Towards Efficient and Effective Self-Supervised Learning of Visual Representations

Sravanti Addepalli*, Kaushal Bhogale*, Priyam Dey, and R.Venkatesh Babu

Video Analytics Lab, Department of Computational and Data Sciences, Indian Institute of Science, Bangalore

Abstract. Self-supervision has emerged as a propitious method for visual representation learning after the recent paradigm shift from handcrafted pretext tasks to instance-similarity based approaches. Most state-of-the-art methods enforce similarity between various augmentations of a given image, while some methods additionally use contrastive approaches to explicitly ensure diverse representations. While these approaches have indeed shown promising direction, they require a significantly larger number of training iterations when compared to the supervised counterparts. In this work, we explore reasons for the slow convergence of these methods, and further propose to strengthen them using well-posed auxiliary tasks that converge significantly faster, and are also useful for representation learning. The proposed method utilizes the task of rotation prediction to improve the efficiency of existing state-of-the-art methods. We demonstrate significant gains in performance using the proposed method on multiple datasets, specifically for lower training epochs.

1 Introduction

The unprecedented progress achieved using Deep Neural Networks over the past decade was fuelled by the availability of large-scale labelled datasets such as ImageNet [9], coupled with a massive increase in computational capabilities. While their initial success was contingent on the availability of annotations in a supervised learning framework [22, 32, 17, 24], recent years have witnessed a surge in self-supervised learning methods, which could achieve comparable performance, albeit using a higher computational budget and larger model capacities [4, 6, 15, 3]. Early self-supervised approaches [34, 26, 13] aimed at learning representations while solving specialized tasks that require a semantic understanding of the content to accomplish. While generative networks such as task-specific encoder-decoder architectures [21, 35, 29] and Generative Adversarial Networks (GANs) [14, 12] could learn useful representations, they were superseded by the use of discriminative tasks such as solving Jigsaw puzzles [26] and rotation prediction [13], as the latter could be achieved using lower model capacities and lesser compute. The surprisingly simple task of rotating every image by

* Equal contribution.
Correspondence to: Sravanti Addepalli <sravantia@iisc.ac.in>

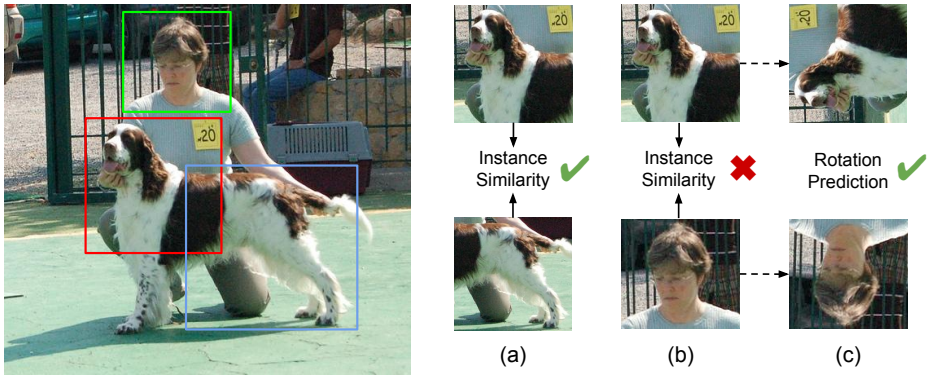


Fig. 1: We demonstrate noise in the training objective of instance-similarity based learning tasks. Consider the three random crops shown in the input image. The two crops in (a) are desirable, while the crops shown in (b) give an incorrect signal to the network. Since the task of rotation prediction shown in (c) aims to predict the rotation angle of each cropped image independently, there is no noise in the training objective.

a random angle from the set $\{0°, 90°, 180°, 270°\}$, and training the network to predict this angle was seen to outperform other handcrafted task based methods with a similar convergence rate as supervised training [13]. Compared to these pretext task based methods, recent approaches have achieved a significant boost in performance by learning similar representations across various augmentations of a given image [16, 15, 4, 3, 6]. While these methods show improvements at a low training budget as well, they achieve a further boost when trained for a larger number of epochs [6], indicating that improving the convergence of such methods can lead to valuable gains at a low computational cost.

In this work, we empirically show that a key reason for the slow convergence of instance-similarity based approaches is the presence of noise in the training objective, owing to the nature of the learning task, as shown in Fig.1. We further propose to strengthen the recent state-of-the-art instance-similarity based self-supervised learning algorithms such as BYOL [15] and SwAV [3] using a noise-free auxiliary training objective such as rotation prediction in a multi-task framework. As shown in Fig.4, this leads to a similar convergence rate as RotNet [13], while also resulting in better representations from the instance-similarity based objective. We further study the invariance of the network to geometric transformations, and show that in natural images, rotation invariance hurts performance and learning covariant representations across multiple rotated views leads to improved results. We demonstrate significant gains in performance across multiple datasets - CIFAR-10, CIFAR-100 [23] and ImageNet-100 [33, 9], and the scalability of the proposed approach to ImageNet-1k [9] as well.

Our code is available here: https://github.com/val-iisc/EffSSL.

2 Related Works

2.1 Handcrafted Pretext task based methods

Discriminative pretext tasks use pseudo-labels that are generated automatically without the need for human annotations. This includes tasks based on spatial context of images such as context prediction [10], image jigsaw puzzle [26] and counting visual primitives [27].

RotNet: Rotation prediction, proposed by Gidaris et al. [13], has been one of the most successful pretext tasks for the learning of useful semantic representations. In this approach, every image is transformed using all four rotation transformations, and the network is trained to predict the corresponding rotation angle used for transforming the image. Due to its simplicity and effectiveness, the rotation task has been used to improve the training of GANs [14, 5] as well.

Multi-task Learning: Doersch and Zisserman [11] investigated methods for combining several pretext tasks in a multi-task learning framework to learn better representations. Contrary to a general multi-task learning setting, in this work we aim to improve instance similarity based tasks such as BYOL [15] and SwAV [3] using handcrafted pretext tasks. We empirically show that the training objective of instance-similarity based tasks is noisy, and combining it with the well-defined objective of rotation prediction leads to improved performance.

2.2 Instance Discriminative approaches

Recent approaches aim to learn similar representations for different augmentations of the same image, while generating diverse representations across different images. Several works achieve this using contrastive learning approaches [28, 18, 4, 16, 25], where multiple augmentations of a given image are considered as positives, and augmentations of other images are considered as negatives. PIRL [25] and MoCo [16] maintain a queue to sample more negatives.

SimCLR: The work by Chen et al. [4] presents a Simple Framework for Contrastive Learning of Visual Representations (SimCLR), that utilizes existing architectures such as ResNet [17], and avoids the need for memory banks. The authors proposed the use of multiple data augmentations and a learnable nonlinear transformation between representations to improve the effectiveness of contrastive learning. Two independent augmentations for every image are considered as positives in the contrastive learning task, while the augmentations of all other images are considered as negatives. The network is trained by minimizing the normalized temperature-scaled cross entropy loss (NT-Xent) loss.

BYOL, SimSiam: While prior approaches relied on the use of negatives for training, Grill et al. [15] proposed Bootstrap Your Own Latent (BYOL), which could achieve state-of-the-art performance without the use of negatives. The two augmentations of a given image are passed through two different networks - the base encoder and the momentum encoder. The base encoder is trained such that the representation at its output can be used to predict the representation at the output of the momentum encoder, using a predictor network. Chen and He [6]

show that it is indeed possible to avoid a collapsed representation even without the momentum encoder using Simple Siamese (SimSiam) networks, and that the stop-gradient operation is crucial for achieving this.

Clustering based methods, SwAV: Clustering-based self-supervised approaches use pseudo-labels from the clustering algorithm to learn representations. DeepCluster [2] alternates between using k-means clustering for producing pseudo-labels, and training the network to predict the same. Asano *et al.*[1] show that degenerate solutions exist in the DeepCluster [2] algorithm. To address this, they cast the pseudo-label assignment problem as an instance of the optimal transport problem and solve it efficiently using a fast variant of the Sinkhorn-Knopp algorithm [7]. SwAV [3] also uses the Sinkhorn-Knoop algorithm for clustering the data while simultaneously enforcing consistency between cluster assignments by Swapping Assignments between Views (SwAV), and using them as targets for training.

2.3 Relation with concurrent works

There has been some recent interest towards improving instance-similarity based approaches by combining them with pretext tasks [20, 8] . In particular, Kinakh et al. [20] show that the use of pretext auxiliary tasks in addition to the contrastive loss can boost the accuracy of models like ScatNet and ResNet-18 on small-scale datasets like STL-10 and CIFAR-100-20. Dangovski et al. [8] claim that learning equivariant representations is better than learning invariant representations, and hence the auxiliary rotation prediction task helps. Our work complements these efforts, and highlights a key issue in the instance-discriminative learning objective: the *impact of noise* in their slow convergence, and shows that combining them with a *noise-free* auxiliary pretext task can significantly improve their efficiency and effectiveness.

3 Motivation

The evolution of self-supervised learning algorithms from handcrafted pretext task-based methods [34, 13, 26] to instance discriminative approaches [16, 15, 4, 3, 6] has indeed led to a significant boost in the performance of downstream tasks. However, as shown in Fig.4, the latter require a larger number of training epochs for convergence. In this section, we show using controlled experiments that the slow convergence of instance-discriminative algorithms can be attributed to a noisy training objective, and eliminating this noise can lead to improved results.

3.1 Impact of False Negatives in SimCLR

The contrastive learning objective in SimCLR [4] considers two augmentations of a given image as positives and the augmentations of all other images in the batch as negatives. These negatives could belong to the same class as the anchor image, and possibly be as similar to the anchor image as the corresponding

Table 1: **Eliminating False Negatives** in contrastive learning across varying levels of supervision (% Labels). Elimination of noise in the training objective leads to higher linear evaluation accuracy (%) within a fixed training budget.

% Labels	SimCLR [4]	Ours	Gain (%)
0	88.77	90.91	2.14
30	92.26 $_{+3.49}$	93.94	1.68
50	92.93 $_{+0.67}$	94.02	1.09
100	93.27 $_{+0.34}$	94.15	0.88

Table 2: **Eliminating False Positives** in BYOL [15] across varying levels of supervision (% Good Crops). Elimination of noise in the training objective leads to higher linear evaluation accuracy (%) within a fixed training budget.

% Good Crops	BYOL [15]	Ours	Gain (%)
0	63.64	68.62	4.98
25	64.50 $_{+0.86}$	68.30	3.80
50	66.30 $_{+1.80}$	68.90	2.60
100	66.72 $_{+0.42}$	70.26	3.54

positive, leading to a noisy training objective. While the probability of same class negatives is higher when batch size is higher than the number of classes, this issue can occur even otherwise, when there exist negative images that are more similar to the anchor when compared to the positive. Khosla et al. [19] use supervision from labels in a Supervised Contrastive (SupCon) framework to convert the same-class false negatives to additional positives, and show an improvement over supervised learning methods.

In order to specifically study the impact of eliminating false-negatives, we first perform experiments by using labels to avoid using the same class samples as negatives. We do not add these eliminated negatives as positives, in order to avoid excessive supervision. In Table-1 we present results of an experiment on the CIFAR-10 dataset, where the same-class negatives in SimCLR are eliminated using a varying fraction of labels. The fraction of labels serves as an upper bound to the amount of noise reduction in the training objective, considering that other sources of noise such as false-positives are still not eliminated. Using 30% labels, we achieve a 3.49% increase in accuracy when compared to the SimCLR baseline (0% labels case). It is also interesting to note that the boost in accuracy is highest for 30% supervision and reduces as the fraction of labels increase. This indicates that the network can possibly overcome the impact of noise more effectively when the amount of noise is lower. Overall, we obtain 4.5% boost in the case where all the labels are used. By jointly training SimCLR with the task of rotation prediction (Ours), we achieve highest gains in the case of 0% labels or the no supervision case, and significantly lower gains as the amount of labels increase. We discuss this in greater detail in Section-5.2.

3.2 Impact of False Positives in BYOL

Since BYOL does not use a contrastive learning objective, it is not directly impacted by noise due to false negatives. However, as shown in Fig.1(b), the augmentations considered may not be similar to each other, leading to false positives. Selvaraju et al. [31] show that unsupervised saliency maps can be

used for the selection of better crops, and also as a supervisory signal in the training objective. This leads to improved performance on scene datasets which contain multiple objects. Inspired by this, we use Grad-CAM [30] based saliency maps from a supervised ImageNet pre-trained network to select crops such that the ratio of mean saliency score of the cropped image and that of the full image is higher than a certain threshold (Details in Sec.2 of the Supplementary). It is to be noted that the only difference with respect to BYOL is in the use of supervised saliency maps for the selection of crops. Alternatively, unsupervised saliency maps could also be used for the same. We demonstrate results of this experiment on the ImageNet-100 dataset in Table-2. We observe that by using saliency-maps for crop selection, the accuracy improves by 3.08% for a fixed training budget. While this experiment shows the impact of reducing the false positives in the BYOL objective, it does not completely eliminate noise in the training objective, since the saliency maps themselves are obtained from a Deep Neural Network, and hence may not be very accurate.

4 Proposed Method

In this section, we examine the advantages of instance-discriminative approaches and handcrafted pretext-task based methods, and further discuss our proposed approach which integrates both methods to overcome their limitations.

The key ingredients for the success of a self-supervised learning algorithm are (i) Well-posedness of the learning task; (ii) Extent of correlation between representations that help accomplish the pretext task, and ideal representations, whose quality is evaluated using downstream tasks.

The success of instance-similarity based approaches in achieving state-of-the-art performance on downstream tasks indeed shows that the representations learnt using such tasks are well correlated with ideal representations. However, these methods require to be trained on a significantly larger number of training epochs when compared to the supervised counterparts. As seen in the previous section, a possible reason for the slow convergence of these methods is the noise in training objective due to the presence of false positives and false negatives. While it is possible to overcome this noise using additional supervision from (unsupervised) pre-trained models, such as the use of saliency maps for crop selection, these methods are not very successful as this supervisory signal is also not perfect in practice. Moreover, this method assumes the availability of a network which is pre-trained on a relevant dataset, which may not always hold true, and hence adds to the training cost. We observe that the boost in performance is not good enough to justify the additional computational overhead. If the same computational budget is invested in the main self-supervised task, it leads to better performance (Details in Sec.2 of the Supplementary).

On the other hand, task-based objectives such as rotation prediction score higher on the well-posedness of the learning task. In this task, a known random rotation transformation is applied to an image, and the task of the network is to predict the angle of rotation. Since the rotation angle is known a priori, there is

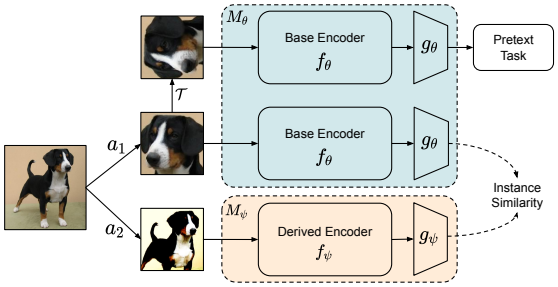


Fig. 2: Schematic diagram illustrating the proposed approach. A pretext task such as rotation prediction is combined with base methods like BYOL and SimCLR. For methods like BYOL and MoCo, the derived network M_ψ is a momentum-averaged version of M_θ, and for methods like SimCLR, M_θ and M_ψ share the same parameters.

very little scope for noise in labels or in the learning objective, leading to faster training convergence.

In this work, we propose to enhance the convergence of instance-similarity based approaches using pretext-task based objectives such as rotation prediction. The proposed approach can be used to enhance many existing instance-discrimination based algorithms (referred to as base algorithm) as shown in Section-5. A schematic diagram of our proposed approach is presented in Fig.2.

We term the main feature extractor to be learned as the base encoder, and denote it as f_θ. Some of the self-supervised learning algorithms use an additional encoder, which is derived from the weights of the base encoder. We call this as a derived encoder and represent it using f_ψ. It is to be noted that the derived encoder may be also be identical to the base encoder, which represents an identity mapping between θ and ψ. As proposed by Chen et al. [4], many of the approaches use a learnable nonlinear transformation between the representations and the final instance-discriminative loss. We denote this projection network and its derived network using g_θ and g_ϕ respectively. We note that the base algorithm may have additional layers between the projection network and the final loss, such as the predictor in BYOL [15] and SimSiam [6], which are not explicitly shown in Fig.2.

An input image x is first subject to two augmentations a_1 and a_2 to generate x^{a_1} and x^{a_2}. We use the augmentation pipeline from the respective base algorithm such as BYOL or SimCLR. These augmented images are passed through the base encoder f_θ and the derived encoder f_ϕ respectively, and the outputs of the projection networks g_θ and g_ϕ are used to compute the training objective of the respective base algorithm. The augmentation x^{a_1} is further transformed using a rotation transformation t which is randomly sampled from the set $\mathcal{T} = \{0°, 90°, 180°, 270°\}$. The rotated image $x^{a_1,t}$ is passed through the base encoder f_θ and projection network g_θ which are shared with the instance-based task. We represent the overall network formed by the composition of f_θ and g_θ by M_θ, and similarly the composition of f_ψ and g_ψ by M_ψ. The representation

$M_\theta(x^{a_1,t})$ is input to a task-specific network h_θ whose output is a 4-dimensional softmax vector over the outputs in the set $\mathcal{T}$. The overall training objective is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda \cdot \frac{1}{2B} \sum_{i=0}^{B-1} \sum_{m=1}^{2} \ell_{CE}(h_\theta(M_\theta(x_i^{a_m,t_k}), t_k)) \tag{1}$$

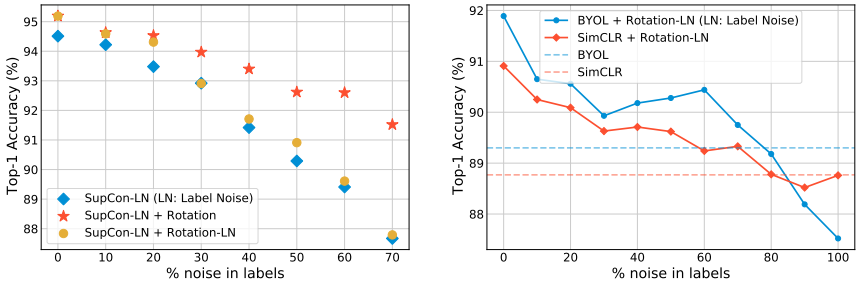
Here t_k is sampled uniformly at random for each image from the set $\mathcal{T}$. $\mathcal{L}_{\text{base}}$ represents the symmetric loss of the base instance-similarity based algorithm used. We describe the base loss for BYOL [15] and SimCLR [4] in Sec.1 of the Supplementary. λ is the weighting factor between rotation task and the instance-similarity objective. While the RotNet algorithm [13] uses all four rotations for every image, we consider only two in the overall symmetric loss. Therefore, when compared to the base algorithm, the computational overhead of the proposed method is limited to one additional forward propagation for every augmentation, which is very low when compared to the other components of training such as data loading and backpropagation. There is no additional overhead in backpropagation since the combined loss (Eq.1) is used for training.

5 Experiments and Analysis

In this section, we first describe our experimental settings (Sec.5.1), following which we present an empirical analysis to highlight the importance of the auxiliary task towards improving the efficiency and effectiveness of the base learning algorithm (Sec.5.2). We further compare the properties of the learned representations using different training methods and show that learning representations that are covariant to rotation also aids in boosting performance (Sec.5.3). We finally compare the results of the proposed method with the state-of-the-art approaches in Sec.5.4.

5.1 Experimental Setup

We run our experiments either on a single 32GB V100 GPU, or across two such GPUs unless specified otherwise. We train our models with ResNet-18 [17] architecture on CIFAR-10 and CIFAR-100 [23] dataset and with ResNet-50 [17] architecture on ImageNet-1k [9] dataset. Our primary evaluations are run for 200 epochs on CIFAR-10 and CIFAR-100, and 100 epochs on ImageNet-100 [9] dataset. We show additional evaluations across varying number of training epochs in Sec.5.4. We describe the training hyperparameters in Sec.4 of the Supplementary. We use the respective base algorithm or the proposed approach to learn the base encoder f_θ, and evaluate its effectiveness by training a linear classifier over this, as is common in prior works [4, 6, 15, 3]. In this step, the weights of the base encoder are frozen. We additionally report results in a semi-supervised (Sec.5.4) and transfer learning setting (Sec.7 of the Supplementary) as well.



(a) SupCon [19] (b) BYOL [15], SimCLR [4]

Fig. 3: The plots demonstrate the impact of label noise in different training objectives on CIFAR-10 dataset. The proposed method (+ Rotation) results in higher performance boost when the amount of label noise in the base method is larger. Addition of label noise to the rotation task reduces the gain in performance.

5.2 Robustness to Noise in the Training objective

As discussed in Section-3, instance-similarity based tasks such as SimCLR [4] and BYOL [15] suffer from noise in the training objective, and eliminating this noise can lead to significant performance gains in a fixed training budget. We additionally report results of the proposed approach integrated with SimCLR and BYOL in Tables-1 and 2 respectively, and obtain gains over the base approach across varying settings of supervision levels. However, as can be seen from the column Gain (%), the gains using the proposed approach reduce with increasing levels of supervision. This is aligned with our hypothesis that the rotation task helps in overcoming the impact of noise in the base instance-similarity task, and therefore, when additional supervision already achieves this objective, gains using the proposed approach are lower.

Label Noise in a Supervised Learning setting: We consider the task of supervised learning using the supervised contrastive (SupCon) learning objective proposed by Khosla et al. [19]. The training objective is similar to that of SimCLR [4] with the exception that same-class negatives are treated as positives. The authors demonstrate that this method outperforms standard supervised training as well. We choose this training objective as this is similar to the instance-similarity based tasks we consider in this paper, while also having significantly lesser noise due to the elimination of false negatives in training. As shown in Fig.3a, even in this setting, the proposed method achieves 0.68% improvement, achieving a new state-of-the-art in supervised learning. In order to highlight the impact of noise in training, we run a controlled set of experiments by adding a fixed amount of label noise in each run. The plot in Fig. 3a shows the trend in accuracy of the SupCon algorithm with increasing label noise. The proposed method achieves a significant boost over the SupCon baseline consistently across different noise levels. Further, as the amount of noise in training

increases, we achieve higher gains using the proposed approach, indicating that the rotation task is indeed helping overcome noise in the training objective.

We also consider a set of experiments where an equal amount of label noise is added to the SupCon training objective and to the rotation prediction task. We note that in majority of the runs (excluding the case of noise above 70%), the accuracy is very similar to the SupCon baseline with the same amount of noise. This indicates that the knowledge of true labels in handcrafted tasks such as rotation prediction is the key factor that contributes to the improvement achieved using the proposed approach.

We perform the experiments of adding label noise to the rotation prediction task when combined with BYOL and SimCLR as well. As shown in Fig.3b we find that the gains with the rotation prediction task drops considerably over $0-20\%$ label noise, indicating that a similar amount of noise ($\sim 20\%$) is present in the BYOL/ SimCLR training objectives as well. Further, addition of rotation prediction task helps marginally ($0.47 - 1.38\%$) even with higher amount of noise ($30-60\%$) in rotation annotations. This indicates that, while the rotation prediction primarily helps by providing a noise-free training objective, it aids the main task in other ways too. We investigate this in the following section.

5.3 Learning rotation-covariant representations

The task of enforcing similarity across various augmentations of a given image yields representations that are invariant to such transformations. In sharp contrast, the representations learned by humans are covariant with respect to factors such as rotation, color and scale, although we are able to still correlate multiple transformations of the same object very well. This hints at the fact that learning covariant representations could help the accuracy of downstream tasks such as object detection and classification.

In Table-3, we compare the rotation sensitivity and contrastive task accuracy of representations at the output of the base encoder f_θ, and the projection network g_θ. We follow the process described by Chen et al. [4] to obtain these results. We freeze the network till the respective layer (f_θ or g_θ) and train a rotation task classifier over this using a 2-layer MLP head. We measure the rotation task accuracy, which serves as an indication of the amount of rotation sensitivity in the base network. We further compute the contrastive task accuracy on the representations learned, by checking whether the two augmentations of a given image are more similar to each other when compared to augmentations of other images in the same batch.

Interestingly, a fully supervised network is more sensitive to rotation (80.54%) when compared to the representations learned using BYOL (73.4%). Chen et al. [4] also show that rotation augmentation hurts performance of SimCLR. These observations indicate that invariance to rotation hurts performance, and reducing this lead to better representations. While RotNet has higher accuracy on the rotation task, it does significantly worse on the instance discrimination task, leading to sub-optimal performance compared to BYOL. In the proposed

Table 3: **Task Performance (%):** Evaluation of representations learned using various algorithms on the task of rotation prediction and instance-discrimination.

Method	Linear	Rotation Acc		Contrastive Acc	
		f(.)	g(f(.))	f(.)	g(f(.))
Supervised	94.03	80.54	-	46.36	-
BYOL	89.30	73.40	58.32	78.53	78.82
Rotation	84.00	93.69	93.46	31.61	1.52
BYOL+Rotation	91.89	93.73	93.54	72.85	67.81

Table 4: BYOL + Rotation with **varying noise in the rotation labels**. Rotation prediction accuracy correlates with linear evaluation accuracy.

Rotation Noise	Linear	Rotation Acc		Contrastive Acc	
		f(.)	g(f(.)	f(.)	g(f(.)
30%	89.93	91.88	91.78	73.42	64.25
50%	90.28	89.95	85.82	78.18	77.39
70%	89.75	80.49	67.26	78.55	77.31
80%	89.18	77.43	63.53	77.26	76.92

method, we achieve better rotation task accuracy with a small drop in the contrastive task accuracy when compared to BYOL. This also results in an overall higher performance after Linear evaluation.

We also investigate rotation invariance for the experiments in Sec.5.2 with BYOL as the base method, where noise is added to the rotation task. As shown in Table-4, we find that as the amount of noise increases in the rotation task, the amount of rotation invariance increases, leading to a drop in accuracy. Even with 50% noise in the rotation task, we achieve 16.55% boost in rotation performance, leading to 0.98% improvement in the accuracy after linear evaluation. Since the BYOL learning task possibly contains lesser noise compared to this, the gain in performance can be justified by the fact that rotation-covariant representations lead to improved performance on natural image datasets.

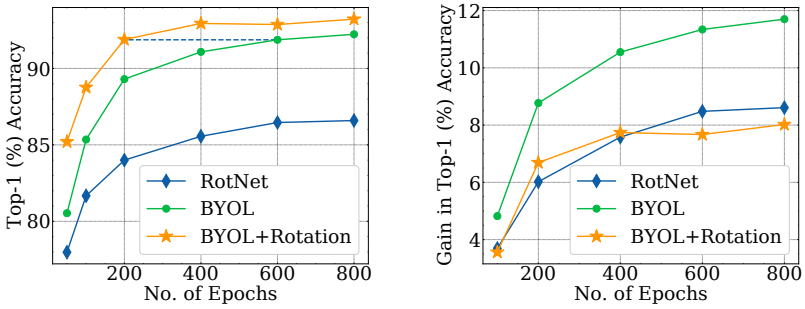
5.4 Comparison with the state-of-the-art

We compare the performance of the proposed method with the respective baselines in the setting of linear evaluation on CIFAR-10, CIFAR-100 (Table-5), ImageNet-100 and ImageNet-1k (Table-6) datasets. We perform extensive hyper-parameter search to obtain reliable results on the baseline methods for CIFAR-10 and CIFAR-100, since most existing works report the optimal settings for ImageNet-1k training alone. As shown in Table-5, although the performance of Rotation prediction [13] itself is significantly worse that other methods, we obtain gains of 2.14%, 2.59%, 3.6% and 2.14% on CIFAR-10 and 2.44%, 6.4%, 7.1% and 3.11% on CIFAR-100 by using the proposed method with SimCLR [4], BYOL [15], SwAV [3] and SimSiam [6] respectively.

We present results on CIFAR-10 dataset with varying number of training epochs in Fig.4a using BYOL as the base approach. Across all settings, we obtain improved results over the BYOL baseline. The proposed method achieves the same accuracy as the baseline in one-third the training time (shown using blue dotted line) as shown in Fig.4a. We show the difference in accuracy with respect to accuracy obtained with 50 epochs of training in Fig.4b, to clearly visualize the convergence rate of different methods. It can be seen that the proposed method has a similar convergence trend as the Rotation task, while

Table 5: **CIFAR-10, CIFAR-100:** Accuracy (%) of the proposed method compared to baselines under two evaluation settings - K-Nearest Neighbor (KNN) classification with K=200 and Linear classifier training. The proposed method achieves significant performance gains across all settings.

Method	CIFAR-10 (200 epochs) KNN	CIFAR-10 (200 epochs) Linear	CIFAR-100 (200 epochs) KNN	CIFAR-100 (200 epochs) Linear
Rotation Pred. [13]	78.01	84.00	36.25	50.87
SimCLR [4]	86.37	88.77	55.10	62.96
SimCLR + Ours	88.69	90.91	57.09	65.40
BYOL [15]	86.56	89.30	54.37	60.67
BYOL + Ours	89.80	91.89	58.41	67.03
SwAV [3]	80.65	83.60	40.35	51.50
SwAV + Ours	85.26	87.20	50.09	58.60
SimSiam [6]	87.05	89.77	56.90	64.27
SimSiam + Ours	**90.35**	**91.91**	**58.92**	**67.38**



(a) Top-1 Accuracy (b) Gain in Top-1 Accuracy

Fig. 4: (a) Accuracy (%) after Linear layer training for BYOL [15], RotNet [13] and the proposed method (BYOL+Rotation) on CIFAR-10. The proposed method achieves the same accuracy as the baseline in one-third the training time (shown using blue dotted line). (b) Gain in Top-1 Accuracy (%), or the difference between accuracy of the current epoch and epoch-50. Plot (a) shows the improvement in effectiveness of the proposed approach and plot (b) shows the improvement in efficiency or convergence rate.

outperforming BYOL in terms of Top-1 Accuracy, highlighting that integrating these methods indeed combines the benefits of both methods.

We present results on ImageNet-100 dataset in Table-6. To limit the computational cost on our ImageNet-100 and ImageNet-1k runs, we either use the tuned hyperparameters from the official repository, or follow the settings from other popular repositories that report competent results. Due to the unavailability of tuned hyperparameters on this dataset for SimSiam, we skip reporting results of this method on ImageNet-100. We achieve gains of 2.58%, 1.22% and 2.2% on BYOL [15], SimCLR [4] and SwAV [3] respectively in Top-1 accuracy.

Table 6: **ImageNet-100 and ImageNet-1k:** Performance (%) of the proposed method when compared to baselines under three evaluation settings - Linear classifier training and Semi-Supervised Learning with 1% and 10% labels. The proposed method achieves significant performance gains.

Method	Linear Acc	Semi-Supervised 1% labels		Semi-Supervised 10% labels	
	Top-1	Top-1	Top-5	Top-1	Top-5
ImageNet-100 (100 epochs, ResNet-18)					
Rotation Prediction [13]	53.86	34.72	65.70	51.18	81.38
BYOL [15]	71.02	46.60	75.50	68.00	89.80
BYOL + Ours	73.60	56.40	83.50	72.30	91.40
SimCLR [4]	72.02	57.28	83.69	71.44	91.72
SimCLR + Ours	73.24	**57.80**	**83.84**	**72.52**	**92.10**
SwAV [3]	72.20	49.38	78.41	67.56	90.78
SwAV + Ours	**74.40**	52.02	80.01	69.68	91.43
ImageNet-1k (30 epochs, ResNet-50)					
SwAV [3]	54.90	32.20	58.20	51.82	77.60
SwAV + Ours	**57.30**	**32.80**	**59.12**	**53.80**	**78.54**

We obtain the best results by integrating the proposed method with SwAV, and hence report ImageNet-1k results on the same method, in order to demonstrate the scalability of the proposed method to a large-scale dataset. We present the result of 30-epochs of training on ImageNet-1k in Table-6. Using the proposed approach, we obtain a boost of 2.4% in Top-1 accuracy over the SwAV baseline. We present additional results on longer training epochs in Sec.8 of the Supplementary.

Furthermore, we present results on ImageNet-100 dataset with varying number of training epochs in Fig.5. Using the proposed method, we achieve gains across all settings with respect to the number of training epochs. We obtain improved results over the base methods in semi-supervised learning (Table-6) and transfer learning settings as well. We discuss the transfer learning results in Sec.7 of the Supplementary.

5.5 Integration with other tasks

In this work, we empirically show that combining instance-discriminative tasks with well-posed handcrafted pretext tasks such as Rotation prediction [13] can indeed lead to more effective and efficient learning of visual representations. While we choose the Rotation prediction task due to its simplicity in implementation, and applicability to low resolution images (such as CIFAR-10), it is indeed possible to achieve gains by using other well-posed tasks as well. In Table-7, we report results on the ImageNet-100 [33] dataset by combining the base BYOL [15] algorithm individually with Rotation prediction [13], Jigsaw puzzle solving

Table 7: **Combining BYOL with handcrafted pretext tasks:** Accuracy in (%) after linear evaluation, of various algorithms on ImageNet-100 dataset.

	Top-1 (%)	**Top-5** (%)
RotNet [13] (R)	53.86	81.26
Jigsaw [26] (J)	42.01	72.10
BYOL [15]	71.02	91.78
BYOL + R	73.60	**92.98**
BYOL + J	73.60	92.72
BYOL + J + R	**74.72**	92.94

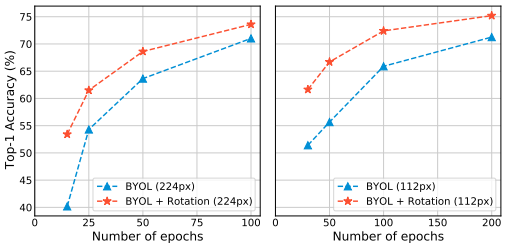


Fig. 5: Accuracy (%) after Linear layer training for BYOL and the proposed method (BYOL+Rotation) for ImageNet-100. The proposed method achieves significant gains over the baseline in all settings.

[26] and both. Although the Jigsaw puzzle solving task is sub-optimal when compared to the Rotation prediction task, we achieve similar gains in performance when these tasks are combined with BYOL. We obtain the best gains (3.7%) when we combine both tasks with BYOL. This shows that the analysis on well-defined tasks being able to aid the learning of instance-discriminative tasks that are noisy is indeed generic, and not specific to the Rotation prediction task alone.

6 Conclusions

In this work, we investigate reasons for the slow convergence of recent instance-similarity based methods, and propose to improve the same by jointly training them with well-posed tasks such as rotation prediction. While instance-discriminative approaches learn better representations, handcrafted tasks have the advantage of faster convergence as the training objective is well defined and there is typically no (or very less) noise in the generated pseudo-labels. The complementary nature of the two kinds of tasks makes it suitable to achieve the gains associated with both by combining them. Using the proposed approach, we show significant gains in performance under a fixed training budget, along with improvements in training efficiency. We show similar gains in performance by combining the base algorithms with the task of Jigsaw puzzle solving as well. We hope that our work will revive research interest in designing specialized tasks, so that they can be help boost the effectiveness and efficiency of state-of-the-art methods.

7 Acknowledgments

This work was supported by the Qualcomm Innovation Fellowship. We are thankful for the support.

Bibliography

[1] Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. In: International Conference on Learning Representations (ICLR) (2020) 4

[2] 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) (2018) 4

[3] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 1, 2, 3, 4, 8, 11, 12, 13

[4] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML) (2020) 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13

[5] Chen, T., Zhai, X., Ritter, M., Lucic, M., Houlsby, N.: Self-supervised gans via auxiliary rotation loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3

[6] Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 1, 2, 3, 4, 7, 8, 11, 12

[7] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems (NeurIPS) (2013) 4

[8] Dangovski, R., Jing, L., Loh, C., Han, S., Srivastava, A., Cheung, B., Agrawal, P., Soljacic, M.: Equivariant self-supervised learning: Encouraging equivariance in representations. In: International Conference on Learning Representations (ICLR) (2022) 4

[9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 1, 2, 8

[10] Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: IEEE International Conference on Computer Vision (ICCV) (2015) 3

[11] Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: IEEE International Conference on Computer Vision (ICCV) (2017) 3

[12] Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: International Conference on Learning Representations (ICLR) (2017) 1

[13] Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (ICLR) (2018) 1, 2, 3, 4, 8, 11, 12, 13, 14

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS) (2014) 1, 3
[15] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 1, 2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 14
[16] 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 (CVPR) (2020) 2, 3, 4
[17] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1, 3, 8
[18] Hénaff, O.J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S.M.A., Van Den Oord, A.: Data-efficient image recognition with contrastive predictive coding. In: Proceedings of the 37th International Conference on Machine Learning (ICML) (2020) 3
[19] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 5, 9
[20] Kinakh, V., Voloshynovskiy, S., Taran, O.: ScatsimCLR: self-supervised contrastive learning with pretext task regularization for small-scale datasets. In: 2nd Visual Inductive Priors for Data-Efficient Deep Learning Workshop (2021) 4
[21] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 1
[22] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2012) 1
[23] Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009) 2, 8
[24] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015) 1
[25] 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 (CVPR) (2020) 3
[26] Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision (ECCV) (2016) 1, 3, 4, 14
[27] Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning to count. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017) 3
[28] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 3

[29] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2016) 1
[30] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2017) 6
[31] Selvaraju, R.R., Desai, K., Johnson, J., Naik, N.: Casting your model: Learning to localize improves self-supervised representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 5
[32] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2015) 1
[33] Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision (ECCV) (2020) 2, 13
[34] Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision (ECCV) (2016) 1, 4
[35] Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 1