RÉNYI SUPERVISED CONTRASTIVE LEARNING FOR TRANSFERABLE REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A mighty goal of representation learning is to train a feature that can transfer to various tasks or datasets. A conventional approach is to pre-train a neural network on a large-scale labeled dataset, e.g., ImageNet, and use its feature for downstream tasks. However, the feature often lacks transferability due to the class-collapse issue; existing supervised losses (such as cross-entropy) restrain the intra-class variation and limit the capability of learning rich representations. This issue becomes more severe when pre-training datasets are class-imbalanced or coarse-labeled. To address the problem, we propose a new representation learning method, named Rényi supervised contrastive learning (RényiSCL), which can effectively learn transferable representation using a labeled dataset. Our main idea is to use the recently proposed self-supervised Rényi contrastive learning in the supervised setup. We show that RényiSCL can mitigate the class-collapse problem by contrasting features with both instance-wise and class-wise information. Through experiments on the ImageNet dataset, we show that RényiSCL outperforms all supervised and self-supervised methods under various transfer learning tasks. In particular, we also validate the effectiveness of RényiSCL under class-imbalanced or coarse-labeled datasets.

1 INTRODUCTION

Deep neural networks' essential and unique property is that they can transfer to other networks of different tasks or datasets. Thus, it has been a common practice to pre-train a deep neural network on a large-scale dataset such as ImageNet (Deng et al., 2009) and transfer it to various downstream tasks (Kornblith et al., 2019; Huh et al., 2016). For example, the ImageNet pre-trained networks have been widely used for fine-grained classification (Kornblith et al., 2019), few-shot learning (Guo et al., 2020), object detection (Huang et al., 2017), or semantic segmentation (He et al., 2017; Chen et al., 2017). Inspired by the success of ImageNet, such pre-training, then transfer learning strategy has also been extensively studied in various other domains such as natural language processing (Sarzynska-Wawer et al., 2021; Devlin et al., 2018; Brown et al., 2020), speech processing (Oord et al., 2018; Schneider et al., 2019; Hsu et al., 2021) and multimodal representation learning (Zhang et al., 2020; Radford et al., 2021; Xu et al., 2021).

A straightforward yet effective way to pre-train a neural network (for transfer learning) is to train a classifier with the standard cross-entropy loss; Kornblith et al. (2019) empirically observed a strong correlation between the ImageNet classification accuracy and the transfer learning performance. On the other hand, Kornblith et al. (2021) claimed that the tactics to improve the classification accuracy can worsen the transfer learning performance, showing the trade-off between generalization and transferability. Meanwhile, several recent works (Ericsson et al., 2021; Sariyildiz et al., 2021) evidenced that supervised methods often lag behind unsupervised or self-supervised methods (Grill et al., 2020; Chen et al., 2021; Caron et al., 2021) for transfer learning. The inferior transfer learning performance of supervised models is often attributed to the limited intra-class variation, often referred to as class-collapse issue (Graf et al., 2021), i.e., the features in the same class concentrate around a single prototypical vector. Graf et al. (2021) showed that the class-collapse issue appears in both cross-entropy and supervised contrastive loss; even though those methods attain high accuracy for the pre-training task, the learned representation might be sub-optimal for transferring to other downstream tasks (e.g., see Table 1).

Contribution. This paper proposes Rényi supervised contrastive learning (RényiSCL), a simple yet effective method to obtain a more transferable representation by mitigating the class-collapse issue. Our approach generalizes the recently proposed Rényi self-supervised contrastive learning (Lee & Shin, 2022) to the supervised case. We rigorously analyze how RényiSCL improves the transferability of representations and balance between generalization and transferability trade-off. In particular, we show that RényiSCL alleviates the class-collapse issue and improves transferability by performing easy positive mining on the intra-class samples: it imposes instance-wise weight on positives rather than following the class prototypes. Moreover, we show that RényiSCL performs hard negative mining on the inter-class samples, increasing the class separability.

Through experiments on ImageNet, we show that RényiSCL outperforms other (supervised or self-supervised) representation learning methods in various transfer learning tasks such as fine-grained object classification and cross-domain few-shot learning. In particular, we empirically find that RényiSCL benefits from sophisticated data augmentations such as multi-crop data augmentation (Caron et al., 2020), typically studied for self-supervised learning (and rarely used for supervised learning). In addition, we demonstrate the effectiveness of RényiSCL on the imbalanced datasets and provide a simple yet effective method to improve the transferability of representation and generalization on minor class samples. Finally, we demonstrate the effectiveness of RényiSCL in coarse-to-fine transfer learning compared to existing supervised baselines.

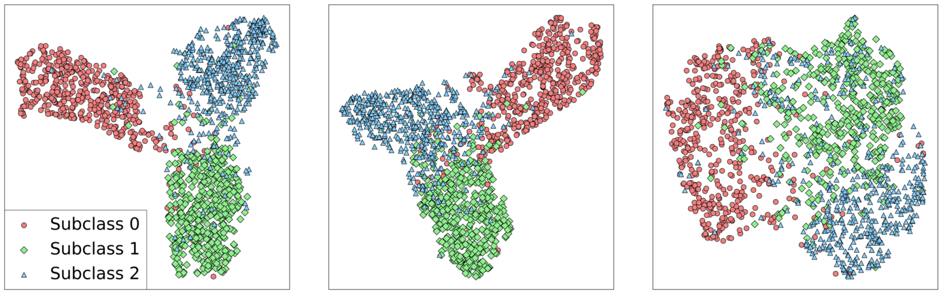
2 Related works

Transfer learning. Pre-training a model on a large-scale dataset such as ImageNet, and transferring to downstream tasks is a classical approach in deep learning. A straightforward approach is to train a classifier with cross-entropy loss and use its feature extractor for transfer learning. Kornblith et al. (2019) provided empirical evidence of a strong correlation between ImageNet accuracy and transfer learning performance. However, they also showed that subtle techniques such as regularizations or minor changes in cross-entropy loss used to improve the ImageNet classification accuracy could lead to inferior transfer learning performance (Kornblith et al., 2019; 2021). On the other hand, Salman et al. (2020) showed that adversarially trained models transfer better than standard cross-entropy-based models, even though they offer inferior performance on ImageNet validation accuracy. Recently, a series of works demonstrated the effectiveness of self-supervised learning (Chen et al., 2020b;a; Grill et al., 2020; Caron et al., 2020; 2021; Dwibedi et al., 2021; Lee & Shin, 2022) in learning transferable representations, outperforming supervised baselines. Inspired by the success of self-supervised methods, many works proposed to improve the transferability of supervised models (Zhao et al., 2020; Feng et al., 2021; Wang et al., 2022). Zhao et al. (2020) presented supervised learning based on exemplar SVM (Malisiewicz et al., 2011) by leveraging the techniques from MoCo (Chen et al., 2020a). Feng et al. (2021) proposed a new supervised pre-training method using a k-nearest neighborhood instead of a prototypical layer as in the usual cross-entropy-based approach. Wang et al. (2022) showed that inserting a projection MLP to the standard cross-entropy loss can further increase the transferability, reducing the gap between self-supervised methods.

Supervised contrastive learning. Supervised contrastive learning (Khosla et al., 2020) is a generalization of self-supervised contrastive learning to the supervised setup. While self-supervised learning contrasts between instances, supervised contrastive learning contrasts between class samples, thus showing outstanding generalization performance (Graf et al., 2021). Also, recent works demonstrated the effectiveness of supervised contrastive learning in long-tailed recognition problems (Cui et al., 2021; Kang et al., 2021; Li et al., 2022). In particular, Kang et al. (2021) proposed balanced supervised contrastive learning that improves the generalization and downstream transfer learning performance on the imbalanced dataset.

Despite its decent generalization performance, recent works noticed the low transferability of supervised contrastive learning models (Islam et al., 2021; Chen et al., 2022). Thus, Islam et al. (2021) proposed to add self-supervised loss to increase the intra-class variation and improve the transfer learning performance on various object classification and few-shot learning tasks. On the other hand, Chen et al. (2022) proposed to use additional class conditional contrastive loss with an autoencoder to improve the coarse-to-fine transfer learning performance. However, those methods often require sensitive hyperparameter search or computational burden, which limits the applicability to a large-scale dataset.

Figure 1: UMAP visualization of coares-to-fine transfer learning with RényiSCL, SupCon (Khosla et al., 2020), and SLMLP (Wang et al., 2022). We visualize the features of a superclass that contains three subclasses of TinyImageNet dataset. See Section 5.3 and Appendix B.3 for details.



(a) RényiSCL (b) SupCon (b) SLMLP

3 Preliminaries

Notation. Let $x \in \mathcal{X}$ be a data point drawn from data distribution $p(x)$ over $\mathcal{X}$. Our goal is to train an encoder $g : \mathcal{X} \to \mathbb{R}^d$ (e.g. ResNet50 (He et al., 2016)) that maps input x into a compact feature space, and use g as a feature extractor for various downstream transfer learning tasks. For notational simplicity, we denote $z = g(x) \in \mathbb{R}^d$ be a feature from input x. Assume we have discrete label space $\mathcal{Y}$ consisting of C classes, and let $L : \mathcal{X} \to \mathcal{Y}$ be the labeling function that assigns data into the ground-truth class. Denote $L(z) = L(x)$ for notational simplicity.

3.1 Supervised learning

Supervised learning with cross-entropy loss. The most common approach in supervised learning is to jointly optimize g and a linear weight $W \in \mathbb{R}^{d\times K}$ by using a cross-entropy loss defined by

$$\ell_{\texttt{CE}}(z, y) = -\log \frac{\exp\left(z^\top w_y\right)}{\sum_{k=1}^{C} \exp\left(z^\top w_k\right)},$$

where $w_k \in \mathbb{R}^d$ is a k-th column of W. Recently, Wang et al. (2022) proposed *Supervised Learning with MLP (SLMLP)* which improves the transferability of supervised models by attaching a projection MLP on z and using cosine cross-entropy loss. Formally, let $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a projection MLP, and $W : \mathbb{R}^{d'} \to \mathbb{R}^C$ be linear weight, then the SLMLP loss is given as follows:

$$\ell_{\texttt{SLMLP}}(z, y) = -\log \frac{\exp\left(\tilde{h}(z)^\top \tilde{w}_y/\tau\right)}{\sum_{k=1}^{C} \exp\left(\tilde{h}(z)^\top \tilde{w}_k/\tau\right)}, \quad \text{with} \quad \tilde{h}(z) = \frac{h(z)}{\|h(z)\|_2}, \tilde{w}_k = \frac{w_k}{\|w_k\|_2},$$

for each $k = 1, \dots, C$ and $\tau > 0$ is a temperature for cosine cross-entropy loss. Adding MLP helps the representation retain the intra-class variation, which is crucial in transfer learning.

Supervised contrastive learning. The supervised contrastive (SupCon) learning (Khosla et al., 2020) is a generalization of self-supervised contrastive learning (Chen et al., 2020b) to a supervised version. In self-supervised learning, positives are defined by instances that are augmented from the same input. In contrast, in SupCon, positives are extended to instances that are in the same class. Let us denote $z_i^+, i = 1, \dots, M$ be M positives of z, i.e., $z^+ \sim p\big(z^+ \mid L(z) = L(z^+)\big)$, and $z_j^-, j = 1, \dots, K$ be K negatives of z, i.e., $z^- \sim p\big(z^- \mid L(z) \neq L(z^-)\big)$. Then the SupCon loss is defined as follows:

$$\ell_{\texttt{SupCon}}\left(z; \{z_i^+\}_{i=1}^M, \{z_j^-\}_{j=1}^K\right) = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{\exp\left(f(z, z_i^+)\right)}{\sum_{i=1}^{M} \exp\left(f(z, z_i^+)\right) + \sum_{j=1}^{K} \exp\left(f(z, z_j^-)\right)},$$

where $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a similarity function that is jointly optimized during training. In particular, one can use cosine-similarity function followed by a projection MLP $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ for f, i.e., $f(z, z') = \frac{h(z)^\top h(z')}{\tau\|h(z)\|\cdot\|h(z')\|}$ with temperature $\tau > 0$.

Class collapse issue for cross-entropy and SupCon. While the models trained by cross-entropy or SupCon losses show decent generalization, they lack transferability due to the class collapse issue: the features lack intra-class variation (Graf et al., 2021). Here, we provide some high-level intuition on how cross-entropy and SupCon losses lead to class collapse by breaking up the loss functions into alignment terms (i.e., loss for intra-class closeness) and uniformity terms (i.e., loss for inter-class repulsion) (Wang & Isola, 2020). By taking off the logarithm term, the cross-entropy loss can be written as $\ell_{\texttt{CE}}(z,y) = -z^\top w_y + \log\sum_{k=1}^C \exp\left(z^\top w_k\right)$. On the other hand, suppose we use an inner product similarity function for f, i.e., $f(z,z') = z^\top z'$. Then the SupCon loss becomes

$$\begin{aligned}\ell_{\texttt{SupCon}}(z) &= -\frac{1}{M}\sum_{i=1}^M \log\frac{\exp\left(z^\top z_i^+\right)}{\sum_{i=1}^M \exp\left(z^\top z_i^+\right) + \sum_{j=1}^K \exp\left(z^\top z_j^-\right)} \\ &= -\frac{1}{M}\sum_{i=1}^M z^\top z_i^+ + \log\left(\sum_{i=1}^M \exp(z^\top z_i^+) + \sum_{j=1}^K \exp\left(z^\top z_j^-\right)\right).\end{aligned}$$

If we have sufficiently large M, then we have $\frac{1}{M}\sum_{i=1}^M z_i^+ \to \bar{w}_y$ for some class prototype vector $\bar{w}_y \in \mathbb{R}^d$. Therefore, the first terms of each cross-entropy and SupCon loss are equivalent in that they aim to maximize the alignment with respect to the class prototype, and the first term enforces the features to concentrate around the class prototype, leading to class collapse.

3.2 Contrastive learning with Rényi divergence

Lee & Shin (2022) proposed self-supervised Rényi contrastive learning, which defines a new contrastive learning objective by variational estimator of the skew Rényi divergence between positives and negatives. Note that the skew divergence allows the variational estimation to have low variance, otherwise the exploding variance disrupts the learning. Formally, let P and Q be distributions with densities p and q, respectively, and suppose P is absolutely continuous with respect to Q, (i.e., $p(x)/q(x) > 0$ for all x). Then, *Rényi divergence* (Rényi, 1961) of order $\gamma \in (0,1)\cup(1,\infty)$ between P and Q is defined by

$$R_\gamma(P \,\|\, Q) \coloneqq \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_P\left[\left(\frac{p(x)}{q(x)}\right)^{\gamma-1}\right],$$

and the α-skew Rényi divergence of order γ between P and Q is defined by the Rényi divergence between P and $\alpha P + (1-\alpha)Q$, i.e., $R_\gamma^{(\alpha)}(P \,\|\, Q) = R_\gamma(P\|\alpha P + (1-\alpha)Q)$. From Birrell et al. (2021); Lee & Shin (2022), the following variational form of skew Rényi divergence holds:

$$R_\gamma^{(\alpha)}(P \,\|\, Q) = \sup_f \frac{1}{\gamma-1}\log\mathbb{E}_P[e^{(\gamma-1)f(x)}] - \frac{1}{\gamma}\log(\alpha\mathbb{E}_P[e^{\gamma f(x)}] + (1-\alpha)\mathbb{E}_Q[e^{\gamma f(x)}]). \quad (1)$$

Then the self-supervised Rényi contrastive learning is derived by applying equation 1 on the skew Rényi divergence between positives and negatives. Remark that Rényi contrastive learning conducts intrinsic easy positive mining and hard negative mining, showing its effectiveness in self-supervised learning with hard data augmentations.

4 Rényi supervised contrastive learning

Now we propose *Rényi supervised contrastive learning (RényiSCL)* by using variational representation of skew Rényi divergence in equation 1. Analogous to the supervised contrastive learning, given an anchor z, let $z^+ \sim p\big(z^+ \mid L(z) = L(z^+)\big)$ be positive of z, i.e., sampled from same class, and let $z^- \sim p\big(z^- \mid L(z) \neq L(z^-)\big)$ be negative of z. Then by using variational equality in equation 1, we define (α,γ)-Rényi supervised contrastive learning (RSCL) loss as following:

$$\ell_{\texttt{RSCL}}^{(\alpha,\gamma)}(z) = -\frac{1}{\gamma-1}\log\mathbb{E}_{z^+}[e^{(\gamma-1)f(z,z^+)}] + \frac{1}{\gamma}\log\left(\alpha\mathbb{E}_{z^+}[e^{\gamma f(z,z^+)}] + (1-\alpha)\mathbb{E}_{z^-}[e^{\gamma f(z,z^-)}]\right).$$

Remark that when $\gamma \to 1$, one can observe that the RSCL loss is equivalent to the SupCon loss by setting α proportional to the number of positive samples. Thus, one can define $\alpha-$SupCon loss similarly and see that RényiSCL is a generalized version of SupCon. Note that RényiSCL is slightly different from Rényi contrastive learning in its derivation. In Appendix A.1, we provide a detailed comparison between supervised and self-supervised Rényi contrastive learning.

4.1 Intuition behind RényiSCL

Here, we analyze how RényiSCL can learn a transferable representation without class collapse. In particular, we show that RényiSCL performs intra-class easy positive and inter-class hard negative mining by controlling the hyperparameter γ.

Recall that the first term of SupCon and cross-entropy losses directly maximize the alignment with the class prototype. On the other hand, we show that RényiSCL prevents the direct alignment by easy positive sampling, i.e., add importance weights proportional to the similarity. Let the similarity function f_θ be a neural network parametrized by θ. Then the gradient of the first term of RényiSCL loss with respect to θ gives us

$$\nabla_\theta \ell^{(\gamma)}_{\mathtt{RSCL-1st}}(z) = -\frac{\mathbb{E}_{z^+}[e^{(\gamma-1)f_\theta(z,z^+)}\nabla_\theta f_\theta(z,z^+)]}{\mathbb{E}_{z^+}[e^{(\gamma-1)f_\theta(z,z^+)}]} = -\mathbb{E}_{\mathtt{sg}(q_\theta(z^+;z))}[\nabla_\theta f_\theta(z,z^+)],$$

where $\mathtt{sg}(q_\theta(z^+;z)) \propto \exp((\gamma-1)f_\theta(z,z^+))$ is a self-normalizing importance weights (Owen, 2013) and $\mathtt{sg}$ is a stop-gradient operator. Thus, it is equivalent to the following in terms of gradient:

$$\ell^{(\gamma)}_{\mathtt{RSCL-1st}}(z) = -\mathbb{E}_{\mathtt{sg}(q_\theta(z^+;z))}[f_\theta(z,z^+)]. \tag{2}$$

The equation 2 shows that in contrast to SupCon or cross-entropy, RényiSCL conducts easy positives mining by imposing more weights on positive instances z^+ that currently have high similarity $f_\theta(z,z^+)$. Remark that when $\gamma \to 1$, the loss in equation 2 is equal to alignment term of cross-entropy and SupCon loss so that the features are pushed to follow the class prototype. On the other hand, when $\gamma \to \infty$, the q_θ assigns weights on only the closest instances of z, which resembles the idea of instance discrimination in self-supervised learning (Wu et al., 2018). Thus, by controlling the value of γ, one can balance the trade-off between generalization and transferability (Kornblith et al., 2021): by letting $\gamma \to 1$, the intra-class features are more tightened, and by letting $\gamma \to \infty$, the intra-class variation increases and helps transferability of representations.

Also, in Appendix A.2, we show that RényiSCL performs hard negative mining (Robinson et al., 2020; Lee & Shin, 2022) on the inter-class features; pushing the feature z^- that currently has the highest similarity $f_\theta(z,z^-)$. Therefore, this helps increasing the separability between classes and especially, it helps when harder data augmentation is applied. In Section 5.1, we empirically validate that RényiSCL exhibits the largest gain on using the sophisticated data augmentation such as multi-crop (Caron et al., 2020), compared to cross-entropy or SupCon.

Class-wise control of γ for class-imbalanced dataset. When the pre-training dataset is class-imbalanced, the easy positive mining on the major group, i.e., the set of classes with many data points, can improve transferability. On the other hand, the easy positive mining on the minor group, i.e., the set of classes with few data points, has only a little effect as there are only a small number of data points. Meanwhile, the effect of γ can hurt the generalization in minor group samples as it interferes with forming a cluster. Therefore, in learning representation on the imbalanced dataset, we propose to use different values of γ for the major and minor groups to balance the generalization and transferability trade-off in minor groups. Our idea is straightforward: we use a higher value of γ for the major group and a smaller value of γ for the minor group. In Section 5.2, we empirically verify that this design choice of γ helps the transfer learning performance on the imbalanced dataset and the generalization performance on minor group samples.

Coarse-to-fine transfer learning. The class collapse issue of SupCon and cross-entropy loss becomes severe when the pre-training dataset is coarse-labeled and the transfer learning dataset is fine-labeled (Chen et al., 2022). For example, when we train the supervised model on a coarse-labeled dataset with cross-entropy or SupCon, the features that share a common superclass tend to entangle in a single cluster. On the other hand, RényiSCL can learn disentangled features even when trained on the superclass. For example, in Figure 1, we visualize the features of each RényiSCL, SupCon, and cross-entropy method trained on coarse-labeled TinyImageNet dataset (see Section 5.3 for details) with UMAP (McInnes et al., 2018). While it is hard to discriminate the features trained with cross-entropy or SupCon loss by subclasses, one can easily distinguish the RényiSCL learned features by subclasses.

Method	IMN	C10	C100	FOOD	PET	FLO	CAL	CAR	AIR	DTD	SUN	Average
Self-supervised learning methods												
BYOL (Grill et al., 2020)	74.3	91.3	78.4	75.3	90.4	96.1	94.2	67.8	60.6	75.5	62.2	79.2
SwAV (Caron et al., 2020)	75.3	94.2	79.8	76.9	87.5	94.8	92.7	62.4	58.0	77.8	65.7	78.9
MoCo v3 (Chen et al., 2021)	74.8	94.8	80.1	73.8	90.6	94.6	94.5	66.0	61.4	75.7	62.6	79.4
DINO (Caron et al., 2021)	75.3	93.9	79.4	78.7	89.3	96.1	92.5	68.0	62.5	77.2	66.0	80.4
NNCLR (Dwibedi et al., 2021)	75.6	93.7	79.0	76.7	91.8	95.1	91.3	67.1	64.1	75.5	62.5	79.7
RényiCL (Lee & Shin, 2022)	76.2	94.0	78.8	78.0	89.5	96.5	93.3	71.5	61.8	77.3	66.1	80.7
Supervised learning methods												
CE (He et al., 2016)	76.5	91.8	74.3	71.1	92.4	90.9	91.0	50.0	48.7	72.0	60.4	74.3
RSB (Wightman et al., 2021)	**80.4**	92.6	75.3	71.4	**92.8**	89.8	93.0	54.3	46.6	73.8	63.1	75.3
SupCon (Khosla et al., 2020)	79.1	93.0	76.3	71.9	92.7	92.5	94.3	61.2	57.4	74.7	62.9	77.7
SupCon+SSL (Guo et al., 2020)	77.1	94.4	79.6	74.7	92.5	94.5	**94.7**	64.0	59.0	74.7	64.1	79.2
RényiSCL (Ours)	78.4	**95.3**	**80.6**	**80.1**	91.5	**97.0**	93.2	**73.6**	**65.6**	**78.9**	**66.9**	**82.3**

Table 1: Transfer learning performance on fine-grained classification benchmark. We compare both supervised and self-supervised methods. For Pets, (PET), Caltech101 (CAL), Aircraft (AIR), and Flowers (FLO), we report mean per-class accuracy (%); otherwise, we report Top-1 classification accuracies (%). IMN denotes the classification accuracy on the ImageNet dataset, and the Average is calculated over 10 downstream datasets. All baseline models are from their official repositories (see Appendix C.2).

5 Experiments

5.1 Main results

We follow the two-stage setup of Khosla et al. (2020): we pre-train an encoder on ImageNet (Deng et al., 2009) train dataset and use frozen encoder as feature extractor. We use ResNet50 (He et al., 2016) as a base encoder g and implement the similarity function f by a temperature-scaled cosine similarity followed by a projection MLP. For data augmentation, we use default data augmentation from (Chen et al., 2020b; Grill et al., 2020) and further use RandAugment (Cubuk et al., 2020) and multi-crop (Caron et al., 2020). Then we pre-train for 200 epochs with RényiSCL loss of $(\alpha, \gamma) = (0.001, 2.0)$. We use linear evaluation protocol, i.e., train a linear classifier at the top of frozen feature, for ImageNet validation accuracy. We compare the transfer learning performance of various open-sourced self-supervised and supervised representation learning methods on 10 fine-grained object classification datasets and 4 cross-domain few-shot learning datasets. See Appendix C.2 for more details.

Transfer learning on fine-grained object classification. Following (Kornblith et al., 2019), we evaluate the transfer learning performance of a representation by linear evaluation on 10 fine-grained object classification datasets. The datasets are consist of CIFAR10&100 (C10&C100) (Krizhevsky et al., 2009), Food101 (FOOD) (Bossard et al., 2014), Oxford Pets (PET) (Parkhi et al., 2012), Flowers (FLO) (Nilsback & Zisserman, 2008), Caltech101 (CAL) (Fei-Fei et al., 2004), Stanford Cars (CAR) (Krause et al., 2013), Aircraft (AIR) (Maji et al., 2013), DTD (Cimpoi et al., 2014), and SUN397 (SUN) (Xiao et al., 2010). See Appendix C.1 for the detailed experimental setup.

In Table 1, we report the Top-1 ImageNet validation accuracy and transfer learning accuracy on each dataset. Remark that RényiSCL achieves the state-of-the-art performance in average transfer learning accuracy, outperforming previous state-of-the-art RényiCL (Lee & Shin, 2022) by **+1.6%**. Note that RényiSCL lags behind by ResNet-Strikes-Back (RSB) (Wightman et al., 2021) and SupCon (Khosla et al., 2020) in ImageNet validation accuracy by 2.0% and 0.7%, respectively. However, RényiSCL shows better overall performance in generalization and transferability, as it outperforms RSB and SupCon in average transfer learning accuracy by **+7.0%** and **+4.6%**, respectively.

Cross-domain few-shot learning. Furthermore, we consider cross-domain few-shot learning tasks to evaluate the capability of learned representations to adapt to unseen tasks. For datasets, we use CUB200 (Cubuk et al., 2020) and FC100 (Oreshkin et al., 2018), which are datasets that have high domain-similarity with ImageNet, and CropDisease (Mohanty et al., 2016) and EuroSAT (Helber et al., 2019), which are datasets that have low domain-similarity with ImageNet (Oh et al., 2022).

Method	CropDisease (5, 1)	CropDisease (5, 5)	EuroSAT (5, 1)	EuroSAT (5, 5)	CUB200 (5, 1)	CUB200 (5, 5)	FC100 (5, 1)	FC100 (5, 5)
MoCo v3 (Chen et al., 2021)	81.18	94.73	72.48	89.68	69.11	89.00	38.87	57.70
DINO (Caron et al., 2021)	82.89	95.60	72.69	89.94	59.10	81.52	38.16	54.70
RényiCL (Lee & Shin, 2022)	**83.80**	95.64	73.45	90.31	57.65	82.42	36.61	53.37
RSB (Wightman et al., 2021)	75.37	92.11	66.92	85.48	**75.75**	91.81	41.28	59.81
SupCon (Khosla et al., 2020)	74.17	89.59	70.58	86.59	70.68	87.28	40.38	57.86
SupCon+SSL (Guo et al., 2020)	79.82	93.20	75.05	89.46	60.90	82.35	**42.76**	59.79
RényiSCL (Ours)	82.97	**95.99**	**75.49**	**91.19**	71.89	**92.32**	39.52	**60.06**

Table 2: The transferability of representation learning methods on few-shot learning tasks. We report the mean few-shot classification accuracy (%) over 600 episodes (with 95% confidence interval) on CropDisease, CUB200, EuroSAT, and FC100 datasets. (N, K) denotes N-way K-shot classification. All baseline models are from their official repositories (see Appendix C.2).

Method	AIR	CAR	DTD	SAT	FLO	ISIC	PET
Exemplar v2 (Zhao et al., 2020)	50.9	48.3	73.5	96.8	90.0	81.4	85.0
LOOK (Feng et al., 2021)	60.0	71.9	72.3	95.0	94.7	75.0	91.0
SLMLP (Wang et al., 2022)	59.2	63.8	72.7	96.7	94.5	80.7	91.2
RényiSCL (Ours)	**65.2**	**72.1**	**76.2**	**97.1**	**96.2**	**81.7**	**91.9**

Table 3: Transfer learning accuracy of supervised learning models. We report Top-1 accuracy (%) for EuroSAT (SAT) and ISIC datasets, and otherwise we use the same metric in Table 1. We train RényiSCL for 100 epochs without using multi-crop for a fair comparison. Exemplar v2 and LOOK are from their official repositories (see Appendix C.2).

Method	Transfer	ImageNet
SLMLP	78.6	76.2
+ MC	78.8(+0.2)	76.8(+0.6)
SupCon	79.6	76.1
+ MC	80.1(+0.5)	76.9(+0.8)
RényiSCL	81.2	76.6
+ MC	**81.7(+0.5)**	**77.8(+1.2)**

Table 4: Effect of multi-crop (MC) data augmentation.

For evaluation, we train a logistic regression on the top of frozen representation, and generate 600 episodes to compute the means of 5-way, 1-shot and 5-shot accuracies with 95% confidence interval.

In Table 2, we report the Top-1 accuracy (%) on each cross-domain few-shot learning task. We observe that RényiSCL shows the best overall performance, outperforming various supervised and self-supervised models. As shown in Oh et al. (2022), the supervised models show better performance on CUB200 and FC100 as these datasets have high domain similarity with respect to ImageNet, while self-supervised models perform better on CropDisease and EuroSAT. Since RényiSCL takes the best of both approaches, it shows the best overall performance.

Comparison with supervised learning methods for transfer learning. We present additional comparison with various supervised representation learning methods that were proposed to improve the transferability. For the baseline, we compare with Exemplar v2 (Zhang et al., 2020), LOOK (Feng et al., 2021), and SLMLP (Wang et al., 2022). For a fair comparison, we do not use multi-crop data augmentation in RényiSCL and trained for 100 epochs. In Table 3, we compare the transfer learning performance of various supervised representation learning methods on the subset of 10 object classification datasets with additional EuroSAT (SAT) (Helber et al., 2019) and ISIC (Codella et al., 2019) datasets. Remark that RényiSCL clearly outperforms other supervised representation learning baselines.

Effect of multi-crop data augmentation. Furthermore, we experiment on the effect of multi-crop data augmentation (Caron et al., 2020) on supervised representation learning methods. In Table 4, we compare the ImageNet validation accuracy (%) and average transfer learning accuracy (%) on object classification datasets. Compared to SLMLP, SupCon and RényiSCL attains larger gain from the usage of multi-crop, and RényiSCL achieves the best results in both ImageNet accuracy and average transfer learning accuracy.

5.2 RÉNYISCL ON CLASS-IMBALANCED DATASET

In this section, we consider ImageNet-LT (Liu et al., 2019) for imbalanced pre-training dataset, which is a long-tailed version of ImageNet (Deng et al., 2009) dataset by sampling with Pareto

Method	IMN	C10	C100	FOOD	PET	FLO	CAL	CAR	AIR	DTD	SUN	Average
τ-norm (Kang et al., 2020)	54.5	86.9	65.8	59.4	82.0	85.2	83.6	33.4	36.3	63.1	49.0	64.5
KCL (Kang et al., 2021)	51.4	86.2	66.0	60.9	83.1	**87.6**	84.1	**38.3**	**40.7**	66.0	**51.7**	66.5
TSC (Li et al., 2022)	51.9	86.5	66.5	60.9	83.3	87.0	83.4	38.0	40.5	66.2	51.6	66.4
PaCo (Cui et al., 2021)	**58.2**	90.1	68.8	59.1	85.3	81.6	86.6	33.1	36.9	64.4	47.9	65.4
RényiSCL (Ours)	57.7	**90.2**	**70.1**	**61.1**	**85.7**	86.8	**87.7**	37.8	38.5	**66.7**	50.3	**67.5**

Table 5: Comparison on the transfer learning performance of various representation learning methods on ImageNet-LT dataset trained with ResNeXt50. We use the same metric as in Table 1. IMN denotes the classification accuracy on ImageNet dataset, and Average is calculated over 10 downstream datasets. All baseline models are from their official repositories (see Appendix C.2).

	ImageNet-LT				iNaturalist			
Method	Many	Medium	Few	All	Many	Medium	Few	All
Balanced SoftMax (Ren et al., 2020)	66.7	52.9	33.0	55.0	72.3	72.6	71.7	71.8
τ-norm (Kang et al., 2020)	65.0	52.2	32.3	54.5	**74.1**	72.1	70.4	71.5
KCL (Kang et al., 2021)	64.8	47.3	27.4	51.4	-	-	-	68.6
TSC (Li et al., 2022)	64.5	48.6	28.0	51.9	72.6	70.6	67.8	69.7
PaCo (Cui et al., 2021)	67.5	**56.9**	36.7	**58.2**	70.3	73.2	73.6	73.2
RényiSCL (Ours)	62.8	55.4	**50.9**	57.7	69.5	**73.6**	**74.6**	**73.5**

Table 6: Top-1 accuracy (%) of various long-tailed recognition methods on ImageNet-LT and iNaturalist datasets. All ImageNet-LT models are ResNeXt50 and iNaturalist models are ResNet50. We report the group-wise accuracy by dividing into Many (>100 shots), Medium ($20-100$ shots), and Few (<20 shots). All baseline models are from their official repositories (see Appendix C.2).

distribution. The number of images in each class varies from 5 to 1,280. We use ResNeXt50 (Xie et al., 2017) for fair comparison with previous studies. Following the best practice (Cui et al., 2021), we introduce class prototypes to supervised contrastive learning (see Appendix C.3 for details). Note that one can divide the classes into 3 groups: Many (>100 shots), Medium (Med) ($20-100$ shots), and Few (<20 shots). Then as explained in Section 4.1, we use $\gamma_{\texttt{many}} = \gamma_{\texttt{med}} = 1.5$, and $\gamma_{\texttt{few}} = 1.0$.

Results. In Table 5, we compare the transfer learning performance of various representation learning methods trained on ImageNet-LT dataset. Remark that RényiSCL attains average transfer accuracy of 67.5%, outperforming other representation learning methods. Moreover, RényiSCL achieves 57.7% in ImageNet-LT classification accuracy, which is comparable to the state-of-the-art method PaCo (Cui et al., 2021). In particular, in Table 6, we compare the group-wise accuracy of various long-tailed recognition methods. Remark that RényiSCL achieves the state-of-the-art performance in few samples by a large margin. Also, we consider iNaturalist (Van Horn et al., 2018) dataset, which is natural class-imbalanced dataset. Similar to ImageNet-LT, RényiSCL outperforms existing methods in few and medium group, and achieves the state-of-the-art performance in overall accuracy.

$(\gamma_{\texttt{many}}, \gamma_{\texttt{med}}, \gamma_{\texttt{few}})$	Many	Med	Few	All	TF
(1.5, 1.5, 1.5)	**63.3**	56.6	47.2	57.9	67.4
(1.5, 1.5, 1.0)	62.8	55.4	**50.9**	57.7	**67.5**
(1.5, 1.2, 1.0)	62.8	57.7	46.9	**58.2**	66.8
(1.5, 1.0, 1.0)	62.2	**58.0**	44.7	57.8	66.6

Table 7: Ablation on the effect of group-wise assignment of γ values. We report group-wise accuracy and TF denotes average transfer accuracy on 10 downstream datasets. Experiments with ResNeXt50 on ImageNet-LT dataset.

Effect of group-wise γ. In Section 4.1, we proposed to use higher value of γ for major group, and smaller value of γ on minor group. In Table 7, we demonstrate the effect of assigning different values of γ. One can observe that using higher value of γ on Few group does not affect the performance of transfer learning. Thus, by using small value of $\gamma_{\texttt{few}}$, one can increase the generalization on few samples. Otherwise, using smaller value of $\gamma_{\texttt{med}}$ degrades the transferability, while increases the generalization on Medium group samples. See Appendix B.2 for more information.

	$N_l = 1000$		$N_l = 918$		$N_l = 753$		$N_l = 486$	
Method	Fine-TF	Fine-IMN	Fine-TF	Fine-IMN	Fine-TF	Fine-IMN	Fine-TF	Fine-IMN
SupCon (Khosla et al., 2020)	79.6	76.1	69.3	58.2	63.8	50.1	61.4	48.4
SLMLP (Wang et al., 2022)	78.6	76.2	77.6	73.1	78.6	70.2	73.7	63.1
RényiSCL (Ours)	81.2	76.6	**80.6**	**76.2**	**79.6**	**74.3**	**77.1**	**70.6**

Table 8: Results of coarse-to-fine transfer learning on coarse-labeled ImageNet dataset with number of labels $N_l = 918, 753$, and 486. We report (average) transfer learning accuracy (%) on the 10 fine-grained object classification benchmark (denoted by Fine-TF), and fine-labeled ImageNet of 1000 classes (denoted by Fine-IMN). For fair comparison, we train each method for 100 epochs without changing the hyperparameters that were used in fine-labeled ImageNet experiments. For comparison, we also report the original results using the fine-labeled ImageNet datatset to pre-train (with colored by gray).

5.3 RÉNYISCL ON COARSE-LABELED DATASET

In this section, we consider coarse-labeled versions of ImageNet as pre-training datasets. Note that the 1000 classes of the ImageNet dataset are generated from the leaves of WordNet (Miller, 1995) hierarchy. Huh et al. (2016) proposed two different approaches to generate superclasses for ImageNet: top-down and bottom-up approaches. The top-down approach generates superclasses by choosing leaves with the same distance from the root. However, this approach can make only 3 different taxnomies, where the number of labels is 127, 10, and 2. On the other hand, the bottom-up approach iteratively clusters the leaf nodes of the same parent nodes as one superclass, and this approach can have 18 different taxonomies. Therefore, we select 3 of them, where the number of labels is 918, 753, and 486, respectively.

Given the coarse-labeled ImageNet datasets, we compare three supervised learning methods: SLMLP (Wang et al., 2022), SupCon (Khosla et al., 2020), and RényiSCL. We train each model for 100 epochs without using multi-crop data augmentation. Specifically, we do not change the hyperparameters used for fine-labeled ImageNet experiments in Section 5.1. After pre-training, we use the linear evaluation protocol on the original fine-labeled (i.e., 1000 classes) ImageNet train dataset, and transfer to 10 fine-grained object classification datasets used in Section 5.1.

Results. In Table 8, we compare the transfer learning performance of three supervised learning methods. Remark that RényiSCL is the most robust method that retains transfer learning accuracy on both 10 fine-grained object classification and fine-labeled ImageNet datasets, despite the scarcity of label information. Especially, when pre-trained with 486 number of labels, RényiSCL outperforms SLMLP and SupCon by **+3.4%** and **+15.7%** for the former, and **+7.5%** and **+22.2%** for the latter, respectively. The performance gap between RényiSCL and baselines is larger when the number of pre-trained coarse-labels is smaller, which indeed confirms that RényiSCL resolves the class-collapse issue (Graf et al., 2021).

Ablation study. We also consider coarse-labeled version of CIFAR10, CIFAR100 (Krizhevsky et al., 2009), and TinyImageNet (Le & Yang, 2015) (see Appendix C.1 for details). We refer to Appendix B.3 for the complete results of our ablative study. In Figure 1, we visualize the features in a superclass of TinyImageNet that has 3 subclasses using UMAP (McInnes et al., 2018). Remark that RényiSCL learned features are distinguishable by subclasses, while the features learned by SupCon or SLMLP are entangled regardless of the subclasses.

6 CONCLUSION

This work presents Rényi supervised contrastive learning, which effectively learns transferable representation. We show that it performs easy positive sampling among intra-class instances and hard negative sampling among inter-class instances and provides empirical evidence supporting our findings. Significantly, the ImageNet pre-trained model with Rényi supervised contrastive learning outperforms other supervised and self-supervised models in various transfer learning tasks. We believe our paper could bring new insights for pre-training a large-scale foundation model, beneficial for various downstream tasks, e.g., supervised or semi-supervised learning might be better than purely self-supervised learning.

Ethics Statement

Since the supervised models highly rely on the label information, recent works noticed the weakness of supervised models on spurious attributes (e.g., hair color and gender attribute (Sagawa et al., 2019)). We believe that RényiSCL could be an alternative approach to handle the bias in the labeled dataset, which we leave for future work.

Reproducibility Statement

We provide all the implementation details to reproduce our experimental results in Section 5 and Appendix B. Also, we attach our codes in supplementary materials.

References

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational representations and neural network estimation of rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3(4):1093–1116, 2021.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3090–3122. PMLR, 2022.

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

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

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

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

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.

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

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

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5414–5423, 2021.

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

Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. Rethinking supervised pre-training for better downstream transferring. *arXiv preprint arXiv:2110.06014*, 2021.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pp. 3821–3830. PMLR, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pp. 124–141. Springer, 2020.

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6533–6537. IEEE, 2021.

Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7310–7311, 2017.

Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8845–8855, 2021.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OqtLIabPTit.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34: 28648–28662, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

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

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Kyungmin Lee and Jinwoo Shin. R\'enyicl: Contrastive representation learning with skew r\'enyi divergence. *arXiv preprint arXiv:2208.06270*, 2022.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6918–6928, 2022.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.

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

Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *2011 International conference on computer vision*, pp. 89–96. IEEE, 2011.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016.

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

Jaehoon Oh, Sungnyun Kim, Namgyu Ho, Jin-Hwa Kim, Hwanjun Song, and Se-Young Yun. Understanding cross-domain few-shot learning: An experimental study. *arXiv preprint arXiv:2202.01339*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

Art B Owen. Monte carlo theory, methods and examples. 2013.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3533–3545. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/24357dd085d2c4b1a88a7e0692e60294-Paper.pdf.

Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. Concept generalization in visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9629–9639, 2021.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, and Wanli Ouyang. Revisiting the transferability of supervised pretraining: an mlp perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9183–9193, 2022.

Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021. URL https://openreview.net/forum?id=NG6MJnVl6M5.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

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

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.

A Detailed Analysis

A.1 Comparison with RényiCL

In self-supervised setup, let Z and Z' be i.i.d random variables of feature $z = g(x)$. Then, a pair of samples $(z, z') \sim P_{Z,Z'}(z, z')$ is a positive pair if z and z' are augmented from same input. Otherwise, a pair of samples $(z, z') \sim P_Z(z)P_{Z'}(z')$ is a negative pair if they are augmented from different input. Then, (Lee & Shin, 2022) showed that the InfoNCE (a.k.a CPC) objective is a variational lower bound of skew KL divergence as following equality holds:

$$D_{\texttt{KL}}^{(\alpha)}(P_{Z,Z'} \| P_Z P_{Z'}) = \sup_f \mathcal{I}_{\texttt{CPC}}(f), \quad \text{where}$$
$$\mathcal{I}_{\texttt{CPC}}(f) = \mathbb{E}_{P_{Z,Z'}}[f(z,z')] - \log\big(\alpha \mathbb{E}_{P_{Z,Z'}}[e^{f(z,z')}] + (1-\alpha)\mathbb{E}_{P_Z P_{Z'}}[e^{f(z,z')}]\big).$$

Then they generalize to Rényi divergence to define RényiCL objective, which satisfies following:

$$R_\gamma^{(\alpha)}(P_{Z,Z'} \| P_Z P_{Z'}) = \sup_f \mathcal{I}_{\texttt{RenyiCL}}(f), \quad \text{where}$$
$$\mathcal{I}_{\texttt{RenyiCL}}(f) = \frac{1}{\gamma-1}\log \mathbb{E}_{P_{Z,Z'}}[e^{(\gamma-1)f(z,z')}] - \frac{1}{\gamma}\log\big(\alpha \mathbb{E}_{P_{Z,Z'}}[e^{\gamma f(z,z')}] + (1-\alpha)\mathbb{E}_{P_Z P_{Z'}}[e^{\gamma f(z,z')}]\big).$$

In contrast to the RényiCL objective, which is done in the full Rényi divergence between $P_{Z,Z'}$ and $P_Z P_{Z'}$, the Rényi supervised contrastive learning we considered in this paper, conducts variational estimation on the conditional Rényi divergence:

$$R_\gamma^{(\alpha)}(P_{Z'|Z} \| P_{Z'} | Z = z) = \sup_f \mathcal{I}_{\texttt{RenyiSCL}}(f, z), \quad \text{where}$$
$$\mathcal{I}_{\texttt{RenyiSCL}}(f) = \frac{1}{\gamma-1}\log \mathbb{E}_{P_{Z'|Z=z}}[e^{(\gamma-1)f(z,z')}] - \frac{1}{\gamma}\log\big(\alpha \mathbb{E}_{P_{Z'|Z=z}}[e^{\gamma f(z,z')}] + (1-\alpha)\mathbb{E}_{P_{Z'}}[e^{\gamma f(z,z')}]\big).$$

Remark that the conditioning does not change the variational objective in KL divergence, i.e., we have

$$D_{\texttt{KL}}^{(\alpha)}(P_{Z,Z'} \| P_Z P_{Z'}) = \mathbb{E}_{z\sim Z}[D_{\texttt{KL}}^{(\alpha)}(P_{Z'|Z=z} \| P_{Z'})],$$

but in general, the above equality does not holds for Rényi divergence (Rényi, 1961; Van Erven & Harremos, 2014), i.e.,

$$R_\gamma^{(\alpha)}(P_{Z,Z'} \| P_Z P_{Z'}) \neq \mathbb{E}_{z\sim Z}[R_\gamma^{(\alpha)}(P_{Z'|Z=z} \| P_{Z'})].$$

Thus, the generalization of CPC objective to SupCon loss is equivalent that they estimate the same objective, however, RényiCL and RényiSCL do not share the common objective due to the difference in their derivation. A straightforward generalization of RényiCL could be an interesting direction, which we leave for future work.

A.2 Hard Negative Mining

Here, we show that RényiSCL also performs hard negative mining (we explained easy positive mining in Section 4.1). Similar to Section 4.1, suppose f_θ is a neural network parametrized by θ. Suppose we have M positives $z_i^+, i = 1, \ldots, M$ such that $z_i^+ \sim p(z_i^+ | L(z) = L(z_i^+))$, and K negatives $z_j^-, j = 1, \ldots, K$ such that $z_j^- \sim p(z_j^- | L(z) \neq L(z_j^-))$. Then the RényiSCL loss is given as follows:

$$\ell_{\texttt{RSCL}}(z;\theta) = -\frac{1}{\gamma-1}\log\sum_{i=1}^M e^{(\gamma-1)f_\theta(z,z_i^+)} + \frac{1}{\gamma}\log\Bigg(\frac{\alpha}{M}\sum_{i=1}^M e^{\gamma f_\theta(z,z_i^+)} + \frac{1-\alpha}{K}\sum_{j=1}^K e^{\gamma f_\theta(z,z_j^-)}\Bigg).$$

Then the gradient of the RényiSCL loss is given as follows:

$$\begin{aligned}\nabla_\theta \ell_{\texttt{RSCL}}(z;\theta) &= -\mathbb{E}_{\texttt{sg}(q_\theta(z^+;z))}[\nabla_\theta f_\theta(z,z_i^+)] + \frac{\frac{\alpha}{M}\sum_{i=1}^M e^{\gamma f(z,z_i^+)}\nabla_\theta f_\theta(z,z_i^+) + \frac{1-\alpha}{K}\sum_{j=1}^K e^{\gamma f(z,z_j^-)}\nabla_\theta f_\theta(z,z_j^-)}{\frac{\alpha}{M}\sum_{i=1}^M e^{\gamma f_\theta(z,z_i^+)} + \frac{1-\alpha}{K}\sum_{j=1}^K e^{\gamma f_\theta(z,z_j^-)}} \\ &= -\mathbb{E}_{\texttt{sg}(q_\theta(z^+;z))}[\nabla_\theta f_\theta(z,z_i^+)] + \mathbb{E}_{\texttt{sg}(r_\theta(z';z))}[\nabla_\theta f_\theta(z,z')],\end{aligned}$$

Method	IMN	C10	C100	FOOD	PET	FLO	CAL	CAR	AIR	DTD	SUN	Average
SLMLP (Wang et al., 2022)	76.2	95.0	80.3	72.9	91.2	94.5	94.6	63.8	59.2	72.7	62.1	78.6
SLMLP †(Wang et al., 2022)	76.8	94.2	79.3	76.1	90.9	94.9	93.3	60.1	57.0	76.8	65.4	78.8
SupCon(Khosla et al., 2020)	76.1	**95.1**	81.1	73.1	90.3	95.3	94.7	65.5	64.3	74.4	62.6	79.6
SupCon † (Khosla et al., 2020)	76.9	94.6	79.9	77.5	90.4	96.1	93.3	64.7	60.8	77.1	**66.4**	80.1
RényiSCL (Ours)	76.6	**95.1**	**81.5**	74.7	**91.9**	96.2	**95.3**	**72.1**	**65.2**	76.2	63.7	81.2
RényiSCL † (Ours)	**77.8**	94.5	80.1	**79.3**	91.6	**96.4**	93.9	70.9	64.7	**78.9**	66.3	**81.7**

Table 9: Extended results of Table 4. We use same metric as in Table 1. † denotes the use of multi-crop data augmentation. All models are trained for 100 epochs.

$(\gamma_{\texttt{many}}, \gamma_{\texttt{medium}}, \gamma_{\texttt{few}})$	C10	C100	FOOD	PET	FLO	CAL	CAR	AIR	DTD	SUN	Average
(1.0, 1.0, 1.0)	90.1	68.8	59.1	85.3	81.6	86.6	33.1	36.9	64.4	47.9	65.4
(1.5, 1.0, 1.0)	90.4	69.5	60.3	85.3	85.1	87.8	36.0	36.9	65.5	49.6	66.6
(1.5, 1.2, 1.0)	90.3	69.8	60.4	85.8	85.3	86.9	35.0	39.0	65.9	49.9	66.8
(1.5, 1.5, 1.0)	90.2	70.1	61.1	85.7	86.8	87.7	37.8	38.5	66.7	50.3	67.5
(1.5, 1.5, 1.5)	90.4	70.0	60.8	85.7	85.6	88.2	35.7	39.8	67.3	50.5	67.4

Table 10: Comparison on the transfer learning performance of transfer learning on ImageNet-LT by using group-wise different values of γ. We use same metric as in Table 1.

where $\texttt{sg}(q_\theta(z^+; z))$ is a self-normalized importance weights defined in equation 2, and $r_\theta(z'; z)$ is also a self-normalized importance weights defined as

$$r_\theta(z'; z) \propto \begin{cases} \alpha \exp\left(\gamma f_\theta(z, z')\right) & z' \sim p(z'|L(z) = L(z')) \\ (1-\alpha)\exp\left(\gamma f_\theta(z, z')\right) & z' \sim p(z'|L(z) \neq L(z')) \end{cases}.$$

Thus, the RényiSCL loss is equivalent to following in terms of gradient:

$$\ell_{\texttt{RSCL}}(z; \theta) \equiv -\mathbb{E}_{\texttt{sg}(q_\theta(z^+;z))}[f_\theta(z, z_i^+)] + \mathbb{E}_{\texttt{sg}(r_\theta(z';z))}[f_\theta(z, z')]$$

Thus, by using higher value of γ, the RSCL loss imposes more weight on hard negative samples, i.e., the negatives that currently have high similarity $f_\theta(z, z^-)$. Therefore, alike RényiCL, RényiSCL performs hard negative mining, but with inter-class instances. Thus, hard negative sampling helps increase the class separability, and show its effectiveness when used with harder data augmentation.

B Extended results

B.1 ImageNet experiments

In Table 9, we report the extended results of Table 4.

B.2 Imbalanced dataset

In Table 10, we report the extended results of Table 7. Additionally, verify the effect of group-wise γ on CIFAR100-LT dataset with imbalance factor of 100 (the description on the dataset is in Table 15). In Table 11, we report the group-wise accuracy on CIFAR100 test dataset. Similarly, we observe that using a smaller value of γ for minor group, and a larger value of γ for major group increases the generalization performance on minor group.

B.3 Coarse-to-fine experiments

Coarse-to-fine transfer learning on ImageNet. In Table 12, we report the extended results of Table 8.

Coarse-to-fine transfer learning on CIFAR10, CIFAR100, TinyImageNet. Here, we show the experimental results on the small scale coarse-to-fine transfer learning tasks with CIFAR10, CIFAR100, and TinyImageNet datasets (see Appendix C.1 for the coarse-labeled dataset and Appendix C.4 for the details of experimental setup). For baselines, we compare with SupCon (Khosla

$(\gamma_{\texttt{many}}, \gamma_{\texttt{medium}}, \gamma_{\texttt{few}})$	Many	Medium	Few	All
(1.5, 1.5, 1.5)	56.1	48.7	43.0	49.6
(1.5, 1.5, 1.0)	57.0	48.6	43.4	50.0
(1.5, 1.2, 1.0)	56.9	51.7	41.6	50.5
(1.5, 1.0, 1.0)	56.9	53.1	40.4	50.6

Table 11: Effect of group-wise γ on CIFAR-100-LT with imbalance factor 100. All models are trained with ResNet-32 as a backbone. We report the group-wise accuracy by dividing into many (>100 shots), medium (20-100 shots), and few (<20 shots).

Method	IMN	C10	C100	FOOD	PET	FLO	CAL	CAR	AIR	DTD	SUN	Average
Coarse-labeled dataset with $N_l = 918$												
SupCon (Khosla et al., 2020)	58.2	90.3	70.9	59.9	77.3	89.3	89.6	45.3	49.5	68.7	52.0	69.3
SLMLP (Wang et al., 2022)	73.1	94.7	80.1	72.3	90.1	94.7	94.0	63.4	60.8	73.7	62.1	78.6
RényiSCL (Ours)	**76.2**	94.8	80.9	73.9	91.5	95.7	94.4	72.6	65.1	75.0	62.6	**80.6**
Coarse-labeled dataset with $N_l = 753$												
SupCon (Khosla et al., 2020)	50.1	88.5	66.8	56.1	60.9	88.1	89.2	33.0	45.5	65.1	45.2	63.8
SLMLP (Wang et al., 2022)	70.2	94.4	80.3	71.5	80.0	94.8	94.4	64.5	60.7	73.6	61.6	77.6
RényiSCL (Ours)	**74.3**	95.0	80.9	73.8	82.4	95.5	95.3	71.8	64.6	73.9	62.8	**79.6**
Coarse-labeled dataset with $N_l = 486$												
SupCon (Khosla et al., 2020)	48.4	87.2	64.7	55.0	56.1	85.8	86.2	28.6	41.5	64.0	44.8	61.4
SLMLP (Wang et al., 2022)	63.0	92.4	76.2	69.0	68.0	94.1	92.9	56.6	56.9	72.1	59.4	73.7
RényiSCL (Ours)	**70.6**	93.5	78.9	72.3	74.5	95.1	94.6	64.1	61.3	74.2	62.3	**77.1**

Table 12: Extended results of Table 8. We use the same metric as in Table 1.

et al., 2020), SLMLP (Wang et al., 2022). In addition, note that Chen et al. (2022) introduced class-conditional InfoNCE (cNCE) loss to prevent the class-collapse problem and improve the transferability. Thus, we show the results on SupCon+cNCE, and since cNCE loss can be used for RényiSCL loss, we also report that one. In case of RényiSCL with cNCE loss, we implement the Rényi divergence variant of cNCE loss. Detailed explanation is provided in section C.3. We use $\gamma = 2.5$ for all experiments conducted for both RenyiSCL and joint training of RenySCL and cNCE objective. In Table 13, we report the results of coarse-to-fine transfer learning experiments. Remark that RényiSCL outperforms other baseline, and for CIFAR100 and TinyImageNet, RényiSCL attains the best performance without using additional cNCE loss. Remark that for CIFAR10, there are only two 2 labels in coarse-labeled dataset, thus the cNCE loss can greatly improve the performance. On the other hand, when there are sufficiently many classes such as CIFAR100 or TinyImageNet, using RényiSCL alone suffices to achieve the good performance.

Method	CIFAR10		CIFAR100		TinyImageNet	
	Fine	Coarse	Fine	Coarse	Fine	Coarse
SLMLP (Wang et al., 2022)	95.5	79.8	74.4	71.2	61.0	41.2
SupCon (Khosla et al., 2020)	95.5	81.2	72.1	70.1	59.2	54.5
SupCon+cNCE (Chen et al., 2022)	95.0	86.7	72.5	70.9	58.0	47.7
RényiSCL (Ours)	95.6	82.6	74.5	**71.4**	60.9	**58.4**
RényiSCL+cNCE	94.3	**89.4**	74.4	71.3	59.5	56.4

Table 13: Comparison with class-conditional InfoNCE objective (Chen et al., 2022). cNCE denotes class-conditional InfoNCE objective. For each dataset and method, we report the Top-1 accuracy of fine-label (Fine) and coarse-label (Coarse). Every experiment use ResNet-18 as backbone.

Dataset	# of classes	Training	Validation	Test	Metric
CIFAR10 (Krizhevsky et al., 2009)	10	45000	5000	10000	Top-1 accuracy
CIFAR100 (Krizhevsky et al., 2009)	100	45000	5000	10000	Top-1 accuracy
Food (Bossard et al., 2014)	101	68175	7575	25250	Top-1 accuracy
Pets (Parkhi et al., 2012)	37	2940	740	3669	Mean per-class accuracy
Flowers (Nilsback & Zisserman, 2008)	102	1020	1020	6149	Mean per-class accuracy
Caltech101 (Fei-Fei et al., 2004)	101	2525	505	5647	Mean Per-class accuracy
Cars (Krause et al., 2013)	196	6494	1650	8041	Top-1 accuracy
Aircraft (Maji et al., 2013)	100	3334	3333	3333	Mean Per-class accuracy
DTD (split 1) (Cimpoi et al., 2014)	47	1880	1880	1880	Top-1 accuracy
SUN397 (split 1) (Xiao et al., 2010)	397	15880	3970	19850	Top-1 accuracy
EuroSAT (Helber et al., 2019)	10	13500	5400	8100	Average accuracy
ISIC (Codella et al., 2019)	7	5007	2003	3005	Average accuracy
FC100 (Oreshkin et al., 2018)	20	-	-	12000	Average accuracy
CUB200 (Cubuk et al., 2020)	200	-	-	11780	Average accuracy
Plant Disease (Mohanty et al., 2016)	38	-	-	54305	Average accuracy
EuroSAT (Helber et al., 2019)	10	-	-	27000	Average accuracy

Table 14: Dataset information for the transfer learning tasks. For FC100, CUB200, Plant Disease, we perform few-shot learning, otherwise, we perform linear evaluation.

C Implementation details

C.1 Dataset information

In Table 14, we list the information on the datasets that we used for transfer learning experiments of fine-grained object classification and cross-domain few-shot learning in Section 5.1. In Table 15, we list the information on the datasets that we used for experiments in Section 5.2. In Table 16, we list the information on the datasets that we used for experiments in Section 5.3.

C.2 Experiment on ImageNet

Implementation details of RényiSCL on ImageNet. We use ResNet50 (He et al., 2016) for all of our experiments. We use two layer projection MLP with dimension 4096-256. For optimization, we use LARS (You et al., 2017) optimizer with base learning rate of 0.8 (i.e., the learning rate is multiplied by base learning rate $\times$ batch size / 256), and decay by cosine learning rate schedule, and the weight decay is 1e-6. For the similarity function, we use cosine-similarity with temperature $\tau = 0.2$. Following (Khosla et al., 2020), we use memory queue of size 65536 (without using momentum encoder). For data augmentation, we use base data augmentation from (Chen et al., 2020b; Grill et al., 2020; Lee & Shin, 2022), and further applied RandAugment (Cubuk et al., 2020) and multi-crop Caron et al. (2020) data augmentation. For the hyperparameters, we use $\alpha = 0.001$ and $\gamma = 2.0$ for all of our experiments.

Re-implementation of SupCon and SLMLP. To reproduce SupCon (Khosla et al., 2020), we use same setting for RényiSCL, and only changed γ to 1.0. To reproduce SLMLP (Wang et al., 2022), we use same setting for RényiSCL, except that we use prototypical layer for cross-entropy based classification. Given that the output of projection MLP is of dimension 256, the prototypical

Dataset	# of training data	# of classes	Max. # sample	Min. # sample
CIFAR-100-LT (Cao et al., 2019)	10.8K	100	500	5
ImageNet-LT (Liu et al., 2019)	115.8K	1000	1280	5
iNaturalist (Van Horn et al., 2018)	437.5K	8142	1000	2

Table 15: Information for long-tailed dataset. Max. # sample and Min. # sample indicate the number of samples in the most frequent and the rarest class, respectively.

Dataset	# of coarse classes	# of fine classes
CIFAR-10 (Krizhevsky et al., 2009)	10	2
CIFAR-100 (Krizhevsky et al., 2009)	100	20
TinyImageNet (Le & Yang, 2015)	200	52
ImageNet (Deng et al., 2009)	1000	918, 753, 486

Table 16: Dataset information for coarse-to-fine transfer experiment.

layer is a linear layer with dimension 256×1000, and is ℓ_2-normalized throughout the training. The temperature for cosine cross-entropy loss is 0.2, and we use same optimizer as in RényiSCL experiment. Lastly, we use all the same data augmentation setup for each RényiSCL, SupCon, and SLMLP.

Baselines. We list the information on the baselines that we compared in Table 1:

- SimCLR (Chen et al., 2020b): results excerpted from their original paper.
- BYOL (Grill et al., 2020): results excerpted from their original paper.
- NNCLR (Dwibedi et al., 2021): results excerpted from their original paper.
- SwAV (Caron et al., 2020): use checkpoints from their official code[1].
- DINO (Caron et al., 2021): use checkpoints from their official code[2].
- MoCo v3 (Chen et al., 2021): use checkpoints from their official code[3].
- ResNet-Strikes-Back (Wightman et al., 2021): use checkpoints from their official code[4].
- RényiCL (Lee & Shin, 2022): we reproduced the results from their original paper.
- SupCon (Khosla et al., 2020): use checkpoints from their official code [5].
- SupCon+SSL (Islam et al., 2021): use checkpoints from their official code [6].
- LOOK (Feng et al., 2021): results excerpted from their original paper.
- Exemplar V2 (Zhao et al., 2020): use checkpoints from their official code [7].

Transfer learning. For linear evaluation on fine-grained object classification, we use the same data augmentation that we used for linear evaluation on the ImageNet dataset. For optimization, we ℓ_2-regularized L-BFGS, where the regularization hyperparameter search is done on logarithmically spaced values 10^{-6} to 10^5, and fine the best hyperparameter by testing on the validation set. Then, we train a linear classifier using both training and validation splits and report the test accuracy using the metric instructed in Table 1. The maximum number of iterations is 5000 and we use the previous solution as an initial point, i.e., a warm start, for the next step. For few-shot learning experiments, we perform logistic regression on the top of frozen representations and use $N \times K$ support samples without fine-tuning and data augmentation in a N-way K-shot episode.

[1] https://github.com/facebookresearch/swav
[2] https://github.com/facebookresearch/dino
[3] https://github.com/facebookresearch/moco-v3
[4] https://github.com/rwightman/pytorch-image-models/
[5] https://github.com/HobbitLong/SupContrast
[6] https://github.com/asrafulashiq/transfer_broad
[7] https://github.com/nanxuanzhao/Good_transfer

C.3 Experiments on Imbalanced dataset

Dataset. ImageNet-LT (Liu et al., 2019) is a long-tailed version of ImageNet (Deng et al., 2009), generated by subsampling ImageNet following Pareto distribution (Reed, 2001) with power value $\alpha = 6$. It consists of 115.8K images and contains 1000 categories. The number of samples per each class varies from 5 to 1280. iNaturalist (Van Horn et al., 2018) is a real-world large-scale dataset, which contains 437.5K images from 8142 classes. CIFAR100-LT (Cao et al., 2019) is a manually crafted imbalanced subset from CIFAR100 (Krizhevsky et al., 2009), which follows the exponential distribution. The severity of the imbalance is controlled by the imbalance factor, which is a ratio between the number of samples of the most frequent and rare classes.

Implementation of RényiSCL. Our implementation is based on PaCo (Cui et al., 2021). PaCo introduced learnable class prototypes $\mathbf{W} = \{w_k\}_{k=1}^{C}$ where $w_k \in \mathbb{R}^d$ stands for the k-th class prototype. Then, let the similarity function defined as follows:

$$f(z, z') = \begin{cases} z^\top z' + \log q_y & z' \in \{w_k\}_{k=1}^{C} \\ \frac{h(z)^\top h(z')}{\tau \|h(z)\| \|h(z')\|} & z' \sim p(z) \end{cases},$$

where $y = L(z)$ be class of z, and q_y is a class rebalancing ratio, i.e., $q_y = \frac{n_y}{\sum_{k=1}^{C} n_k}$, where n_k is a number of data points in k-th class. Then for M positives $z_i^+ \sim p(z_i^+ | L(z) = L(z_i^+)), i = 1, \ldots, M$, and K negatives $z_j^- \sim p(z_j^- | L(z) \neq L(z_j^-)), j = 1, \ldots, K$, the original PaCo objective is defined by the supervised contrastive learning with consideration of class prototypes:

$$\ell_{\texttt{PaCo}}(z) = -\frac{1}{\beta M + 1}\left(\sum_{i=1}^{M} \beta f(z, z_i^+) + f(z, w_y)\right) + \log\left(\sum_{i=1}^{M} e^{f(z, z_i^+)} + \sum_{j=1}^{K} e^{f(z, z_j^-)} + \sum_{k=1}^{C} e^{f(z, w_k)}\right),$$

where $y = L(z)$ is a ground-truth class of z, and β is a hyperparameter that balances the power of contrast between instances (i.e., z with $z' \sim p(z)$), and class prototypes w_k. Then, we adapt RényiSCL loss to the PaCo objective as follows:

$$\begin{aligned} \ell_{\texttt{RSCL}}(z) = &-\frac{1}{\gamma - 1} \log \frac{1}{\beta M + 1}\left(\sum_{i=1}^{M} \beta e^{(\gamma-1) f(z, z_i^+)} + e^{(\gamma-1) f(z, w_y)}\right) \\ &+ \frac{1}{\gamma} \log\left(\sum_{i=1}^{M} e^{\gamma f(z, z_i^+)} + \sum_{j=1}^{K} e^{\gamma f(z, z_j^-)} + \sum_{k=1}^{C} e^{\gamma f(z, w_k)}\right) \end{aligned}$$

We set γ as $(\gamma_{\texttt{many}}, \gamma_{\texttt{med}}, \gamma_{\texttt{few}}) = (1.5, 1.5, 1.0)$ for ImageNet-LT and 1.1 on every group for iNaturalist dataset. We experiment with $\gamma \in \{1.1, 1.2, 1.5\}$, and we choose the best hyperparameter. We use $\beta = 0.05$ for both dataset, following Cui et al. (2021).

Model. We use ResNet-32, ResNeXt-50, and ResNet-50 backbone for each CIFAR100-LT, ImageNet-LT, and iNaturalist dataset, respectively. For CIFAR experiment, we train the model for 400 epochs with SGD optimizer with learning rate 0.05, momentum 0.9, and weight decay 5e-4. We decrease the learning rate with factor of 0.1 at epoch 320 and 360. For ImageNet-LT experiment, we also use SGD optimizer for 400 epochs. The initial learning rate is 0.02, is decayed by cosine learning rate schedule.The weight decay is 5e-4. We use the same configuration for iNaturalist with ImageNet-LT, except that weight decay is 1e-4.

Baselines. We list the baseline models that we compared in Table 5:

- τ-norm (Kang et al., 2020): use checkpoints from their official code[8].
- KCL (Kang et al., 2021), TSC (Li et al., 2022) : we reproduce the ResNeXt-50 model with the official TSC impementation[9].
- PaCo (Cui et al., 2021) : use checkpoints from their official code[10].

[8] https://github.com/facebookresearch/classifier-balancing
[9] https://github.com/LTH14/targeted-supcon
[10] https://github.com/dvlab-research/Parametric-Contrastive-Learning

C.4 COARSE-TO-FINE TRANSFER LEARNING

Implementation details on ImageNet For each RényiSCL, SupCon, and SLMLP experiment, we use the exactly same hyperparameters that we used in Section C.2.

Coarse-labeled TinyImageNet Following Chen et al. (2022), we use the coarse-labeled dataset of TinyImageNet, which is composed of 52 superclasses. The list of 52 superclasses are following:

- 'arachnid', 'bear', 'bird', 'bug', 'butterfly', 'cat', 'coral', 'crocodile', 'crustacean', 'dog', 'echinoderms', 'fish', 'frog', 'fruit', 'fungus', 'hog', 'marine mammals', 'marsupial', 'mollusk', 'plant', 'primate', 'rodent', 'salamander', 'snake', 'trilobite', 'ungulate', 'vegetable', 'wild cat', 'accessory', 'ball', 'boat', 'building', 'clothing', 'container', 'cooking', 'decor', 'electronics', 'fence', 'food', 'furniture', 'hat', 'instrument', 'lab equipment', 'outdoor scene', 'paper', 'sports equipment', 'technology', 'tool', 'toy', 'train', 'vehicle', 'weapon'.

Note that while Chen et al. (2022) originally presented 67 super-classes, we found out some of these suggested categories do not contain any subclasses. Thus, the actual number of super-classes are 52.

Coarse-labeled CIFAR10 and CIFAR100 The coarse label of CIFAR10 dataset is composed with 2 super-classes; 'animals' and 'vehicles', and each has 6 and 4 subclasses, respectively. CIFAR100 has 20 superclasses, and each of them contains 5 subclasses. For instance, super-class 'fish' includes 'aquarium fish', 'flatfish', 'ray', 'shark', and 'trout' as its subclasses.

Implementation details on CIFAR10/100 and TinyImageNet datasets. We use ResNet-18 architecture (He et al., 2016) adjusted for CIFAR dataset as the backbone; in particular, the kernel size of the first convolutional layer of 7×7 is replaced by 3×3, and the max pooling layer is omitted. We use SGD optimizer with learning rate 0.5, cosine learning rate scheduling, batch size 512, momentum 0.9, and weight decay 1e-4. We train the models for 400 epochs. Data augmentation methods used in pre-training stage are random resized cropping, horizontal flipping, color jittering, and grayscale conversion. For evaluation, we freeze the backbone and train the linear layer upon the learned representation for 100 epochs with learning rate 3.0, cosine learning rate scheduling, batch size 256, and momentum 0.9 without weight decay. We use random resized cropping and horizontal flipping for training the linear layer.

For the class-conditional infoNCE (Chen et al., 2022) (cNCE) experiments, we implement the objective as below:

$$\ell_{\texttt{SupCon}}(z) = -\frac{1}{M}\sum_{i=1}^{M} f(z, z_i^+) + \log\left(\sum_{i=1}^{M}\exp\left(f(z, z_i^+)\right) + \sum_{j=1}^{K}\exp\left(f(z, z_j^-)\right)\right).$$

$$\ell_{\texttt{cNCE}}(z) = -f(z, z^{++}) + \log\left(\sum_{i=1}^{M}\exp\left(f(z, z^{++})\right)\right).$$

$$\ell_{\texttt{SupCon+cNCE}}(z) = (1-\lambda)\ell_{\texttt{SupCon}}(z) + \lambda\ell_{\texttt{cNCE}}(z),$$

where z^{++} is a positive instance that is augmented from the same input of z.

Also, for joint training of RényiSCL and cNCE loss experiment, we implement the self-supervised Rényi contrastive learning (RényiCL) (Lee & Shin, 2022) variant of cNCE, i.e., Rényi-cNCE loss, which is defined as follows:

$$\ell_{\texttt{RSCL}}^{(\alpha,\gamma)}(z) = -\frac{1}{\gamma-1}\log\sum_{i=1}^{M} e^{(\gamma-1)f(z,z_i^+)} + \frac{1}{\gamma}\log\left(\sum_{i=1}^{M} e^{\gamma f(z,z_i^+)} + \sum_{j=1}^{K} e^{\gamma f(z,z_j^-)}\right)$$

$$\ell_{\texttt{Renyi-cNCE}}(z) = -f(z, z^{++}) + \frac{1}{\gamma}\log\left(e^{\gamma f(z,z^{++})} + \sum_{j=1}^{K} e^{\gamma f(z,z_j^-)}\right)$$

$$\ell_{\texttt{RSCL+R-cNCE}}(z) = (1-\lambda)\ell_{\texttt{RSCL}}(z) + \lambda\ell_{\texttt{Renyi-cNCE}}(z).$$

We set $\lambda = 0.5$ for every experiment following Chen et al. (2022).