in Neural Information Processing Systems*, 2020.

[66] Xiaohang Zhan, Ziwei Liu, Ping Luo, Xiaoou Tang, and Chen Change Loy. Mix-and-match tuning for self-supervised semantic segmentation. In *AAAI Conference on Artificial Intelligence*, 2018.

[67] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

[68] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[69] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021.

[70] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? In *International Conference on Learning Representations*, 2021.

[71] Yifan Zhang, Hanbo Chen, et al. From whole slide imaging to microscopy: Deep microscopy adaptation network for histopathology cancer image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 360–368, 2019.

[72] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *ArXiv*, 2021.

[73] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *ArXiv*, 2021.

[74] Yifan Zhang, Ying Wei, et al. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 2020.

[75] Yifan Zhang, Peilin Zhao, et al. Online adaptive asymmetric active learning for budgeted imbalanced data. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[76] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? In *International Conference on Learning Representations*, 2021.

[77] Peilin Zhao, Yifan Zhang, Min Wu, Steven CH Hoi, Mingkui Tan, and Junzhou Huang. Adaptive cost-sensitive online classification. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[78] Jincheng Zhong, Ximei Wang, Zhi Kou, Jianmin Wang, and Mingsheng Long. Bi-tuning of pre-trained representations. *Arxiv*, 2021.

Supplementary Materials:
Unleashing the Power of Contrastive Self-Supervised Visual Models via Contrast-Regularized Fine-Tuning

This supplementary material provides proofs for the analysis of the contrastive loss (cf. Appendix A), the pseudo-code of the proposed method (cf. Appendix B), more implementation details (cf. Appendix C), and more empirical results and analysis (cf. Appendices D and E).

A Proof of Theoretical Analysis

This appendix provides proofs for both Theorems 1 and 2.

A.1 Proof for Theorem 1

Proof We follow the notations in the main paper and further denote the sample set of the class k by $\mathcal{Z}_k$. Moreover, we assume the classes of samples are balanced so that the sample number of each class is constant $|\mathcal{Z}_k| = \frac{n}{K}$, where n denotes the total number of samples and K indicates the number of classes. Let us start by splitting the contrastive loss into two terms.

$$
\begin{aligned}
\mathcal{L}_{con} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|P_i|}\sum_{z_j\in P_i}\log\frac{e^{(v_i^\top v_j/\tau)}}{\sum_{v_k\in A_i}e^{(v_i^\top v_k/\tau)}} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|P_i|}\sum_{z_j\in P_i}\frac{z_i^\top z_j}{\tau} + \frac{1}{n}\sum_{i=1}^{n}\log\sum_{z_k\in A_i}e^{(\frac{z_i^\top z_k}{\tau})}.
\end{aligned}
\tag{2}
$$

Let $c_k = \frac{1}{|\mathcal{Z}_k|}\sum_{z\in\mathcal{Z}_k}z$ denote the hard mean of all features from the class k, and let the symbol $\overset{c}{=}$ indicate equality up to a multiplicative and/or additive constant. We first analyze the first term in Eq. (2) by connecting it to a tightness term of the center loss, *i.e.,* $\sum_{z_i\in\mathcal{Z}_k}\|z_i-c_k\|^2$ [59]:

$$
\begin{aligned}
\sum_{z_i,z_j\in\mathcal{Z}_k}-\frac{z_i^\top z_j}{\tau} &\overset{c}{=} \frac{1}{|\mathcal{Z}_k|}\sum_{z_i,z_j\in\mathcal{Z}_k}-z_i^\top z_j \\
&\overset{c}{=} \frac{1}{|\mathcal{Z}_k|}\sum_{z_i,z_j\in\mathcal{Z}_k}\|z_i\|^2 - z_i^\top z_j \\
&= \sum_{z_i\in\mathcal{Z}_k}\|z_i\|^2 - \frac{1}{|\mathcal{Z}_k|}\sum_{z_i\in\mathcal{Z}_k}\sum_{z_j\in\mathcal{Z}_k}z_i^\top z_j \\
&= \sum_{z_i\in\mathcal{Z}_k}\|z_i\|^2 - 2\frac{1}{|\mathcal{Z}_k|}\sum_{z_i\in\mathcal{Z}_k}\sum_{z_j\in\mathcal{Z}_k}z_i^\top z_j \\
&\quad + \frac{1}{|\mathcal{Z}_k|}\sum_{z_i\in\mathcal{Z}_k}\sum_{z_j\in\mathcal{Z}_k}z_i^\top z_j \\
&= \sum_{z_i\in\mathcal{Z}_k}\|z_i\|^2 - 2z_i^\top c_k + \|c_k\|^2 \\
&= \sum_{z_i\in\mathcal{Z}_k}\|z_i-c_k\|^2,
\end{aligned}
$$

where we use the property of ℓ_2-normalized features that $\|z_i\|^2 = \|z_j\|^2 = 1$ and the definition of the class hard mean $c_k = \frac{1}{|\mathcal{Z}_k|}\sum_{z\in\mathcal{Z}_k}z$.

By summing over all classes k, we obtain:

$$\sum_{i=1}^{n} \sum_{z_j \in P_i} -\frac{z_i^\top z_j}{\tau} \overset{c}{=} \sum_{i=1}^{n} \|z_i - c_{y_i}\|^2.$$

Based on this equation, following [1], we can interpret the first term in Eq. (2) as a conditional cross-entropy between Z and another random variable $\bar{Z}$, whose conditional distribution given Y is a standard Gaussian centered around c_Y: $\bar{Z}|Y \sim \mathcal{N}(c_y, i)$:

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{1}{|P_i|} \sum_{z_j \in P_i} \frac{z_i^\top z_j}{\tau} \overset{c}{=} \mathcal{H}(Z; \bar{Z}|Y) = \mathcal{H}(Z|Y) + \mathcal{D}_{KL}(Z||\bar{Z}|Y).$$

Based on this, we know that the first term in Eq. (2) is an upper bound on the conditional entropy of features Z given labels Y:

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{1}{|P_i|} \sum_{z_j \in P_i} \frac{z_i^\top z_j}{\tau} \overset{c}{\geq} \mathcal{H}(Z|Y).$$

where the symbol $\overset{c}{\geq}$ indicates "larger than" up to a multiplicative and/or an additive constant. When $Z|Y \sim \mathcal{N}(c_y, i)$, the bound is tight. As a result, minimizing the first term in Eq. (2) is equivalent to minimizing $\mathcal{H}(Z|Y)$:

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{1}{|P_i|} \sum_{z_j \in P_i} \frac{z_i^\top z_j}{\tau} \propto \mathcal{H}(Z|Y). \tag{3}$$

This concludes the proof for the relationship of the first term in Eq. (2).

We then analyze the second term in Eq. (2), which has the following relationship:

$$\frac{1}{n} \sum_{i=1}^{n} \log \sum_{z_k \in A_i} e^{(\frac{z_i^\top z_k}{\tau})}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{k:y_i=y_k} e^{(\frac{z_i^\top z_k}{\tau})} + \sum_{k:y_i \neq y_k} e^{(\frac{z_i^\top z_k}{\tau})} \right)$$

$$\geq \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{k:y_i \neq y_k} e^{(\frac{z_i^\top z_k}{\tau})} \right)$$

$$\overset{c}{\geq} \frac{1}{n} \sum_{i=1}^{n} \sum_{k:y_i \neq y_k} \frac{z_i^\top z_k}{\tau}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \frac{z_i^\top z_k}{\tau} - \frac{1}{n} \sum_{i=1}^{n} \sum_{k:y_i=y_k} \frac{z_i^\top z_k}{\tau}$$

$$\overset{c}{=} -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \|z_i - z_k\|^2 - \frac{1}{n} \sum_{i=1}^{n} \sum_{k:y_i=y_k} \frac{z_i^\top z_k}{\tau}, \tag{4}$$

where we use Jensen's inequality in the fourth line. The first term in Eq. (4) is close to the differential entropy estimator of features Z provided by [56]:

$$\hat{\mathcal{H}}(Z) = \frac{d}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{n} \log \|z_i - z_k\|^2 \overset{c}{=} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \log \|z_i - z_k\|^2 \propto \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \|z_i - z_k\|^2, \tag{5}$$

where d is the dimension of features. Combining Eq. (4) and Eq. (5) leads to:

$$\frac{1}{n} \sum_{i=1}^{n} \log \sum_{z_k \in A_i} e^{(\frac{z_i^\top z_k}{\tau})} \overset{c}{\geq} -\mathcal{H}(Z) - \frac{1}{n} \sum_{i=1}^{n} \sum_{k:y_i=y_k} \frac{z_i^\top z_k}{\tau}. \tag{6}$$

The second term in the right side of Eq. (6) is essentially a redundant term with the first term in Eq. (2), so we ignore it here. Then, we know that minimizing the second term in Eq. (2) is equivalent to maximizing $\mathcal{H}(Z)$:

$$\frac{1}{n} \sum_{i=1}^{n} \log \sum_{z_k \in A_i} e^{(\frac{z_i^\top z_k}{\tau})} \propto -\mathcal{H}(Z). \tag{7}$$

Combining Eq. (3) and Eq. (7), we conclude the proof of Theorem 1. $\square$

A.2 Proof for Theorem 2

Proof The mutual information between features Z and labels Y can be defined in two ways:

$$\mathcal{I}(Z;Y) = \mathcal{H}(Y) - \mathcal{H}(Y|Z) = \mathcal{H}(Z) - \mathcal{H}(Z|Y). \tag{8}$$

Based on Theorem 1, we know that:

$$\mathcal{L}_{con} \propto \mathcal{H}(Z|Y) - \mathcal{H}(Z) = -\mathcal{I}(Z;Y). \tag{9}$$

Combining Eq. (8) and Eq. (9), we have:

$$\mathcal{L}_{con} \propto \mathcal{H}(Y|Z) - \mathcal{H}(Y). \tag{10}$$

Then, we relate the conditional entropy $\mathcal{H}(Y|Z)$ to the cross entropy loss:

$$\mathcal{H}(Y;\hat{Y}|Z) = \mathcal{H}(Y|Z) + \mathcal{D}_{KL}(Y\|\hat{Y}|Z). \tag{11}$$

According to Eq. (11), when we minimize cross-entropy $\mathcal{H}(Y;\hat{Y}|Z)$, we implicitly minimize both $\mathcal{H}(Y|Z)$ and $\mathcal{D}_{KL}(Y\|\hat{Y}|Z)$. In fact, the optimization could be decoupled into 2 steps in a maximize-minimize (or bound-optimization) way [1]. The first step fixes the parameters of the network encoder and only minimizes Eq. (11) with respect to the parameters of the network classifier. As this step, $\mathcal{H}(Y|Z)$ is fixed and the predictions $\hat{Y}$ are adjusted to minimize $\mathcal{D}_{KL}(Y\|\hat{Y}|Z)$. Ideally, $\mathcal{D}_{KL}(Y\|\hat{Y}|Z)$ would vanish at the end of this step [1]. In this sense, we know that:

$$\mathcal{H}(Y|Z) = \inf \mathcal{H}(Y;\hat{Y}|Z). \tag{12}$$

The second step fixes the classifier and minimizes Eq. (11) with respect to the encoder. By combining Eq. (10) and Eq. (12), we conclude the proof of Theorem 2. $\square$

B Pseudo-code of Core-tuning

We summarize the scheme of Core-tuning in Algorithm 1. Here, all hard pair generation is conducted within each sample batch.

Algorithm 1 The training scheme of Core-tuning.

Require: Pre-trained encoder G_e; Loss factor η; Mixup factor α; Batch size B; Epoch number T.
Initialize: Classifier G_y; Projection head G_c.
  1: **for** t=1,...,T **do**
  2:    Sample a batch of training data $\{(x_i, y_i)\}_{i=1}^{B}$;
  3:    Obtain features $z_i = G_e(x_i)$ for each sample;
  4:    **for** i=1,...,B **do**
  5:      Construct positive pair set P_i and full pair set A_i for z_i;
  6:      Generate hard positive pair (z_i^+, y_i^+) and add it to P_i, A_i;
  7:      Generate hard negative pair (z_i^-, y_i^-) and add it to A_i;
  8:    **end for**
  9:    Obtain contrastive features $v_i = G_c(z_i)$ for all features;
10:    Compute the focal contrastive loss $\mathcal{L}_{con}^{f}$;
11:    Predict $\hat{y}_i = G_y(z_i)$ for the original and generated samples;
12:    Compute the cross-entropy loss $\mathcal{L}_{ce}^{m}$;
13:    loss.backward();   // loss $= \mathcal{L}_{ce}^{m} + \eta \mathcal{L}_{con}^{f}$.
14: **end for**

C More Experimental Details

C.1 Implementation Details of Feature Visualization

In the feature visualization, we train ResNet-18 on CIFAR10 with two kinds of losses, *i.e.,* (1) cross-entropy $\mathcal{L}_{ce}$; (2) cross-entropy and the contrastive loss $\mathcal{L}_{ce}+\mathcal{L}_{con}$. For better visualization, following [38], we add two fully connected layers before the classifier. The two layers first map the 512-dimensional features to a 3-dimensional feature sphere and then map back to the 10-dimensional feature space for prediction. The contrastive loss $\mathcal{L}_{con}$ is enforced on the 3-dimensional features. After training, we visualize the 3-dimensional features learned by ResNet-18 in MATLAB.

C.2 More Details of Image Classification

Dataset details. Following [28], we test on 9 natural image datasets, including ImageNet20 (a subset of ImageNet with 20 classes) [11], CIFAR10, CIFAR100 [30], Caltech-101 [15], DTD [10], FGVC Aircraft [41], Standard Cars [29], Oxford-IIIT Pets [46] and Oxford 102 Flowers [44]. In addition, considering real-world datasets may be class-imbalanced [72, 73, 75, 77], we also evaluate Core-tuning on the iNaturalist18 dataset [52]. Most datasets are obtained from their official websites, except ImageNet20 and Oxford 102 Flowers. The ImageNet20 dataset is obtained by combining two open-source ImageNet subsets with 10 classes, *i.e.,* ImaegNette and ImageWoof [22]. Moreover, Oxford 102 Flowers is obtained from Kaggle[5]. These datasets cover a wide range of classification tasks, including coarse-grained object classification (*i.e.,* ImageNet20, CIFAR, Caltech-101), fine-grained object classification (*i.e.,* Cars, Aircraft, Pets) and texture classification (*i.e.,* DTD). The statistics of all datasets are reported in Table 10.

Table 10: Statistics of datasets.

DataSet	#Classes	# Training	# Test
ImageNet20 [22, 11]	20	18,494	7,854
CIFAR10 [30]	10	50,000	10,000
CIFAR100 [30]	100	50,000	10,000
Caltech-101 [15]	102	3,060	6,084
Describable Textures (DTD) [10]	47	3,760	1,880
FGVG Aircraft [41]	100	6,667	3,333
Standard Cars [29]	196	8,144	8,041
Oxford-IIIT Pets [46]	37	3,680	3,369
Oxford 102 Flowers [44]	102	6,552	818
iNaturalist18 [52]	8,142	437,513	24,426

Implementation details. We implement all methods in PyTorch. All checkpoints of self-supervised models are provided by the authors or by the PyContrast GitHub repository[6]. For most datasets, following [6, 28], we preprocess images via random resized crops to 224×224 and flips. At the test time, we resize images to 256×256 and then take a 224×224 center crop. In such a setting, however, we find it difficult to reproduce the performance of some CSL models [6]. Therefore, for some datasets (*e.g.,* CIFAR10 and Aircraft), we resize images to different scales and use rotation augmentations. Although the preprocessing of some datasets is slightly different from [6], the results in this paper are obtained with the same preprocessing method *w.r.t.* each dataset and thus are fair.

Following [28], we initialize networks with the checkpoints of contrastive self-supervised models. For most datasets, we fine-tune networks for 100 epochs using Nesterov momentum via the cosine learning rate schedule. For ImageNet20, we fine-tune networks using stochastic gradient descent via the linear learning rate decay. For iNaturalist18, we fine-tune networks for 160 epochs. For all datasets, the momentum parameter is set to 0.9, while the factor of weight decay is set to 10^{-4}. As for Core-tuning, we set the clipping thresholds of hard negative generation to be λ_n=0.8 and the temperature τ=0.07. The dimension of the contrastive features is 256 and the depth of non-linear projection is 2 layers. Following [6], we perform hyper-parameter tuning for each dataset. Specifically, we select the batch size from $\{64, 128, 256\}$, the initial learning rate from $\{0.01, 0.1\}$ and η/α from $\{0.1, 1, 10\}$. The experiments are conducted on 4 TITAN RTX 2080 GPUs for iNaturalist18, and 1 GPU for all other datasets. All results are averaged over 3 runs. We adopt the top-1 accuracy as the metric. The statistics of the used hyper-parameters are provided in Table 11. For other baselines, we use the same training setting for each dataset, and tune their hyper-parameters as best as possible.

---

[5]https://www.kaggle.com/c/oxford-102-flower-pytorch.
[6]https://github.com/HobbitLong/PyContrast.

Table 11: Statistics of the used hyper-parameters in Core-tuning.

Hyper-parameter	ImageNet20	CIFAR10	CIFAR100	Caltech101	DTD	Aircraft	Cars	Pets	Flowers	iNarutalist18
epochs	100									160
batch size	256	256	256	256	256	64	64	64	64	128
loss trade-off factor η	0.1	0.1	1	1	0.1	0.1	0.1	0.1	1	10
mixup factor α	1	1	0.1	0.1	1	0.1	0.1	1	0.1	1
learning rate (lr)	0.1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.1
lr schedule	linear	cosine decay								
temperature τ	0.07									
threshold λ_n	0.8									
weight decay factor	10^{-4}									
momentum factor	0.9									
projection dimension	256									
projection depth	2 layers									

C.3 More Details of Domain Generalization

Dataset details. We use 3 benchmark datasets, *i.e.,* PACS [33], VLCS [14] and Office-Home [53]. The data statistics are shown in Table 12, where each dataset has 4 domains. In each setting, we select 3 domains to fine-tune the networks and then test on the rest of the unseen domains. The key challenge is the distribution discrepancies among domains, leading to poor performance of neural networks on the target domain [71, 74].

Table 12: Statistics of datasets.

DataSet	#Domains	#Classes	#Samples	Size of images
PACS	4	7	9,991	(3,224,224)
VLCS	4	5	10,729	(3,224,224)
Office-Home	4	65	15,588	(3,224,224)

Implementation details. The overall scheme of Core-tuning for domain generalization is shown in Figure 3. The experiments are implemented based on the DomainBed repository [18] in PyTorch. During fine-tuning, we preprocess images through random resized crops to 224×224, horizon flips, color jitter and random gray scale. At the test time, we directly resize images to 224×224. We initialize ResNet-50 with the weights of the MoCo-v2 pre-trained model, and fine-tune it for 20,000 steps at a batch size of 32 using the Adam optimizer on a single TITAN RTX 2080 GPU. We set the initial learning rate to 5×10^{-5} and adjust it via the exponential learning rate decay. All other hyper-parameters of Core-tuning are the same as image classification. Besides, we use Accuracy as the metric in domain generalization.
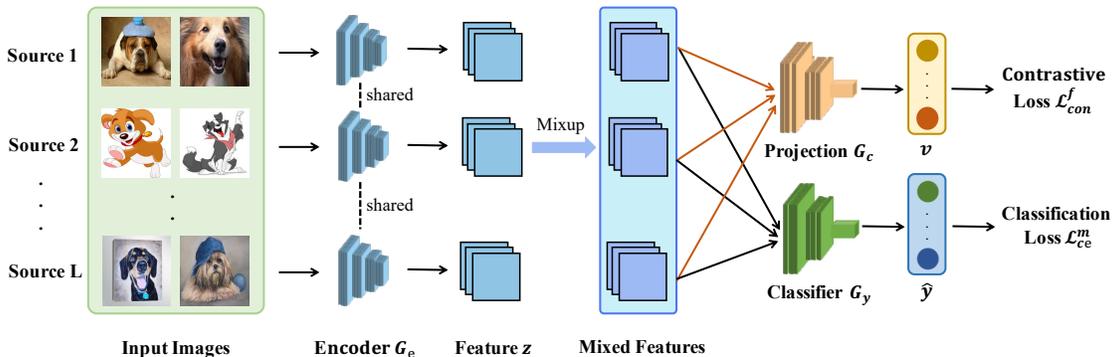

Figure 3: The overall scheme of Core-tuning in the setting of cross-domain generalization.

C.4 Implementation Details of Robustness Training

We conduct this experiment in PyTorch. We take Caltech-101, DTD, Pets, and CIFAR10 as datasets, whose preprocessing are the same as the ones in image classification. We use MoCo-v2 pre-trained ResNet-50 as the backbone, and use Projected Gradient Descent (PGD) [40] to generate adversarial samples. During adversarial training (AT), we use both clean and adversarial samples for training with various fine-tuning methods on a single TITAN RTX 2080 GPU. Other training schemes (*e.g.,* the optimizer, the hyper-parameters, the learning rate scheme) are the same as image classification.

D More Experimental Results

D.1 More Results on Domain Generalization

This appendix further reports the results of domain generalization on OfficeHome. The observations from Table 13 are same to the main text. First, when fine-tuning with cross-entropy, the contrastive self-supervised model performs worse than the supervised pre-trained model. This results from the relatively worse discriminative abilities of the contrastive self-supervised model, which can also be found in Table 1. Second, enforcing contrastive regularizer during fine-tuning improves domain generalization performance, since the contrastive regularizer helps to learn more discriminative features (cf. Theorem 1) and also helps to alleviate distribution shifts among domains [25], hence leading to better performance. Last, Core-tuning further improves the generalization performance of models on all datasets. This is because hard pair generation further boosts contrastive learning, while smooth classifier learning also benefits model generalizability. We thus conclude that Core-tuning improves model generalization on downstream tasks.

Table 13: Domain generalization accuracies of various fine-tuning methods for MoCo-v2 pre-trained ResNet-50 the on Office-Home dataset. CE means cross-entropy; CE-Con enhances CE with the contrastive loss. Here, A/C/P/R are four domains in Office-Home.

Pre-training	Fine-tuning	Office-Home				
		A	C	P	R	Avg.
Supervised	CE	56.08	50.83	72.49	75.21	63.82
MoCo-v2	CE	50.31	48.91	64.72	68.76	58.18
	CE-Con	55.87	50.23	71.51	74.99	63.15
	ours	**58.70**	**52.43**	**72.89**	**75.36**	**64.85**

D.2 More Results on Adversarial Training

In the main paper, we apply Core-tuning to adversarial training on CIFAR10, while this appendix further provides the results of adversarial training on three other natural image datasets, *i.e.,* Caltech-101, DTD and Pets. We draw several observations based on the results on 3 image datasets in Table 14. First, despite good clean accuracy, standard fine-tuning with cross-entropy cannot defend against adversarial attack, leading to poor robust accuracy. Second, AT with cross-entropy improves the robust accuracy significantly, but it inevitably degrades the clean accuracy due to the accuracy-robustness trade-off [51]. In contrast, the contrastive regularizer improves both robust and clean accuracies. This is because contrastive learning helps to improve robustness generalization (*i.e.,* alleviating the distribution shifts between clean samples and adversarial samples), thus leading to better performance. Last, Core-tuning further boosts AT and, surprisingly, even achieves better clean accuracy than the standard fine-tuning under the ℓ_2 attack. To our knowledge, this is quite promising since even the most advanced AT methods [65, 69] find it difficult to conquer the accuracy-robustness trade-off [67]. The improvement is mainly derived from that both contrastive learning and smooth classifier learning boost the robustness generalization. We thus conclude that Core-tuning is beneficial to model robustness. We also hope that Core-tuning can motivate people to rethink the accuracy-robustness trade-off in adversarial training in the future.

Table 14: Adversarial training performance of MoCo-v2 pre-trained ResNet-50 under the attack of PGD-10 in terms of robust and clean accuracies. CE indicates cross-entropy; AT-CE indicates adversarial training (AT) with CE; AT-CE-Con enhances AT-CE with the contrastive loss; AT-ours uses Core-tuning for AT.

Method	PGD - ℓ_2 attack ($\epsilon = 0.5$)						PGD - ℓ_∞ attack ($\epsilon = 4/255$)					
	Caltech101		DTD		Pets		Caltech101		DTD		Pets	
	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean
CE	55.69	91.87	42.25	71.68	30.94	89.05	27.03	**91.87**	18.37	**71.68**	4.63	**89.05**
AT-CE	87.35	91.61	61.93	68.81	78.67	86.25	78.61	90.65	47.27	67.13	63.59	84.21
AT-CE-Con	88.67	92.61	64.75	71.24	79.53	87.01	79.87	91.08	48.95	69.07	65.60	86.85
AT-ours	**89.21**	**92.83**	**66.49**	**72.94**	**82.54**	**89.22**	**80.73**	91.64	**49.43**	70.65	**67.98**	87.20

D.3   More Results on Image Classification

The results with standard errors. In the main paper, we report the results of image classification and ablations studies on 9 natural image datasets in terms of the average accuracy. To make the results more complete, this appendix further reports the results with their standard errors (cf. Tables 15-16).

Table 15: Comparisons of various fine-tuning methods for MoCo-v2 pre-trained ResNet-50 on image classification in terms of top-1 accuracy. Here, "Avg." indicates the average accuracy over 9 datasets. SL-CE-tuning denotes supervised pre-training on ImageNet and then fine-tuning with cross-entropy.

Algorithm	ImageNet20	CIFAR10	CIFAR100	Caltech101	DTD
SL-CE-tuning	91.01+/-1.27	94.23+/-0.07	83.40+/-0.12	93.65+/-0.21	74.40+/-0.45
CE-tuning	88.28+/-0.47	94.70+/-0.39	80.27+/-0.60	91.87+/-0.18	71.68+/-0.53
L2SP [36]	88.49+/-0.40	95.14+/-0.22	81.43+/-0.22	91.98+/-0.07	72.18+/-0.61
M&M [66]	88.53+/-0.21	95.02+/-0.07	80.58+/-0.19	92.91+/-0.08	72.43+/-0.43
DELTA [34]	88.35+/-0.41	94.76+/-0.05	80.39+/-0.41	92.19+/-0.45	72.23+/-0.23
BSS [9]	88.34+/-0.62	94.84+/-0.21	80.40+/-0.30	91.95+/-0.12	72.22+/-0.17
RIFLE [35]	89.06+/-0.28	94.71+/-0.13	80.36+/-0.07	91.94+/-0.23	72.45+/-0.30
SCL [19]	89.29+/-0.07	95.33+/-0.09	81.49+/-0.27	92.84+/-0.03	72.73+/-0.31
Bi-tuning [78]	89.06+/-0.08	95.12+/-0.15	81.42+/-0.01	92.83+/-0.06	73.53+/-0.37
Core-tuning	**92.73+/-0.17**	**97.31+/-0.10**	**84.13+/-0.27**	**93.46+/-0.06**	**75.37+/-0.37**

Algorithm	Aircraft	Cars	Pets	Flowers	Avg.
SL-CE-tuning	87.03+/-0.02	89.77+/-0.11	92.17+/-0.12	98.78+/-0.10	89.35
CE-tuning	86.87+/-0.18	88.61+/-0.43	89.05+/-0.01	98.49+/-0.06	87.76
L2SP [36]	86.55+/-0.30	89.00+/-0.23	89.43+/-0.27	98.66+/-0.20	88.10
M&M [66]	87.45+/-0.28	88.90+/-0.70	89.60+/-0.09	98.57+/-0.15	88.22
DELTA [34]	87.05+/-0.37	88.73+/-0.05	89.54+/-0.48	98.65+/-0.17	87.99
BSS [9]	87.18+/-0.71	88.50+/-0.02	89.50+/-0.42	98.57+/-0.15	87.94
RIFLE [35]	87.60+/-0.50	89.72+/-0.11	90.05+/-0.26	98.70+/-0.06	88.29
SCL [19]	87.44+/-0.31	89.37+/-0.13	89.71+/-0.20	98.65+/-0.10	88.54
Bi-tuning [78]	87.39+/-0.01	89.41+/-0.28	89.90+/-0.06	98.57+/-0.10	88.58
Core-tuning	**89.48+/-0.17**	**90.17+/-0.03**	**92.36+/-0.14**	**99.18+/-0.15**	**90.47**

Table 16: Ablation studies of Core-tuning (Row 5) for fine-tuning MoCo-v2 pre-trained ResNet-50 on 9 natural image datasets in terms of top-1 accuracy. Here, "Avg." indicates the average accuracy over the 9 datasets. Besides, $\mathcal{L}_{con}$ is the original supervised contrastive loss, while $\mathcal{L}_{con}^f$ is our focal contrastive loss. Moreover, "mix" denotes the manifold mix, while "mix-H" indicates the proposed hardness-directed mixup strategy in our method.

$\mathcal{L}_{ce}$	$\mathcal{L}_{con}$	$\mathcal{L}_{con}^f$	mix	mix-H	ImageNet20	CIFAR10	CIFAR100	Caltech101	DTD
√					88.28+/-0.47	94.70+/-0.39	80.27+/-0.60	91.87+/-0.18	71.68+/-0.53
√	√				89.29+/-0.07	95.33+/-0.09	81.49+/-0.27	92.84+/-0.03	72.73+/-0.31
√		√			90.67+/-0.09	95.43+/-0.06	81.03+/-0.11	92.68+/-0.06	73.31+/-0.40
√	√			√	92.20+/-0.15	97.01+/-0.10	83.89+/-0.20	93.22+/-0.18	74.78+/-0.31
√		√		√	**92.73+/-0.17**	**97.31+/-0.10**	**84.13+/-0.27**	**93.46+/-0.06**	**75.37+/-0.37**

$\mathcal{L}_{ce}$	$\mathcal{L}_{con}$	$\mathcal{L}_{con}^f$	mix	mix-H	Aircraft	Cars	Pets	Flowers	Avg.
√					86.87+/-0.18	88.61+/-0.43	89.05+/-0.01	98.49+/-0.06	87.76
√	√				87.44+/-0.31	89.37+/-0.13	89.71+/-0.20	98.65+/-0.10	88.54
√		√			88.37+/-0.14	89.06+/-0.14	91.37+/-0.03	98.74+/-0.11	88.96
√	√			√	88.88+/-0.34	89.79+/-0.12	91.95+/-0.33	98.94+/-0.12	90.07
√		√		√	**89.48+/-0.17**	**90.17+/-0.03**	**92.36+/-0.14**	**99.18+/-0.15**	**90.47**

The fine-tuning results on ImageNet. Since ImageNet has rich labeled samples for fine-tuning and the CSL models are also pre-trained on ImageNet, the performance gain of different fine-tuning methods may not vary as significantly as on the small-scale target datasets. Even so, the results in Table 17 also demonstrate the effectiveness of Core-tuning on very large-scale data.

Table 17: Fine-tuning results of the MoCo-v2 ResNet-50 fine-tuned by various methods, on ImageNet.

Pre-training	Fine-tuning	Top-1 accuracy
MoCo-v2 [8]	CE-tuning	76.82
MoCo-v2 [8]	CE-Contrastive-tuning	77.13
MoCo-v2 [8]	Core-tuning (ours)	**77.43**

More results on different pre-training methods. This appendix provides the fine-tuning results of Core-tuning for the SimCLR pre-trained models. Since the official checkpoints of SimCLR-v1 [6] and SimCLR-v2 [7] are based on Tensorflow, we convert them to the PyTorch and try to reproduce cross-entropy tuning (CE-tuning) in our experimental settings. Note that although the reproduction performance of CE-tuning is slightly worse than the original paper [6, 7], the results in this paper are obtained with the same preprocessing method *w.r.t.* each dataset and thus are fair. As shown in Table 18, Core-tuning consistently outperforms CE-tuning for SimCLR pre-trained models.

Table 18: Fine-tuning results of ResNet-50, pre-trained by various methods.

Pre-training	Caltech101		DTD		Pets	
	CE-tuning	ours	CE-tuning	ours	CE-tuning	ours
SimCLR-v1 [6]	90.53+/-0.06	**92.40+/-0.06**	90.53+/-0.06	**71.26+/-0.05**	89.34+/-0.46	**90.89+/-0.09**
SimCLR-v2 [7]	92.44+/-0.18	**93.46+/-0.02**	71.26+/-0.26	**74.75+/-0.41**	88.28+/-0.26	**90.64+/-0.31**

The results on linear evaluation. This appendix provides linear evaluation for Core-tuning. Specifically, we first fine-tune the MoCo-v2 pre-trained ResNet-50 with Core-tuning and then train a linear classifier for prediction. As shown in Table 19, Core-tuning performs better than CE-tuning.

Table 19: Results of linear evaluation for the ResNet-50 fine-tuned by various methods, on CIFAR10.

Pre-training	Fine-tuning	Top-1 accuracy
MoCo-v2 [8]	CE-tuning	94.78+/-0.28
MoCo-v2 [8]	Core-tuning (ours)	**97.09+/-0.14**

The results on KNN evaluation. This appendix provides the KNN evaluation for Core-tuning. To be specific, we first fine-tune the MoCo-v2 pre-trained ResNet-50 with Core-tuning and then use KNN for prediction. As shown in Table 20, Core-tuning also outperforms CE-tuning.

Table 20: Results of KNN evaluation for the ResNet-50 fine-tuned by various methods, on CIFAR10.

Pre-training	Fine-tuning	Top-1 accuracy
MoCo-v2 [8]	CE-tuning	94.63+/-0.32
MoCo-v2 [8]	Core-tuning (ours)	**96.65+/-0.06**

D.4 The Results with Standard Errors on Semantic Segmentation

In the main paper, we report the average results of semantic segmentation on PASCAL VOC. This appendix further reports the results with their standard errors (cf. Table 21).

Table 21: Fine-tuning performance on PASCAL VOC semantic segmentation based on DeepLab-V3 with ResNet-50, pre-trained by various CSL methods. CE indicates cross-entropy.

Pre-training	Fine-tuning	MPA	FWIoU	MIoU
Supervised	CE	87.10+/-0.20	89.12+/-0.17	76.52+/-0.34
InsDis [62]	CE	83.64+/-0.12	88.23+/-0.08	74.14+/-0.21
	ours	**84.53+/-0.31**	**88.67+/-0.07**	**74.81+/-0.13**
PIRL [43]	CE	83.16+/-0.26	88.22+/-0.24	73.99+/-0.42
	ours	**85.30+/-0.24**	**88.95+/-0.08**	**75.49+/-0.36**
MoCo-v1 [21]	CE	84.71+/-0.56	88.75+/-0.04	74.94+/-0.12
	ours	**85.70+/-0.32**	**89.19+/-0.02**	**75.94+/-0.23**
MoCo-v2 [8]	CE	87.31+/-0.31	90.26+/-0.12	78.42+/-0.28
	ours	**88.76+/-0.34**	**90.75+/-0.04**	**79.62+/-0.12**
SimCLR-v2 [7]	CE	87.37+/-0.48	90.27+/-0.12	78.16+/-0.19
	ours	**87.95+/-0.20**	**90.71+/-0.13**	**79.15+/-0.33**
InfoMin [49]	CE	87.17+/-0.20	89.84+/-0.09	77.84+/-0.24
	ours	**88.92+/-0.36**	**90.65+/-0.09**	**79.48+/-0.30**

E   More Analysis of Core-tuning

E.1   Analysis of Projection Dimension and Depth

In previous experiments, we use a 2-layer MLP to extract contrastive features with dimension 256. Here, we further analyze how the dimension and the depth influence Core-tuning. The results on ImageNet20 are reported in Figure 4, where the fine-tuning performance of Core-tuning can be further improved by changing the feature dimension to 128 and the depth to 3. Note that the best dimension and depth of the projection head may vary on different datasets, but the default setting (*i.e.,* dimension 256 and depth 2) is enough to obtain consistently good performance.
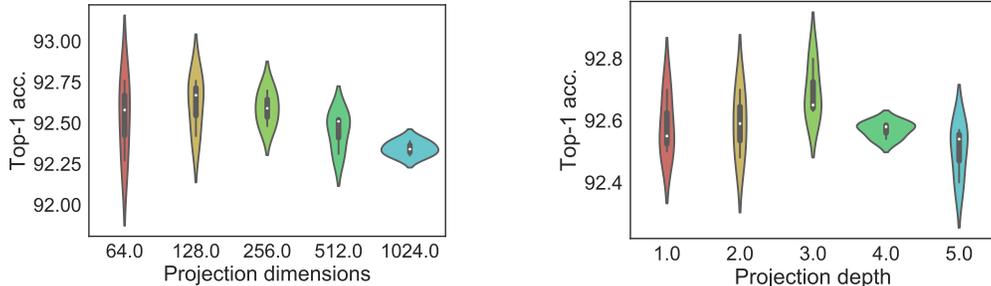

Figure 4: Analysis of the projection dimension and the projection depth in Core-tuning on ImageNet20 based on MoCo-v2 pre-trained ResNet-50. Each run tests one parameter and fixes others. Best viewed in color.

E.2   Analysis of Loss and Mixup Hyper-Parameters

This appendix discusses the influence of the loss trade-off parameter η and the mixup sampling factor α on Core-tuning based on the ImageNet20 dataset. Each run tests one parameter and fixes others. As shown in Figure 5, when η=0.1 and α=1, Core-tuning performs slightly better on ImageNet20. Note that the best η and α can be different on diverse datasets.

E.3   Analysis of Temperature Factor

Following the implementation of the supervised contrastive loss [26], we set the temperature factor τ to 0.07 for Core-tuning by default. In this section, we further analyze the influence of τ on Core-tuning when fine-tuning MoCo-v2 pre-trained models on ImageNet20. As shown in Figure 5, when τ is small (*e.g.,* 0.01 or 0.07), Core-tuning performs slightly better on ImageNet20. The potential reason is that a small temperature parameter implicitly helps the method to learn hard positive/negative pairs [55], which are more informative and beneficial to contrastive learning. Note that the best τ can be different on different datasets, but the default setting (*i.e.,* $\tau = 0.07$) is enough to achieve comparable performance.
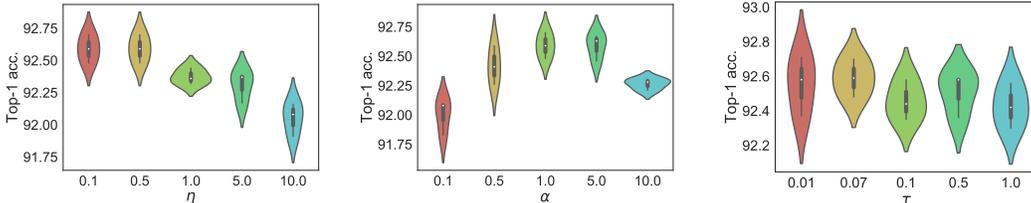

Figure 5: Analysis of η, α and the temperature factor in Core-tuning on ImageNet20 based on MoCo-v2 pre-trained ResNet-50. Each run tests one factor and fixes others. Best viewed in color.

E.4 Analysis of Hard Pair Thresholds

In our hardness-directed mixup strategy, to make the generated negative pairs closer to negative pairs, we clip $\lambda \sim \text{Beta}(\alpha, \alpha)$ by $\lambda \geq \lambda_n$ when generating hard negative pairs. In our experiments, we set the threshold $\lambda_n = 0.8$. In this appendix, we analyze the influences of the negative pair threshold λ_n. Meanwhile, although we do not constrain hard positive generation, we also analyze the potential positive pair threshold λ_p. The results on ImageNet20 are reported in Table 5. On the one hand, λ_n satisfies our expectation that the generated hard negative pairs should be closer to negatives, *i.e.,* a larger λ_n can lead to better performance. On the other hand, we find when no crop is conducted for hard positive generation (*i.e.,* $\lambda_p=0$), the performance is slightly better. We conjecture that since the generated hard positives are located in the borderline area between positives and negatives, allowing the generated hard positives to close to negatives may have a margin effect on contrastive learning and thus boosts performance. Despite this, Core-tuning with a large λ_p performs similarly well.

Table 22: Threshold analysis for hard pair generation in Core-tuning on ImageNet20 based on MoCo-v2 pre-trained ResNet-50. Each run tests one parameter and fixes another one to 0.8.

Thresholds	0	0.2	0.4	0.6	0.8
Negative pair threshold λ_n	91.55	91.94	92.19	92.36	92.59
Positive pair threshold λ_p	92.73	92.68	92.64	92.60	92.59

E.5 Relationship Between Pre-Training and Fine-Tuning Accuracies

We further explore the relationship between ImageNet performance and Core-tuning fine-tuning performance on Caltech-101 for various contrastive self-supervised models. Here, the ImageNet performance of a contrastive self-supervised model is obtained by training a new linear classifier on the frozen pre-trained representation and then evaluate the model on the ImageNet test set. For convenience, we directly follow the ImageNet performance reported in the original paper of the corresponding methods. As shown in Figure 6, the fine-tuning result of each contrastive self-supervised model on Caltech-101 is highly correlated with the model result on ImageNet. This implies that the ImageNet performance can be a good predictor for the fine-tuning performance of contrastive self-supervised models. Such a finding is consistent with supervised pre-trained models [28]. Even so, note that the correlation is not perfect, where a contrastive pre-trained model with better ImageNet performance does not necessarily mean better fine-tuning performance, *e.g.,* SimCLR-v2 vs MoCo-v2.
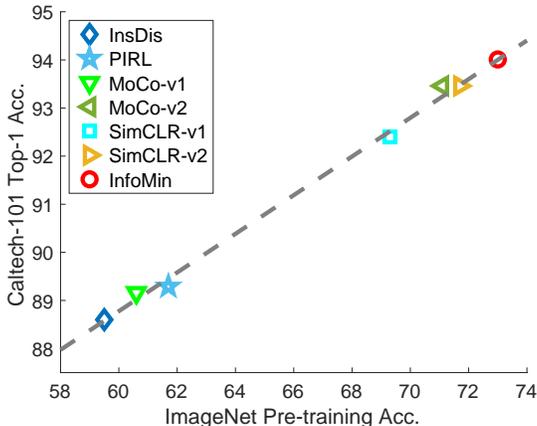

Figure 6: The relationship between ImageNet performance and Core-tuning fine-tuning performance on Caltech-101 for contrastive self-supervised ResNet-50 models. Better viewed in color.

E.6 Effectiveness of Hard Pair Generation for Contrastive Fine-Tuning

In our proposed Core-tuning, we use all the generated positive sample pairs and the original samples as positive pairs for contrastive fine-tuning. In this appendix, to better evaluate the effectiveness of hard pair generation, we do not use original data as positive pairs but only use the generated hard positive pairs for contrastive learning. As shown in Table 23, only using the generated hard positive pairs for contrastive learning is enough to obtain comparable performance. Such results further verify the effectiveness of our hardness-directed mixup strategy as well as the importance of hard positive pairs for contrastive fine-tuning.

Table 23: Comparisons with only using the generating hard positive pairs for contrast on CIFAR10.

Pre-training	Fine-tuning	The used positive pairs for contrast?	Top-1 accuracy
MoCo-v2 [8]	CE-tuning	$\times$	94.70+/-0.39
MoCo-v2 [8]	Core-tuning	only the generated hard positive pairs	97.31+/-0.09
MoCo-v2 [8]	Core-tuning	all positive pairs	97.31+/-0.10

E.7 Effectiveness of Smooth Classifier Learning

In Core-tuning, to better exploit the learned discriminative feature space by contrastive fine-tuning, we use the mixed samples for classifier training, so that the classifier can be more smooth and far away from the original training data. In this appendix, to better evaluate the effectiveness of smooth classifier learning, we compare Core-tuning with a variant that does not use the mixed data for classifier learning. As shown in Table 24, smooth classifier learning contributes to the fine-tuning performance of contrastive self-supervised models on downstream tasks. The results demonstrate the effectiveness of smooth classifier learning and also show its importance in Core-tuning.

Table 24: Influence of smooth classifier learning on CIFAR10.

Pre-training	Fine-tuning	Smooth classifier learning?	Top-1 accuracy
MoCo-v2 [8]	CE-tuning	$\times$	94.70+/-0.39
MoCo-v2 [8]	CE-tuning	$\sqrt{}$	95.43+/-0.20
MoCo-v2 [8]	Core-tuning	$\times$	96.13+/-0.11
MoCo-v2 [8]	Core-tuning	$\sqrt{}$	97.31+/-0.10