Enhancing Visual Representation Learning Through Original Image Utilization in Contrastive Learning

Abstract

Contrastive instance discrimination techniques exhibit superior performance in downstream tasks, including image classification and object detection, compared to supervised learning. However, a strong reliance on data augmentation during representation learning is a hallmark of these methods, potentially causing suboptimal outcomes if not meticulously executed. A prevalent data augmentation approach in contrastive learning involves random cropping followed by resizing. This practice might diminish the quality of representation learning when two random crops encompass disparate semantic information. To counter this, we propose an innovative framework termed LeOCLR (Leveraging Original Images for Contrastive Learning of Visual Representations). This framework integrates a novel instance discrimination strategy and a refined loss function, effectively mitigating the loss of crucial semantic features that may arise from mapping different object segments during representation learning. Our empirical evaluations reveal that LeOCLR consistently enhances representation learning across a spectrum of datasets, surpassing baseline models. Notably, LeOCLR exhibits a 5.1% improvement over MoCo-v2 on ImageNet-1K in linear evaluation and demonstrates superior performance in transfer learning and object detection tasks compared to several other techniques.

1 Introduction

Self-supervised learning (SSL) methods based on instance discrimination are heavily dependent on data augmentations, like random cropping, rotation, and color jitter, to construct invariant representations for all instances within a dataset. These augmentations are used to generate two altered views (positive pairs) of the same instance, which are subsequently drawn closer in the latent space. Simultaneously, strategies are employed to prevent a collapse to a trivial solution, commonly referred to as representation collapse. The efficacy of these methods in acquiring meaningful representations has been demonstrated through various downstream tasks, such as image classification and object detection, serving as proxies for evaluating representation learning. However, these techniques often overlook the crucial aspect that augmented views may diverge in semantic content because of random cropping, potentially degrading the quality of visual representation learning. Creating positive pairs via random cropping and subsequently prompting the model to align them based on shared information in both views poses an increased challenge to the SSL task, ultimately leading to an enhancement in representation quality. Moreover, random cropping followed by resizing guides the model's representation to encompass object-related information across diverse aspect ratios, thereby promoting invariance to occlusions. Conversely, minimizing the feature distance in the latent space, which equates to maximizing similarity, between views that encompass distinct semantic concepts may inadvertently discard valuable image information.

Instances of incorrect semantic positive pairs, which are pairs containing mismatched semantic information about the same object, might arise from random cropping. When the model is compelled to align the representations of different parts of an object closer in the latent space, it may discard crucial semantic features. This occurs because the model's representations are based on the shared area between the two views. If this shared region lacks semantically consistent information, the

.

representations become trivial. For random cropping to be effective and achieve occlusion invariance, the shared area must convey the same semantic meaning in both views. Nevertheless, contrasting pairs that might include diverse semantic information about the same object can be valuable, as it can facilitate learning global features.

The creation of random crops for a one-centric object does not ensure the acquisition of accurate semantic pairs. This observation holds significant importance for the enhancement of representation learning. Instance discrimination SSL techniques encourage the model to approximate positive pairs, i.e., two views of the same instance, in the latent space, irrespective of their semantic content. This limitation might hinder the model's ability to learn representations of different object components and could potentially impair its capability to learn semantic feature representations (see Figure 2 (left) in the original paper).

Undesirable views containing different semantic content may be unavoidable when employing random cropping. Therefore, a method is needed to train the model on different parts of an object, developing robust representations against natural transformations like scale and occlusion, rather than merely pulling augmented views together indiscriminately. Addressing this issue is vital, as downstream task performance relies on high-quality visual representations learned through self-supervised learning.

Our work presents a new instance discrimination SSL approach designed to avoid compelling the model to create similar representations for two positive views, irrespective of their semantic content. As shown in Figure 2 (right) of the original paper, we incorporate the original image X into the training process, since it contains all the semantic features of the views X1 and X2. In our method, the positive pairs (i.e., X1 and X2) are drawn towards the original image X in the latent space, in contrast to contrastive state-of-the-art (SOTA) approaches like SimCLR and MoCo-v2, which draw the two views towards each other. This training method guarantees that the information in the shared region between the attracted views (X, X1) and (X, X2) is semantically accurate. Consequently, the model acquires enhanced semantic features by aligning with the appropriate semantic content, rather than matching random views that might contain disparate semantic information. In essence, the model learns representations of various object parts because the shared region encompasses correct semantic components of the object. This contrasts with other methods that may discard vital semantic features by incorrectly mapping object parts in positive pairs. Our contributions are outlined as follows:

- We present a new contrastive instance discrimination SSL method, LeOCLR, created to minimize the loss of semantic features caused by mapping two semantically inconsistent random views.
- We establish that our method enhances visual representation learning in contrastive instance discrimination SSL, surpassing state-of-the-art techniques across a variety of downstream tasks.
- We show that our method consistently improves visual representation learning for contrastive instance discrimination across multiple datasets and contrastive mechanisms.

2 Related Work

Self-supervised learning (SSL) techniques are categorized into two primary groups: contrastive and non-contrastive learning. While all these techniques endeavor to approximate positive pairs in the latent space, they employ distinct strategies to circumvent representation collapse.

Contrastive Learning: Instance discrimination techniques, such as SimCLR, MoCo, and PIRL, employ a similar concept. These methods bring the positive pairs closer while driving the negative pairs apart in the embedding space, albeit through different mechanisms. SimCLR employs an end-to-end strategy where a large batch size is utilized for negative examples, and the parameters of both encoders in the Siamese network are updated simultaneously. PIRL uses a memory bank for negative examples, and both encoders' parameters are updated together. MoCo adopts a momentum contrastive approach where the query encoder is updated during backpropagation, which subsequently updates the key encoder. Negative examples are maintained in a separate dictionary, facilitating the use of large batch sizes.

Non-Contrastive Learning: Non-contrastive techniques utilize solely positive pairs to learn visual representations, employing a variety of strategies to prevent representation collapse. The

initial category encompasses clustering-based techniques, where samples exhibiting similar features are assigned to the same cluster. DeepCluster employs pseudo-labels from the previous iteration, rendering it computationally demanding and challenging to scale. SWAV addresses this challenge by implementing online clustering, though it necessitates determining the correct number of prototypes. The second category involves knowledge distillation. Techniques like BYOL and SimSiam utilize knowledge distillation methods, where a Siamese network comprises an online encoder and a target encoder. The target network's parameters are not updated during backpropagation. Instead, solely the online network's parameters are updated while being encouraged to predict the representation of the target network. Despite the encouraging results, the mechanism by which these methods prevent collapse remains not fully understood. Inspired by BYOL, Self-distillation with no labels (DINO) employs centering and sharpening, along with a distinct backbone (ViT), enabling it to surpass other self-supervised techniques while maintaining computational efficiency. Another method, Bag of visual words (BoW), employs a teacher-student framework inspired by natural language processing (NLP) to avert representation collapse. The student network predicts a histogram of the features for augmented images, analogous to the teacher network's histogram. The final category is information maximization. Methods like Barlow twins and VICReg eschew negative examples, stop gradient, or clustering. Instead, they utilize regularization to avoid representation collapse. The objective function of these techniques seeks to eliminate redundant information in the embeddings by aligning the correlation of the embedding vectors closer to the identity matrix. While these techniques exhibit encouraging results, they possess limitations, including the sensitivity of representation learning to regularization and reduced effectiveness if certain statistical properties are absent in the data.

Instance Discrimination With Multi-Crops: Various SSL techniques introduce multi-crop strategies to enable models to learn visual representations of objects from diverse perspectives. However, when generating multiple cropped views from the same object instance, these views might contain disparate semantic information. To tackle this issue, LoGo generates two random global crops and N local views. They posit that global and local views of an object share similar semantic content, enhancing similarity between these views. Simultaneously, they contend that different local views possess distinct semantic content, thus diminishing similarity among them. SCFS proposes a different approach for managing unmatched semantic views by searching for semantically consistent features between the contrasted views. CLSA generates multiple crops and applies both strong and weak augmentations, using distance divergence loss to enhance instance discrimination in representation learning. Prior methods assume that global views contain similar semantic content and treat them indiscriminately as positive pairs. However, our technique suggests that global views might contain incorrect semantic pairs due to random cropping, as illustrated in Figure 1 in the original paper. Therefore, we aim to attract the two global views to the original (intact and uncropped) image, which fully encapsulates the semantic features of the crops.

3 Methodology

The mapping of incorrect semantic positive pairs, specifically those containing different semantic views, results in the loss of semantic features, which in turn degrades the model's representation learning. To address this, we propose a novel contrastive instance discrimination SSL strategy called LeOCLR. Our approach is designed to capture meaningful features from two random positive pairs, even when they encompass different semantic content, thereby improving representation learning. Achieving this necessitates ensuring the semantic correctness of the information within the shared region between the attracted views. This is crucial because the selection of views dictates the information captured by the representations learned in contrastive learning. Given that we cannot guarantee the inclusion of correct semantic parts of the object within the shared region between the two views, we propose the inclusion of the original image in the training process. The original image X, which is not subjected to random cropping, encompasses all the semantic features of the two cropped views, X1 and X2.

Our method, illustrated in Figure 3 (left) in the original paper, generates three views (X, X1, and X2). The original image (X) is resized without cropping, while the other views (X1 and X2) undergo random cropping and resizing. All views are then randomly augmented to prevent the model from learning trivial features. We employ data augmentations akin to those used in MoCo-v2. The original image (X) is encoded by the encoder fq, while the two views (X1, X2) are encoded by a momentum encoder fk. The parameters of fk are updated using the formula:

$\theta_k \leftarrow m\theta_k + (1-m)\theta_q$ (1)

where m is a coefficient set to 0.999, θ_q represents the encoder parameters of fq updated through backpropagation, and θ_k denotes the momentum encoder parameters of fk updated by θ_q. Ultimately, the objective function compels the model to draw both views (X1, X2) closer to the original image (X) in the embedding space while simultaneously pushing apart all other instances, as depicted in Figure 3 (right) in the original paper.

3.1 Loss function

Initially, we briefly outline the loss function of MoCo-v2, given our utilization of momentum contrastive learning. Subsequently, we will detail our modification to the loss function.

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

where similarity is quantified by the dot product. The objective function amplifies the similarity between the positive pairs (u . v+) by drawing them closer in the embedding space, while simultaneously driving apart all the negative samples (vn) in the dictionary to prevent representation collapse. τ denotes the temperature hyperparameter of the softmax function. In our method, we augment the similarity between the original image's feature representation, u = fq(x), and the positive pair's feature representation, v+ = fk(xi) (i = 1, 2), while driving apart all the negative examples (vn). Consequently, the total loss for the mini-batch is:

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

where sg(.) denotes the stop-gradient operation, which is vital for averting representation collapse. As depicted in Equation 3, the total loss l_t attracts the two views (v_i^1 and v_i^2) to their original instance u_i. This enables the model to capture semantic features from the two random views, even if they contain different semantic information. Our technique captures improved semantic features compared to prior contrastive methods, as we ensure that the shared region between the attracted views contains accurate semantic information. Since the original image contains all segments of the object, any part contained in the random crop is also present in the original image. Thus, when we draw the original image and the two random views closer in the embedding space, the model learns representations of the different parts, creating an occlusion-invariant representation of the object across various scales and angles. This contrasts with earlier techniques, which draw the two views together in the embedding space regardless of their semantic content, leading to the loss of semantic features.

Equation 3 and Algorithm 1 in the original paper highlight the primary distinctions between our method and prior multi-crop techniques, such as CLSA, SCFC, and DINO. The key differences are as follows:

- Previous methods assume that two global views contain identical semantic information, encouraging the model to concentrate on similarities and generate similar representations for both views. In contrast, our method utilizes the original images instead of global views, as we contend that global views may contain incorrect semantic information for the same object. While they may aid in capturing certain global features, this could restrict the model's capacity to learn more universally applicable semantic features, ultimately impacting performance.

- Prior methods employ several local random crops, which might be time- and memory-intensive, while our method utilizes only two random crops.

- Our objective function employs different strategies to enhance the model's visual representation learning. We encourage the model to align the two random crops with the original image, which encompasses the semantic information for all random crops while avoiding compelling the two crops to have similar representations if they do not share similar semantic information. This approach differs from prior methods, which encourage all crops (global and local) to have similar representations, regardless of their semantic content. Consequently, although useful for learning certain global features, those methods may discard pertinent semantic information, potentially hindering the transferability of the resulting representations to downstream tasks.

4 Experiments

We executed multiple experiments on three datasets: STL-10 "unlabeled", comprising 100,000 training images, CIFAR-10, containing 50,000 training images, and ImageNet-1K, with 1.28 million training images.

Training Setup: We employed ResNet50 as the backbone architecture. The model was trained using the SGD optimizer, with a weight decay set to 0.0001, momentum at 0.9, and an initial learning rate of 0.03. The mini-batch size was configured to 256, and the model underwent training for up to 800 epochs on the ImageNet-1K dataset.

Evaluation: We employed diverse downstream tasks to assess LeOCLR's representation learning against leading SOTA approaches on ImageNet-1K: linear evaluation, semi-supervised learning, transfer learning, and object detection. For linear evaluation, we adhered to the standard evaluation protocol, where a linear classifier was trained for 100 epochs on top of a frozen backbone pre-trained with LeOCLR. The ImageNet-1K training set was used to train the linear classifier from scratch, with random cropping and left-to-right flipping augmentations. Results are presented on the ImageNet-1K validation set using a center crop (224 x 224). In the semi-supervised setting, we fine-tuned the network for 60 epochs using 1% of labeled data and 30 epochs using 10% of labeled data. Additionally, we evaluated the learned features on smaller datasets, such as CIFAR, and fine-grained datasets, using transfer learning. Lastly, we utilized the PASCAL VOC dataset for object detection.

Comparing with SOTA Approaches: We employed vanilla MoCo-v2 as a baseline for comparison with our method across various benchmark datasets, considering our use of a momentum contrastive learning framework. Furthermore, we benchmarked our method against other SOTA techniques on the ImageNet-1K dataset.

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

Approach	Epochs	Batch	Accuracy
MoCo-v2	800	256	71.1%
BYOL	1000	4096	74.4%
SWAV	800	4096	75.3%
SimCLR	1000	4096	69.3%
HEXA	800	256	71.7%
SimSiam	800	512	71.3%
VICReg	1000	2048	73.2%
MixSiam	800	128	72.3%
OBoW	200	256	73.8%
DINO	800	1024	75.3%
Barlow Twins	1000	2048	73.2%
CLSA	800	256	76.2%
RegionCL-M	800	256	73.9%
UnMix	800	256	71.8%
HCSC	200	256	73.3%
UniVIP	300	4096	74.2%
HAIEV	200	256	70.1%
SCFS	800	1024	75.7%
LeOCLR (ours)	800	256	76.2%

Table 1 presents the linear evaluation of our method in comparison to other SOTA techniques. As shown, our method surpasses all others, outperforming the baseline (i.e., vanilla MoCo-v2) by 5.1%. This lends credence to our hypothesis that while two global views can capture certain global features, they may also encompass distinct semantic information for the same object (e.g., a dog's head versus its leg), which should be taken into account to enhance representation learning. The observed performance gap (i.e., the difference between vanilla MoCo-v2 and LeOCLR) demonstrates that mapping pairs with divergent semantic content impedes representation learning and impacts the model's performance in downstream tasks.

Semi-Supervised Learning on ImageNet-1K: In this section, we assess the performance of LeOCLR under a semi-supervised setting. Specifically, we utilize 1% and 10% of the labeled training

data from ImageNet-1K for fine-tuning, adhering to the semi-supervised protocol introduced in SimCLR. The top-1 accuracy, presented in Table 2 after fine-tuning with 1% and 10% of the training data, demonstrates LeOCLR's superiority over all compared techniques. This can be attributed to LeOCLR's enhanced representation learning capabilities, particularly in comparison to other SOTA methods.

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

Approach	ImageNet-1K 1%	ImageNet-1K 10%
MoCo-v2 *	47.6%	64.8%
SimCLR	48.3%	65.6%
BYOL	53.2%	68.8%
SWAV	53.9%	70.2%
DINO	50.2%	69.3%
RegionCL-M	46.1%	60.4%
SCFS	54.3%	70.5%
LeOCLR (ours)	62.8%	71.5%

Transfer Learning on Downstream Tasks: We evaluate our self-supervised pretrained model using transfer learning by fine-tuning it on small datasets such as CIFAR, Stanford Cars, Oxford-IIIT Pets, and Birdsnap. We adhere to the transfer learning procedures to identify optimal hyperparameters for each downstream task. As shown in Table 3, our method, LeOCLR, surpasses all compared approaches on a variety of downstream tasks. This demonstrates that our model acquires valuable semantic features, enabling it to generalize more effectively to unseen data in different downstream tasks compared to other techniques. Our method preserves the semantic features of the given objects, thereby enhancing the model's representation learning capabilities. Consequently, it is more effective at extracting crucial features and predicting correct classes on transferred tasks.

Table 3: Transfer learning results from ImageNet-1K with the standard ResNet-50 architecture. * denotes the results are reproduced in this study.

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

Object Detection Task: To further assess the transferability of the learned representation, we compare our method with other SOTA techniques using object detection on the PASCAL VOC. We follow the same settings as MoCo-v2, fine-tuning on the VOC07+12 trainval dataset using Faster R-CNN with an R50-C4 backbone, and evaluating on the VOC07 test dataset. The model is fine-tuned for 24k iterations (ˇ2248 23 epochs). As shown in Table 4, our method surpasses all compared techniques. This superior performance can be attributed to our model's ability to capture richer semantic features compared to the baseline (MoCo-v2) and other techniques, leading to improved results in object detection and related tasks.

5 Ablation Studies

In the subsequent subsections, we further analyze our approach using a different contrastive instance discrimination technique (i.e., an end-to-end mechanism) to investigate how our method performs within this framework. Moreover, we conduct studies on the benchmark datasets STL-10 and CIFAR-10 using a distinct backbone (ResNet-18) to assess the consistency of our approach across various datasets and backbones. Additionally, we employ a random crop test to simulate natural

Table 4: Results (Average Precision) for PASCAL VOC object detection using Faster R-CNN with ResNet-50-C4.

Approach	AP50	AP	AP75
MoCo-v2	82.5%	57.4%	64%
CLSA	83.2%	-	-
SCFS	83%	57.4%	63.6%
LeOCLR (ours)	83.2%	57.5%	64.2%

transformations, such as variations in scale or occlusion of objects in the image, to analyze the robustness of the features learned by our approach, LeOCLR. We also compare our approach with vanilla MoCo-v2 by manipulating their data augmentation techniques to determine which model's performance is more significantly affected by the removal of certain augmentations. In addition, we experiment with different fine-tuning settings to evaluate which model learns better and faster. Furthermore, we adapt the attraction strategy and cropping method of the original image, as well as compute the running time of our approach. Lastly, we examine our approach on a non-centric object dataset where the probability of mapping two views containing distinct information is higher.

5.1 Different Contrastive Instance Discrimination Framework

We utilize an end-to-end framework in which the two encoders fq and fk are updated through backpropagation to train a model with our approach for 200 epochs with a batch size of 256. Subsequently, we conduct a linear evaluation of our model against SimCLR, which also employs an end-to-end mechanism. As presented in Table 5, our approach outperforms vanilla SimCLR by a substantial margin of 3.5%, demonstrating its suitability for integration with various contrastive learning frameworks.

Table 5: Comparing vanilla SimCLR with LeOCLR after training our approach 200 epochs on ImageNet-1K.

Approach	ImageNet-1K
SimCLR	62%
LeOCLR (ours)	65.5%

5.2 Scalability

In Table 6, we evaluate our approach on different datasets (STL-10 and CIFAR-10) using a ResNet-18 backbone to ensure its consistency across various backbones and datasets (i.e., scalability). We pre-trained all the approaches for 800 epochs with a batch size of 256 on both datasets and then conducted a linear evaluation. Our approach demonstrates superior performance on both datasets compared to all approaches. For instance, our approach outperforms vanilla MoCo-v2, achieving accuracies of 5.12% and 5.71% on STL-10 and CIFAR-10, respectively.

Table 6: SOTA approaches versus LeOCLR on CIFAR-10 and STL-10 with ResNet-18.

Approach	STL-10	CIFAR-10
MoCo-v2	80.08%	73.88%
DINO	84.30%	78.50%
CLSA	82.62%	77.20%
BYOL	79.90%	73.00%
LeOCLR (ours)	85.20%	79.59%

5.3 Center and Random Crop Test

In Table 7, we report the top-1 accuracy for vanilla MoCo-v2 and our approach after 200 epochs on ImageNet-1K, concentrating on two tasks: a) center crop test, where images are resized to 256

pixels along the shorter side using bicubic resampling, followed by a 224 x 224 center crop; and b) random crop, where images are resized to 256 x 256 and then randomly cropped and resized to 224 x 224. 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 improved semantic features, demonstrating greater invariance to natural transformations like occlusion and variations in object scales. Additionally, we compare the performance of CLSA with our approach, given that both perform similarly 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 employs only 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 address concerns about the increased computational cost associated with training LeOCLR compared to MoCo V2, we include the training time for both approaches in Table 7. We trained both models on three A100 GPUs with 80GB for 200 epochs. Our approach took an additional 13 hours to train over the same number of epochs, but it delivers significantly better performance than the baseline.

Table 7: Comparing LeOCLR with vanilla MoCo-v2 and CLSA after training 200 epochs on ImageNet-1K.

Approach	Center Crop	Random Crop	Time
MoCo-v2	67.5%	63.2%	68h
CLSA	69.4%	-	-
LeOCLR (ours)	71.7%	68.9%	81h

graph1.pdf

Figure 1: *
(a) Top-1 accuracy

graph2.pdf

Figure 2: *
(b) Top-5 accuracy

Figure 3: Semi-supervised training with a fraction of ImageNet-1K labels on a ResNet-50.

5.4 Augmentation and Fine-tuning

Contrastive instance discrimination techniques are sensitive to the choice of image augmentations. This sensitivity necessitates further analysis comparing our approach to Moco-v2. These experiments aim to explore which model learns better semantic features and produces more robust representations under different data augmentations. As shown in Figure 4, both models are affected by the removal of certain data augmentations. However, our approach shows a more invariant representation and exhibits less performance degradation due to transformation manipulation compared to vanilla MoCo-v2. For instance, when we apply only random cropping augmentation, the performance of vanilla MoCo-v2 drops by 28 percentage points (from a baseline of 67.5% to 39.5% with only random cropping). In contrast, our approach experiences a decrease of only 25 percentage points (from a baseline of 71.7% to 46.6% with only random cropping). This indicates that our approach learns

improved semantic features and produces more effective representations for the given objects than vanilla MoCo-v2.

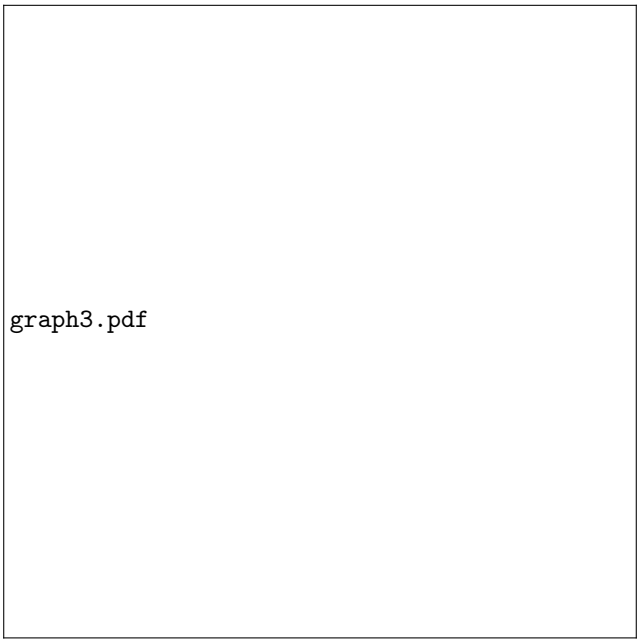

Figure 4: Decrease in top-1 accuracy (in % points) of LeOCLR and our reproduction of vanilla MoCo-v2 after 200 epochs, under linear evaluation on ImageNet-1K. R$_{Grayscalereferstoresultswithoutgrayscaleaugmentations, whileR_{c}olorreferstoresultswithoutcolorjitterbutwith}$

In Table 2, presented in Section 4, we fine-tune the representations over the 1% and 10% ImageNet-1K splits using the ResNet-50 architecture. In the ablation study, we compare the fine-tuned representations of our approach with the reproduced vanilla MoCo-v2 across 1%, 2%, 5%, 10%, 20%, 50%, and 100% of the ImageNet-1K dataset. In this setting, we observe that tuning a LeOCLR representation consistently outperforms vanilla MoCo-v2. For instance, Figure 3 (a) demonstrates that LeOCLR fine-tuned with 10% of ImageNet-1K labeled data outperforms vanilla Moco-v2 fine-tuned with 20% of labeled data. This indicates that our approach is advantageous when the labeled data for downstream tasks is limited.

5.5 Attraction Strategy

In this subsection, we apply a random crop to the original image (x) and attract the two views (x1, x2) toward it to evaluate its impact on our approach's performance. We also conducted an experiment where all views were attracted to each other. However, in our method, we avoid attracting the two views to each other, enforcing the model to draw the two views toward the original image only (i.e., the uncropped image containing semantic features for all crops). For these experiments, we pre-trained the model on ImageNet-1K for 200 epochs using the same hyperparameters employed in the main experiment. The experiments in Table 8 underscore the significance of the information shared between the two views. They also highlight the importance of leveraging the original image and avoiding the attraction of views containing varied semantic information to preserve the semantic features of the objects. When we create a random crop of the original image (x) and force the model to make the two views similar to the original image (i.e., LeOCLR(Random original image)), the model performance decreases by 2.4%.

This performance reduction occurs because cropping the original image and compelling the model to attract the two views towards it increases the probability of having two views with differing semantic information, resulting in a loss of semantic features of the objects. The situation deteriorates when we attract all views (x, x1, x2) to each other in LeOCLR (attract all crops), causing performance to drop closer to that of vanilla MoCo-v2 (67.5%). This decline is attributed to the high likelihood of attracting two views containing distinct semantic information.

Table 8: Comparisons of augmentation 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%

5.6 Non-Object-Centric Tasks

Non-object-centric datasets, like COCO, depict real-world scenes where the objects of interest are not centered or prominently positioned, unlike object-centric datasets such as ImageNet-1K. In this scenario, the chance of generating two views containing distinct semantic information for the object is elevated, thus exacerbating the issue of losing semantic features. Therefore, we train both our approach and the MoCo-v2 baseline from scratch on the COCO dataset to evaluate how our method manages the discarding of semantic features in such datasets. We utilized identical hyperparameters as for ImageNet-1K, training the models with a batch size of 256 over 500 epochs. Subsequently, we fine-tuned these pre-trained models on the COCO dataset for object detection.

Table 9: Results for pre-training followed by fine-tuning on COCO for object detection using Faster R-CNN with ResNet-50-C4.

Approach	AP50	AP	AP75
MoCo-v2	57.2%	37.6%	41.5%
LeOCLR (ours)	59.3%	39.1%	43.0%

Table 9 reveals that our approach captured enhanced semantic features for the given object compared to the baseline. This emphasizes that our method of avoiding the attraction of two distinct views is more effective at preserving semantic features, even in a non-object-centric dataset.

6 Conclusion

This paper presents a new contrastive instance discrimination approach for SSL to improve representation learning. Our method reduces the loss of semantic features by including the original image during training, even when the two views contain different semantic content. We show that our approach consistently enhances the representation learning of contrastive instance discrimination across various benchmark datasets, backbones, and mechanisms, including momentum contrast and end-to-end methods. In linear evaluation, we achieved an accuracy of 76.2% on ImageNet-1K after 800 epochs, surpassing several SOTA instance discrimination SSL methods. Furthermore, we demonstrated the invariance and robustness of our approach across different downstream tasks, such as transfer learning and semi-supervised fine-tuning.