LeOCLR: Leveraging Original Images for Contrastive Learning of Visual Representations

Anonymous authors
Paper under double-blind review

Abstract

Contrastive instance discrimination approaches outperform supervised learning in downstream tasks like image classification and object detection. However, these approaches heavily rely on data augmentation during representation learning, which may result in inferior results if not properly implemented. Random cropping followed by resizing is a common form of data augmentation used in contrastive learning, but it can lead to degraded representation learning if the two random crops contain distinct semantic content. To address this issue, this paper introduces LeOCLR (Leveraging Original Images for Contrastive Learning of Visual Representations), a framework that employs a new instance discrimination approach and an adapted loss function to alleviate discarding semantic features caused by mapping different object parts during representation learning. The experimental results show that our approach consistently improves representation learning across different datasets compared to baseline models. For example, our approach outperforms MoCo-v2 by 5.1% on ImageNet-1K in linear evaluation and several other methods on transfer learning tasks.

1 Introduction

Self-supervised learning (SSL) approaches based on instance discrimination (Chen et al., 2020b; Chen & He, 2021; Chen et al., 2020a; Misra & Maaten, 2020; Grill et al., 2020) heavily rely on data augmentations such as (random cropping, rotation, and colour Jitter) to build invariant representation for all the instances in the dataset. To do so, the two augmented views (i.e., positive pairs) for the same instance are attracted in the latent space while avoiding collapse to the trivial solution (i.e., representation collapse). These approaches have proven efficient in learning useful representations by using different downstream tasks (i.e., image classification and object detection) as a proxy evaluation for representation learning. However, these strategies ignore the important fact that the augmented views may have different semantic content because of random cropping and thus tend to degenerate visual representation learning (Song et al., 2023a; Zhang et al., 2022; Liu et al., 2020; Mishra et al., 2021). On the one hand, creating positive pairs by random cropping and encouraging the model to make them similar based on the information in the shared region between the two views makes the SSL model task harder and improves representation quality (Mishra et al., 2021; Chen et al., 2020a). In addition, random cropping followed by resizing leads model representation to capture information for the object from varying aspect ratios and induce occlusion invariance (Purushwalkam & Gupta, 2020). Conversely, minimizing the feature distance in the latent space (i.e., maximizing similarity) between views containing distinct semantic concepts tends to result in the loss of valuable image information (Purushwalkam & Gupta, 2020; Zhang et al., 2022; Song et al., 2023a).

Figure 1 (a and b) show examples of wrong semantic positive pairs (i.e., positive pairs contain wrong semantic information for the same object) that might be created by random cropping. In case (a), when the model is forced to bring the two representations of the dog's head and leg closer in the latent space, it will discard important semantic features. This is because the model makes the representations of the views similar based on the information in the shared region between the two views. Thus, the representation will be trivial if the shared region between the two views is not semantically matched. The shared region between the views must encompass the same semantic information to obtain the advantage of random cropping and achieve

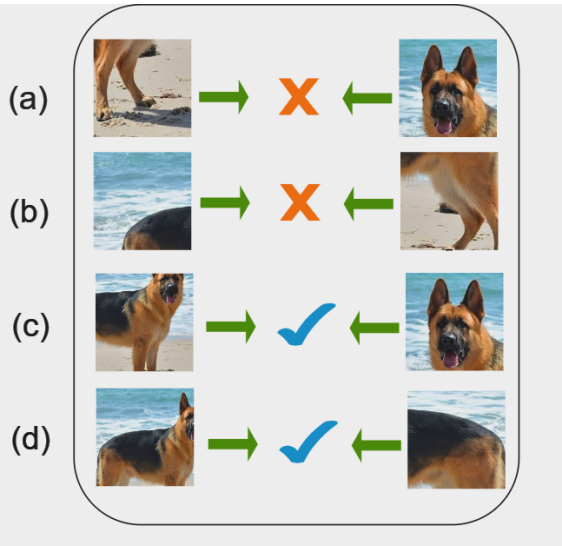


Figure 1: Examples of positive pairs that might be created by random cropping and resizing.

occlusion invariance. In Figure 1 (c and d), the information in the shared region between the two views contains similar semantic content. The dog's head is presented in the two views of positive pairs (c), which facilitates the model capturing the dog's head features on variant scales and angles.

As the examples show, creating random crops for one-centric object does not guarantee obtaining correct semantic pairs. This fact should be considered to improve representation learning. The instance discrimination SSL approaches such as MoCo-v2 (Chen et al., 2020b) and SimCLR (Chen et al., 2020a) encourage the model to bring the positive pairs (i.e., two views for the same instance) closer in the latent space regardless of their semantic content (Xiao et al., 2021; Zhang et al., 2022). This may restrain the model from learning the representation of different object parts and damage semantic features representation (Song et al., 2023a; Zhang et al., 2022) (see Figure 2 (left)).

It has been shown that undesirable views containing different semantic content may be unavoidable when employing random cropping (Song et al., 2023a). Therefore, we need a method to train the model on different object parts to make a robust representation against natural transformations such as scale and occlusion rather than just pulling the augmented views together indiscriminately (Mishra et al., 2021). This issue should be mitigated because the performance of downstream tasks depends on high-quality visual representation learned by self-supervised learning (Alkhalefi et al., 2024; Donahue et al., 2014; Manová et al., 2023; Girshick et al., 2014; Zeiler & Fergus, 2014; Kim & Walter, 2017; Zhang & Ma, 2022; Xiao et al., 2020).

This study introduces a new instance discrimination SSL approach to avoid forcing the model to make similar representations for the two positive views regardless of their semantic content. As shown in Figure 2 (right), we include the original image X in the training process because it encompasses all the semantic features of the views X^1 and X^2. In our approach, the positive pairs (i.e., X^1 and X^2) are pulled to the original image X in the latent space in contrast to the contrastive SOTA approach simCLR (Chen et al., 2020a), and MoCo-v2 (Chen et al., 2020b) which attracted the two views to each other. This training method ensures that the information in the shared region between the attracted views (X,X^1) and (X,X^2) is semantically correct. Therefore, the model representation learning is improved because the model captures better semantic features from the correct semantic positive pairs rather than just matching two random views that might depict different semantic information. In other words, the model learns the representation of diverse parts of the object because the shared region includes correct semantic parts of the object. This is contrary to other approaches, which discard important semantic features due to incorrectly mapping object parts in positive pairs. Our contributions are as follows:

- We introduce a new contrastive instance discrimination SSL method called LeOCLR to alleviate discarding semantic features caused by mapping two random views that are semantically not correct.
- We demonstrate that our approach enhances visual representation learning in Contrastive instance discrimination SSL compared to state-of-the-art (SOTA) approaches using variant downstream tasks.
- We demonstrate that our approach consistently enhances visual representation learning for contrastive instance discrimination across different datasets and contrastive mechanisms.

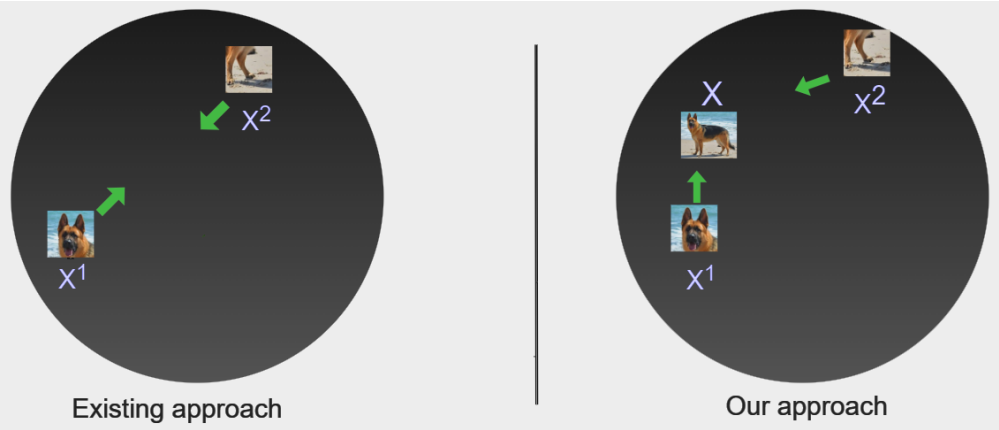


Figure 2: On the left, an existing approach shows the embedding space of the SOTA approaches (Chen et al., 2020a;b) where the two views are attracted to each other regardless of their content. Conversely, the figure on the right depicts our approach, which clusters the two random views together with the original image in the embedding space.

2 Related Work

SSL approaches are divided into two broad categories: contrastive and non-contrastive learning. Broadly speaking, all these approaches aim to attract the positive pairs closer in latent space, but each has a different method to avoid representation collapse. This section provides a brief overview of some of these approaches, but we would like to encourage readers to read the respective papers for more details.

Contrastive Learning: Instance discrimination, such as SimCLR, MoCo, and PIRL (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Misra & Maaten, 2020) employ a similar idea. They attract the positive pairs together and push the negative pairs apart in the embedding space albeit through a different mechanism. SimCLR (Chen et al., 2020a) uses an end-to-end approach where a large batch size is used for the negative examples, and both encoders' parameters in the Siamese network are updated together. PIRL (Misra & Maaten, 2020) uses a memory bank for negative examples, and both encoders' parameters are updated together. MoCo (Chen et al., 2020b; He et al., 2020) uses a momentum contrastive approach whereby the query encoder is updated during backpropagation, and the query encoder updates the key encoder. The negative examples are located in a dictionary separate from the mini-batch, which enables holding large batch sizes.

Non-Contrastive Learning: Non-contrastive approaches use only positive pairs to learn the visual representation with different methods to avoid representation collapse. The first approach is clustering-based methods, where samples with similar features are assigned to the same cluster. DeepCluster (Caron et al., 2018) obtains the pseudo-label from the previous iteration, which makes it computationally expensive and hard to scale. SWAV (Caron et al., 2020) solved this issue by using online clustering, but it needs to determine the correct number of prototypes. The second approach is Knowledge distillation. BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) use techniques inspired by knowledge distillation where a Siamese network has an online encoder and a target encoder. The target network parameters are not updated during

backpropagation. Instead, the online network parameters are updated while being encouraged to predict the representation of the target network. Although these methods have produced promising results, how they avoid collapse has yet to be fully understood. Self-distillation with no labels (DINO) (Caron et al., 2021) was inspired by BYOL, but they use centring with sharping and different backbone (ViT), which enables it to achieve better results than other self-supervised methods while being more computationally efficient. Bag of visual words (Gidaris et al., 2020; 2021) also uses a teacher-student scheme inspired by natural language processing (NLP) to avoid representation collapse. The student network is encouraged to predict the features' histogram for the augmented images, similar to the teacher network's histogram. The last approach is information maximisation. Barlow twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2021) do not require negative examples, stop gradient or clustering. Instead, they use regularisation to avoid representation collapse. The objective function of these methods aims to reduce the redundant information in the embeddings by making the correlation of the embedding vectors closer to the identity matrix. Though these methods provide promising results, they have limitations, such as the representation learning being sensitive to regularisation. The effectiveness of these methods is also reduced if certain statistical properties are not available in the data.

Instance Discrimination With Multi-Crops: Different SSL approaches introduce multi-crop methods to enable the model to learn the visual representation of the object from various aspects. However, creating multi-crop views from the same instance might cause two map views containing distinct semantic information. To solve this issue, LoGo (Zhang et al., 2022) creates two random global crops and N local views. They assume that the global and local views share similar semantic content, thus increasing their similarity, while decreasing the similarity between the local views due to their presumed distinct semantic content. SCFS (Song et al., 2023a) introduces a different solution to solve the unmatched semantic views. They search for semantic-consistent features between the contrasted views. CLSA (Wang & Qi, 2022) creates multi-crops, then applies strong and weak augmentations to the crops. After that, they use distance divergence loss to improve the representation learning of the instance discrimination. The prior approaches assume that the global views contain similar semantic content and treat them indiscriminately as positive pairs. However, our approach argues that the global views may contain incorrect semantic pairs due to random cropping, as illustrated in Figure 1. Therefore, we aim to attract the two global views to the original image (i.e., intact image and not cropped) because it encompasses the semantic features of the crops.

3 Methodology

Mapping incorrect semantic positive pairs (i.e., positive pairs containing different semantic views) results in the discarding of semantic features, degrading the learning of model representations (Mishra et al., 2021; Purushwalkam & Gupta, 2020; Song et al., 2023b). To overcome this, we introduce a new contrastive instance discrimination SSL strategy called LeOCLR. Our approach aims to capture meaningful features from two random positive pairs, even if they contain different semantic content, to enhance representation learning. To achieve this, it is essential to ensure that the information in the shared region between the attracted views is semantically correct. This is because the choice of views controls the information captured by the representations learned in contrastive learning (Tian et al., 2020). Since we cannot guarantee that the shared region between the two views includes correct semantic parts of the object, we propose to involve the original image in the training process. The original image X is intact from cropping (i.e., no random crop), so it encompasses all the semantic features of the two cropped views X^1 and X^2.

As shown in Figure 3 (left), our methodology creates three views (X, X^1, and X^2). The original image (i.e., X) is resized without cropping, while the other views (X^1 and X^2) are randomly cropped and resized. After that, all the views are randomly augmented to avoid the model learning trivial features. We use similar data augmentations that are used in MoCo-v2 (Chen et al., 2020b). Then the original image (i.e., X) is encoded by the encoder f_q and the two views (i.e.,X^1,X^2) are encoded by a momentum encoder f_k which is parameters are updated by the following formula:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{1}$$

where m is the coefficient set to 0.999, (θ_q) are encoder parameters of (f_q) which are updated by the backpropagation and (θ_k) momentum encoder parameters (i.e., f_k) are updated by (θ_q). Finally, the objective

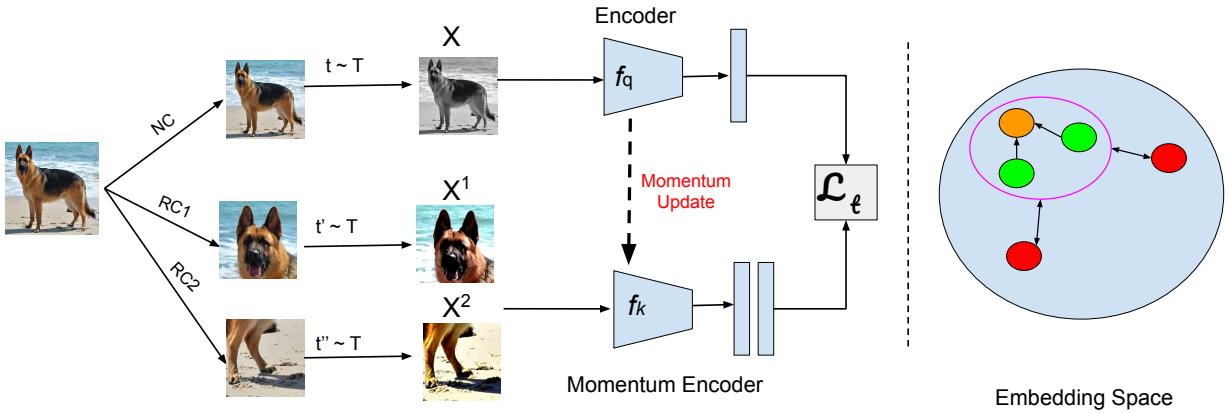


Figure 3: LeOCLR: the concept of the proposed approach. The left part shows that the original image X is not cropped (i.e., NC), just resized to (224×224), and then transformations are applied. The other views (X^1 and X^2) are randomly cropped (i.e., RC1 and RC2) and resized to 224×224. After that, transformations are applied to them. The embedding space of our approach is shown on the right of the Figure.

function forces the model to pull both views (i.e., X^1, X^2) toward the original image (X) in the embedding space and push apart all other instances (as shown in Figure 3 (right)).

3.1 Loss function

Firstly, we briefly describe the loss function of MoCo-v2 (Chen et al., 2020b) since we are using momentum contrastive learning for our approach, and then we will explain our modification to the loss function.

$$\ell(u, v^+) = -\log \frac{\exp(\mathrm{u} \cdot v^+/\tau)}{\sum_{n=0}^{N} \exp(\mathrm{u} \cdot v_n/\tau)}, \tag{2}$$

where the similarity is measured by the dot product. The objective function increases the similarity between the positive pairs ($u \cdot v^+$) by bringing them closer in the embedding space and pushing apart all the negative samples (v_n) in the dictionary to avoid representation collapse. τ is a temperature hyperparameter of softmax. In our approach, we increase the similarity between the original image (i.e., query's feature representation) $u = f_q(x)$ with the positive pair (i.e., key's feature representation) $v^+ = f_k(x^i)$ $(i = 1, 2)$ and push apart all the negative examples(v_n). Therefore the total loss for the mini-batch is:

$$l_t = \sum_{i=1}^{N} \ell(\mathrm{u}_i, sg(v_i^1)) + \ell(\mathrm{u}_i, sg(v_i^2)) \tag{3}$$

Note: $sg(.)$ denotes the stop-gradient trick that is crucial to avoid representation collapse. As shown in Table 3, the l_t (i.e., Total loss) attracts the two views (v_i^1 and v_i^2) to their original instance u_i. This facilitates the model to capture the semantic features from the two random views even though they have distinct semantic information. Our approach captures better semantic features than the prior contrastive approaches (Chen et al., 2020a;b; He et al., 2020) because we ensure that the shared region between the attracted views contains correct semantic information. In other words, the original image contains all the parts of the object, so whatever the object's part contained in the random crop, this part is certainly present in the original image. Thus, when we bring the original image with the two random views closer in the embedding space, the model learns the representation of the different parts and creates an occlusion invariant representation for the object from different scales and angles. This is contrary to the prior approaches, which attract the two views in the embedding space regardless of their semantic content, which leads to discarding semantic features (Liu

et al., 2020; Purushwalkam & Gupta, 2020; Song et al., 2023a) (see Algorithm 1 for the implementation of our approach).

Algorithm 1 Proposed Approach

for X in *dataloader* **do**
  X^1, X^2= rc(X) ▷ random crop first and second views
  X,X^1,X^2 = augment(X,X^1,X^2) ▷ apply random augmentation for all the views
  $X = f_q(X)$ ▷ encode the original image
  $X^1 = f_k(X^1)$ ▷ encode the first view by momentum encoder
  $X^2 = f_k(X^2)$ ▷ encode the second view by momentum encoder
  $loss1 = \ell(X,X^1)$ ▷ computed as shown in $eq.1$
  $loss2 = \ell(X,X^2)$ ▷ computed as shown in $eq.1$
  $l_t = loss1 + loss2$ ▷ computed the total loss as shown in $eq.2$
end for

def rc(x):
  x= T.RandomResizedCrop(224,224) ▷ T is transformation from torchvision module
  return x

Equation 3 and Algorithm 1, illustrate the key differences between our approach and prior multi-crop approaches such as CLSA (Wang & Qi, 2022), SCFC (Song et al., 2023a) , and DINO (Caron et al., 2021). The key differences are as follows:

- The prior approaches assume that the two global views contain the same semantic information; therefore, they encourage the model to capture the similar information between them and create a similar representation for the two views. In our approach, we use original image instead of global views because we assume they might contain wrong semantic information for the same object, which might constrain the model from learning useful semantic features.
- The prior approaches use several local random crops, which might be time and memory-consuming, whereas our approach uses only two random crops (Caron et al., 2020; Wang & Qi, 2022).
- Our objective function uses different methods to facilitate the model's visual representation learning. We encourage the model to make the two random crops similar to the original image, which contains the semantic information for all the random crops while avoiding making the two crops have similar representations because they might not have similar semantic information. This differs from previous approaches, which encourage all crops (i.e., global and local) to have similar representations regardless of their semantic information. This leads to discarding relevant semantic information and subsequently affects the ability to transfer the resulting representations to downstream tasks.

4 Experiments and Results

Datasets: We run multiple experiments on three datasets, i.e., STL-10 "unlabeled" with 100K training images (Coates & Ng, 2011), CIFAR-10 with 50K training images (Krizhevsky, 2009), and ImageNet-1K with 1.28M training images (Russakovsky et al., 2015).
Training Setup: We use ResNet50 as a backbone, and the model is trained with SGD optimizer, weight decay 0.0001, momentum 0.9 and initial learning rate of 0.03. The mini-batch size is 256, and the model is trained for up to 800 epochs on ImageNet-1K.
Evaluation: We evaluated LeOCLR by using linear evaluation and semi-supervised setting against leading SOTA approaches on ImageNet-1K. In linear evaluation, we followed the standard evaluation protocol (Chen et al., 2020a; He et al., 2020; Huynh et al., 2022; Dwibedi et al., 2021). We trained a linear classifier for 100 epochs on top of a frozen backbone pre-trained with LeOCLR. We used the ImageNet training set with random cropping and random left-to-right flipping augmentations to train the linear classifier from scratch.

Table 1: Comparisons between our approach LeOCLR and SOTA approaches on ImageNet.

Approach	Epochs	Batch	Accuracy
MoCo-v2 (Chen et al., 2020b)	800	256	71.1%
BYOL (Grill et al., 2020)	1000	4096	74.4%
SWAV (Caron et al., 2020)	800	4096	75.3%
SimCLR (Chen et al., 2020a)	1000	4096	69.3%
HEXA (Li et al., 2020)	800	256	71.7%
SimSiam (Chen & He, 2021)	800	512	71.3%
VICReg (Bardes et al., 2021)	1000	2048	73.2%
MixSiam (Guo et al., 2021)	800	128	72.3%
OBoW (Gidaris et al., 2021)	200	256	73.8%
DINO (Caron et al., 2021)	800	1024	75.3%
Barlow Twins (Zbontar et al., 2021)	1000	2048	73.2%
CLSA (Wang & Qi, 2022)	800	256	76.2%
RegionCL-M (Xu et al., 2022)	800	256	73.9%
UnMix (Shen et al., 2022)	800	256	71.8%
HCSC (Guo et al., 2022)	200	256	73.3%
UniVIP (Li et al., 2022)	300	4096	74.2%
HAIEV (Zhang & Ma, 2022)	200	256	70.1%
SCFS (Song et al., 2023a)	800	1024	75.7%
LeOCLR(*ours*)	800	256	**76.2%**

The results are reported on the ImageNet validation set with center crop (224×224). In a semi-supervised setting, we fine-tune the network with 60 epochs using 1% labeled data and 30 epochs using 10% labeled data. Finally, we assess the learned features from the ImageNet dataset on small datasets CIFAR (Krizhevsky, 2009) and fine-grained datasets (Krause et al., 2013; Parkhi et al., 2012; Berg et al., 2014) using transfer learning.

Comparing with SOTA Approaches: We use vanilla MoCo-v2 (Chen et al., 2020b) as a baseline to compare it with our approach on different benchmark datasets, given our utilization of a momentum contrastive learning framework. Additionally, we compare our approach with other state-of-the-art (SOTA) methods on the ImageNet-1K dataset.

Table 1 presents the linear evaluation of our approach compared to other SOTA methods. As depicted, our approach outperforms all others. For instance, it surpasses the baseline (i.e., vanilla MoCo-v2) by 5.1%. This highlights our hypothesis that two global views may encapsulate different semantic information for the same object (e.g., a dog's head and leg), which warrants consideration for enhancing representation learning. The observed performance gap (i.e., the difference between vanilla MoCo-v2 and LeOCLR) illustrates that mapping pairs with divergent semantic content hampers representation learning and impedes the model's effectiveness in downstream tasks.

Semi-Supervised Learning on ImageNet: In this part, we evaluate the performance of LeOCLR under the semi-supervised setting. Specifically, we use 1% and 10% of the labeled training data from ImageNet-1K for fine-tuning, which follows the semi-supervised protocol introduced in SimCLR (Chen et al., 2020a). The top-1 accuracy, reported in Table 2 after fine-tuning with 1% and 10% of the training data, showcases LeOCLR's superiority over all compared methods. This can be attributed to LeOCLR's representation learning capabilities especially compared to the other SOTA methods.

Transfer Learning on Downstream Tasks: We evaluate our self-supervised pretrained model using transfer learning when fine-tuned on small datasets such as CIFAR (Krizhevsky, 2009), Stanford Cars (Krause et al., 2013), Oxford-IIIT Pets (Parkhi et al., 2012), and Birdsnap (Berg et al., 2014). We follow similar procedures for transfer learning as in (Chen et al., 2020a; Grill et al., 2020) to find optimal hyperparameters for each downstream task. Table 3 shows that our approach, LeOCLR, outperforms all compared approaches on various downstream tasks. This demonstrates that our model learns useful semantic features, enabling it to generalize better to unseen data in different downstream tasks than other counterpart approaches. Our

Table 2: Semi-supervised training results on ImageNet: Top-1 performances are reported for fine-tuning a pre-trained ResNet-50 with the ImageNet 1% and 10% datasets.
** denotes the results are reproduced in this study.*

Approach\Fraction	ImageNet 1%	ImageNet 10%
MoCo-v2 (Chen et al., 2020b) *	47.6%	64.8%
SimCLR (Chen et al., 2020a)	48.3%	65.6%
BYOL(Grill et al., 2020)	53.2%	68.8%
SWAV (Caron et al., 2020)	53.9%	70.2%
DINO (Dwibedi et al., 2021)	50.2%	69.3%
RegionCL-M (Xu et al., 2022)	46.1%	60.4%
SCFS (Song et al., 2023a)	54.3%	70.5%
LeOCLR(*ours*)	**62.8**%	**71.5**%

Table 3: Transfer learning results from ImageNet with the standard ResNet-50 architecture.
** denotes the results are reproduced in this study.*

Approach	CIFAR-10	CIFAR-100	Car	Birdsnap	Pets
MoCo-v2 (Chen et al., 2020b)*	97.2%	85.6%	91.2%	75.6%	90.3%
SimCLR (Chen et al., 2020a)	97.7%	85.9%	91.3%	75.9%	89.2%
BYOL(Grill et al., 2020)	97.8%	86.1%	91.6%	76.3%	91.7%
DINO (Song et al., 2023a)	97.7%	86.6%	91.1%	-	91.5%
SCFS (Song et al., 2023a)	97.8%	86.7%	91.6%	-	91.9%
LeOCLR(*ours*)	**98.1**%	**86.9**%	**91.6**%	**76.8**%	**92.1**%

method preserves the semantic features of the given objects, thereby improving the model's representation learning ability. As a result, it becomes more effective at extracting important features and predicting correct classes on transferred tasks.

Object Detection Task: To further evaluate the transferability of the learned representation, we fine-tune the model on PASCAL VOC object detection. We use similar settings as in MoCo (Chen et al., 2020b), where we finetune on the VOC07+12 trainval dataset using Faster R-CNN with a R50-C4 backbone and evaluate on VOC07 test dataset. We fine-tuned for 24k iterations ($\approx$ 23 epochs). Our method outperforms all compared approaches, as shown in Table 4. This is because it excels at capturing semantic features compared to the baseline (MoCo-v2) and other solutions tackling the same problem, which leads to superior performance on object detection and other related tasks.

Table 4: Results (Average Precision) for PASCAL VOC object detection using Faster R-CNN with ResNet50-C4.

Approach	AP_{50}	AP	AP_{75}
MoCo-v2 (Chen et al., 2020b)	82.5%	57.4%	64%
CLSA (Wang & Qi, 2022)	83.2%	-	-
SCFS (Wang & Qi, 2022)	83%	57.4%	63.6%
LeOCLR(*ours*)	**83.2%**	**57.5**%	**64.2%**

5 Ablation Studies

In this section, we conduct further analysis of our approach using another contrastive instance discrimination approach, SimCLR(Chen et al., 2020a), to explore how our approach will perform within this end-to-end framework. Also, we conduct studies on the benchmark datasets STL-10 and CIFAR-10 with a different backbone (Resnet18) to check the consistency of our approach with other datasets and backbones. Furthermore, we employ a random crop test to simulate natural transformations, such as variations in scale

Table 5: Comparing vanilla SimCLR with LeOCLR after training our approach 200 epochs on ImageNet

Approach	ImageNet
SimCLR (Chen et al., 2020a)	62%
LeOCLR(*ours*)	**65.5**%

Table 6: Vanilla MoCo-v2 versus LeOCLR on CIFAR-10 and STL-10 with ResNet18.

Approach	STL-10	CIFAR-10
MoCo-v2	80.08%	73.88%
LeOCLR(*ours*)	**85.20**%	**79.59**%

or occlusion of objects appearing in the image, in order to conduct further analysis on the robustness of features learned by our approach, LeOCLR. In addition, we compare our approach with vanilla MoCo when manipulating their data augmentation to see which model's performance is more affected by removing some of the data augmentation. Also, we use different fine-tuning settings to check which model learns better and faster. Finally, we manipulate the attraction strategy and crop of the original image as well as compute the running time of our approach.

We use an end-to-end framework, where the two encoders f_q and f_k are updated via backpropagation, to train a model with our approach for 200 epochs and 256 batch size. Subsequently, we perform a linear evaluation of our model against SimCLR, which uses an end-to-end mechanism. As shown in Table 5, our approach outperforms vanilla SimCLR by a significant margin of 3.5%, demonstrating its suitability for integration with various contrastive learning frameworks.

In Table 6, we evaluate our approach on different datasets (STL-10 and CIFAR-10) using another backbone, namely ResNet18, to ensure its consistency across various backbones and datasets (i.e., scalability). We pre-trained both models (Vanilla MoCo-v2 and LeOCLR) for 800 epochs on both datasets and then conducted a linear evaluation for both models. Our approach demonstrates superior performance on both datasets compared to vanilla MoCo-v2, achieving accuracies of 5.12% and 5.71% on STL-10 and CIFAR-10, respectively.

In Table 7, we reported the top-1 accuracy for vanilla MoCo-v2 and our approach after 200 epochs on ImageNet. Table 7 shows two testing methods: center crop test similar to (Chen et al., 2020a;b): images are resized to 256 pixels along the shorter side using bicubic resampling, after which a 224×224 center crop is applied. The second test is a random crop, where the image is resized to 256×256 but randomly cropped and resized to 224×224. We took the MoCo-v2 center crop result directly from (Chen et al., 2020b), while the random crop result was not reported. Therefore, we replicated the MoCo-v2 with the same hyperparameters used in the original paper to report the center crop, ensuring a fair comparison. According to the results, the performance of MoCo-v2 dropped by 4.3% with random cropping, whereas our approach experienced a smaller drop of 2.8%. This suggests that our approach learns better semantic features, demonstrating greater invariance to natural transformations such as occlusion and variations in object scales. Also, we compare the performance of CLSA (Wang & Qi, 2022) with our approach because we have the same performance after 800 epochs (see Table 1. Note that the CLSA approach uses multi-crop (i.e., five strong and two weak augmentations), while our approach only uses two random crops and the original image. As shown in Table 7 LeOCLR outperforms the CLSA approach by 2.3% after 200 epochs on ImageNet-1K. To alleviate concerns about the increased computational cost associated with training LeOCLR relative to MoCo V2,

Table 7: Comparing LeOCLR with vanilla MoCo-v2 and CLSA after training 200 epochs on ImageNet.

Approach	Center Crop	Random Crop	Time
MoCo-v2 (Chen et al., 2020b)	67.5%	63.2%	68h
CLSA (Wang & Qi, 2022)	69.4%	-	-
LeOCLR(*ours*)	**71.7**%	**68.9**%	81h

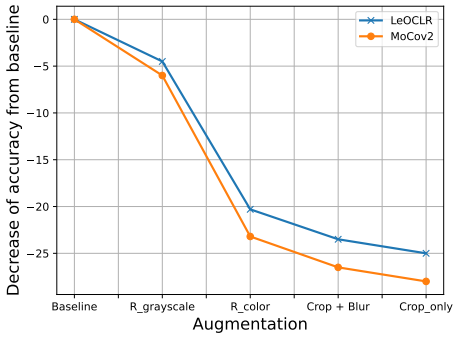
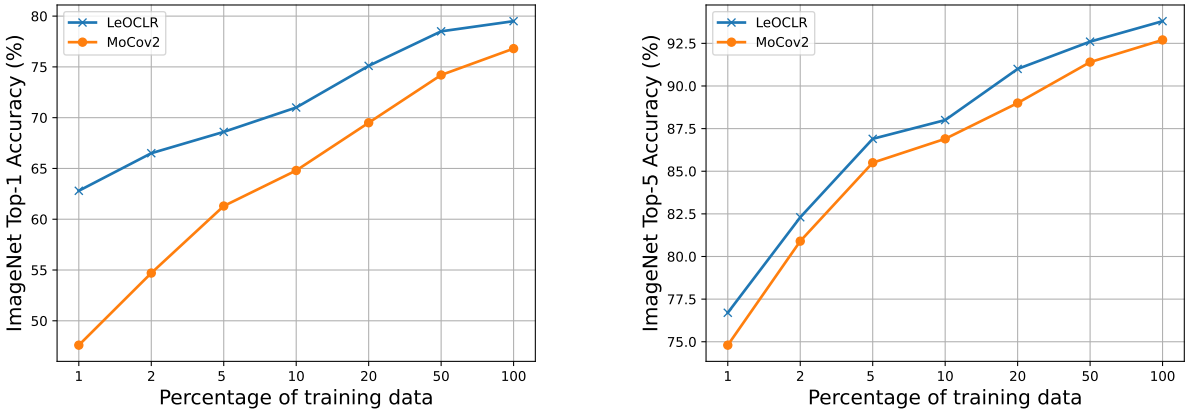


Figure 4: Decrease in top-1 accuracy (in % points) of LeOCLR and our own reproduction of Vanilla MoCo-v2 at 200 epochs, under linear evaluation on ImageNet. *R_Grayscale* means to remove the grayscale augmentations, and *R_color* removes color jitter with grayscale augmentations.



(a) Top-1 accuracy

(b) Top-5 accuracy

Figure 5: Semi-supervised training with a fraction of ImageNet labels on a ResNet-50.

we add the training time of our approach and vanilla MoCo-v2 to Table 7. We trained both on three A100 GPUs 80GB for 200 epochs. Our approach takes 13 hours more to train on the same number of epochs, but our performance is significantly better than the baseline.

Contrastive instance discrimination approaches are sensitive to the choice of image augmentations (Grill et al., 2020). Thus, we do further analysis of our approach against Moco-v2 (Chen et al., 2020b). These experiments aim to see which model learns better semantic features and creates robust representation under different data augmentations. As shown in Figure 4, both models are affected by removing some data augmentations. However, our approach shows a more invariant representation and less performance impact due to transformation manipulation than vanilla MoCo-v2. For example, when we apply only random cropping augmentation, the performance of vanilla MoCo-v2 is reduced by 28 points (i.e., from 67.5% baseline to 39.5% only random cropping), while our approach reduces only 25 points (i.e., from 71.7% baseline to 46.6% only random cropping). This means our approach learns better semantic features and creates better representation for the given objects than vanilla MoCo-v2.

Table 2 presented in Section 4, we fine-tune the representation over the 1% and 10% ImageNet splits from (Chen et al., 2020a) with ResNet-50 architecture. In the ablation study, we compare the fine-tuned representation of our approach and reproduced vanilla MoCo-v2(Chen et al., 2020b) over 1%, 2%, 5%, 10%, 20%, 50%, and 100% of the ImageNet dataset as in (Henaff, 2020; Grill et al., 2020). In this setting, we observed

that tuning a LeOCLR representation always outperforms vanilla MoCo-v2. For instance, Figure 5 (a) shows that LeOCLR fine-tuned with 10% of ImageNet labeled data performed better than Vanilla Moco-v2 (Chen et al., 2020b) fine-tuned with 20% of labeled data. This means that our approach is suitable in case we have smaller labeled data for downstream tasks than vanilla MoCo-v2.

5.1 Augmentations:

In this subsection, we apply a random crop to the original image (x) and attract the two views (x^1,x^2) toward it to see how that will affect the performance of our approach. Also, we conducted another experiment where all the views were attracted to each other. Please note that in our approach, we avoid the two views attracted to each other and enforce the model to attract the two views toward the original image only (i.e., the image without cropping and contains semantic features for all the crops). In these experiments, we pre-trained the model on ImageNet for 200 epochs using the same hyperparameters used with the main experiment.

Table 8: Comparisons across augmentations strategies using our proposed approach after 200 epochs.

Approach	Accuracy
LeOCLR(*Random original image*)	69.3%
LeOCLR(*attract all crops*)	67.7%
LeOCLR(*ours*)	**71.7**%

The experiments in Table 8 Show the importance of the information shared between the two views. In addition, it illustrates the importance of leveraging the original image as well as avoiding attracting views with variant semantic information to preserve the semantic features of the objects. When we create a random crop for the original image (x) and enforce the model to make the two views similar to the original image (i.e., LeOCLR(*Random original image*)), the model performance reduces by 2.4%. This is because when we crop the original image and enforce the model to attract the two views toward it, there is a high probability of having two views containing variant semantic information, which leads to discarding the semantic features of the objects. The problem becomes worse when we attract all the views (x,x^1,x^2) to each other in LeOCLR (*attract all crops*), where the performance becomes closer to the Vanilla MoCo-v2 (67.5%). This is because attracting two views containing distinct semantic information is highly possible.

6 Conclusion

This paper introduces a new contrastive instance discrimination approach for SSL to enhance representation learning. Our approach alleviates discarding semantic features while attracting two views containing distinct semantic content by incorporating the original image in training. Our method consistently improves the representation learning of contrastive instance discrimination across different benchmark datasets, backbone, and various mechanisms, such as momentum contrast and end-to-end methods. In linear evaluation, we achieved an accuracy of 76.2% on ImageNet after 800 epochs, outperforming several instance discrimination SOTA SSL approaches. Also, we showed the invariance and robustness of our approach through different downstream tasks (i.e., transfer learning and semi-supervised fine-tuning).

References

Mohammad Alkhalefi, Georgios Leontidis, and Mingjun Zhong. Semantic positive pairs for enhancing visual representation learning of instance discrimination methods. *Transactions on Machine Learning Research*, 2024.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2018, 2014.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

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

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

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

Adam Coates and Andrew Y Ng. Analysis of large-scale visual recognition. In *Advances in neural information processing systems*, pp. 284–292, 2011.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pp. 647–655. PMLR, 2014.

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

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6928–6938, 2020.

Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Perez. Obow: Online bag-of-visual-words generation for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6840, 2021.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Xiaoyang Guo, Tianhao Zhao, Yutian Lin, and Bo Du. Mixsiam: a mixture-based approach to self-supervised representation learning. *arXiv preprint arXiv:2111.02679*, 2021.

Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9706–9715, June 2022.

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 Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pp. 4182–4192. PMLR, 2020.

Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2785–2795, 2022.

Dong-Ki Kim and Matthew R Walter. Satellite image-based localization via learned embeddings. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2073–2080. IEEE, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Chunyuan Li, Xiujun Li, Lei Zhang, Baolin Peng, Mingyuan Zhou, and Jianfeng Gao. Self-supervised pre-training with hard examples improves visual representations. *arXiv preprint arXiv:2012.13493*, 2020.

Zhaowen Li, Yousong Zhu, Fan Yang, Wei Li, Chaoyang Zhao, Yingying Chen, Zhiyang Chen, Jiahao Xie, Liwei Wu, Rui Zhao, et al. Univip: A unified framework for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14627–14636, 2022.

Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020.

Alžběta Manová, Aiden Durrant, and Georgios Leontidis. S-jea: Stacked joint embedding architectures for self-supervised visual representation learning. *arXiv preprint arXiv:2305.11701*, 2023.

Shlok Mishra, Anshul Shah, Ankan Bansal, Abhyuday Jagannatha, Abhishek Sharma, David Jacobs, and Dilip Krishnan. Object-aware cropping for self-supervised learning. *arXiv preprint arXiv:2112.00319*, 2021.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6707–6717, 2020.

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

Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2216–2224, 2022.

Kaiyou Song, Shan Zhang, Zimeng Luo, Tong Wang, and Jin Xie. Semantics-consistent feature search for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16099–16108, October 2023a.

Kaiyou Song, Shan Zhang, Zimeng Luo, Tong Wang, and Jin Xie. Semantics-consistent feature search for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16099–16108, 2023b.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5549–5560, 2022.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10539–10548, 2021.

Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Regioncl: exploring contrastive region pairs for self-supervised representation learning. In *European Conference on Computer Vision*, pp. 477–494. Springer, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Junbo Zhang and Kaisheng Ma. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16650–16659, 2022.

Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16580–16589, 2022.