EXPLORING BALANCED FEATURE SPACES FOR REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing self-supervised learning (SSL) methods are mostly applied for training representation models from artificially balanced datasets (*e.g.* ImageNet). It is unclear how well they will perform in the practical scenarios where datasets are often imbalanced w.r.t. the classes. Motivated by this question, we conduct a series of studies on the performance of self-supervised contrastive learning and supervised learning methods over multiple datasets where training instance distributions vary from a balanced one to a long-tailed one. Our findings are quite intriguing. Different from supervised methods with large performance drop, the self-supervised contrastive learning methods perform stably well even when the datasets are heavily imbalanced. This motivates us to explore the *balanced feature spaces* learned by contrastive learning, where the feature representations present similar linear separability w.r.t. all the classes. Our further experiments reveal that a representation model generating a balanced feature space can generalize better than that yielding an imbalanced one across multiple settings. Inspired by these insights, we develop a novel representation learning method, called k-positive contrastive learning. It effectively combines strengths of the supervised method and the contrastive learning method to learn representations that are both discriminative and balanced. Extensive experiments demonstrate its superiority on multiple recognition tasks. Remarkably, it achieves new state-of-the-art on challenging long-tailed recognition benchmarks. Code and models will be released.

1 INTRODUCTION

Self-supervised learning (SSL) has been popularly explored as it can learn data representations without requiring manual annotations and offer attractive potential of leveraging the vast amount of unlabeled data in the wild to obtain strong representation models (Gidaris et al., 2018; Noroozi & Favaro, 2016; He et al., 2020; Chen et al., 2020a; Wu et al., 2018). For instance, some recent SSL methods (Hénaff et al., 2019; Oord et al., 2018; Hjelm et al., 2018; He et al., 2020) use the unsupervised contrastive loss (Hadsell et al., 2006) to train the representation models by maximizing the instance discriminativeness, which are shown to generalize well across various downstream tasks, and even surpass the supervised learning counterparts in some cases (He et al., 2020; Chen et al., 2020a).

Despite the great success, existing SSL methods focus on learning data representations from the artificially balanced datasets (*e.g.* ImageNet (Deng et al., 2009)) where all the classes have similar numbers of training instances. However in reality, since the classes in natural images follow the Zipfian distribution, the datasets are usually imbalanced and show a long-tailed distribution (Zipf, 1999; Spain & Perona, 2007), *i.e.*, some classes involving significantly fewer training instances than others. Such imbalanced datasets are very challenging for supervised learning methods to model, leading to noticeable performance drop (Wang et al., 2017; Mahajan et al., 2018; Zhong et al., 2019). Thus several interesting questions arise: How well will SSL methods perform on imbalanced datasets? Will the quality of their learned representations deteriorate as the supervised learning methods? Or can they perform stably well? Answering these questions is important for understanding the behavior of SSL in practice. But these questions remain open as no research investigations have been conducted along this direction so far.

Our work is motivated by the above questions to study the properties of data representations learned with supervised/self-supervised methods in a practical scenario. We start with two representative

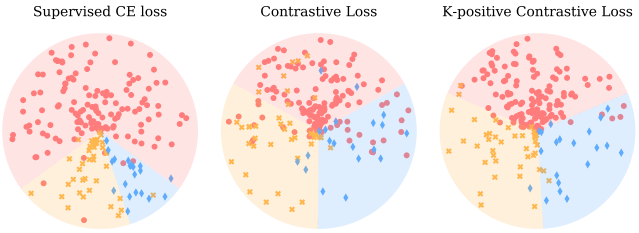

Figure 1: *Feature spaces learned with different losses given an imbalanced dataset.* The supervised cross-entropy (CE) learns a space biased to the dominant class. The space learned by unsupervised contrastive loss is balanced but less semantically discriminative. Our proposed k-positive contrastive loss learns a balanced and discriminative feature space. The shadow area ( ) indicates the decision boundary of each class.

losses used by these methods, *i.e.*, the supervised cross-entropy and the unsupervised contrastive losses (Hadsell et al., 2006; Oord et al., 2018), and investigate the classification performance of their trained representation models from multiple training datasets where the instance distribution gradually varies from a balanced one to a long-tailed one. We surprisingly observe that, different from the ones learned from supervised cross-entropy loss where performance drops quickly, the representation models learned from the unsupervised contrastive loss perform stably well, no matter how much the training instance distribution is skewed to be imbalanced. Such a stark difference between the two representation learning methods drives us to explore why SSL performs so stably. We find that **using the contrastive loss can obtain representation models generating *a balanced feature space*** that has similar separability (and classification performance) for all the classes, as illustrated in Figure 1.

Such a balanced property of the feature spaces from SSL is intriguing and provides a new perspective to understand the behavior of SSL methods. We dig deeper into its benefits via a systematic study. In particular, since a pre-trained representation model is often used as initialization for downstream tasks (He et al., 2020; Newell & Deng, 2020; Hénaff et al., 2019), we evaluate and compare the generalization ability of the models that produce feature spaces of different balanced levels (or 'balancedness'). We find that **a more balanced model tends to generalize better across a variety of settings**, including the out-of-distribution recognition as well as the cross-domain and cross-task applications. These studies imply that feature space balancedness is an important but often neglected factor for learning high-quality representations.

Inspired by the above insights, we propose a new representation learning method, the k-positive constrastive learning, which inherits the strength of constrastive learning in learning balanced feature spaces and meanwhile improves the feature spaces' discriminative capability. Specifically, different from the contrastive learning methods lacking semantic discriminativeness, the proposed k-positive constrastive method leverages the available instance semantic labels by taking k instances of the same label with the anchor instance to embed semantics into the contrastive loss. As such, it can learn representations with desirable balancedness and discriminativeness (Figure 1). Extensive experiments and analyses clearly demonstrate its superiority over the supervised learning and latest contrastive learning methods (He et al., 2020) for various recognition tasks. In particular, the proposed method establishes new state-of-the-art on the ImageNet-LT (Liu et al., 2019) and iNaturalist (iNatrualist, 2018) benchmarks for long-tailed recognition.

This work makes the following important observations and contributions. (1) We present the first systematic studies on the performance of self-supervised contrastive learning on imbalanced datasets which are helpful to understanding the merits and limitations of SSL in practice. (2) Our studies reveal an intriguing property of the model trained by contrastive learning—the model can robustly learn balanced feature spaces—that has never been discussed before. (3) Our empirical analysis demonstrates that learning balanced feature spaces benefits the generalization of representation models and offer a new perspective for understanding deep model generalizability. (4) We develop a new method to explicitly pursue balanced feature spaces for representation learning and it outperforms the popular cross-entropy and contrastive losses based methods. We believe our findings and the novel k-positive contrastive method are inspiring for future research on representation learning.

2 Related Works

Self-supervised learning Self-supervised learning is a form of unsupervised learning. Recently there has been a surge of self-supervised data representation learning methods developed to allevi-

ate the demand for manual annotations by mining free supervision information through specifically designed loss functions and pretext tasks. The contrastive loss measures the similarities of sample pairs in a feature space and is at the core of several recent SSL methods (Chen et al., 2020a;b; He et al., 2020; Chen et al., 2020c). Adversarial losses that measure the distribution difference are also exploited for self-supervised representation learning (Donahue et al., 2016; Doersch & Zisserman, 2017). A wide range of pretext tasks have been developed including image inpainting (Jenni & Favaro, 2018; Pathak et al., 2016), image colorization (Larsson et al., 2016; 2017), context prediction (Doersch et al., 2015), jigsaw puzzles (Carlucci et al., 2019; Noroozi & Favaro, 2016; Wei et al., 2019), rotation prediction (Gidaris et al., 2018). Though very successful, the behavior of SSL largely remains a mystery. Recently Wang & Isola (2020) analyze contrastive learning from the perspective of uniformity and alignment of learned representations. However investigations on the behavior of contrastive learning on imbalanced datasets are still absent. We present the first study on this problem and our investigation methodology is also applicable to other SSL methods.

Long-tailed recognition In practice, the visual data usually follow a long-tailed distribution (Zipf, 1999; Spain & Perona, 2007), challenging supervised learning methods. Due to the imbalance in training instance number for the classes, they tend to perform much more poorly on instance-rare classes than on instance-rich ones. To alleviate this performance bias, existing approaches either re-balance the data distribution through sampling (Chawla et al., 2002; Han et al., 2005; Shen et al., 2016; Mahajan et al., 2018) or the loss for each class (Cui et al., 2019; Khan et al., 2017; Cao et al., 2019; Khan et al., 2019). Recently, Kang et al. (2020) demonstrate that learning good feature spaces is crucial for long-tailed recognition. Inspired by our findings on the benefits of a balanced feature space for generalization, we introduce the k-positive contrastive learning method to solve the long-tailed recognition problem by learning more balanced feature spaces.

3 Balanced Feature Spaces from Contrastive Learning

In this section, we systematically study the performance of representation models trained by SSL from a collection of training datasets with varying instance number distributions, in contrast with the models learned by supervised learning methods, to explore how SSL performs when the training datasets are not artificially balanced. Furthermore, we investigate the generalization performance of these learned representation models under multiple settings, in order to explore the relationship between the representation model's generalizability and the property of its learned feature space.

Notations We define the notations used in this paper. Representation learning aims to obtain a representation model f_θ that maps a sample x_i into a feature space V such that its corresponding representation $v_i \in V$ encapsulates desired features for target applications. Let $\mathcal{D}_{\text{rep-train}} = \{x_i, y_i\}$, $i = 1, \ldots, N$ be the dataset for training the representation model, where y_i is the class label for sample x_i. Let C denote the number of total classes and n_j denote the number of instances within class j. We use $\{q_1, \ldots, q_C\}$ with $q_j = n_j/N$ to denote the discrete instance distribution over the C classes. An *imbalanced* dataset has significant difference in the class instance numbers, *e.g.*, $q_1 \gg q_C$. We use a multi-layer convolutional neural network $f_\theta(\cdot) : x_i \mapsto v_i$ to implement the representation model. The final classification prediction $\hat{y}$ is given by a linear classifier $\hat{y} = \arg\max[W^\top v + b]$, where W denotes the classifier weight matrix and b denotes the bias term.

3.1 Methodology of Our Study

Representation learning methods Various loss functions have been developed for learning the representation model f_θ on the training dataset. Among them, the most popular one is the supervised cross-entropy (CE) loss:

$$\mathcal{L}_{\text{CE}} = \frac{1}{N} \sum_{i=1}^{N} -\log p_{y_i}, \tag{1}$$

where $p_{y_i} = \text{softmax}(W_{y_i}^\top v_i + b)$ is the normalized probability prediction of sample i belonging to its ground truth class y_i. Using the semantic labels directly as supervision signal (y_i in Equation 1), the representation model trained by the CE loss can have strong semantic discrimination ability but its generated feature space is easily biased by the imbalance of the training instance distribution—if some classes have significantly more training instances than the others, their data representations

will occupy dominant portion of the feature space (Figure 1) and get higher classification accuracy than the instance-rare classes (Kang et al., 2020; Wang et al., 2017).

Different from the supervised learning ones, self-supervised learning methods adopt semantic-free loss functions to learn representations from unlabeled data (He et al., 2020; Gidaris et al., 2018). For example, the contrastive loss[1] (CL) (Oord et al., 2018) learns representations via maximizing the instance-wise discriminativeness:

$$\mathcal{L}_{\text{CL}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(v_i \cdot v_i^+/\tau)}{\exp(v_i \cdot v_i^+/\tau) + \sum_{v_i^- \in V^-} \exp(v_i \cdot v_i^-/\tau)}, \tag{2}$$

where τ is a temperature hyper-parameter, v_i^+ is a positive sample for the anchor instance i (typically produced by data augmentation), $v_i^- \in V^-$ is the negative sample randomly drawn from the training samples excluding instance i. This contrastive loss encourages the feature representations from positive pairs to be similar, while pushing features from the sampled negative pairs apart.

We take the two loss functions ($\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{CL}}$) as representatives to study how the representation models (and the corresponding feature spaces) trained with supervised/self-supervised methods are affected by the training instance distribution $\{q_1, \dots, q_C\}$.

Balancedness of feature spaces Since semantic labels are not involved in the contrastive loss (Equation 2), we hypothesize it may lead to representation models yielding feature spaces that are less biased by the imbalance of the training dataset, compared with the ones from the supervised loss (Equation 1). To verify this, we introduce a metric to characterize such an "unbiased" or "balanced" property of a feature space at first. A feature space V is *balanced* if the representations $\{v_i\}$ from different classes within it have similar degrees of linear separability. As the linear separability degree of the representations is usually evaluated by the accuracy of a linear classifier over them (Vapnik, 2013), we follow this criterion to develop the balancedness metric. Specifically, let $a_1, \dots, a_C$ denote the classification accuracy of a linear classifier (W, b) over the representations $\{v_i\} \subset V$ from C classes. We take the following uniformity of these accuracies as the *balancedness* of the feature space V:

$$\beta(V) \triangleq \frac{1}{C^2}\sum_{i,j}^{C} \exp\left(-\frac{|a_i - a_j|^2}{\sigma}\right), \text{ where } a_j = \frac{\#\{v_i | \hat{y}_i = j, y_i = j, v_i \in V\}}{\#\{v_i | y_i = j, v_i \in V\}}. \tag{3}$$

Here σ is a fixed scaling parameter. This metric achieves its maximum when all the class-wise accuracies are equal, *i.e.*, there being no separability bias of the learned representations to any class.

Experimental protocol We adopt a multi-stage protocol for learning and evaluating the feature spaces. (1) *Representation learning*: pre-train the representation model f_θ on the provided training set $\mathcal{D}_{\text{rep-train}}$ using the above training losses $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{CL}}$; (2) *Classifier learning*: train a linear classifier (W, b) on top of f_θ with θ fixed using another training dataset $\mathcal{D}_{\text{train}}$[2] and the supervised CE loss; (3) *Representation evaluation:* evaluate the classification accuracy of the learned classifier on the test dataset $\mathcal{D}_{\text{test}}$ with the representations from f_θ and compute the above balancedness $\beta(V)$.

To thoroughly investigate sensitiveness of different representation learning methods to the imbalance level of training datasets, we construct six datasets from the long-tailed benchmark ImageNet-LT (Liu et al., 2019) ($\mathcal{D}_{\text{LT}}$) by varying its instance distribution $\{q_1, \dots, q_C\}$ from a long-tailed one to a uniform one gradually, while keeping the total instance number similar. The generated datasets, denoted as $\mathcal{D}_{\text{LT0}}, \dots, \mathcal{D}_{\text{LT8}}, \mathcal{D}_{\text{LT}}$ (which are increasingly more imbalanced), are used as $\mathcal{D}_{\text{rep-train}}$ for representation learning in the following experiments. See appendix for their details.

3.2 Contrastive Loss Helps Learn Balanced Feature Spaces

We first investigate classification performance of the representation models trained with the CE and CL losses on the above six datasets $\mathcal{D}_{\text{LT0}}, \dots, \mathcal{D}_{\text{LT8}}, \mathcal{D}_{\text{LT}}$ that are increasingly more imbalanced.

[1] The term of *contrastive loss* has been used to refer to various loss functions over positive and negative samples. This work focuses on the specific form in Equation 2 that is widely used in modern SSL methods.

[2] Note $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{rep-train}}$ can be the same dataset as in recent SSL works (*e.g.*, He et al. (2020)).

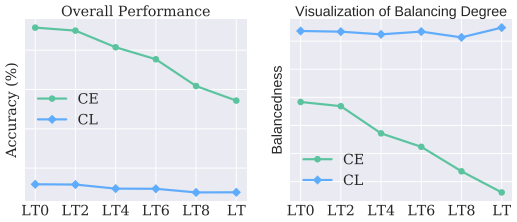

Figure 2: Classification accuracy (left) and balancedness (right) of the representations learned from cross-entropy (CE) loss and contrastive loss (CL) on datasets (LT0 to LT) with increasing imbalance.

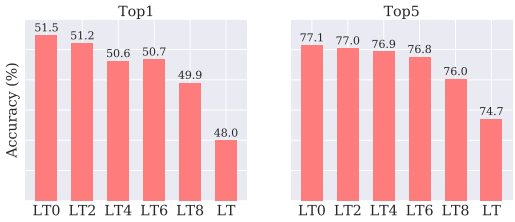

Figure 3: Out-of-distribution generalization on ImageNet. Top 1 and Top 5 testing accuracy of the model are learned from datasets LT0 to LT that are increasingly more imbalanced.

Since linear classifiers are easily biased by skewed training dataset distribution (Kang et al., 2020), it is necessary to eliminate the imblancedness of the evaluation datasets for reliable representation evaluation. Thus, we use the (balanced) training and test sets of ImageNet as $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ to learn classifiers and evaluate their classification accuracy, following the above protocol.

The results are summarized in Figure 2, from which we make an important observation: compared with the supervised cross-entropy loss, **the model trained with the unsupervised contrastive loss generates a more balanced feature space**, even in presence of highly imbalanced training instance distribution. As shown in Figure 2 (left), the classification accuracy of representation models learned with the CE loss drops quickly when the dataset becomes more imbalanced—the quality of representations from these models is very sensitive to the imbalance of training datasets. In contrast, the classification accuracy of CL-trained models remains stable even when the training dataset transits to a heavily long-tailed one. Such surprising performance robustness to imbalance of the training datasets implies that using contrastive learning can consistently learn balanced feature spaces. To see this, we also visualize the balancedness scores (Equation 3) of the learned feature spaces from CE and CL in Figure 2 (right). Even when the training set is heavily long-tailed, the feature spaces learned with CL loss are as highly balanced as the ones learned from a uniform training distribution, while the blancedness score of the feature spaces from CE loss is lower and drops quickly. Such a balanced feature space offered by CL loss is much desired for the representation learning in practice, where the training instance distribution is usually long-tailed. Certainly, using an unsupervised loss will sacrifice semantic discriminativeness of the representations, leading to the accuracy gap between the CE and CL-trained models.

3.3 More Balanced Representation Models Generalize Better

The above studies reveal that the representation models trained with the contrastive loss can produce more balanced feature spaces. A natural question is what are the benefits from a balanced model for recognition? Since a pre-trained representation model is often used to facilitate downstream tasks (He et al., 2020; Newell & Deng, 2020; Hénaff et al., 2019), we here conduct extensive experiments to study its potential benefits on model generalization performance under the following settings.

Out-of-distribution generalization We first study the relationship between the balancedness of representation models and their generalizability for recognizing new classes. To thoroughly evaluate performance of representation models with different balancedness, we evenly divide the 1,000 classes into two splits (500 vs. 500 classes) on ImageNet, referred to as the *source* and *target* split respectively. We use the subsets (corresponding to the source class split) of the above $\mathcal{D}_{\text{LT0}}, \ldots, \mathcal{D}_{\text{LT8}}, \mathcal{D}_{\text{LT}}$ datasets to construct six different $\mathcal{D}_{\text{rep-train}}$ for training the representation model f_θ. To obtain models with different balancedness, we use the CE loss for training, since the above studies reveal using the CL loss will always produce models with similar balancedness (Figure 2). We use the subsets (corresponding to the target class split) of the training and test sets of ImageNet as $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ for classifier learning and evaluation, with the representation model fixed.

The testing performance of the models with different balancedness on the target classes is presented in Figure 3. It is observed that as the source dataset becomes increasingly more imbalanced (from $\mathcal{D}_{\text{LT0}}$ to $\mathcal{D}_{\text{LT}}$) and the corresponding models become more imbalanced, their generalization performance degrades correspondingly. Such a positive correlation between balancedness of the models and testing accuracy on the target classes clearly demonstrate that **more balanced representation**

Table 1: Results on Places365, VOC and COCO. AP_{50} is the default metric for VOC, while AP^{bb} and AP^{mk} denote the bounding-box and mask AP for COCO respectively. Black / gray numbers correspond to results of the representation models trained on ImageNet-LT / ImageNet respectively. See appendix for complete results.

	cross-domain	*cross-task*		
	Places365 (Top1)	VOC (AP_{50})	COCO (AP^{bb})	COCO (AP^{mk})
CE	38.50 / 46.06	76.45 / 81.26	38.13 / 40.08	33.29 / 34.85
CL	41.24 / 46.16	78.19 / 82.28	39.67 / 40.41	34.73 / 35.14
$\Delta_{\text{CL,CE}}$	**+2.74** / +0.10	**+1.64** / +0.02	**+1.54** / +0.33	**+1.44** / +0.29

models tend to generalize better for recognizing unseen classes. More details and results about the out-of-distribution generalization studies are deferred to the appendix.

Cross-domain and cross-task generalization We then explore whether learning balanced representation models is able to benefit model's generalizability to new domains and tasks. We use the ImageNet-LT as $\mathcal{D}_{\text{rep-train}}$ to train the models with the CE and CL losses, obtaining imbalanced and balanced representation models respectively. For the cross-domain setting, we train a linear classifier on the Places365 dataset (Zhou et al., 2017) with the representation model fixed. From the results in Table 1, it can be clearly observed that the balanced representation model (from CL) surpasses the less balanced one (from CE) significantly, in terms of the top-1 accuracy (by 2.74%). For the cross-task setting, we take train/test splits of the PASCAL VOC (Everingham et al., 2010) and COCO (Lin et al., 2014) datasets as $\mathcal{D}_{\text{train}}/\mathcal{D}_{\text{test}}$ for evaluating detection performance. The results are given in Table 1. Again, the balanced model (from the CL loss) outperforms the less balanced one (from the CE loss) significantly (up to 1.64%). In comparison, the improvement from the CL-trained model over the CE-trained model is moderate (around 0.3%) when using the full ImageNet for training, as CE can learn a relatively balanced feature space from a balanced dataset. This clearly shows that the generalization performance boost for the cross-domain and cross-task settings brought from CL-trained models does not simply stem from using self-supervised pre-training, but indeed come from learning more balanced feature spaces.

4 Learning Balanced Feature Spaces for Recognition

The above studies demonstrate the representation models trained with the contrastive loss can generate balanced feature spaces showing strong generalizability. Here we explore how to effectively leverage these findings in practice. We introduce a new method that inherits the strength of the contrastive loss in learning balanced feature spaces and enhances the feature spaces' semantic discrimination capability simultaneously. We thoroughly study its superiority via two application cases, *i.e.*, the long-tailed recognition and pre-training representation models for downstream tasks.

4.1 K-Positive Constrastive Loss

Though balanced, the feature spaces from contrastive learning have limited capability of semantic discrimination, as shown in Figure 2 (left). This is because the contrastive loss blindly encourages instance-level discrimiantiveness. Every two instances, even if they are from the same class, are forced to be apart from each other in the learned feature space. To embed semantic discriminativeness into the representations while maintaining desired balancedness, we develop a new method to leverage the provided semantic labels to adaptively compute the instance contrastive loss.

Concretely, given an anchor training instance x_i with its semantic label y_i, our proposed method draws k instances from the same class to form the positive sample set $V^+_{i,k}$, instead of only using its augmentation as in Equation 2. Thus, it gives a new loss called k-positive contrastive loss (KCL):

$$\mathcal{L}_{\text{KCL}} = \frac{1}{N(k+1)} \sum_{i=1}^{N} \sum_{v_j^+ \in \{\tilde{v}_i\} \cup V^+_{i,k}} -\log \frac{\exp(v_i \cdot v_j^+/\tau)}{\exp(v_i \cdot \tilde{v}_i/\tau) + \sum_{v_j \in V_i} \exp(v_i \cdot v_j/\tau)}, \tag{4}$$

where $\tilde{v}_i$ is generated by augmenting v_i, V_i is the current batch of examples excluding v_i, and $V^+_{i,k} \subset V_i$ is a positive set containing k instances randomly drawn from the same class as v_i. The proposed KCL loss purposely keeps the number of positive instances equal, which is crucial for balancing the learned feature spaces. It brings two benefits. First, it helps learn representations

Table 2: ImageNet-LT results

Method	Many	Medium	Few	All
OLTR (Liu et al., 2019)[a]	35.8	32.3	21.5	32.2
Joint (SL-i) (Kang et al., 2020)	64.9	35.2	6.8	42.5
τ-norm (Kang et al., 2020)	56.6	44.2	27.4	46.7
cRT (Kang et al., 2020)	58.8	44.0	26.1	47.3
FCL	61.4	47.0	28.2	49.8
KCL	61.8	49.4	30.9	**51.5**

[a]Reproduced by re-running their code with ResNet50.

Table 3: iNaturalist 2018 results

Method	Top1
CB-Focal (Cui et al., 2019)	61.1
LDAM (Cao et al., 2019)	64.6
LDAM+DRW (Cao et al., 2019)	68.0
cRT (Kang et al., 2020)	65.2
τ-norm (Kang et al., 2020)	65.6
BBN (Zhou et al., 2020)	66.3
FCL	66.4
KCL	**68.6**

with stronger discriminative ability as it leverages the label information as supervised learning. Secondly, it uses the same number of instances (*i.e.*, k) for all the classes in positive pair construction which further balances the learned feature space. Note our proposed KCL is different from the supervised contrastive learning (Khosla et al., 2020) that leverages all the instances from the same class to construct the positive pairs, which cannot avoid the dominance of instance-rich classes in the representation learning. This is also evidenced by our following experiments on long-tailed recognition. In the following experiments, we choose $k = 6$ via validation and use ResNet50 as the backbone. Other hyper-parameter choices and implementation details are given in the appendix.

4.2 Long-Tailed Recognition

KCL provides feature spaces with desirable balancedness and semantic discriminativeness, which makes it naturally fit for addressing the challenges of long-tailed recognition, *i.e.*, severe performance bias to the instance-rich classes and poor generalization to the instance-rare classes (Mahajan et al., 2018; Zhong et al., 2019). Here we implement and evaluate KCL for long-tailed recognition, following the two-stage training strategy from Kang et al. (2020): 1) train the representation model with the KCL loss; 2) learn a linear classifier with cross-entropy loss and class-balanced sampling.

Baselines Besides well established state-of-the-arts, we consider following three kinds of baselines for justifying the advantages of KCL. (1) *Classifier balancing* methods, *i.e.*, τ-norm and cRT (Kang et al., 2020), that re-train classifiers with class-balanced sampling as KCL but learn the representation models by supervised cross-entropy loss. Comparison with them helps understand the effectiveness of learning balanced features in long-tailed recognition. (2) Methods that train the representation model and classifier jointly with cross-entropy loss (SL) and various data re-sampling strategies, including instance-balanced (SL-i), class-balanced (SL-c), progressively-balanced (SL-p) and square-root re-sampling (SL-s) (Kang et al., 2020). Comparison with them will show advantages of KCL over these data-enriching strategies in feature space balancing. (3) *A full-positive variant of KCL*, named full-positive contrastive learning (FCL), that uses all the available same-class samples in the current batch to construct positive pairs for computing the contrastive loss, which is similar to the supervised contrastive learning (Khosla et al., 2020). Comparing KCL with FCL will show the benefits of keeping the number of positive samples equal for all the anchor instances in KCL.

Results We evaluate KCL and compare it with the above strong baselines on two large-scale benchmark datasets, ImageNet-LT (Liu et al., 2019) and iNaturalist 2018 (iNatrualist, 2018). For comprehensive evaluation, following (Liu et al., 2019), we split the classes of ImageNet-LT into many-shot (>100 images), medium-shot (20∼100 images) and few-shot (<20 images) groups. The results are summarized in Tables 2 and 3 respectively, along with the balancedness of these methods on ImageNet-LT in Figure 4. We make the following observations.

More balanced feature spaces give better performance. We first compare KCL with cRT and τ-norm, the latest state-of-the-arts with feature spaces learned by supervised cross-entropy loss and thus less balanced as demonstrated in Sec. 3. Compared with them, KCL improves the overall accuracy by a large margin (4.2% on ImageNet-LT and 3% on iNaturalist), demonstrating the importance of learning more balanced feature spaces for long-tailed recognition.

KCL is more effective at learning balanced and discriminative feature spaces. Data re-sampling is widely used as a straightforward approach to alleviate performance bias for long-tailed recognition (Kang et al., 2020). We compare the feature space balancedness of KCL and SL methods with different data re-sampling strategies on ImageNet-LT in Figure 4. Clearly, data re-sampling cannot

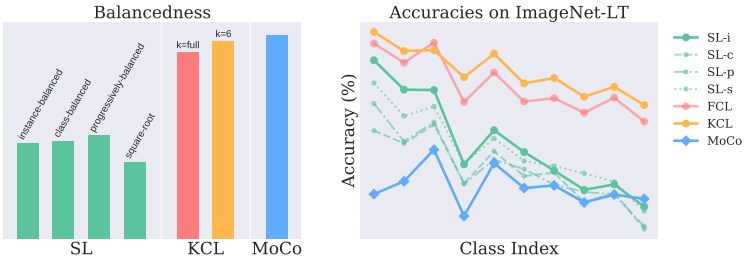

Figure 4: Comparison of different methods on their feature space balancedness (left) and class-wise accuracy (right). Here the linear classifier is fine-tuned on the full ImageNet for representation evaluation (see Sec. 3.2).

effectively improve balancedness of the feature space as KCL. Besides data re-sampling, Figure 4 also shows the balancedness of the feature spaces learned by the latest contrastive learning method MoCo (He et al., 2020) on ImageNet-LT. MoCo can balance the feature space but has lower accuracy, due to the lack of semantic discriminativeness in the learned feature space. KCL performs the best, demonstrating its effectiveness at learning both balanced and discriminative feature spaces.

Equalizing the number of positive instances in KCL is important. To further justify the design of KCL loss in keeping the number of positive instances to be equal, we compare it with its variant FCL. From Tables 2, 3 and Figure 4, though FCL outperforms other baselines, its performance is inferior to KCL, in terms of both the overall accuracy and the balancedness of the learned feature spaces. Equalizing the number of positive instances as KCL is crucial for learning balanced feature spaces and improving recognition performance.

4.3 Pre-training Representation Models for Downstream Tasks

The effectiveness of KCL is not limited to the cases where training datasets are imbalanced. Here we apply KCL for pre-training the representation model on balanced datasets which is later fine-tuned for downstream tasks, including the out-of-distribution (OOD) recognition (the source/target split[3] of ImageNet for pre-training/evaluation) and detection (ImageNet: pre-training; PASCAL VOC and COCO: evaluation). Note we aim to study the generalizability of KCL as a representation learning method, rather than aiming at state-of-the-art performance. Hence we compare it with the vanilla supervised cross-entropy loss (SL) and MoCo (which the KCL is built on) (He et al., 2020). The results are summarized in Table 4. We also evaluate the discrminativeness of the learned representations (the "repr" in the table) from their classification accuracy by learning a linear classifier on the pretraining datasets. Clearly, KCL outperforms SL and MoCo for both downstream tasks. This is because KCL learns more balanced feature spaces than SL with similar discriminativeness, and learns more discriminative features than MoCo.

Table 4: Comparison of different representation learning methods for the downstream tasks.

	OOD Recognition		Object Detection			
	repr	Top1	repr	VOC (AP_{50})	COCO (AP^{bb})	COCO (AP^{mk})
SL	81.2	70.7	76.6	81.26	40.08	34.85
MoCo	67.0	60.1	60.6	81.28	40.41	35.15
KCL	81.4	**74.8**	76.8	**82.32**	**40.79**	**35.45**

5 Conclusions

This work piloted studies on performance of the self-supervised learning methods for imbalanced datasets, and made several intriguing findings. At the heart of these findings is the balanced feature space, which is identified to be an inherent property of the representations learned by the contrastive learning and bring stronger generalizability. It provides a new perspective for understanding the behavior of the contrastive learning. This work further developed a new representation learning method to leverage the benefits of balanced feature spaces. We believe the findings and method developed here are inspiring for the future research on representation learning. However, theoretical understandings on balanced feature spaces are not mature yet and worthy of future exploration.

[3]Details of the class split are given in Sec. 3.3.

REFERENCES

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

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

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

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.

Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060, 2017.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer, 2005.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

iNatrualist. The inaturalist 2018 competition dataset. https://github.com/visipedia/inat_comp/tree/master/2018, 2018.

Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2733–2742, 2018.

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.

Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pp. 577–593. Springer, 2016.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6874–6883, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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 Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.

Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7345–7354, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

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

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pp. 467–482. Springer, 2016.

Merrielle Spain and Pietro Perona. Measuring and predicting importance of objects in our visual world. 2007.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pp. 7029–7039, 2017.

Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1910–1919, 2019.

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.

Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9719–9728, 2020.

George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*, volume 21. Psychology Press, 1999.

A IMPLEMENTATION DETAILS

Representation Learning For supervised cross-entropy (CE) loss, we adopt the standard PyTorch distributed training implementation[4]. For unsupervised contrastive loss (CL), we use the official implementation[5] of MoCo (He et al., 2020) with default hyper-parameters. Our k-positive contrastive loss is implemented based on MoCo by randomly selecting k positive examples from the key memory for each of the query example. When KCL is applied for a long-tailed dataset, there might be less than k positive examples in the memory for some of the classes. In such cases, we use all the positive examples. Besides, we keep all the hyper-parameters (*e.g.*, data augmentation, learning rate and batch size) the same as supervised learning. We only carefully tune the number of training epochs for KCL to make it achieve similar performance as supervised learning on a balanced dataset (full Imagenet) for fair comparison (see Table 5). As a result, throughout the paper our KCL is trained for 200 epochs while its supervised counterpart is trained for 90 epochs. This is reasonable as contrastive learning usually takes much longer to converge (He et al., 2020). We are using $k = 6$ for KCL throughout the paper, which is carefully tuned on the validation set of ImageNet-LT, as shown in Fig. 5. The detailed hyper-parameters of different loss functions are given in Table 6.

Table 5: Results on full ImageNet.

	Top1	Top5
CE (90 epochs)	76.616	93.090
SCL (200 epochs)	76.976	92.972
KCL (200 epochs)	76.814	92.936

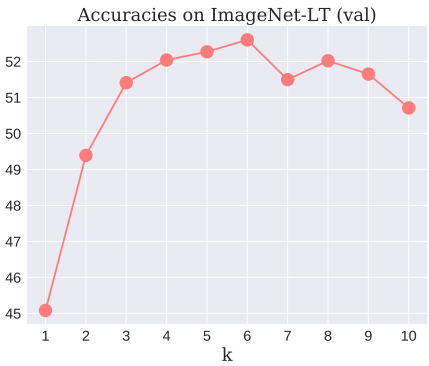

Figure 5: Validation accuracies of KCL on ImageNet-LT as the value of k varies.

hps	CE	CL	KCL
epochs	90	200	200
batch size	256	256	256
learning rate	0.1	0.03	0.1
learning rate schedule	cosine	step	cosine
data augmentation	default	moco v1	default
memory size	-	65536	65536
encoder momentum	-	0.999	0.999
feature dimension	-	128	128
softmax temperature	-	0.07	0.07
k	-	-	6

Table 6: Hyper-parameters used by different loss functions. "default" means the standard data augmentation strategies used by supervised learning.

Classifier Learning We need to train linear classifiers in two cases. 1) *representation evalauation* (Sec. 3 and Figure 4). For CE-learned representations, we train a linear classifier using the same parameters as representation learning in Table 6 with a smaller number of epochs (10). For CL and KCL, we adopt the classifier training protocol introduced by MoCo (He et al., 2020) with default hyper-parameters (*i.e.*, learning rate 30 and weight decay 0). 2) *Long-tailed recognition* (Sec. 4.2). We obtain a re-balanced classifier for CE representations following Kang et al. (2020), and adopt the MoCo classifier learning strategy with class-balanced sampling by setting the learning rate to 10 on ImageNet-LT and 30 on iNaturalist 2018.

Detection model training We use exactly the same setting and evaluation metrics as He et al. (2020). R50-C4 backbone is used with BN tuned. The image is rescaled to [640, 800] during training and 800 at inference. All layers are fine-tuned end-to-end with batch size = 16. For Pascal VOC, we train Faster R-CNN (Ren et al., 2015) on `trainval07+12` set with 24k schedule and evaluate on `test07` set. For COCO, we train Mask R-CNN (He et al., 2017) on `train2017` set with $\times 2$ schedule and evaluate on `val2017` set.

[4] https://github.com/pytorch/examples/tree/master/imagenet
[5] https://github.com/facebookresearch/moco

B DATASET CONSTRUCTION

Datasets for studying balancedness of feature spaces We here explain the details on the construction of the series of datasets used in our study in Sec. 3.

In particular, we take the standard long-tailed training set from ImageNet-LT (Liu et al., 2019) whose instances follow the Pareto distribution as the base dataset, denoted as $\mathcal{D}_{\text{LT}}$. We vary its training instance distribution $\{q_1, \dots, q_C\}$ gradually to obtain different datasets as follows,

$$n_j = \left\lfloor N_{\mathcal{D}} \times \frac{q_j^{\alpha}}{\sum_k q_k^{\alpha}} + \frac{1}{2} \right\rfloor, \tag{5}$$

where $N_{\mathcal{D}}$ is the total number of training instances in $\mathcal{D}_{\text{LT}}$, $\alpha \in [0, 1]$ controls the dataset balancedness. When $\alpha = 0$, it corresponds to a fully balanced dataset; when $\alpha = 1$, it becomes a heavy long-tailed ones. In total, we generated 6 datasets with $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, denoted as $\mathcal{D}_{\text{LT0}}, \dots, \mathcal{D}_{\text{LT8}}, \mathcal{D}_{\text{LT}}$ respectively, as different examples of $\mathcal{D}_{\text{rep-train}}$ for representation learning. The detailed statistics and visualization of the datasets $\mathcal{D}_{\text{LT0}}, \dots, \mathcal{D}_{\text{LT8}}, \mathcal{D}_{\text{LT}}$ are summarized in Table 7 and Fig. 6.

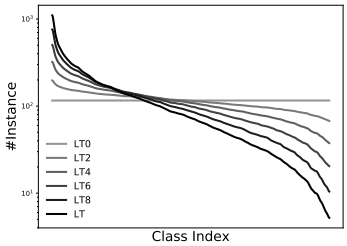

Figure 6: Training instance number distributions of the datasets we use in our empirical studies.

Dataset	Max	Min	Total
$\mathcal{D}_{\text{LT0}}$	115	115	115,000
$\mathcal{D}_{\text{LT2}}$	204	67	115,885
$\mathcal{D}_{\text{LT4}}$	343	37	115,801
$\mathcal{D}_{\text{LT6}}$	553	20	115,836
$\mathcal{D}_{\text{LT8}}$	857	10	115,852
$\mathcal{D}_{\text{LT}}$	1,280	5	115,846

Table 7: Dataset statistics on training instance numbers, including the maximal and minimal instance number per class and the total number.

Datasets for generalizability studies We carefully choose the proper datasets to construct the $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ for evaluating generalizability of the representation models under multiple settings. The choices are summarized in Table 8.

Table 8: Summary on used datasets for representation model pre-training and evaluation in our studies.

study	$\mathcal{D}_{\text{rep-train}}$	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{test}}$
Balancedness (Sec. 3.2)	$\mathcal{D}_{\text{LT0}}, \dots, \mathcal{D}_{\text{LT8}}, \mathcal{D}_{\text{LT}}$	ImageNet (train)	ImageNet (val)
Out-of-distribution (Sec. 3.3)	source split of above	target split of above	
Cross-domain (Sec. 3.3)	ImageNet-LT	Places 365 (train)	Places 365 (test)
Cross-task (Sec. 3.3)	ImageNet-LT	VOC/MSCOCO (train)	VOC/MSCOCO (test)

C ADDITIONAL RESULTS ON MODEL GENERALIZATION PERFORMANCE

We now provide more detailed generalization results of models learned with different loss functions, involving supervised cross-entropy (CE) loss, unsupervised contrastive loss (CL) and k-positive contrastive loss (KCL).

C.1 OUT-OF-DISTRIBUTION GENERALIZATION

We first consider the OOD generalization from one *imbalanced* dataset to another dataset without overlapping class between them, as discussed in Sec. 4.3. In the experiments, the *Split-overlap* (split with semantic overlap) scheme is adopted that allows the classes within the two splits to share the same super class (*e.g.*, *dog* and *wolf* from *canidae* are put into different splits) in the ImageNet ontology. As such, though the target classes are all novel to the representation model, some of

Table 9: *OOD generalization results (top-1 accuracy) on balanced datasets.* We use the *source* classes of origninal ImageNet to learn representation network (ResNet50), and use the *target* classes and *all* classes respectively to learn linear classifiers for evaluation.

	Split-overlap			*Split-independent*		
	source	target	all	source	target	all
CE	81.2	70.7	67.2	82.8	50.3	62.4
CL	67.0	60.1	58.3	68.2	54.8	58.2
KCL	81.4	74.8 (+4.1)	70.8 (+3.6)	83.2	58.1 (+3.3)	67.2 (+4.8)

their attributes have been seen by the model before from the source classes. The results for the split-overlap are presented in the main text (Table 4).

We have shown that a balanced feature space gives better generalization ability when when representation learning is performed on imbalanced datasets. A natural question is whether the conclusion still holds for large-scale balanced datasets, such as ImageNet? To answer this, we split the whole ImageNet into two halves using the above protocol and train the representation model on a balanced source half of ImageNet and evaluate it on the target half.

In addition to the *Split-overlap*, we consider another split scheme *Split-independent* (split without semantic overlap) strictly avoids classes from the same super classes to be distributed into different splits. *Split-independent* presents a more challenging case for model's generalization ability as all the target classes (and attributes) to recognize are novel.

The generalization performance comparsion of the representation models with different methods is given in Table 9. When the source classes and target classes share similar semantics (on *split-overlap*), the CE-trained model surpasses the CL-trained model on both the source and target classes. But when looking into the generalization gap (*i.e.*, the difference between the source and target accuracy), the CL-trained model suffers larger generalization gap than the CL-trained model (10.5 vs. 1.8). When there is not semantic overlap between source and target (on *split-independent*), the CL-model outperforms CE-model on the target classes by 4.3% with much smaller generalization gap (13.4 v.s. 32.5). By comparing the "full" performance from *split-overlap* to *split-independent*, one can observe that CL loss performs consistently well (58.3 and 58.1), but CE drops as large as 5%. This implies that CL is robust to imbalanced training distribution used for representation learning, while CE is extremely sensitive to it. These results clearly demonstrate *the consistent superiority of balanced representation learning in terms of generalization* for various training dataset distribution. Notably, our proposed KCL loss surpasses both CE and CL loss on all the four different settings by a large margin (more than 3 points). These results clearly demonstrate that KCL is able to learn a balanced and discriminative feature space, and balancedness is a general property that benefits both balanced and imbalanced datasets.

C.2 Cross-domain and Cross-task Generalization

We also evaluate the generalization ability of the representation models trained on the balanced full ImageNet datasets, for cross-domain and cross-task applications. The results are given in Table 10 (cross-domain) and Tables 11 and 12 (for detection) respectively. From Table 10, when the training datasets are balanced, the models trained with CL and CE achieve comparable performance. While when the training datasets are not balanced, the CL model significantly outperforms the CE model (Table 1). This demonstrates that the CL loss can consistently produce balanced representation models and the model generalization performance can indeed benefit from being balanced.

Similar conclusion can be drawn for the cross-task generalization. From Tables 11 and 12, when training the model on the full ImageNet dataset, using self-supervised CL loss can produce the model performing slightly better than using the supervised CE loss. On PASCAL VOC, the performance advantage is as marginal as 0.02% in AP_{50}. In contrast, when training the model on the ImageNet-LT dataset, using CL loss can boost the model performance over using the CE loss much more significantly. The improvement is as large as 1.64% in AP_{50}. Thus the performance benefit on the generalization to detection brought by CL does not simple stem from using self-supervised pre-training, but indeed come from learning more balanced feature spaces.

Similar to OOD generalization, the model learned with KCL gives better downstream performance on both VOC and COCO, which mean enforcing feature space balancedness with KCL is indeed able to help learning better representation.

Table 10: Results on Places 365. The encoder is trained on ImageNet.

	Top1	Top5
CE	46.06	77.11
CL	46.16 (+0.1)	76.27 (-0.84)

Table 11: Object detection Results on PASCAL VOC. The representation model is trained on ImageNet and ImageNet-LT. We report results in AP_{50}: VOC metric; AP: COCO-style metric.

	ImageNet			ImageNet-LT		
	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}
CE	81.26	53.66	59.19	76.45	48.53	51.01
CL	81.28 (+0.02)	56.10 (+2.44)	62.71 (+3.52)	78.19 (**+1.64**)	51.52 (**+2.99**)	56.48 (**+5.47**)
KCL	82.32 (**+1.06**)	55.51 (**+1.85**)	62.05 (**+2.86**)	79.70 (**+3.25**)	52.63 (**+4.10**)	57.89 (**+6.88**)

Table 12: Object detection Results on COCO. The representation model is trained on ImageNet and ImageNet-LT. We report results in bounding-box AP (AP^{bb}) and mask AP (AP^{mk}).

		ImageNet			ImageNet-LT		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
AP^{bb}	CE	40.08	59.76	43.29	38.13	57.38	41.15
	CL	40.41 (+0.33)	60.05 (+0.29)	44.09 (+0.80)	39.67 (**+1.54**)	59.40 (**+2.02**)	42.73 (**+1.58**)
	KCL	40.79 (**+0.78**)	60.63 (**+0.87**)	43.99 (**+0.70**)	39.43 (**+1.30**)	59.08 (**+1.70**)	42.56 (**+1.41**)
AP^{mk}	CE	34.85	56.60	37.02	33.29	54.24	35.38
	CL	35.14 (+0.29)	56.88 (+0.28)	37.56 (+0.54)	34.73 (**+1.44**)	56.07 (**+1.83**)	37.13 (**+1.75**)
	KCL	35.45 (**+0.60**)	57.40 (**+0.80**)	37.80 (**+0.78**)	34.38 (**+1.09**)	55.81 (**+1.57**)	36.39 (**+1.01**)