Beyond Instance Discrimination: Relation-Aware Contrastive Self-Supervised Learning

Yifei Zhang, Chang Liu, Yu Zhou, Weiping Wang, Qixiang Ye, *Senior Member, IEEE*, and Xiangyang Ji, *Member, IEEE*

***Abstract*—Contrastive self-supervised learning (CSL) based on instance discrimination typically attracts positive samples while repelling negatives to learn representations with pre-defined binary self-supervision. However, vanilla CSL is inadequate in modeling sophisticated instance relations, limiting the learned model to retain fine semantic structure. On the one hand, samples with the same semantic category are inevitably pushed away as negatives. On the other hand, differences among samples cannot be captured. In this paper, we present relation-aware contrastive self-supervised learning (ReCo) to integrate instance relations, *i.e.*, global distribution relation and local interpolation relation, into the CSL framework in a plug-and-play fashion. Specifically, we align similarity distributions calculated between the positive anchor views and the negatives at the global level to exploit diverse similarity relations among instances. Local-level interpolation consistency between the pixel space and the feature space is applied to quantitatively model the feature differences of samples with distinct apparent similarities. Through explicitly instance relation modeling, our ReCo avoids irrationally pushing away semantically identical samples and carves a well-structured feature space. Extensive experiments conducted on commonly used benchmarks justify that our ReCo consistently gains remarkable performance improvements.**

***Index Terms*—Global distribution relation, local interpolation relation, relation-aware contrastive self-supervised learning, self-supervised learning.**

I. Introduction

In THE deep learning era, large-scale pre-training [1], [2] then downstream fine-tuning has become a dominant learning paradigm [3], [4], [5], [6]. However, supervised pre-training typically focuses on task-specific features, resulting in limited model generalization. Building finely annotated large-scale datasets is also laborious, expensive, and sometimes impractical. Inspired by human cognition from unlabeled data, unsupervised visual representation learning is attracting growing attention [7], [8], [9], [10], [11].

Mainstream approaches either manually design specific pretext tasks to assimilate the intrinsic data structure [8], [12], [13], [14], [15], or encode data similarities with a contrastive self-supervised learning (CSL) paradigm [9], [16], [17], [18]. Unlike handcrafted pretext tasks that are limited in exhausting correlating human priori, CSL with instance discrimination aims at learning view-invariant representation, which presents superior performance and great potential [9], [18], [19], [20], [21]. Based on InfoNCE loss [16], ISIF [19], MoCo [9] and SimCLR [18] introduce siamese networks to attract different instance views as positives while repelling other instances in a mini-batch or a memory bank as negatives. However, since negative samples are naively defined as different images, false negatives with the same semantic content inevitably occur, and their specific similarity relations are also not taken into account. Models learned with "hard" binary positive and negative assignments are apparently limited by biased and incomplete semantic structure learning of the data.

In this article, we propose a simple yet effective **re**lation-aware **co**ntrastive self-supervised learning (ReCo) approach to concurrently explore "soft" instance relations of global distribution and local interpolation, Fig. 1. Specifically, in the global perspective, we enrich positive sample pairs with positive distribution pairs by calculating similarity distributions of augmented input views to their negative samples. Feature representation can be significantly improved by explicitly coupling complex similarity information between the positive augmented samples and the negative samples with distribution alignment, Fig. 2(b). In the local perspective, we interpolate randomly selected images in a mini-batch with a typical data mixture strategy, e.g., cutmix [22]. The interpolation ratio can quantitatively control the apparent similarity of the synthetic image to the original image pair. Meanwhile, we interpolate features of the image pair with the same ratio to obtain the feature as the self-supervision signal of the interpolated image. Attracting corresponding features in the feature space, the consistency of local interpolation relation can be assimilated, Fig. 2(c).

By incorporating the global distribution and local interpolation relations in a plug-and-play fashion, the proposed ReCo

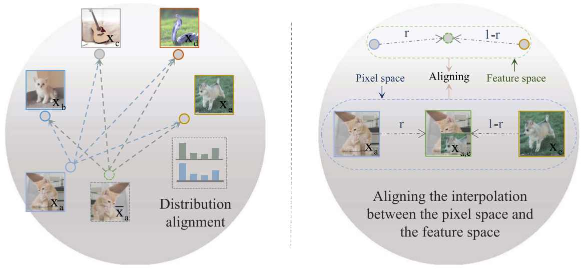
Manuscript received 1 September 2022; revised 6 June 2023; accepted 30 September 2023. Date of publication 13 October 2023; date of current version 8 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62376266, and in part by the Key Research Program of Frontier Sciences, CAS under Grant ZDBS-LY-7024. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Guo-Jun Qi. *(Yifei Zhang and Chang Liu contributed equally to this work.) (Corresponding author: Yu Zhou.)*

Yifei Zhang, Yu Zhou, and Weiping Wang are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100089, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100089, China (e-mail: zhangyifei0115@iie.ac.cn; zhouyu@iie.ac.cn; wangweiping@iie.ac.cn).

Chang Liu and Xiangyang Ji are with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: liuchang2022@tsinghua.edu.cn; xyji@tsinghua.edu.cn).

Qixiang Ye is with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qxye@ucas.ac.cn).

Digital Object Identifier 10.1109/TMM.2023.3324588



(a) Global distribution relation

(b) Local interpolation relation

Fig. 1. Instance relation illustration: (a) Global distribution relation enriches the view-invariant representation from the instance pair level to the dataset level by aligning similarity distributions, where specific similarities of negative pairs are well-exploited. For example, the similarity between cat x_a and cat x_b is higher than cat x_a and guitar x_c. (b) Local interpolation relation quantitatively controls the apparent similarity by utilizing a data mixture technique and exploits the interpolation consistency by aligning interpolations between the pixel and the feature spaces. The alignment is achieved by maximizing the similarity of the features of the interpolated image $X_{a,e}$ to the interpolated features of the original images X_a and X_e. The color-circled points in the figure indicate the features corresponding to the images.

takes full use of specific similarities of diverse sample pairs to relax the constraint that all positives/negatives should be equally attracted/repelled. Extensive experiments justify the effectiveness of ReCo, which produces a locally aggregated yet globally uniform feature space, Fig. 6. Specifically, ReCo achieves state-of-the-art performance with 71.3% top-1 accuracy for linear classification and 78.9% and 87.9% top-5 accuracies for semi-supervised classification with 1% and 10% labeled data. Transferring to the VOC [23] dataset, ReCo improves MoCo-v2 [24] by at least 6.3% mAP for low-shot classification with k=1,2,4,8,16 and 0.9% AP_{50} for object detection.

The contributions are summarized as follows:

1) We propose relation-aware contrastive self-supervised learning (ReCo) to effectively retain the data semantic structures by exploring instance relations from both global and local perspectives. It is a novel attempt to break through the limitation of the error-prone binary label assignment of vanilla CSL.
2) We exploit the global distribution relation to explicitly constrain the specific similarity of different samples other than repelling all negative samples equally.
3) We exploit the local interpolation relation to carve the semantic structure of the feature space with quantitative appearance similarity retention.
4) The proposed ReCo outperforms existing CSL works on multiple benchmarks and shows better generalization ability, especially for insufficient supervision regimes, e.g., it significantly exceeds MoCo-v2 in semi-supervised learning with 1%/10% labeled data and low-shot classification with 1/2/4/8/16 samples.

II. Related Works

A. Unsupervised Visual Representation Learning

Unsupervised visual representation learning aims at utilizing unlabeled data to learn transferable feature representations to initialize downstream tasks, such as image classification [4], object detection [25], and semantic segmentation [26], [27], which can be roughly divided into handcrafted pretext tasks and contrastive self-supervised learning.

Handcrafted Pretext Tasks: Such methods typically assimilate common sense through self-supervision signals generated based on the inherent structure of the data. Specifically, some works aim at recovering input images under pre-defined corruptions, such as colorization [13], inpainting [14], and split-brain autoencoding [28]. Some works generate self-supervision via specific transformations, such as context prediction [12], solving jigsaw puzzle [8], rotation prediction [15], *etc*. Developing sophisticated pretext tasks largely depends on human prior knowledge, which limits their rapid evolution.

Contrastive Self-supervised Learning: With InfoNCE loss [16] and its variants, CSL methods typically construct informative positive and negative sets to encode similarities of positive instance pairs and differences of negative ones. NPID [17] introduces a memory bank to store features of the whole dataset and formulates instance discrimination [7] as a non-parametric classification problem. MoCo [9] proposes a moving-average encoder and a dynamic queue to build positive and negative pairs effectively and efficiently. SimCLR [18] fulfills the contrast procedure in the current mini-batch and introduces more data augmentations to report impressive performance. Interestingly, some researches justify that augmentation invariant representations can also be well learned without negative samples, such as BYOL [29], SimSiam [30], SwAV [31], BarlowTwins [32], *etc*. Moreover, some works [33], [34] attempt to combine contrastive loss with handcrafted pretext tasks, which demonstrate their complementary nature.

To better explore class boundary information, some recent works delve into positive sample discovery. Clustering-based methods [35], [36], [37], [38], [39], [40] target at iteratively grouping instances for reliable pseudo label assignment. Neighbour-discovery-based methods [41], [42], [43], [44], [45] usually set specific rules to select reliable positive samples in the local neighbourhood. The contrastive learning paradigm is also applicable to multimodal scenarios [46], [47], [48]. For instance, CLIP [47] trains a powerful foundation model and demonstrates remarkable zero-shot transfer [49], [50] capabilities. Despite the effectiveness of contrastive learning, as a strong addition to CSL, relation-aware contrastive learning based on soft instance relations [51], [52], [53] of similarity distribution at the global level and interpolation consistency at the local level has not been fully exploited, which hinders the development of CSL.

B. Instance Relations Exploration

The informative data semantic structure can be captured via instance relation exploration, which is usually established in terms of similarity distribution and data interpolation [52], [53], [54], [55], [56], [57], [58], [59], [60], [61]. The distribution depicts unique similarities of diverse sample pairs and the interpolation consistency models relations between synthetic images and original inputs. Their complementary nature appears under-studied.

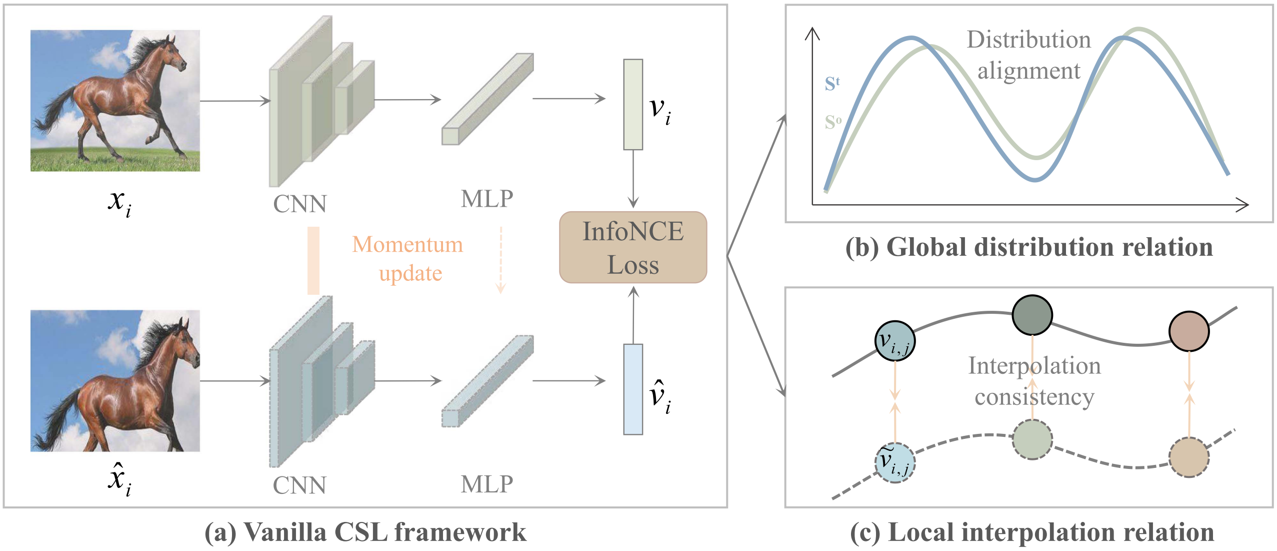


Fig. 2. Architecture overview of our ReCo, which consists of (a) a vanilla CSL framework, e.g., MoCo-v2, (b) global distribution relation, and (c) local interpolation relation. v_i, $\hat{v}_i$, and $v_{i,j}$ are features of input views x_i and $\hat{x}_i$, and the interpolated image, respectively. $\tilde{v}_{i,j}$ denotes the interpolated feature.

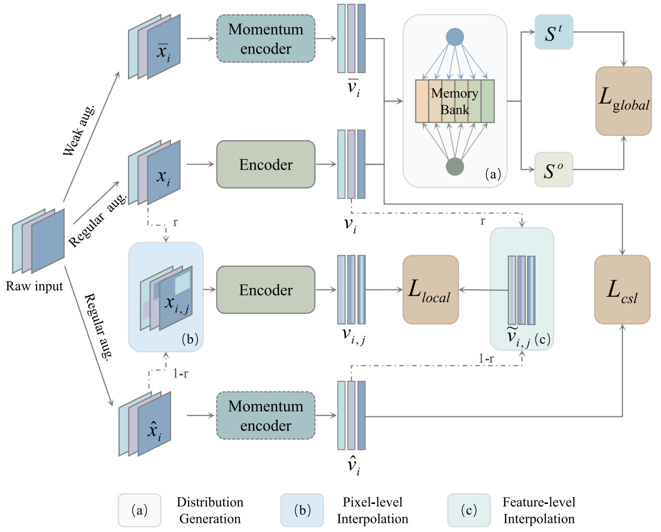


Fig. 3. Detailed implementation framework of our ReCo, in which three modules are jointly optimized.

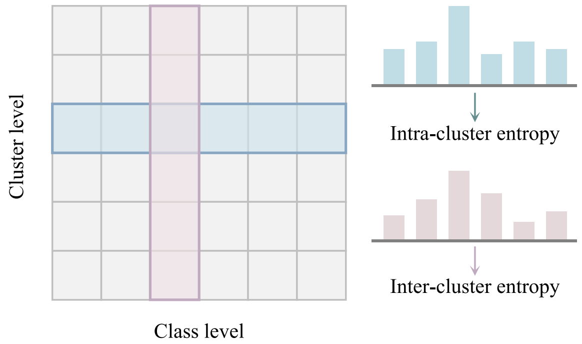
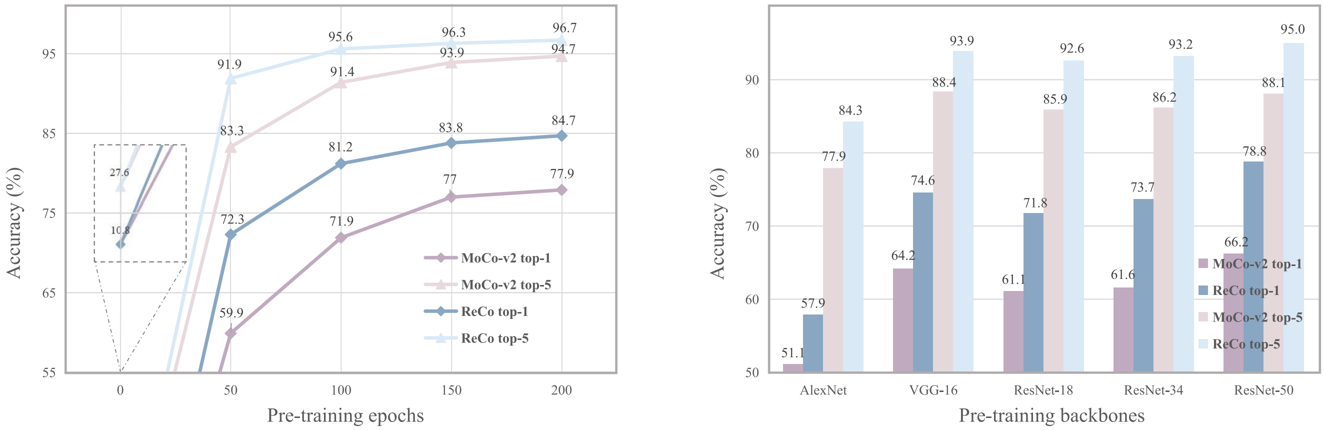


Fig. 4. Illustration of the calculation of intra-cluster entropy and inter-cluster entropy for the analysis of complex relations.

Similarity Distribution: Similarity distribution is typically exploited in knowledge distillation [52] and consistency regularization in semi-supervised learning [62], [63]. Logit-based knowledge distillation [52] proposes to use the output of the softmax layer of the teacher model as soft labels to train the student model. Its effectiveness lies in the fact that the soft labels depict the relation between different classes. After that, some methods explicitly establish the structural relation between the outputs of different samples rather than individual outputs themselves, e.g., relational knowledge distillation [64], similarity-preserving knowledge distillation [65], and self-supervised distillation [66], *etc.*

Consistency regularization in semi-supervised learning [63], [67] insists that the output of the model should be similar before and after perturbing the input data, which is achieved by distribution alignment. A lot of semi-supervised learning works are devoted to how to generate better target distribution, e.g., Mean Teacher [63], MixMatch [67], SsCL [68], *etc*. Some current self-supervised learning methods [54], [55], [69], [70] are exploring the utilization of similarity distribution and have achieved remarkable results. Typically, CO2 [54] improves MoCo-v2 by additionally aligning the similarity distribution of two views to negative samples. ISD [69] and ReSSL [55] utilize weak data augmentation to optimize the distribution alignment term without explicitly pushing away negative samples. CLSA [70] matches the distribution obtained from stronger and regular augmentations to explore new patterns ignored in MoCo-v2. However, the local-level relation that apparently similar inputs should be close in feature space is not explicitly considered.

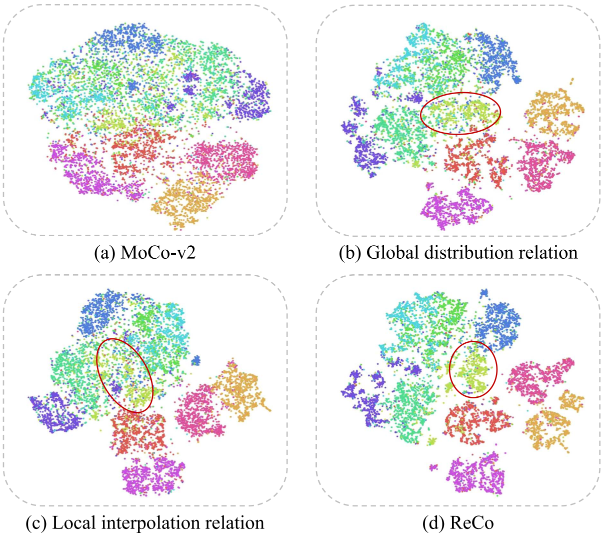
Data Mixture: Data mixture typically targets at augmenting the sample space to reduce incompatibilities during inference. The model generalization ability can be enhanced by exploring relations between synthetic and raw data. Mixup [53] performs the corresponding pixel-weighted summation of the input image



(a) Scalability of epochs

(b) Scalability of backbones

Fig. 5. Scalability of different pre-training epochs and backbones on ImageNet-100. Note that 0-th epoch in (a) denotes random initialization.



(a) MoCo-v2

(b) Global distribution relation

(c) Local interpolation relation

(d) ReCo

Fig. 6. Visualization of 2-D t-SNE distributions of the embedding space. The area circled in red is more discriminative. (Best viewed in color)

pairs, and the label is also linearly interpolated. CutMix [22] replaces the removed regions with a patch from another image. Beyond supervised scenarios, data mixture is also applied in semi-supervised [71] and unsupervised [56] learning. Specifically, UnMix [56] and MixCo [72] perform data mixing in the input space, and then weight the loss with the interpolation ratio. Deviating from merely local-level interpolation, ReCo further exploits the similarity distribution to delineate global-level relations.

III. Methodology

CSL methods based on instance discrimination typically rely on predefined hard binary assignments, which are error-prone and ignore the exploitation of different relations among instances. To retain the semantic structure of the data and produce a locally aggregated and globally uniform feature space [21], [73], we propose relation-aware contrastive self-supervised learning (ReCo) which simultaneously explores soft instance relations of similarity distribution at the global level and interpolation consistency at the local level, Fig. 2.

A. Overview

Baseline: We choose the seminal work MoCo-v2 [24] to clarify the implementation details of our ReCo which can also be applied on common CSL frameworks. It takes two views of the i-th instance x_i and $\hat{x}_i$ as input, which are generated from the same image through a combination of data augmentations. The corresponding features v_i and $\hat{v}_i$ are extracted by an online encoder f_θ and a momentum encoder $f_{\hat{\theta}}$ as $v_i = f_\theta(x_i)$ and $\hat{v}_i = f_{\hat{\theta}}(\hat{x}_i)$, where the encoder consists of a backbone network ($e.g.$ ResNet-50 [4]) and an MLP head, Fig. 2(a). The feature $\hat{v}_i$ from the momentum branch is stored in a queue (memory bank) with the size of K. v_i and $\hat{v}_i$ are defined as positive sample pairs that attract each other in the feature space while staying away from the negative samples in the queue. The learning objective is to minimize the InfoNCE [16] loss:

$$\mathcal{L}_{csl} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(v_i \cdot \hat{v}_i/\tau)}{exp(v_i \cdot \hat{v}_i/\tau) + \sum_j^K exp(v_i \cdot \tilde{v}_j/\tau)}, \tag{1}$$

where N is the training set size, τ is the temperature [52], and $\tilde{v}_j$ is the j-th sample in the queue.

Pipeline: As illustrated in Fig. 2, ReCo consists of three modules : (a) a vanilla CSL framework (MoCo-v2), (b) global distribution relation, and (c) local interpolation relation. The global distribution relation utilizes distribution alignment to fully use the specific similarities of diverse samples to relax the constraint that all negatives are equally repelled. The local interpolation relation applies image interpolation between random sample pairs. It explores the interpolation consistency relation between pixel and feature space to quantitatively model samples' apparent similarity.

The overall loss function of ReCo is a combination of the infoNCE loss $\mathcal{L}_{csl}$, the global distribution relation loss $\mathcal{L}_{global}$, and the local interpolation relation loss $\mathcal{L}_{local}$, which can be formulated as

$$\mathcal{L} = \mathcal{L}_{csl} + \lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{local}, \tag{2}$$

where λ_1 and λ_2 are balance weights of $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$.

B. Global Distribution Relation

We extend view-invariant representation learning from the instance level to the distribution level, which is inspired by the consistency regularization in semi-supervised learning that the output of the model (probability distribution) should be similar under variations in the input space [63], [67]. The distribution depicts the specific similarities between different classes and therefore retains rich global relations. Concretely, the global distribution relation is materialized with distribution generation and distribution alignment, Fig. 3.

Distribution Generation: We calculate the similarity distribution of each input view to its negative samples based on the embedding features extracted by the encoder. To obtain a stable target distribution, we employ weak data augmentation that does not introduce severe variation [67]. Therefore, we utilize a new branch with weak augmentation to obtain $\bar{x}_i$ (top of Fig. 3), which is augmented by only randomly resized cropping and random horizontal flipping. To make the differences between samples more transparent, we use a smaller temperature to sharpen the distribution. Specifically, the distribution obtained by v_i is regarded as the online distribution for gradient back-propagation while the distribution obtained by $\bar{v}_i$ is used as the target distribution. Accordingly, for the i-th instance sampled from the min-batch, the online distribution $S^o(i)$ and target distribution $S^t(i)$ can be calculated by $S^o(i) = \{s_j = v_i \cdot \tilde{v}_j^T / \tau_{ot} | j = 1, 2, 3, \ldots, N\}$ and $S^t(i) = \{s_j = \bar{v}_i \cdot \tilde{v}_j^T / \tau_{tt} | j = 1, 2, 3, \ldots, N\}$ respectively. Note that $\tilde{v}_j$ denotes the feature stored in the memory bank, τ_{ot} and τ_{tt} are temperature parameters that control the degree of sharpening of the distribution.

Distribution Alignment: For the two generated distributions, our goal is to align them with a given objective function. Since Kullback–Leibler (KL) divergence [74] is often used in statistics to measure the degree of difference between two distributions, we use KL divergence by default as the objective function to align the two distributions $S^o(i)$ and $S^t(i)$. In this way, the objective function of global distribution relation is formed by

$$\mathcal{L}_{global} = \frac{1}{N} \sum_{i=1}^{N} D_{KL}(S^t(i) || S^o(i)). \tag{3}$$

Note that N refers to the size of the training set, and $S^t(i)$ does not perform the gradient back-propagation.

C. Local Interpolation Relation

We utilize image mixture strategy to quantitatively simulate apparently similar images. Existing data mixture strategy [22], [53] forces the model to behave linearly when dealing with in-between training examples, that is, the image and target are the corresponding linear interpolation. We exploit this linearity to model local interpolation relation. Specifically, we interpolate image pairs and their features with the same ratio, and then pull the extracted features of the interpolated images and the corresponding interpolated features as close as possible in the feature space. The interpolation consistency relation can be well assimilated by transferring the interpolation ratio from pixel space to feature space. The procedure of local interpolation relation can be detailed as three steps: pixel-level interpolation, feature-level interpolation, and interpolation consistency, Fig. 3.

Pixel-level Interpolation: For each mini-batch, we first sample an interpolation ratio r from the beta distribution as $r \in Beta(\alpha, \alpha)$, where α is a hyper-parameter set to 1.0 by default. Then, for two selected instances x_i and x_j in the mini-batch with size N_b, they are interpolated with the ratio r to form the synthetic image $x_{i,j}$. The embedding feature of the interpolated image is defined as

$$v_{i,j} = f_\theta(r \cdot x_i \oplus (1 - r) \cdot x_j), \tag{4}$$

where $\oplus$ denotes image interpolation operation. Specifically, to randomly select two images for interpolating, the index i is sampled from an ordered set $\mathcal{N}_{order}^{N_b} = \{0, 1, 2, \ldots, N_b - 1\}$ and j is sampled from a random-arrangement set $\mathcal{N}_{rand}^{N_b} = randperm(\mathcal{N}_{order}^{N_b})$, where $randperm()$ denotes shuffle the order randomly.

Feature-level Interpolation: To correspond to the feature of interpolated image $v_{i,j}$ under the simple linearization constraint [22], [53], we generate interpolated feature $\tilde{v}_{i,j}$ according to the ratio r, which is regarded as the pseudo "ground-truth" feature of $v_{i,j}$. The normalized feature interpolation in the embedding space can be obtained by

$$\tilde{v}_{i,j} = \ell_2(r \cdot f_\theta(x_i) + (1 - r) \cdot f_\theta(x_j)), \tag{5}$$

where ℓ_2 denotes normalization.

Interpolation Consistency: To assimilate interpolation consistency relations, the feature of the interpolated image $v_{i,j}$ and the interpolated feature $\tilde{v}_{i,j}$ should be attracted to each other, that is transferring interpolation ratio from pixel space to feature space. It can be achieved using contrastive loss. Accordingly, the loss function of the local interpolation relation is formulated as

$$\mathcal{L}_{local} = \frac{1}{N} \sum_{i \in \mathcal{N}_{order}^{N}, j \in \mathcal{N}_{rand}^{N}} -log P(i, j),$$
$$P(i, j) = \frac{exp(v_{i,j} \cdot \tilde{v}_{i,j/\tau})}{\sum_{k \in [0, K)} exp(v_{i,j} \cdot \tilde{v}_k / \tau)}, \tag{6}$$

where $\tilde{v}_k$ is the feature stored in the memory bank and $\tilde{v}_{i,j}$ conducts stop-gradient operation.

D. Discussion

We detail the differences between our ReCo and existing distribution-based and interpolation-based CSL methods in the aspect of exploiting instance relations. Besides, ReCo further pursues the complementary nature of these two relations in retaining semantic structure, Table V and Fig. 6.

Distribution-based Methods: Distribution-based methods utilize different ways to calculate the similarity distribution and then align the distributions to explore global-level relations. In specific, CO2 [54] utilizes the features of the two branches of MoCo-v2 to obtain the similarity distribution and then uses distribution alignment as a regularization term. ReSSL [55] utilizes weak data augmentation to obtain the target distribution and uses a single distribution alignment loss as the optimization objective. CLSA [70] utilizes stronger and regular data augmentation to obtain two distributions as online distribution and target distribution, respectively. Instead, ReCo uses weak augmentation to obtain the target distribution for distribution alignment, which is used to constrain the InfoNCE loss. Moreover, ReCo uses interpolation to explicitly model the local relation that apparently similar inputs are close in feature space, which is not considered in existing distribution-based methods. More details can be referred in Table VII.

Interpolation-based Methods: ReCo quantitatively models the relation of the interpolated data to original inputs in the feature space. Un-Mix [56] and MixCo [72] interpolate in the input space and then weight the loss corresponding to the interpolation ratio. In contrast, we directly interpolate the features according to the interpolation ratio instead of weighting the loss. In specific, there are 4 options for interpolation: q and $randperm$(q), q and $randperm$(k), k and $randperm$(q), and k and $randperm$(k), where $randperm$() denotes randomly shuffle the order of the batch. Since different views may differ greatly in the early training stages, we claim that the choice of image pairs and feature pairs for interpolation has a large impact on the interpolation consistency relation, which has been completely ignored in previous methods. More importantly, ReCo not only considers the local interpolation relation, but also further explores the global distribution relation. More comparisons are shown in Table VIII.

IV. Experiments

A. Pre-Training Settings

ResNet-50 [4] is set as the backbone network by default. The size of each view is set to 224 × 224 for ImageNet pre-training. We use the SGD optimizer with the momentum of 0.9 and weight decay of 0.0005. A cosine learning rate scheduler is employed with a base learning rate of 0.03, and the batch size is 256. The temperature τ of InfoNCE loss is 0.2. The size of the memory bank is 65536 and the momentum encoder is updated with a parameter of 0.999.

B. Ablation Study

Setup: To quickly verify the effectiveness under different parameter settings, we conduct ablation experiments on ImageNet-100 [75] with ResNet-50 [4] architecture and train for 100 epochs. We set the batch size to 128 with the base learning rate of 0.03. Other experimental settings are the same as those of ImageNet-1 K.

Temperature Parameters: Since the temperature parameter in the InfoNCE loss is crucial to balance uniformity and tolerance of the learned embedding space [21], we tune the temperatures τ_{ot} and τ_{tt} carefully. First, we empirically set an approximate range of their values referring to the settings in MoCo (τ=0.07) and MoCo-v2 (τ=0.2). Then, we finely adjust them for the best performance. Results of different temperature parameters are shown in Table I. In general, ce is better when τ_{ot} is larger than τ_{tt}. Especially, the performance drops dramatically when τ_{tt} is larger than τ_{ot} (the last line). This is because a smaller τ_{tt} can sharpen the target distribution and make the difference between various sample pairs more obvious. We set τ_{tt}=0.04 and τ_{ot}=0.1 by default for the best top-1 accuracy of 73.6%.

TABLE I
Ablation of Temperature Parameters in Distribution Generation

τ_{tt}	τ_{ot}	LC Top-1	LC Top-5
0.2	0.2	70.7	90.6
0.1	0.2	70.6	90.6
0.1	0.1	72.5	91.1
0.07	0.1	73.1	92.5
0.04	0.1	73.6	92.4
0.01	0.1	72.1	91.6
0.2	0.1	68.9	89.4

TABLE II
Comparison of Different Settings for the Coefficients in the Loss Function

λ_1	λ_2	LC Top-1	LC Top-5
0.0	0.0	66.2	88.1
0.5	0.0	72.8	91.5
1.0	0.0	73.6	92.4
2.0	0.0	70.5	91.3
1.0	1.0	78.1	94.4
1.0	2.0	78.8	95.0
1.0	3.0	78.5	94.7

TABLE III
Comparison of intra-/ Inter-Class Similarity. (×100)

Methods	s_{intra} (↑)	s_{inter} (↓)	ϕ (↑)
MoCo-v2 [24]	35.5	0.5	134.9
ReCo	40.1 (+4.6)	0.5	139.4 (+4.5)

Coefficients: We study how the global distribution relation term $\mathcal{L}_{global}$ and the local interpolation relation term $\mathcal{L}_{local}$ in Eq. (2) affect the feature representation by using different values of λ_1 and λ_2. We first determine λ_1 based on $\mathcal{L}_{csl}$ and $\mathcal{L}_{global}$ (2 losses), then fix the obtained λ_1 to adjust the coefficient λ_2 (3 losses). As shown in Table II, when λ_1=1.0 and λ_2=2.0, the best top-1 accuracy is 78.8%.

Intra-/ Inter-Class Similarity: To quantitatively verify the semantic structure of the feature space [21], [73], [76], we define the intra-class similarity as $s_{intra} = \frac{1}{N}\sum_i^N \sum_{x^+ \in \mathcal{S}_p} \frac{x_i \cdot x^+}{||\mathcal{S}_p||}$, the inter-class similarity as $s_{inter} = \frac{1}{N}\sum_i^N \sum_{x^- \in \mathcal{S}_n} \frac{x_i \cdot x^-}{||\mathcal{S}_n||}$, and the discriminative index as $\phi = \frac{1}{N}\sum_i^N \frac{\sum_{x^+ \in \mathcal{S}_p} \frac{x_i \cdot x^+}{||\mathcal{S}_p||} + 1}{\sum_{x^- \in \mathcal{S}_n} \frac{x_i \cdot x^-}{||\mathcal{S}_n||} + 1}$, where N is the number of samples, $\mathcal{S}_p$ is the set of all samples that belong to the same semantic class as x_i based on the ground truth, and $\mathcal{S}_n$ is the set of samples of other different classes. Table III reports the results on the ImageNet-100 val set. Experimental results show that ReCo presents higher intra-class similarity and discriminative index than MoCo-v2, which demonstrates that a better semantic structure is obtained [21], [76], [77].

TABLE IV
SEMANTIC RELATION EVALUATION FOR THE LEARNED FEATURE SPACE OF OUR RECO ON CIFAR-10

Methods	IntraEN ↓	InterEN ↓
RandInit	2.99	2.97
SupCon	0.43	0.43
MoCo-v2	1.75	1.76
ReCo (Ours)	0.95	1.02

Note that ↓ indicates that smaller values are better.

TABLE V
EVALUATION ON THE IMAGENET-100 DATASET WITH RESNET-50 BY PERFORMING LINEAR CLASSIFICATION ACCURACY

Methods	Epochs	LC Top-1	LC Top-5
MoCo-v2 [24]	100	66.2	88.1
BYOL [29]	100	76.9	93.8
SimSiam [30]	100	74.2	92.8
MoCo-v2+Local	100	74.9 (+6.4)	92.9 (+3.7)
MoCo-v2+Global	100	73.6 (+5.6)	92.4 (+2.5)
ReCo	100	78.8 (+8.4)	95.0 (+4.5)
BYOL+Local	100	83.0	96.0
BYOL+Global	100	81.2	95.8
BYOL+ReCo	100	83.9	96.7
MoCo-v2 [24]	200	77.9	94.7
MoCo-v2+Global	160	79.2 (+1.3)	95.1 (+0.4)
MoCo-v2+Local	160	81.1 (+3.2)	95.5 (+0.8)
ReCo	135	82.6 (+4.7)	95.9 (+1.2)
ReCo	200	84.0	96.3

TABLE VI
EVALUATION OF OBJECT DETECTION ON VOC 07+12 WITH THE MODEL PRE-TRAINED ON IMAGENET-100 FOR 100 EPOCHS

Methods	AP	AP$_{50}$	AP$_{75}$
MoCo-v2 [24]	48.7	76.1	52.4
MoCo-v2+Local	51.1 (+2.4)	77.6 (+1.5)	55.3 (+2.9)
MoCo-v2+Global	49.0 (+0.3)	76.7 (+0.6)	52.4 (+0.0)
ReCo	51.0 (+2.3)	78.0 (+1.9)	55.3 (+2.9)
ReCo†	52.0 (+3.3)	78.8 (+2.7)	56.8 (+4.4)

† Denotes the modified pre-training settings.

TABLE VII
COMPARISON OF THE DIFFERENCES BETWEEN DISTRIBUTION-BASED METHODS UNDER TRAINING IMAGENET-100 FOR 100 EPOCHS

Methods	Dim.	Encod.	Aug.	Sharp.	Contra.	Decoup.	Acc.
MoCo-v2 [24]	128	o/t	r/r	-	✓	-	66.2
CO2 [54]	128	o/t	r/r	-	✓	-	67.5
CLSA [70]	128	o/o	s/r	-	✓	✓	71.1
ReSSL [55]	512	o/t	r/w	✓	-	-	72.7
ReCo-Global	128	o/t	r/w	✓	✓	✓	73.6

Complex Relations: Inspired by [78], we investigate complex relations in the high-dimensional feature spaces with two metrics: average intra-cluster entropy (IntraEN) and average inter-cluster entropy (InterEN). IntraEN measures class dominance within a cluster, while InterEN measures sample dispersion of the same class across different clusters. Intra-cluster and inter-cluster entropy are calculated by measuring the frequency distribution of different categories within the same cluster and the distribution of the same category among different clusters, Fig. 4. Smaller IntraEN and InterEN values indicate better semantic structure. We train models for 200 epochs on the training set of CIFAR-10 and evaluate on its test set with K=10. Table IV shows that ReCo outperforms MoCo-v2 on both metrics and achieves results closer to those of SupCon [79], demonstrating that ReCo captures better semantic structure in high-dimensional feature spaces.

Scalability: The scalability of our model is verified by training with different epochs and backbones. Fig. 5(a) shows the linear classification accuracies of the pre-trained model under different epochs, which shows that higher performance can be obtained with longer training iterations. Moreover, ReCo with 100 epochs can significantly outperform MoCo-v2 with 200 epochs, which demonstrates the pre-training efficiency of ReCo. Fig. 5(b) further verifies that ReCo can effectively improve the performance of the baseline under various backbones including AlexNet [3], VGG-16 [80] and ResNet-18/34/50 [4].

Module Efficacy: We quantitatively demonstrate the effectiveness of the global distribution relation and local interpolation relation in our ReCo based on MoCo-v2 and BYOL. To ensure a fair comparison, we conducted experiments under the same running time. Specifically, the training time of MoCo-v2 for 200 epochs is close to that of MoCo-v2+Global (*i.e.*, with only global distribution alignment branch) and MoCo-v2+Local (*i.e.*, with only local interpolation consistency branch) for 160 epochs, and the time of ReCo training for 135 epochs. In Table V, both modules significantly improve the baseline performance, and the combination proves their complementarity for semantic structure retention. In Table VI, the VOC object detection results show that global distribution relation has no obvious advantage in precise location (AP$_{75}$). We simply set the parameters of the global distribution relation τ_{tt} and τ_{ot} to 0.1, AP can be improved by 1.0%, and AP$_{75}$ by 1.5%. This also shows that the pre-trained model performs well on classification do not necessarily perform well on object detection [20].

Distribution-based Methods: To demonstrate the difference from existing distribution-based methods, we compare them in detail in Table VII. The differences in related works are reflected in the feature embedding dimension (Dim.), the encoder used to generate the distribution (Encod.), the type of data augmentation (Aug.), whether the distribution is sharpened (Sharp.), whether there is a contrastive learning loss to assist (Contra.), and whether the distribution alignment and contrastive loss are decoupled (Decoup.). Note that "o" and "t" denote the online encoder and target encoder, and "r", "w" and "s" denote regular augmentation, weak augmentation, and strong augmentation respectively. We reimplement the related methods, and the experimental results show that our global distribution relation (ReCo-Global) achieves the highest performance of 73.6% on ImageNet-100.

Interpolation-based Methods: In Table VIII, our method differs from related works in the implementation of the interpolation method and interpolation ratio correspondence. In particular, previous methods interpolate at the loss level (Loss Inter.), while we interpolate at the feature level (Feature Inter.). In addition, we also compare the impact of different interpolation methods in our local interpolation relation (ReCo-Local). The

TABLE VIII
COMPARISON OF THE DIFFERENCES BETWEEN INTERPOLATION-BASED METHODS UNDER TRAINING IMAGENET-100 FOR 100 EPOCHS

Methods	Image Inter.		Feature Inter.				Loss Inter.	Acc.
	q/q	q/k	q/k	k/q	q/q	k/k		
MoCo-v2 [24]	-	-	-	-	-	-	-	66.2
MixCo [72]	✓	-	-	-	-	-	✓	69.4
Un-Mix [56]	✓	-	-	-	-	-	✓	69.5
ReCo-Local	✓	-	-	-	✓	-	-	72.4
ReCo-Local	-	✓	✓	-	-	-	-	**74.9**
ReCo-Local	-	✓	-	✓	-	-	-	71.9
ReCo-Local	-	✓	-	-	-	✓	-	72.0

TABLE IX
COMPARISON OF CLSA+$\mathcal{L}_{local}$ AND RECO

Methods	Type	Dataset	Epochs	Top-1	Top-5
CLSA	Baseline	IN-100	100	71.1	91.9
CLSA+$\mathcal{L}_{local}$	Ours	IN-100	100	77.5	94.0
ReCo	Ours	IN-100	100	78.8	95.0
CLSA	Baseline	IN-1K	200	69.2	89.1
CLSA+$\mathcal{L}_{local}$	Ours	IN-1K	200	70.5	90.2
ReCo	Ours	IN-1K	200	71.3	90.5

All experiments are re-run by us.

q/k interpolation in the corresponding pixel space and feature space obtains the best performance of 74.9%.

Combining CLSA and $\mathcal{L}_{local}$: Since CLSA utilizes stronger augmentation as a query branch and has similar $\mathcal{L}_{global}$ loss, we combine CLSA ($\mathcal{L}_C$ + $\beta * \mathcal{L}_D$) with $\mathcal{L}_{local}$ to verify the effectiveness of our method. The overall loss function is formulated as $\mathcal{L}_C + \beta * \mathcal{L}_D + \gamma * \mathcal{L}_{local}$, where the coefficients β and γ are set to 1 by default. As shown in Table IX, ReCo outperforms CLSA+$\mathcal{L}_{local}$ on both IN-100 (ImageNet-100) and IN-1 K (ImageNet-1 K).

C. Performance and Comparison

Comparisons are mainly listed on 4 downstream tasks: linear classification, semi-supervised classification, low-shot classification, and VOC object detection. More experiments (e.g. kNN classification, COCO object detection and instance segmentation, and Cityscapes semantic segmentation) can be found in supplementary materials.

1) Linear Classification: Convolutional layers initialized by the pre-trained model are frozen while a fully connected linear classifier is initialized from scratch. Its results represent the discriminative ability of the learned representation.

Setup: We use a LARS optimizer for training 90 epochs with an initial learning rate of $0.1 * batch/256$. We incorporate 4 additional views with 4 different sizes (192×192, 160×160, 128×128, and 96×96) as queries to implement the multi-crop augmentation strategy. More details can be found in supplementary materials.

Results: Table X reports the top-1 and top-5 accuracies of SOTA methods on ImageNet-1 K, where our re-implementation of MoCo-v2 achieves 67.6% top-1 accuracy (0.1% higher than the official result). By incorporating instance relations exploration, our ReCo achieves a new SOTA top-1 accuracy of 71.3%, which improves the baseline MoCo-v2 by 3.7%. For the same running time (*i.e.*, MoCo-v2 trains for 200 epochs while ReCo trains for 135 epochs), ReCo outperforms MoCo-v2 by 2.0% (69.6% *vs* 67.6%). These results demonstrate that ReCo retains data semantic structures via exploring instance relations to enhance the feature discriminative capabilities. Trained with merely 200 epochs, ReCo even exceeds that of MoCo-v2 with 800 epochs, which proves that ReCo can also improve the pre-training efficiency. After training for 800 epochs, ReCo with multi-crop augmentation strategy achieves a top-1 accuracy of 75.4%, which presents a 1.7% improvement over the results obtained after 200 epochs.

2) Semi-Supervised Classification: Semi-supervised classification first learns from large-scale unlabeled data and then fine-tunes on small labeled data.

Setup: The backbone and linear layer are fine-tuned on ImageNet with 1% and 10% labeled data. The SGD optimizer is used to train 20 epochs with a batch size of 256. The learning rate is set to 0.01 for the backbone and 1.0 for the linear layer.

Results: Experimental results in Table XI show that ReCo consistently achieves the best performance under different label fractions. Specifically, ReCo surpasses MoCo-v2 by 13.4% top-1 accuracy with 1% labeled data, which demonstrates that the semantic structure learned by exploring instance relations can be more advantageous under insufficient data settings.

3) Transferring to Low-Shot Classification: To verify the discrimination capability of learned features, we train linear SVM using fixed features of $conv5$ under low-shot settings.

Setup: Following PCL [39], linear SVM is trained on the VOC [23] 2007 $trainval$ set and tested on the $test$ set. We select k (k=1,2,4,8,16) samples from each class for training. Performance is evaluated by mean average precision (mAP).

Results: As shown in Table XII, ReCo improves MoCo-v2 by 7.7/7.5/8.3/7.1/6.3 under 1/2/4/8/16 shots. In particular, the performance of ReCo is comparable to the supervised trained model. These indicate that the features learned by ReCo are sufficiently discriminative and representative.

4) Transferring to Object Detection: To verify the transferability and generalization capacity of the learned representation, we transfer the trained model to object detection.

Setup: We fine-tune the Faster R-CNN [25] with ResNet50-C4 architecture on VOC $trainval$ 07+12 and evaluate the results on $test$ 2007. The detailed experimental setup can be found in supplementary materials.

Results: In Table XIII, all state-of-the-art CSL methods outperform supervised pre-training on the object detection task, which demonstrates the advantage of CSL for transfer learning. With τ_{tt}/τ_{ot}=0.1/0.1, ReCo presents 0.6/0.9/0.6 gains over MoCo-v2 under AP/AP$_{50}$/AP$_{75}$. These results demonstrate that exploring instance relations improves the transferability and generalization of the model.

D. Visualization

Embedding Space: We simply use the CIFAR-10 [92] val set with 10 categories for feature space visualization. The t-

TABLE X
COMPARISONS ON IMAGENET-1 K UNDER LINEAR CLASSIFICATION (LC) EVALUATION

Methods	Publisher	Source	Baseline	Architecture	Batch Size	Epochs	LC Top-1	LC Top-5
Supervised	-	[24]	-	R50	-	90	76.5	-
NPID [17]	CVPR18	[17]	-	R50	256	200	54.0	-
LA [76]	ICCV19	[76]	-	R50	128	200	60.2	-
MoCo [9]	CVPR20	[9]	-	R50	256	200	60.6	-
MoCo-v2 [24]	arXiv20	[24]	-	R50-MLP	256	200	67.5	-
SimCLR [18]	ICML20	[24]	-	R50-MLP	8192	200	66.6	-
BYOL [29]	NeurIPS20	[30]	-	R50-MLP	4096	200	70.6	-
MoCHi [81]	NeurIPS20	[81]	MoCo-v2	R50-MLP	256	200	68.0	-
MixCo [72]	NeurIPSW20	[72]	MoCo-v2	R50-MLP	256	200	68.4	-
PCL v2 [39]	ICLR21	[39]	MoCo-v2	R50-MLP	256	200	67.6	-
CO2 [54]	ICLR21	[54]	MoCo-v2	R50-MLP	256	200	68.0	-
SimSiam [30]	CVPR21	[30]	-	R50-MLP	256	200	70.0	-
JigClu [34]	CVPR21	[34]	-	R50-MLP	256	200	66.4	-
PSL [82]	CVPR21	[82]	MoCo-v2	R50-MLP	256	200	68.1	-
ISD [69]	ICCV21	[69]	BYOL	R50-MLP	256	200	69.8	-
VFT [83]	ICCV21	[83]	MoCo-v2	R50-MLP	256	200	69.6	-
TKC [84]	ICCV21	[84]	MoCo-v2	R50-MLP	256	200	69.0	88.7
ISL [85]	ICCV21	[85]	MoCo-v2	R50-MLP	256	200	68.6	-
ReSSL [55]	NeurIPS21	[55]	MoCo-v2	R50-MLP	256	200	69.9	-
CLSA [70]	TPAMI23	[70]	MoCo-v2	R50-MLP	256	200	69.4	-
Un-Mix [56]	AAAI22	[56]	MoCo-v2	R50-MLP	256	200	68.6	-
HCSC [86]	CVPR22	[86]	MoCo-v2	R50-MLP	256	200	69.2	-
MoCo-v2*	arXiv20	Ours	-	R50-MLP	256	200	67.6	88.0
ReCo	-	Ours	MoCo-v2	R50-MLP	256	135	69.6	89.6
ReCo	-	Ours	MoCo-v2	R50-MLP	256	200	71.3	90.5
ReCo†	-	Ours	MoCo-v2	R50-MLP	256	200	73.7	91.9
SimCLR [18]	ICML20	[18]	-	R50-MLP	4096	1000	69.3	89.0
MoCo-v2 [24]	arXiv20	[24]	-	R50-MLP	256	800	71.1	90.1
SimSiam [30]	CVPR21	[30]	-	R50-MLP	256	800	71.3	-
InfoMin [87]	NeurIPS20	[87]	-	R50-MLP	256	800	73.0	91.1
BYOL [29]	NeurIPS20	[29]	-	R50-MLP	4096	800	74.3	91.6
SwAV† [31]	NeurIPS20	[31]	-	R50-MLP	4096	800	75.3	-
Barlow Twins [32]	ICML21	[32]	-	R50-MLP	2048	1000	73.2	91.0
NNCLR† [45]	ICCV21	[45]	-	R50-MLP	4096	1000	75.6	92.4
VICReg [88]	ICLR22	[88]	-	R50-MLP	2048	1000	73.2	91.1
MoCo-v2*	arXiv20	Ours	-	R50-MLP	256	800	70.8	89.9
ReCo	-	Ours	MoCo-v2	R50-MLP	256	800	73.0	91.5
ReCo†	-	Ours	MoCo-v2	R50-MLP	256	800	75.4	92.7
ReCo†‡	-	Ours	MoCo-v2	R50-MLP	256	800	75.9	92.8

"Source" refers to which reference the value comes from. † Denotes multi-crop augmentation. ‡ Means adding an additional fully connected layer (2048-D, with BN) before the 2-layer MLP. *Denotes our re-implementation.

TABLE XI
COMPARISON OF SEMI-SUPERVISED CLASSIFICATION

Methods	Publisher	Source	1% label		10% label	
			Top-1	Top-5	Top-1	Top-5
NPID [17]	CVPR18	[86]	-	39.2	-	77.4
MoCo-v2 [24]	arXiv20	[86]	36.7	64.4	60.7	83.4
SimCLR [18]	ICML20	[86]	46.8	74.2	63.6	86.0
MoCHi [81]	NeurIPS20	[86]	38.2	65.4	61.1	83.5
PCL-v2 [39]	ICLR21	[39]	-	73.9	-	85.0
CO2 [54]	ICLR21	[54]	-	71.0	-	85.7
AdCo [89]	CVPR21	[86]	43.6	71.6	61.8	84.2
HCSC [86]	CVPR22	[86]	48.0	75.6	64.3	86.0
HCSC† [86]	CVPR22	Ours	48.4	75.2	64.0	86.0
MoCo-v2* [24]	arXiv20	Ours	39.4	67.8	61.9	85.0
ReCo	-	Ours	52.8	78.9	66.8	87.9

† Denotes official released model. *Denotes our re-implementation.

SNE [93] technique is utilized to map the feature space onto a 2D plane. Fig. 6 shows intra- and inter-class variation, which reflects the semantic structure of the data. With InfoNCE loss, MoCo-v2 can learn semantic structure to a certain extent, but the interstice between different categories is not clear enough. By considering global distribution relations or local interpolation relations separately, the degree of discrimination of different categories is more obvious than MoCo-v2. In particular, ReCo can obtain a feature space with better semantic structure by combining these two.

Activation Map: We use Grad-CAM [94] to visualize activation map. As shown in Fig. 7, the supervised pre-trained model focuses on the entire object or discriminative regions of the object, while the model learned by MoCo-v2 is more distracted and even focuses on non-foreground object regions. This is because instance discrimination approaches aim at learning sample-specific features while supervised training exploits semantic label information to learn class-specific discriminative features. Compared with MoCo-v2, the model learned by ReCo pays more attention to foreground objects, which is more similar to the supervised training model. This demonstrates the advantage of ReCo in retaining semantic structure.

TABLE XII
EVALUATION OF LOW-SHOT CLASSIFICATION ON VOC2007 USING LINEAR SVMS TRAINED ON FIXED REPRESENTATIONS

Methods	Architecture	k=1	k=2	k=4	k=8	k=16
Random	R50	8.0	8.2	8.2	8.2	8.5
Supervised	R50	54.3	67.8	73.9	79.6	82.3
MoCo [9]	R50	31.4	42.0	49.5	60.0	65.9
PCL [39]	R50	46.9	56.4	62.8	70.2	74.3
SimCLR [18]	R50-MLP	32.7	43.1	52.5	61.0	67.1
MoCo-v2 [24]	R50-MLP	46.3	58.3	64.9	72.5	76.1
PCL-v2 [39]	R50-MLP	47.9	59.6	66.2	74.5	78.3
Supervised†	R50	54.0	67.9	73.8	79.7	82.3
ReSSL† [55]	R50-MLP	45.3	58.1	66.4	74.5	79.3
HCSC† [86]	R50-MLP	47.9	59.6	66.3	74.4	78.4
MoCo-v2* [24]	R50-MLP	47.1	58.3	65.1	72.4	76.3
ReCo	R50-MLP	54.8	65.8	73.4	79.5	82.6

† Denotes our evaluation of the officially released model. *Denotes full re-implementation. Other results are adopted from PCL [39].

TABLE XIII
FINE-TUNING OBJECT DETECTION ON PASCAL VOC

Methods	Publisher	Source	AP	AP_{50}	AP_{75}
Rand Init	-	[30]	33.8	60.2	33.1
Supervised	-	[24]	53.5	81.3	58.8
MoCo [9]	CVPR20	[24]	55.9	81.5	62.6
MoCo-v2 [24]	arXiv20	[24]	57.0	82.4	63.6
CO2 [54]	ICLR21	[54]	57.2	82.7	64.1
BarlowTwins [32]	ICML21	[32]	56.8	82.6	63.4
MaskCo [90]	ICCV21	[90]	56.7	82.1	63.9
Un-Mix [56]	AAAI22	[56]	57.7	83.0	64.3
HCSC [86]	CVPR22	[86]	-	82.5	-
ContrastiveCrop [91]	CVPR22	[91]	57.3	82.5	63.8
ReSSL* [55]	NeurIPS21	Ours	55.6	82.2	61.6
MoCo-v2* [24]	arXiv20	Ours	57.1	82.3	64.1
ReCo	-	Ours	57.7	83.2	64.7

*Denotes our re-implementation.

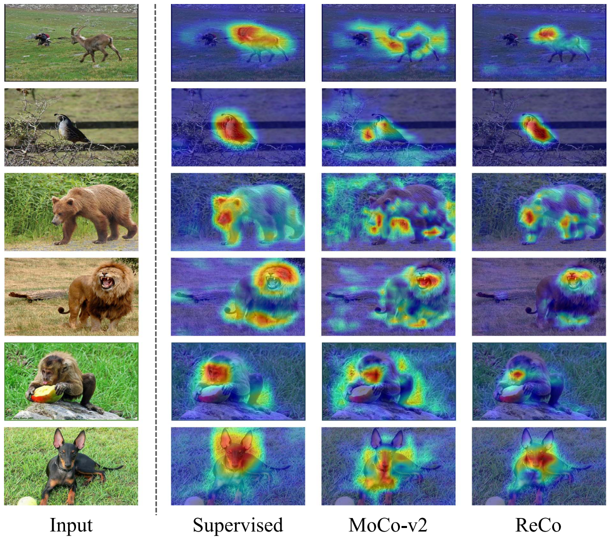


Fig. 7. Activation maps of different pre-trained models using Grad-CAM. Redder colors represent areas that the network pays more attention to.

V. CONCLUSION

In this article, we explicitly exploit semantic relations among instances for relation-aware contrastive self-supervised learning (ReCo). Unlike previous instance discrimination-based CSL methods that only contrast samples with pre-defined hard binary error-prone assignments, ReCo simultaneously explores the soft relation in instance similarity distributions at the global level and interpolation consistency at the local level. With a better semantic structure, the learned feature space appears to be locally aggregated yet globally uniform. It is worth noting that, similar to commonly employed self-supervised learning approaches for object-centric images, our ReCo also faces challenges in avoiding semantic inconsistency across multiple augmented views, particularly in complex images with multiple objects. Still, we expect that ReCo can provide fresh insights into the CSL community, e.g., introducing neighborhood discovery or clustering techniques for better semantic-aware instance relation exploration, extended to various data of different formats and modalities fueled by specialized similarity distributions and data interpolations.

REFERENCES

[1] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[2] D. Mahajan et al., "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 185–201.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[5] C. Zhang, J. Cheng, and Q. Tian, "Multiview label sharing for visual representations and classifications," *IEEE Trans. Multimedia*, vol. 20, pp. 903–913, 2018.

[6] G. Song, S. Wang, Q. Huang, and Q. Tian, "Learning feature representation and partial correlation for multimodal multi-label data," *IEEE Trans. Multimedia*, vol. 23, pp. 1882–1894, 2021.

[7] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, Sep. 2016.

[8] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[10] X. Huo et al., "Heterogeneous contrastive learning: Encoding spatial information for compact visual representations," *IEEE Trans. Multimedia*, vol. 24, pp. 4224–4235, 2022.

[11] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[12] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1422–1430.

[13] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 577–593.

[14] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.

[15] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018.

[16] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[17] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[19] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6210–6219.

[20] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5414–5423.

[21] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2495–2504.

[22] S. Yun et al., "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6022–6031.

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.

[24] X. Chen, H. Fan, R. B. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[26] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[27] L. Ma, H. Xie, C. Liu, and Y. Zhang, "Learning cross-channel representations for semantic segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 2774–2787, 2023.

[28] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1058–1067.

[29] J. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[30] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15750–15758.

[31] M. Caron et al., "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.

[32] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[33] Z. Feng, C. Xu, and D. Tao, "Self-supervised representation learning by rotation feature decoupling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10364–10374.

[34] P. Chen, S. Liu, and J. Jia, "Jigsaw clustering for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11526–11535.

[35] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5147–5156.

[36] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 132–149.

[37] X. Zhan, J. Xie, Z. Liu, Y. S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6688–6697.

[38] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," in *Proc. Int. Conf. Learn. Representations*, 2020.

[39] J. Li et al., "Prototypical contrastive learning of unsupervised representations," in *Proc. Int. Conf. Learn. Representations*, 2021.

[40] C. Zhang, J. Cheng, and Q. Tian, "Unsupervised and semi-supervised image classification with weak semantic consistency," *IEEE Trans. Multimedia*, vol. 21, pp. 2482–2491, 2019.

[41] J. Huang, Q. Dong, S. Gong, and X. Zhu, "Unsupervised deep learning by neighbourhood discovery," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2849–2858.

[42] T. Milbich, O. Ghori, F. Diego, and B. Ommer, "Unsupervised representation learning by discovering reliable image relations," *Pattern Recognit.*, vol. 102, 2020, Art. no. 107107.

[43] H. Fan, P. Liu, M. Xu, and Y. Yang, "Unsupervised visual representation learning via dual-level progressive similar instance selection," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 8851–8861, Sep. 2022.

[44] F. Wang, H. Liu, D. Guo, and F. Sun, "Unsupervised representation learning by invariance propagation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3510–3520.

[45] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9588–9597.

[46] M. Li et al., "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, Mar. 2023.

[47] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[48] X. Yuan et al., "Multimodal contrastive training for visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6995–7004.

[49] C. Yan et al., "ZeroNAS: Differentiable generative adversarial networks search for zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9733–9740, Dec. 2022.

[50] L. Zhang et al., "TN-ZSTAD: Transferable network for zero-shot temporal activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3848–3861, Mar. 2023.

[51] X. Chang et al., "A comprehensive survey of scene graphs: Generation and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, Jan. 2023.

[52] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[53] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.

[54] C. Wei, H. Wang, W. Shen, and A. L. Yuille, "CO2: Consistent contrast for unsupervised visual representation learning," in *Proc. Int. Conf. Learn. Representations*, 2021.

[55] M. Zheng et al., "ReSSL: Relational self-supervised learning with weak augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 2543–2555.

[56] Z. Shen, Z. Liu, Z. Liu, M. Savvides, and T. Darrell, "Un-mix: Rethinking image mixtures for unsupervised visual representation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2216–2224.

[57] C. Cao, F. Zhou, Y. Dai, and J. Wang, "A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability," 2022, *arXiv:2212.10888*.

[58] S. Yun, S. J. Oh, B. Heo, D. Han, and J. Kim, "Videomix: Rethinking data augmentation for video classification," 2020, *arXiv:2012.03457*.

[59] S. Yoon, G. Kim, and K. Park, "SSmix: Saliency-based span mixup for text classification," in *Proc. Findings ACL*, 2021, pp. 3225–3234.

[60] X. Hao et al., "MixGen: A new multi-modal data augmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 379–389.

[61] Q. Fang, R. Ye, L. Li, Y. Feng, and M. Wang, "STEMM: Self-learning with speech-text manifold mixup for speech translation," in *Proc. Assoc. Comput. Linguistics*, 2022, pp. 7050–7062.

[62] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2017.

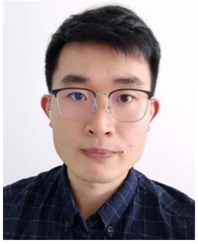
[63] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[64] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3967–3976.

[65] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1365–1374.

[66] Z. Fang et al., "SEED: Self-supervised distillation for visual representation," in *Proc. Int. Conf. Learn. Representations*, 2021.

[67] D. Berthelot et al., “Mixmatch: A holistic approach to semi-supervised learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5050–5060.
[68] Y. Zhang et al., “Semi-supervised contrastive learning with similarity co-calibration,” *IEEE Trans. Multimedia*, vol. 25, pp. 1749–1759, 2023.
[69] A. Tejankar, S. A. Koohpayegani, V. Pillai, P. Favaro, and H. Pirsiavash, “ISD: Self-supervised learning by iterative similarity distillation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9609–9618.
[70] X. Wang and G. J. Qi, “Contrastive learning with stronger augmentations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5549–5560, May 2023.
[71] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, “Interpolation consistency training for semi-supervised learning,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3635–3641.
[72] S. Kim, G. Lee, S. Bae, and S. Yun, “Mixco: Mix-up contrastive learning for visual representation,” 2020, *arXiv:2010.06300*.
[73] T. Wang and P. Isola, “Understanding contrastive representation learning through alignment and uniformity on the hypersphere,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9929–9939.
[74] S. Kullback, *Information Theory and Statistics*. Hoboken, NJ, USA: Wiley, 1959.
[75] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.
[76] C. Zhuang, A. L. Zhai, and D. Yamins, “Local aggregation for unsupervised learning of visual embeddings,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6002–6012.
[77] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, “A theoretical analysis of contrastive unsupervised representation learning,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5628–5637.
[78] T. Li, S. Ma, and M. Ogihara, “Entropy-based criterion in categorical clustering,” in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 536–543.
[79] P. Khosla et al., “Supervised contrastive learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.
[80] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Representations*, 2015.
[81] Y. Kalantidis et al., “Hard negative mixing for contrastive learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21798–21809.
[82] Z. Li et al., “Progressive stage-wise learning for unsupervised feature representation enhancement,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9767–9776.
[83] R. Zhu, B. Zhao, J. Liu, Z. Sun, and C. W. Chen, “Improving contrastive learning by visualizing feature transformation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10286–10295.
[84] W. Feng, Y. Wang, L. Ma, Y. Yuan, and C. Zhang, “Temporal knowledge consistency for unsupervised visual representation learning,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10150–10160.
[85] Z. Wang, Y. Wang, Z. Wu, J. Lu, and J. Zhou, “Instance similarity learning for unsupervised feature representation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10316–10325.
[86] Y. Guo et al., “HCSC: Hierarchical contrastive selective coding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9706–9715.
[87] Y. Tian et al., “What makes for good views for contrastive learning?,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6827–6839.
[88] A. Bardes, J. Ponce, and Y. LeCun, “VICReg: Variance-invariance-covariance regularization for self-supervised learning,” in *Proc. Int. Conf. Learn. Representations*, 2022.
[89] Q. Hu, X. Wang, W. Hu, and G. J. Qi, “AdCo: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1074–1083.
[90] Y. Zhao, G. Wang, C. Luo, W. Zeng, and Z. Zha, “Self-supervised visual representations learning by contrastive mask prediction,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10160–10169.
[91] X. Peng, K. Wang, Z. Zhu, and Y. You, “Crafting better contrastive views for siamese representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16010–16019.
[92] A. Krizhevsky and G. E. Hinton, “Learning multiple layers of features from tiny images,” Univ. Toronto, Toronto, Ontario, 2009.
[93] L. van der Maaten, “Learning a parametric embedding by preserving local structure,” *J. Mach. Learn. Res.*, vol. 5, pp. 384–391, 2009.
[94] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.



Yifei Zhang received the B.S. degree from Huazhong Agriculture University, Wuhan, China, in 2018 and the M.S. degree in 2021 from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, where he is currently working toward the Ph.D. degree with the Institute of Information Engineering. His research interests include pattern recognition and computer vision, specifically for unsupervised visual representation learning.

Chang Liu received the B.S. degree from Jilin University, Jilin, China, in 2012 and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2022. He is currently a Postdoctoral Researcher with the Department of Automation, School of Information Science and Technology, Tsinghua University, Beijing. He has authored or coauthored more than 50 papers in referred conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, AAAI, ACM MM, IJCAI, PR, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. His research interests include computer vision and machine learning, especially for self-supervised learning.

Yu Zhou received the the B.Sc., M.Sc. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China. He is currently an Associate Professor and Ph.D. supervisor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 60 papers in peer-reviewed journals and conferences including CVPR/ICCV/AAAI/IJCAI/ACM MM, and the article PIMNet has been selected as the best paper candidate in ACM MM 2021. His research interests include computer vision and deep learning. He was a AC/SPC/PC members of ICCV, CVPR, IJCAI, AAAI, and ICME, and Reviewers of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, TOMM, and PR.

Weiping Wang is currently a Professor and Director of the Big Data Research Laboratory with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He has undertaken more than 30 national projects, and authored or coauthored more than 100 papers in journals and conferences including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, NeurIPS, AAAI, and IJCAI. His research interests include Big Data and artificial intelligence.

Qixiang Ye (Senior Member, IEEE) received the B.S. and M.S. degrees in mechanical and electrical engineering from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006. He has been a Professor with the University of Chinese Academy of Sciences since 2009, and was a visiting Assistant Professor with the Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD, USA, till 2013. He has authored or coauthored more than 100 papers in referred conferences and journals including IEEE CVPR, ICCV, ECCV, NeurIPS, TNNLS, TIP, and PAMI. His research interests include image processing, visual object detection and machine learning. He is on the Editorial boards of IEEE TRANSACTIONS ON CIRCUIT AND SYSTEMS ON VIDEO TECHNOLOGY and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Xiangyang Ji (Member, IEEE) received the B.S. degree in materials science and the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008. In 2008, he joined Tsinghua University, Beijing, where he is currently a Professor with the Department of Automation, School of Information Science and Technology. He has authored more than 100 referred conference and journal articles. His research interests include signal processing, image/video compressing, and intelligent imaging.