LeOCLR: Leveraging Original Images for Contrastive Learning of Visual Representations

Mohammad Alkhalefi, Georgios Leontidis, and Mingjun Zhong

Department of Computing Science, University of Aberdeen, UK
{m.alkhalefi1.21, georgios.leontidis, mingjun.zhong}@abdn.ac.uk

Abstract. Contrastive instance discrimination outperforms supervised learning in downstream tasks like image classification and object detection. However, this approach heavily relies 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 that ensures the shared region between positive pairs is semantically correct. 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.

Keywords: Self-supervised · Visual representation · Semantic features · Contrastive learning · Instance discrimination

1 Introduction

Self-supervised learning (SSL) approaches based on instance discrimination [7–9,16,30] 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 representation 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 [27, 29, 33, 36]. On the one hand, creating positive pairs by random cropping and encouraging the model to bring these two views closer in the latent space based on the information in the shared region between the two views makes the SSL model task

harder and improves representation quality [7,29]. In addition, random cropping followed by resizing leads model representation to capture information for the object from varying aspect ratios and induce occlusion invariance [32]. Conversely, minimizing the feature distance (i.e., maximizing similarity) between views containing distinct semantic concepts tends to result in the loss of valuable image information [32,33,36].

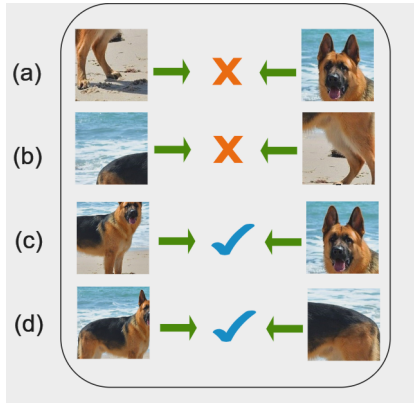


Fig. 1: Examples of positive pairs that might be created by random cropping and resizing.

Fig. 1 (a and b) show examples of wrong semantic positive pairs 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 into 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 occlusion invariance. In Fig. 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 [8] and SimCLR [7] encourage the model to bring the positive pairs closer in the latent space regardless of their semantic content [33,40]. This may restrain the model from learning the representation of different object parts and damage semantic features representation [33,36] (see Fig. 2 (left)).

It has been shown that undesirable views containing different semantic content may be unavoidable when employing random cropping [36]. 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 [29]. This issue should be mitigated because the performance of downstream tasks depends on high-quality visual representation learnt by self-supervised learning [1, 11, 15, 22, 28, 43]

This study introduces a new SSL training approach to solve the issue of existing approaches [7, 8, 19], which attract two random views regardless of the distinct in their semantic content. As shown in Fig. 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 rather than attracted 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.
- We demonstrate that our approach consistently enhances visual representation learning for contrastive instance discrimination across different datasets and transfer learning scenarios.

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 [7, 8, 19, 30] employ a similar idea. They attract the positive pairs together and push the negative pairs apart in the embedding space albeit through

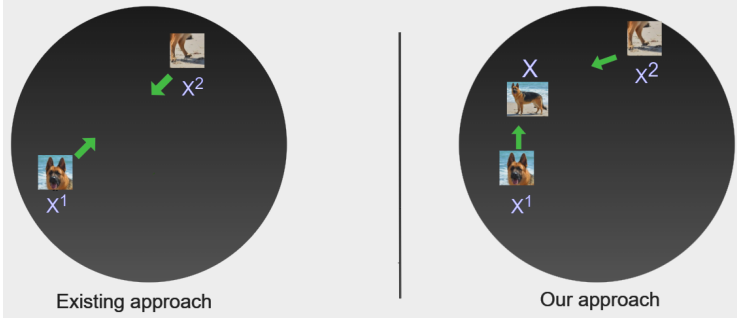


Fig. 2: On the left, an existing approach shows the embedding space of the SOTA approaches [7, 8] where the two views are attracted to each other regardless of their content. Conversely, the figure on the right depicts our approach, which clusters the two random views together with the original image in the embedding space.

a different mechanism. SimCLR [7] 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 [30] uses a memory bank for negative examples, and both encoders' parameters are updated together. MoCo [8, 19] 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 [4] obtains the pseudo-label from the previous iteration, which makes it computationally expensive and hard to scale. SWAV [5] solved this issue by using online clustering, but it needs to determine the correct number of prototypes. The second approach is Knowledge distillation. BYOL [16] and SimSiam [9] 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) [6] 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 [13, 14] 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 [42] and VICReg [2] 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 [33] 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 [36] introduces a different solution to solve the unmatched semantic views. They search for semantic-consistent features between the contrasted views. CLSA [39] 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 Fig. 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. 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 that 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 learnt in contrastive learning [38]. 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.

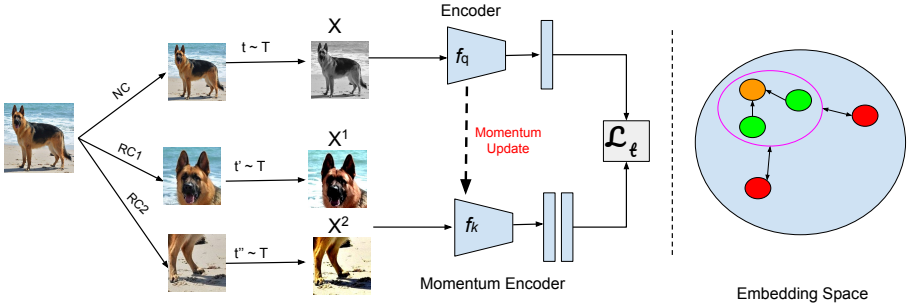


Fig. 3: LeOCLR: the concept of the proposed approach. The left part shows that the original image X is not cropped (i.e., NC), just resized to (224×224), and then transformations are applied. The other views (X^1 and X^2) are randomly cropped (i.e., RC1 and RC2) and resized to 224×224. After that, transformations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

As shown in Fig. 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 [8]. 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 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 Fig. 3 (right)).

3.1 Loss function

Firstly, we briefly describe the loss function of MoCo-v2 [8] 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:

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

Note: $sg(.)$ denotes the stop-gradient trick that is crucial to avoid representation collapse. As shown in Eq. (3), the $TotalLoss$ 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 [7, 8, 19] 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 [27, 32, 36] (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 $eq.1$
  $TotalLoss = loss1 + loss2$ ▷ 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

4 Experiments

Datasets: We run multiple experiments on three datasets, i.e., STL-10 "unlabeled" with 100K training images [10], CIFAR-10 with 50K training images [24], and ImageNet-1K with 1.28M training images [34].
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 [7, 12, 19, 21]. 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. 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 [24] and fine-grained datasets [3, 23, 31] using transfer learning.
Comparing with SOTA Approaches: We use vanilla MoCo-v2 [8] 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.

Tab. 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 [7]. The top-1 accuracy, reported in Tab. 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 [24], Stanford Cars [23], Oxford-IIIT Pets [31], and Birdsnap [3]. We follow similar procedures for transfer learning as in [7, 16] to find optimal

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

Approach	Epochs	Batch	Accuracy
MoCo-v2 [8]	800	256	71.1%
BYOL [16]	1000	4096	74.4%
SimCLR [7]	1000	4096	69.3%
SimSiam [9]	800	512	71.3%
VICReg [2]	1000	2048	73.2%
SWAV [5]	800	4096	75.3%
OBoW [14]	200	256	73.8%
DINO [6]	800	1024	75.3%
Barlow Twins [42]	1000	2048	73.2%
CLSA [39]	800	256	76.2%
HCSC [18]	200	256	73.3%
UniVIP [26]	300	4096	74.2%
SCFS [37]	800	1024	75.7%
RegionCL-M [41]	800	256	73.9%
UnMix [35]	800	256	71.8%
HEXA [25]	800	256	71.7%
MixSiam [17]	800	128	72.3%
LeOCLR(*ours*)	800	256	**76.2%**

hyperparameters for each downstream task. Tab. 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 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.

5 Ablation Studies

In this section, we conduct further analysis of our approach using another contrastive instance discrimination approach SimCLR [7] to explore how our approach will perform within this end-to-end framework. Also, we perform 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 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. Fi-

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 [8] *	47.6%	64.8%
SimCLR [7]	48.3%	65.6%
BYOL [16]	53.2%	68.8%
SWAV [5]	53.9%	70.2%
DINO [12]	50.2%	69.3%
SCFS [37]	54.3%	70.5%
RegionCL-M [41]	46.1%	60.4%
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 [8]*	97.2%	85.6%	91.2%	75.6%	90.3%
SimCLR [7]	97.7%	85.9%	91.3%	75.9%	89.2%
BYOL [16]	97.8%	86.1%	91.6%	76.3%	91.7%
DINO [37]	97.7%	86.6%	91.1%	-	91.5%
SCFS [37]	97.8%	86.7%	91.6%	-	91.9%
LeOCLR(*ours*)	**98.1**%	**86.9**%	**91.6**%	**76.8**%	**92.1**%

nally, we use different fine-tuning settings to check which model learns better and faster.

Table 4: Comparing vanilla SimCLR with LeOCLR after training our approach 200 epochs on ImageNet

Approach	ImageNet
SimCLR [7]	62%
LeOCLR(*ours*)	65.5%

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 Tab. 4, our approach outperforms vanilla SimCLR by a significant margin of 3.5%, demonstrating its suitability for integration with various contrastive learning frameworks.

In Tab. 5, we evaluate our approach on different datasets (STL-10 and CIFAR-10) using another backbone, namely ResNet18, to ensure its consistency

Table 5: 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%

across various backbones and datasets. 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.

Table 6: Comparing LeOCLR with vanilla MoCo-v2 and CLSA after training 200 epochs on ImageNet.

Approach	Center Crop	Random Crop
MoCo-v2 [8]	67.5%	63.2%
CLSA [39]	69.4%	-
LeOCLR(*ours*)	71.7%	68.9%

In Tab. 6, we reported the top-1 accuracy for vanilla MoCo-v2 and our approach after 200 epochs on ImageNet. Tab. 6 shows two testing methods: center crop test similar to [7,8]: 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 [8], 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, as it demonstrates greater invariance to natural transformations such as occlusion and variations in object scales. Also, we compare the performance of CLSA [39] with our approach because we have the same performance after 800 epochs (see Tab. 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 Tab. 5 LeOCLR outperforms the CLSA approach by 2.3% after 200 epochs on ImageNet-1K.

Contrastive instance discrimination approaches are sensitive to the choice of image augmentations [16]. Thus, we do further analysis of our approach against Moco-v2 [8]. These experiments aim to see which model learns better semantic features and creates robust representation under different data augmentations. As shown in Fig. 4, both models are affected by removing some data augmen-

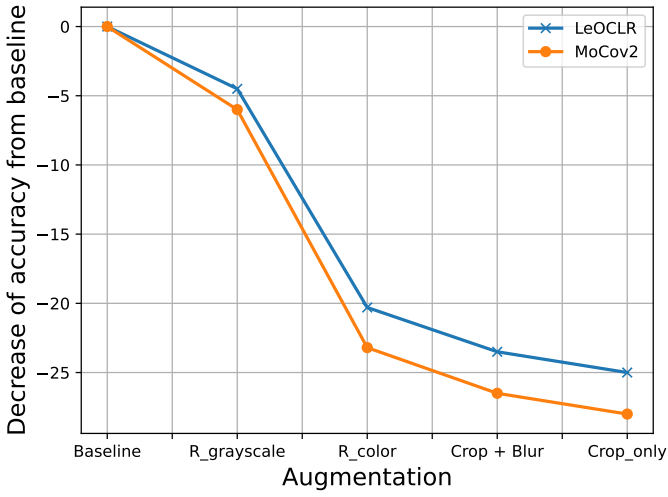
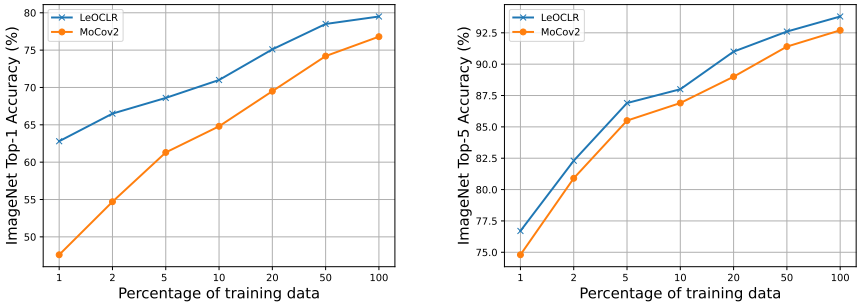


Fig. 4: Decrease in top-1 accuracy (in % points) of LeOCLR and our own reproduction of Vanilla MoCo at 200 epochs, under linear evaluation on ImageNet. *R_Grayscale* means to remove the grayscale augmentations, and *R_color* removes color jitter with grayscale augmentations.

tations. 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.

Tab. 2 presented in Sec. 4, we fine-tune the representation over the 1% and 10% ImageNet splits from [7] with ResNet-50 architecture. In the ablation study, we compare the fine-tuned representation of our approach and reproduced vanilla MoCo-v2 [8] over 1%, 2%, 5%, 10%, 20%, 50%, and 100% of the ImageNet dataset as in [16, 20]. In this setting, we observed that tuning a LeOCLR representation always outperforms vanilla MoCo-v2. For instance, Fig. 5 (a) shows that LeOCLR fine-tuned with 10% of ImageNet labeled data performed better than Vanilla Moco-v2 [8] fine-tuned with 20% of labeled data. This means that our approach is suitable in case we have small labeled data for downstream task than vanilla MoCo-v2.



(a) Top-1 accuracy (b) Top-5 accuracy

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

6 Conclusion

In this paper, we introduce a new SSL approach to improve contrastive instance discrimination representation learning. Our approach alleviates discarding semantic features while attracting two views containing distinct semantic content by incorporating the original image in training. Our approach consistently enhances representation learning for contrastive instance discrimination across different benchmark datasets 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 SOTA SSL approaches. Through extensive ablation studies and experiments, we have demonstrated the robustness and invariance of our method to different backbones and datasets. These findings suggest that our method could be a promising candidate for adoption in various settings where semi-supervised learning methods are employed, extending beyond the contexts considered in this paper.

Acknowledgments

We would like to thank University of Aberdeen's High Performance Computing facility for enabling this work.

References

1. Alkhalefi, M., Leontidis, G., Zhong, M.: Semantic positive pairs for enhancing contrastive instance discrimination. arXiv preprint arXiv:2306.16122 (2023)
2. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)

3. Berg, T., Liu, J., Woo Lee, S., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: 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)
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
9. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
10. Coates, A., Ng, A.Y.: Analysis of large-scale visual recognition. In: Advances in neural information processing systems. pp. 284–292 (2011)
11. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655. PMLR (2014)
12. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: 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)
13. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: 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)
14. Gidaris, S., Bursuc, A., Puy, G., Komodakis, N., Cord, M., Perez, P.: 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)
15. Girshick, R., Donahue, J., Darrell, T., Malik, J.: 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)
16. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
17. Guo, X., Zhao, T., Lin, Y., Du, B.: Mixsiam: a mixture-based approach to self-supervised representation learning. arXiv preprint arXiv:2111.02679 (2021)
18. Guo, Y., Xu, M., Li, J., Ni, B., Zhu, X., Sun, Z., Xu, Y.: Hcsc: Hierarchical contrastive selective coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9706–9715 (June 2022)

19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
20. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: International conference on machine learning. pp. 4182–4192. PMLR (2020)
21. Huynh, T., Kornblith, S., Walter, M.R., Maire, M., Khademi, M.: 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)
22. Kim, D.K., Walter, M.R.: Satellite image-based localization via learned embeddings. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 2073–2080. IEEE (2017)
23. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 2013 IEEE International Conference on Computer Vision Workshops. pp. 554–561 (2013). https://doi.org/10.1109/ICCVW.2013.77
24. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
25. Li, C., Li, X., Zhang, L., Peng, B., Zhou, M., Gao, J.: Self-supervised pre-training with hard examples improves visual representations. arXiv preprint arXiv:2012.13493 (2020)
26. Li, Z., Zhu, Y., Yang, F., Li, W., Zhao, C., Chen, Y., Chen, Z., Xie, J., Wu, L., Zhao, R., 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)
27. Liu, S., Li, Z., Sun, J.: Self-emd: Self-supervised object detection without imagenet. arXiv preprint arXiv:2011.13677 (2020)
28. Manová, A., Durrant, A., Leontidis, G.: S-jea: Stacked joint embedding architectures for self-supervised visual representation learning. arXiv preprint arXiv:2305.11701 (2023)
29. Mishra, S., Shah, A., Bansal, A., Jagannatha, A., Sharma, A., Jacobs, D., Krishnan, D.: Object-aware cropping for self-supervised learning. arXiv preprint arXiv:2112.00319 (2021)
30. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6707–6717 (2020)
31. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
32. Purushwalkam, S., Gupta, A.: Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. Advances in Neural Information Processing Systems **33**, 3407–3418 (2020)
33. Qiu, T.Z.C., Süsstrunk, W.K.S., Salzmann, M.: Leverage your local and global representations: A new self-supervised learning strategy supplementary material
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
35. Shen, Z., Liu, Z., Liu, Z., Savvides, M., Darrell, T., Xing, E.: Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2216–2224 (2022)

36. Song, K., Zhang, S., Luo, Z., Wang, T., Xie, J.: Semantics-consistent feature search for self-supervised visual representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16099–16108 (2023)
37. Song, K., Zhang, S., Luo, Z., Wang, T., Xie, J.: 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 2023)
38. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? Advances in Neural Information Processing Systems **33**, 6827–6839 (2020)
39. Wang, X., Qi, G.J.: Contrastive learning with stronger augmentations. IEEE transactions on pattern analysis and machine intelligence **45**(5), 5549–5560 (2022)
40. Xiao, T., Reed, C.J., Wang, X., Keutzer, K., Darrell, T.: Region similarity representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10539–10548 (2021)
41. Xu, Y., Zhang, Q., Zhang, J., Tao, D.: Regioncl: exploring contrastive region pairs for self-supervised representation learning. In: European Conference on Computer Vision. pp. 477–494. Springer (2022)
42. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
43. Zeiler, M.D., Fergus, R.: 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)