




# Efficient Unsupervised Visual Representation Learning with Explicit Cluster Balancing

Ioannis Maniadis Metaxas<sup>1</sup><sup>\*</sup>, Georgios Tzimiropoulos<sup>1</sup>, and Ioannis Patras<sup>1</sup>

Queen Mary University of London

`{i.maniadis, g.tzimiropoulos, i.pstras}@qmul.ac.uk`

**Abstract.** Self-supervised learning has recently emerged as the preeminent pretraining paradigm across and between modalities, with remarkable results. In the image domain specifically, group (or cluster) discrimination has been one of the most successful methods. However, such frameworks need to guard against heavily imbalanced cluster assignments to prevent collapse to trivial solutions. Existing works typically solve this by reweighing cluster assignments to promote balance, or with offline operations (e.g. regular re-clustering) that prevent collapse. However, the former typically requires large batch sizes, which leads to increased resource requirements, and the latter introduces scalability issues with regard to large datasets. In this work, we propose ExCB, a framework that tackles this problem with a novel cluster balancing method. ExCB estimates the relative size of the clusters across batches and balances them by adjusting cluster assignments, proportionately to their relative size and in an online manner. Thereby, it overcomes previous methods’ dependence on large batch sizes and is fully online, and therefore scalable to any dataset. We conduct extensive experiments to evaluate our approach and demonstrate that ExCB: **a)** achieves state-of-the-art results with significantly reduced resource requirements compared to previous works, **b)** is fully online, and therefore scalable to large datasets, and **c)** is stable and effective even with very small batch sizes. Code and models will be made available here.

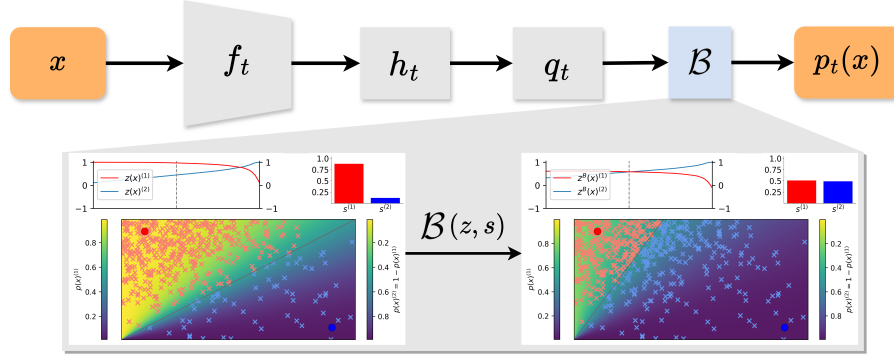
**Keywords:** Self-supervised learning · Representation learning

## 1 Introduction

Unsupervised, or self-supervised, representation learning has recently emerged as a dominant training paradigm to leverage vast amounts of unlabeled data in order to train powerful models. That is typically done by utilizing natural supervisory signals inherent in each domain [5, 16, 30, 35, 37] or between domains [1, 4, 41] in order to craft training objectives, thereby overcoming the need for extensive training data annotations. In the visual domain, various types of self-supervised objectives have been proposed, such as transformation prediction [23], reconstruction [25], instance discrimination [11, 12] and group discrimination [8].

---

<sup>\*</sup> Corresponding author.



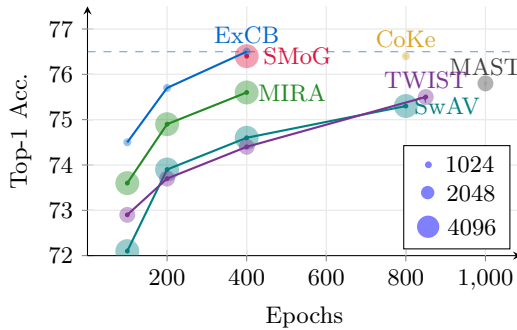
**Fig. 1:** Illustration of ExCB’s balancing operator  $\mathcal{B}$  for two clusters  $c_1$  (red) and  $c_2$  (blue).  $\mathcal{B}(z; s)$  adjusts sample-cluster cosine similarities  $z$  according the relative cluster sizes, as measured in  $s$ . For smaller clusters the similarities are increased ( $z^B > z$ ), whereas for larger clusters the similarities are decreased ( $z^B < z$ ). The impact, as seen in the figure, is that the boundary between clusters shifts, undersized (oversized) clusters are assigned more (fewer) samples, and clusters become more balanced.

Group (or cluster) discrimination methods train models with pseudo-labels (or cluster assignments), so that representations from similar images (that are assigned to the same cluster) are pulled together, and different images (assigned to different clusters) are pushed away. This approach has been very effective, as it alleviates the false-negatives issue that plagues instance discrimination-based methods [44]. A major distinction within clustering-based methods is whether the pseudo-labels are produced online or offline. Offline methods [2, 8] periodically operate on the entire dataset to cluster the data and produce pseudo-labels and/or centroids. These methods are straightforward and stable, but cannot easily scale to large datasets, as they require storing and operating on features from the entire dataset. Online methods [9, 10, 31, 40, 43] produce pseudo-labels for each batch during training and do not require access to the entire dataset at any one time. They can, therefore, scale to any number of samples. However, online methods need to apply balancing constraints to the cluster assignments to avoid collapse (i.e. all samples being assigned to the same cluster). Typically, these constraints are applied using in-batch statistics: cluster assignments within each batch are adjusted to approximate balanced clusters, often by solving optimization problems [9, 31]. This approach requires large batches to work effectively, as the assumption of in-batch balance stands only if the number of samples in the batch is comparable to the number of clusters. In turn, this leads to increased resource requirements and, possibly, training instability.

In this work we propose ExCB, a self-supervised framework that utilizes a novel online cluster balancing method and achieves state-of-the-art performance with remarkable efficiency in terms of training time and batch size requirements. ExCB relies on two simple components: i) We measure the relative size of clusters

over multiple steps using hard assignments. Thereby, ExCB approximates the cluster distribution across the dataset accurately and reliably, without depending on volatile in-batch statistics and the assignments’ confidence, and without requiring a large batch size. **ii)** We adjust cluster assignments according to each cluster’s size, as measured in **(i)**. Specifically, we increase (decrease) the sample-cluster similarities of small (large) clusters to assign them more (fewer) samples, thereby ensuring that no cluster deviates significantly in terms of their size in either direction. Together, the two components enforce an explicit soft balancing constraint on clusters that: **a)** is fully online without requiring a large batch size, **b)** has negligible computational cost, as it does not require solving optimization problems, and **c)** is, conceptually, very intuitive and straightforward.

We conduct extensive experiments and demonstrate that ExCB achieves state-of-the-art results with both convolutional and transformer backbones. Importantly, that is achieved with a much smaller batch size and/or fewer training epochs compared to previous works. Finally, we show that, out of the box, ExCB is stable and achieves strong performance even with very small batch sizes. Jointly, our findings show that ExCB achieves state-of-the-art performance in self-supervised representation learning while also being remarkably efficient, which has the additional benefit of significantly lowering the resource-wise barrier of entry for self-supervised pretraining.



**Fig. 2:** Linear classification accuracy on ImageNet with ResNet50 for different self-supervised methods. Circles indicate pretraining batch size. ExCB achieves state-of-the-art results with the most efficient combination of few epochs and small batch size.

## 2 Related Works

### 2.1 Self-supervised learning for visual data

The objective of self-supervised learning is to leverage *pretext* tasks in order to learn robust representations from non-annotated data, such that they can be used to effectively solve other, *downstream* tasks with minimal supervision. Pretext tasks in the visual domain typically leverage strong data augmentations to formulate training objectives, such as identifying transformations [23], patch permutation [38] and instance discrimination [35]. Recent self-supervised representation learning methods can be roughly grouped based on the objectives they use: a) instance-discrimination [11, 12, 14, 18–21, 36, 49], where samples’ views

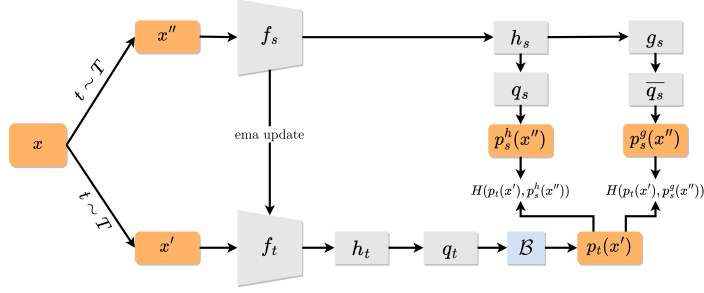
must be discernible from other samples, b) non-contrastive methods, where transformation invariance is imposed via self-distillation [13, 24, 29] or by enforcing specific properties to batch feature statistics [6, 7, 33, 50, 52], c) reconstruction-based methods [3, 25, 48], where the model predicts missing image patches, and d) clustering-based methods, where supervision is provided by pseudo-labels, typically produced via clustering [8–10, 22, 31, 32, 39, 40, 43]. We emphasize that this taxonomy is indicative, and is neither exact nor strict, as the various methods have individual features that could place them in more than one or in distinct groups. Furthermore, we note that, whereas self-supervised representation learning is primarily focused on classification-based downstream tasks, a distinct line of research is focused on self-supervised learning frameworks that prioritize dense prediction tasks such as detection and segmentation [7, 27, 28, 42, 45–47].

## 2.2 Clustering-based self-supervised learning

Self-supervised frameworks based on group discrimination generate sample pseudo-labels (typically via clustering) and use them as supervision to train the model. A key distinction among these methods is whether they are online or offline.

Offline methods [8, 32] typically produce pseudo-labels by regularly clustering samples across the dataset. Distinctly, [2] produces pseudo-labels by regularly extracting cluster assignments for the entire dataset and, after balancing them with the Sinkhorn-Knopp algorithm [15], uses them to further train the model. This approach guarantees training stability but introduces additional computational cost. More importantly, however, offline methods are very hard to scale to large datasets, as they require the storing and processing (i.e. clustering) of representations for the entire dataset, which in turn increases resource requirements and may be intractable for web-scale data.

Online methods overcome this limitation by generating pseudo-labels (or cluster assignments) in-batch during training. These methods seamlessly scale to large datasets, but suffer the risk of samples being assigned to only a few (or even one) clusters, in which case they collapse to trivial solutions and fail to learn meaningful representations. To prevent collapse, online clustering-based frameworks promote balanced clusters by applying appropriate balancing constraints on cluster assignments with methods such as centering [10], optimization [9, 31] or auxiliary training objectives [43]. Crucially, [9, 31, 43] rely on in-batch statistics; that is, for each batch, only its own assignments are considered in applying the respective method’s constraint. As a result, these methods are highly reliant on large batch sizes, otherwise they face poor performance and training instability. Furthermore, [10, 31, 43] rely on soft cluster assignment statistics (e.g. the aggregate assignment confidence for each cluster) to define their constraints. This undermines the effectiveness of the constraints, as these metrics are only proxies for the actual size of the clusters. We note that there are also hybrid methods such as [39, 40]. These methods regulate cluster sizes in an online manner, but, for optimal performance, they both require storing and utilizing representations [39] or cluster assignments [40] for the entire dataset for each epoch, which again raises scalability issues with larger datasets.



**Fig. 3:** Overview of ExCB. The student is trained so that  $\mathbf{p}_s(x'')$  matches  $\mathbf{p}_t(x')$ , where  $x'$  and  $x''$  are transformed views of  $x$ . The balancing module  $\mathcal{B}$  adjusts cluster assignments to promote uniform distribution between the clusters *across the dataset*.

In contrast to previous works, ExCB combines the best of both worlds: it is fully online, and therefore scalable to any dataset size, *and* it does not require large batch sizes. The latter, which is achieved by measuring cluster sizes over multiple batches explicitly with hard assignments, results in more accurate and reliable approximations of the cluster distribution across the dataset and improved training stability, even with a very small batch size.

### 3 Method

We present an overview of ExCB’s architecture and objective in Sec. 3.1. We then present our proposed, novel module for balancing cluster assignments in Sec. 3.2.

#### 3.1 Overview

ExCB utilizes a teacher-student framework, where the student is trained to match the cluster assignments of the teacher. The student’s architecture consists of a backbone model  $f_s$ , a projection MLP head  $h_s$ , a prediction MLP head  $g_s$ , and a layer  $q_s$ , which measures the cosine similarity between an input vector  $\mathbf{v}$  and a set of learned cluster centroids  $\mathbf{C}_s = [\mathbf{c}_1, \dots, \mathbf{c}_K] \in \mathbb{R}^{K \times D}$ , where  $K$  is the number of clusters and  $D$  the output dimension of  $h_s$  and  $g_s$ :

$$q_s(\mathbf{v}, \mathbf{c}) = \frac{\mathbf{v} \cdot \mathbf{c}}{\|\mathbf{v}\| \|\mathbf{c}\|} \quad (1)$$

For each sample  $x$ , we then define the cosine similarity of its projection/prediction features with each cluster centroid as:

$$\mathbf{z}_s^h(x) = (q_s \circ h_s \circ f_s)(x) \quad , \quad \mathbf{z}_s^g(x) = (\overline{q_s} \circ g_s \circ h_s \circ f_s)(x), \quad (2)$$

where  $\mathbf{z}_s^h(x), \mathbf{z}_s^g(x) \in \mathbb{R}^K$  and  $\overline{q_s}$  indicates that we apply stop-gradients to the centroid layer (i.e. the predictor considers the centroids as fixed). The teacher has the same architecture but without a predictor, so that:

$$\mathbf{z}_t(x) = (q_t \circ h_t \circ f_t)(x). \quad (3)$$

We further define the adjusted sample-cluster cosine similarity  $\mathbf{z}_t^B$ :

$$\mathbf{z}_t^B(x) = \mathcal{B}(\mathbf{z}_t(x), \mathbf{s}), \quad (4)$$

where  $\mathcal{B}$  is ExCB’s proposed balancing operator and  $\mathbf{s}$  is a vector measuring relative cluster size. This operator, presented in Sec. 3.2, adjusts sample-cluster similarities to promote balanced clusters in terms of size and prevent collapse.

Finally, the outputs of the student and teacher are the probability assignment vectors  $\mathbf{p}_s^h$ ,  $\mathbf{p}_s^g$  and  $\mathbf{p}_t \in \mathbb{R}^K$ , mapping sample  $x$  to each cluster  $k \in K$ :

$$\mathbf{p}_s^h(x)^{(k)} = \frac{\exp(z_s^h(x)^{(k)}/\tau_s)}{\sum_{i=1}^K \exp(z_s^h(x)^{(i)}/\tau_s)} \quad , \quad \mathbf{p}_s^g(x)^{(k)} = \frac{\exp(z_s^g(x)^{(k)}/\tau_s)}{\sum_{i=1}^K \exp(z_s^g(x)^{(i)}/\tau_s)}, \quad (5)$$

$$\mathbf{p}_t(x)^{(k)} = \frac{\exp((\mathbf{z}_t^B(x)^{(k)})/\tau_t)}{\sum_{i=1}^K \exp((\mathbf{z}_t^B(x)^{(i)})/\tau_t)}, \quad (6)$$

with  $\tau_s$  and  $\tau_t$  being temperature hyperparameters.

Following previous works [10, 12, 13, 31], the teacher’s weights  $\theta_t$  are updated in every training step following an exponential moving average (EMA) of the trained student’s weights  $\theta_s$ , with update rule  $\theta_t = m\theta_t + (1 - m)\theta_s$ . The training objective of ExCB is then formally defined as maximizing the agreement between the student’s and teacher’s outputs across views, by minimizing the loss  $L$ :

$$L = \frac{1}{2} \sum_{x' \in G} \sum_{\substack{x'' \in G \cup L \\ x'' \neq x'}} H(\mathbf{p}_t(x'), \mathbf{p}_s^h(x'')) + \frac{1}{2} \sum_{x' \in G} \sum_{x'' \in G \cup L} H(\mathbf{p}_t(x'), \mathbf{p}_s^g(x'')), \quad (7)$$

where  $H(a, b) = -a^T \log b$  and  $G, L$  represent global and local views produced by random transformations  $t \sim T$  applied to the original image  $x$  [9].

### 3.2 Online Cluster Balancing

The balancing operator  $\mathcal{B}$  is applied to the teacher, as seen in Eq. (4), in order to adjust sample-cluster similarities and promote clusters of equal size. Balancing operators are critical for online clustering-based self-supervised frameworks, as without it they are prone to collapsing to trivial solutions (i.e. the teacher assigning all samples to the same cluster) [9, 10, 31]. In ExCB,  $\mathcal{B}$  consists of two components: a) keeping track of cluster assignments over multiple batches to estimate the clusters’ relative size, and b) adjusting sample-cluster similarity  $z$  accordingly, in order to regulate cluster assignments, and thereby promote balanced clusters and avoid collapse.

*Measuring relative cluster sizes.* We define the relative cluster size vector  $\mathbf{s} \in \mathbb{R}^K$ . For each batch of  $N_B$  samples  $X$  we obtain the teacher’s cluster assignments  $\mathbf{P}_t(X) = [\mathbf{p}_t(x_1), \dots, \mathbf{p}_t(x_{N_B})] \in \mathbb{R}^{N_B \times K}$ , and calculate the in-batch relative cluster size vector  $\mathbf{s}_B \in \mathbb{R}^K$  as the proportion of samples assigned to each cluster:

$$\mathbf{s}_B^{(k)} = \frac{1}{N_B} \sum_{n=1}^{N_B} \mathbf{1}_{\arg\max_{k \in K} (\mathbf{p}_t(x_n))=k}. \quad (8)$$

The vector  $\mathbf{s}$  is then updated for each batch as:

$$\mathbf{s} = m_s \mathbf{s} + \mathbf{s}_B (1 - m_s), \quad (9)$$

where  $m_s$  is a momentum hyperparameter. Essentially,  $\mathbf{s} \in [0, 1]$  measures the proportion of samples assigned to each cluster over multiple batches with an exponential moving average whose window length is determined by  $m_s$ . This approach yields an accurate estimate of cluster sizes across the dataset, without requiring a large batch size. If samples are distributed among clusters with absolute uniformity, then  $\mathbf{s}^{(k)} \rightarrow \frac{1}{K} \forall k \in K$ , whereas  $\mathbf{s}^{(k)} < \frac{1}{K}$  for undersized clusters and  $\mathbf{s}^{(k)} > \frac{1}{K}$  for oversized clusters.

Our approach improves on previous works in two key respects. Firstly, we measure cluster sizes over multiple batches, as opposed to using in-batch statistics. Secondly, we explicitly measure cluster sizes with hard assignments, rather than implicitly, through proxy metrics (e.g. aggregate sample-cluster similarity [10], assignment confidence [43]). This way, ExCB measures cluster sizes **a)** more accurately, as measurements across batches are closer to the dataset-wide cluster distribution, **b)** with lower resource requirements, as a large batch size is not necessary for an accurate estimate, and **c)** more reliably, as using hard cluster assignments means the measurement is independent of factors such as the assignments’ confidence, which fluctuates during training.

*Adjusting sample-cluster similarities.* The cluster assignment tracking vector  $\mathbf{s}$  provides an estimate of relative cluster sizes at any given step. We then use it to adjust cluster assignments so that the teacher will assign more (fewer) samples to undersized (oversized) clusters. To that end, we define the operator  $\mathcal{B}$ :

$$z^B = \mathcal{B}(z; s) = \begin{cases} 1 - [1 - z] s K & , \text{ if } s < \frac{1}{K} \\ [1 + z] \frac{1}{s K} - 1 & , \text{ if } s > \frac{1}{K} \\ z & , \text{ otherwise} \end{cases} \quad (10)$$

where  $z = z_t(x)^{(k)}$  and  $s = \mathbf{s}^{(k)}$  for a given cluster  $k \in K$ .

For any cluster  $k$ ,  $\mathcal{B}$  increases sample-cluster similarity if  $k$  is undersized ( $z^B > z$  for  $s < \frac{1}{K}$ ), and decreases it if  $k$  is oversized ( $z^B < z$  for  $s > \frac{1}{K}$ ). In this way, undersized (oversized) clusters are assigned more (fewer) samples, in an effort to approximate evenly sized clusters ( $\mathbf{s}^{(k)} \rightarrow 1, \forall k \in K$ ). We expand on the proposed balancing operator and the intuition behind it in Supplementary Sec. 2.

*Summary.* Combined, the two components formulate a soft cluster balancing constraint that: a) prevents imbalanced cluster assignments that may lead to collapse and b) does not require a large batch size to be effective, as relative cluster sizes are measured over multiple batches. These advantages are validated through extensive experiments in Sec. 4, where we demonstrate that ExCB achieves state-of-the-art performance in representation learning benchmarks and is stable even when training with very small batch sizes.

## 4 Experiments

### 4.1 Implementation details

*Architecture & Hyperparameters.* Following [10],  $h_s$  and  $h_t$  are 2-layer MLPs with hidden dimension 2,048 and output dimension 256.  $g_s$  is a 1-layer MLP with the same hidden/output dimensions. The temperature parameters  $\tau_t$  and  $\tau_s$  are set to 0.04 and 0.1 respectively. The teacher’s update momentum  $m$  is 0.996 and follows a cosine schedule to 1, and the momentum parameter  $m_s$  of Eq. (9) is set to 0.999. When training without multi-crop, we extract two 224x224 views with scale range [0.14, 1.]. For multi-crop training, we extract two global 224x224 and six local 96x96 views with scale ranges [0.2, 1.] and [0.05, 0.2] respectively. Unless stated otherwise, we set the number of clusters  $K$  to 65,536, the batch size to 1,024 and use the same image augmentations as [24]. In our experiments, we primarily use a ResNet50 backbone, which is most frequently used in this task. We further conduct experiments with a ViT-S/16 [17] backbone, to examine how ExCB performs with transformer-based architectures. When pretraining with a ResNet50 backbone, we use the SGD optimizer with momentum 0.9 and weight decay  $10^{-4}$ . The learning rate is set to  $0.15 \times \text{batch size} / 256$ , and is linearly scaled for 10 epochs followed by a cosine decay schedule. When pretraining with ViT, given limited resources, we **do not tune training hyperparameters** and instead follow the exact configuration used in [10]. Experiments are conducted on 4 A100 GPUs.

*Learning cluster centroids.* It is well established that local views do not produce reliable labels, which is why target assignments are only produced by global views [9]. We extend this approach and only update cluster centroids through global views. This is equivalent to stop-gradient being applied to the centroid layer  $q_s$  when processing local crops. We find that this leads to more stable training and, ultimately, better downstream performance. Furthermore, for ResNet50 pretraining, we use a distinct weight decay for the centroid layer  $q_s$ , following a cosine decay schedule from  $10^{-3}$  to  $10^{-4}$ . The positive impact of these choices is examined in Tab. 8, as well as our use of a predictor, similar to [40].

### 4.2 Results

In this section, we present results across standard evaluation benchmarks for self-supervised representation learning. Unless stated otherwise ExCB is pretrained



**Table 1: Linear & k-NN classification on ImageNet.** We report linear and k-NN classification accuracy on ImageNet, along with each method’s pretraining batch size and epochs. \*TWIST follows standard pretraining with filtered self-labelled training.

Method	Batch Size	Epochs	Linear	k-NN
Supervised	-	-	75.6	-
SimSiam [13]	256	800	71.3	-
SimCLR [11]	4096	800	71.7	-
BYOL [24]	4096	1000	74.4	64.8
MoCo-v3 [14]	4096	1000	74.6	-
DeepCluster v2 [9]	4096	800	75.2	-
Barlow Twins [50]	2048	1000	73.2	66.0
VICReg [6]	2048	1000	73.2	-
SwAV [9]	4096	800	75.3	65.7
DINO [10]	4096	800	75.3	67.5
NNCLR [18]	4096	1000	75.4	-
TWIST* [43]	2048	800+50	75.5	-
MIRA [31]	4096	800	75.7	68.8
MAST [29]	2048	1000	75.8	-
CoKe [40]	1024	800	<u>76.4</u>	-
SMoG [39]	4096	400	<u>76.4</u>	-
<b>ExCB</b>	<b>1024</b>	<b>400</b>	<b>76.5</b>	<b>71.0</b>

**Table 2: Linear classification with ViT.** We report linear classification accuracy on ImageNet for various epochs.

Method	Batch Size	Epochs		
		100	300	800
MoCo-v3 [14]	4096	-	72.5	-
DINO [10]	1024	73.8	75.9	77.0
TWIST [43]	1024	-	<u>76.3</u>	-
<b>ExCB</b>	<b>1024</b>	<b>73.9</b>	<b>76.4</b>	<b>77.1</b>

on ImageNet’s train set for 400 epochs with multi-crop and batch size 1,024. In all cases, the best results are presented in **bold** and the second best are underlined. Results for other methods are taken from [29, 31, 39, 40, 43].

*Linear & k-NN Classifier.* We train a linear classifier on top of a frozen backbone on ImageNet’s train set and present results on the validation set. For ResNet we follow the standard protocol [12] and train the classifier for 100 epochs with the SGD optimizer, batch size 256 and learning rate 0.3 with cosine decay. We also present results for a k-NN classifier, where, following [10], we set the number of neighbours to 20. For the ViT backbone, we again follow [10] and train for 100 epochs, with batch size 1,024, the SGD optimizer, and sweep the learning rate.

The main results for this setting are presented in Tabs. 1 and 2 for ResNet50 and ViT backbones respectively. Additional, more extensive, comparisons, including results without multi-crop and for various pretraining epochs, are presented in Tab. 3 for ResNet50. Across settings, ExCB consistently outperforms previous

**Table 3: Linear classification on ImageNet for various settings.** We report linear classification accuracy on ImageNet for various epochs and with/without multi-crop.

Method	Batch Size	Epochs		
		100	200	400
Without Multi-Crop				
BYOL	4096	66.5	70.6	73.2
SwAV	4096	66.5	69.1	70.7
VicReg	2048	68.6	70.2	72.3
MIRA	4096	69.4	72.3	73.3
MAST	2048	-	70.9	73.5
TWIST	2048	70.4	70.9	71.8
SMoG	2048	67.2	-	73.6
ExCB	1024	70.7	72.7	73.9
With Multi-Crop				
SwAV	4096	72.1	73.9	74.6
TWIST	2048	72.9	73.7	74.4
MIRA	4096	73.6	74.9	75.6
SMoG	4096	-	-	76.4
ExCB	1024	74.5	75.7	76.5

works, achieving state-of-the-art performance. Importantly, this holds across pretraining epochs, as shown in Tab. 3. Furthermore, ExCB outperforms previous works with a ViT backbone, despite using the hyperparameters suggested by [10] without hyperparameter tuning. These results highlight ExCB’s efficiency in terms of training time, its effectiveness across architectures, and its robustness with regard to batch size and training hyperparameters.

*Semi-supervised learning.* Having evaluated ExCB in a frozen backbone setting, we now present results for semi-supervised fine-tuning in Tab. 4, where a linear classifier and the backbone are trained on ImageNet’s train set with limited labels. Following previous works, we train with 1% and 10% of labels using the splits specified in [11], and report top-1 and top-5 accuracy on ImageNet’s validation set. We fine-tune for 50 epochs with batch size 512, and use a backbone learning rate of 0.00008 and 0.0003 and a classification head learning rate of 1. and 0.2 for the 1% and 10% settings respectively. In this setting as well, ExCB achieves state-of-the-art results. We stress that this is achieved in a much more resource-efficient way relative to other works, and highlight that the two most competitive methods in Tab. 4 are pretrained with a larger batch size [39] and for more epochs [29].

*Object Detection & Instance Segmentation.* Finally, we evaluate ExCB on dense prediction tasks, specifically object detection and instance segmentation. Following previous works, we train Mask R-CNN [26] with a C4 backbone on MS COCO [34] **train2017** for the standard  $1\times$  schedule and present results for detection ( $AP^b$ ) and segmentation ( $AP^m$ ) on MS COCO **val2017** in Tab. 5. We find that ExCB performs competitively with previous works, achieving best or second

**Table 4: Semi-supervised finetuning on ImageNet.** We finetune on ImageNet with 1% and 10% of labels and report top-1 and top-5 accuracy on the validation set.

Method	Batch Size	Epochs	1% Labels		10% Labels	
			Top-1	Top-5	Top-1	Top-5
Supervised	-	-	25.4	48.4	56.4	80.4
BYOL [24]	4096	1000	53.2	78.4	68.8	89.0
SwAV [9]	4096	800	53.9	78.5	70.2	89.9
Barlow Twins [50]	2048	1000	55.0	79.2	69.7	89.3
DINO [10]	4096	800	52.2	78.2	68.2	89.1
NNCLR [18]	4096	1000	56.4	80.7	69.8	89.3
MIRA [31]	4096	400	55.6	80.5	69.9	90.0
MAST [29]	2048	1000	55.8	81.0	<u>71.4</u>	<b>90.9</b>
SMoG [39]	4096	400	<b>58.0</b>	<u>81.6</u>	71.2	90.5
<b>ExCB</b>	<b>1024</b>	<b>400</b>	<u>57.8</u>	<b>81.8</b>	<b>71.5</b>	<u>90.7</u>

**Table 5: Object Detection & Segmentation.** We use a ResNet50 backbone pre-trained on ImageNet to initialize a Mask R-CNN [26] detector. We train it on MS COCO **train2017** and present results on **val2017**.

Method	Batch Size	Epochs	$AP^b$	$AP_{50}^b$	$AP_{75}^b$	$AP^m$	$AP_{50}^m$	$AP_{75}^m$
Supervised	-	-	38.2	58.2	41.2	33.3	54.7	35.2
MoCo-v2 [12]	256	800	<b>39.3</b>	58.9	<b>42.5</b>	<b>34.4</b>	55.8	<u>36.5</u>
SwAV [9]	4096	800	38.4	58.6	41.3	33.8	55.2	35.9
DINO [10]	4096	800	37.4	57.8	40.0	33.0	54.3	34.9
SimSiam [13]	256	800	<u>39.2</u>	<u>59.3</u>	42.1	<b>34.4</b>	<u>56.0</u>	<b>36.7</b>
BarlowTwins [50]	2048	1000	39.2	59.0	<b>42.5</b>	34.3	<u>56.0</u>	<u>36.5</u>
TWIST* [43]	2048	800+50	38.0	58.4	40.8	33.5	54.9	35.5
<b>ExCB</b>	<b>1024</b>	<b>400</b>	<u>39.2</u>	<b>59.4</b>	42.3	34.3	<b>56.2</b>	36.3

best outcomes for most metrics, and note that ExCB achieves this performance while having been trained for much fewer epochs than other methods.

### 4.3 Ablations & Analysis

In this section, we analyze the impact of ExCB’s key components and its training properties. Unless stated otherwise, ExCB is pretrained with ResNet50, for 100 epochs and without multi-crop.

*Batch size.* A key advantage of ExCB over previous methods is that it achieves state-of-the-art results with a much smaller batch size (1,024) than is typically used for pretraining. To further examine this property, we train ExCB with even smaller batch sizes, and present results in Tab. 6. We observe that, even for a batch size of 256 and without any hyperparameter tuning, ExCB suffers minimal performance drop, maintaining state-of-the-art performance for 100 epochs pretraining without multi-crop. This is a strong indication that ExCB is indeed remarkably effective and stable with small batch training.

**Table 6:** Linear evaluation accuracy on Imagenet set for varying batch sizes.

Batch Size	256	512	1024
Linear Acc.	70.5	70.6	<b>70.7</b>

**Table 7:** Linear evaluation accuracy on Imagenet for varying cluster numbers  $K$ .

Number of clusters $K$	32768	49152	65536	98304
Linear Acc.	70.3	70.5	<b>70.7</b>	70.6

*Number of clusters.* We ablate the number of clusters  $K$  used to pretrain ExCB and present results in Tab. 7. We observe that the optimal number is  $K = 65,536$ , similarly with [10], with minor performance drops for smaller and larger  $K$  values.

*Architecture & hyperparameters.* In Tab. 8 we present the impact of individual components of ExCB. Specifically, we examine the impact of including a predictor, not updating  $q_s$  for local crops, and having a separate weight decay schedule for  $q_s$ . We observe that each component in turn improves performance, even though, notably, ExCB’s baseline performance is already high and competitive with other works in the 100 pretraining epochs setting (see Tab. 3).

*Training statistics.* To provide insights into ExCB and facilitate future research, we complement our experimental results with information about ExCB’s training. We analyze a 400-epoch training run with multi-crop, and present in Fig. 4: **a)** ExCB’s training loss (Fig. 4a). **b)** The cluster assignments’ confidence  $c_{avg}$  (Fig. 4b), measured as in Eq. (11). **c)** The relative minimum/maximum cluster sizes for each epoch (Fig. 4c), measured as the ratio of the biggest/smallest cluster to the "optimal" size of  $1/K$  samples. **d)** The average agreement of the teacher’s assignments between views  $a_{avg}$  (Fig. 4d), measured as in Eq. (12). **e)** The purity [51] of the clusters with regard to the ground truth labels (Fig. 4e), which indicates how semantically meaningful they are.

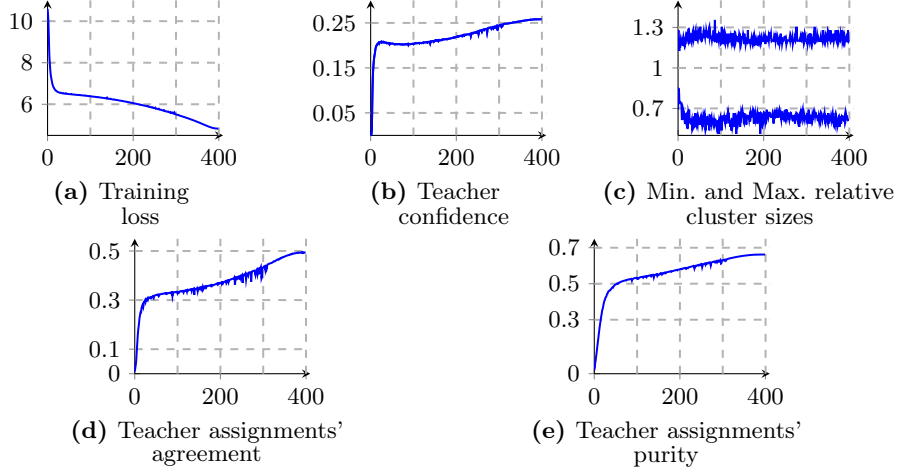
**Table 8:** Linear evaluation accuracy on Imagenet for ExCB *with* multi-crop, ablating on various components.

Predictor	$q_s$ Stop-gradient	$q_s$ WD	Linear Acc.
$\times$	$\times$	0.0001	73.8
$\checkmark$	$\times$	0.0001	74.0
$\checkmark$	$\checkmark$	0.0001	74.3
$\checkmark$	$\checkmark$	0.001 $\rightarrow$ 0.0001	<b>74.5</b>

$$c_{avg} = \frac{1}{NG} \sum_{n=1}^N \sum_{x'_n \in G} \max(\mathbf{p}_t(x'_n)) \quad (11)$$

$$a_{avg} = \frac{1}{NG} \sum_{n=1}^N \sum_{x'_n \in G} \sum_{\substack{x''_n \in G \\ x''_n \neq x'_n}} \mathbf{1}_{\arg\max_{k \in K}(\mathbf{p}_t(x'_n)) = \arg\max_{k \in K}(\mathbf{p}_t(x''_n))} \quad (12)$$

As seen in Figs. 4a and 4b, ExCB converges smoothly and without volatility, which highlights its stability. Furthermore, we observe in Figs. 4d and 4e that, during training, cluster assignments are increasingly reliable and semantically meaningful, as they become more consistent between views and increasingly include images with the same ground truth label. Finally, Fig. 4c demonstrates the effectiveness of our balancing method. Throughout the training, clusters remain between 70% and 130% of the optimal size, which, for ImageNet and  $K=65,536$ , translates to clusters being assigned approximately between 13 and 25 samples. The fact that ExCB operates consistently within such narrow margins is a strong indicator of its effectiveness and stability in terms of cluster balancing. Furthermore, we observe that, in practice, ExCB softly enforces lower and upper bounds on cluster sizes rather than absolute uniformity. This is a positive property, as [40] observed that enforcing balancing too strictly is detrimental to representation learning. Notably, in ExCB the upper/lower bounds emerge dynamically during training, as opposed to [40], where they are key hyperparameters that need to be defined by users.



**Fig. 4:** ExCB training statistics.

## 5 Discussion

As was outlined in Secs. 2 to 4, ExCB combines the best features of previous methods: **a)** it is online, therefore scalable to large datasets, **b)** it measures hard cluster assignments over multiple batches, which leads to reliable cluster size estimates even with small batch sizes, and **c)** it effectively balances clusters throughout training. This leads to state-of-the-art results in the primary, classification-related representation learning benchmarks, which include linear classification and semi-supervised learning. Additionally, ExCB demonstrates competitive performance in dense prediction tasks (i.e. object detection and segmentation), despite training for fewer epochs compared to other works. Furthermore, we apply ExCB to pretraining a ViT backbone and obtain excellent performance, even though no hyperparameter tuning was applied. These results indicate that ExCB is highly effective and versatile, and can be reliably used for pretraining with different architectures and for various downstream tasks.

Beyond its performance, a key feature of ExCB is its remarkable training efficiency in terms of the resources it requires. Regarding VRAM, our experiments demonstrate that ExCB is stable and effective with very small batch sizes, achieving state-of-the-art results with a batch size of 1,024 – by comparison, most other methods use a batch size 2X or 4X as large (with a corresponding increase in GPU memory utilization). With regard to training time, ExCB again demonstrates remarkable efficiency, outperforming methods that were trained for at least 2X as many epochs. We attribute this to our novel balancing module  $\mathcal{B}$ : due to its more accurate and reliable estimation of cluster sizes and its smoother method of regulating them by adjusting their assignments, the model’s supervision is much more stable throughout training. This facilitates convergence and results in better performance with less training time.

## 6 Conclusion

We present ExCB, a novel clustering-based framework for self-supervised representation learning. ExCB relies on a novel cluster balancing method that explicitly measures their sizes across multiple batches, and adjusts their assignments to promote evenly sized clusters. We conduct extensive experiments and find that ExCB achieves state-of-the-art results across benchmarks and backbone architectures. However, crucially, our experiments demonstrate that ExCB is also remarkably efficient, as it achieves the strong performance reported in this paper with less training and a much smaller batch size than most other frameworks. Overall, we believe that the proposed framework is not only significant in terms of its performance, but also as a step toward decreasing the resources required for self-supervised pretraining with visual data.

**Acknowledgements.** This work was supported by the EU H2020 AI4Media No.951911 project. This research utilised Queen Mary’s Apocrita HPC facility, supported by QMUL Research-IT. <http://doi.org/10.5281/zenodo.438045>.

## References

1. Afouras, T., Owens, A., Chung, J.S., Zisserman, A.: Self-supervised learning of audio-visual objects from video. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII* 16. pp. 208–224. Springer (2020)
2. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. In: *International Conference on Learning Representations (ICLR)* (2020)
3. Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N.: Masked siamese networks for label-efficient learning. In: *European Conference on Computer Vision*. pp. 456–473. Springer (2022)
4. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: *International Conference on Machine Learning*. pp. 1298–1312. PMLR (2022)
5. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* **33**, 12449–12460 (2020)
6. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In: *ICLR* (2022)
7. Bardes, A., Ponce, J., LeCun, Y.: Vicregl: Self-supervised learning of local visual features. *Advances in Neural Information Processing Systems* **35**, 8799–8810 (2022)
8. 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)
9. 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)
10. 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)
11. 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)
12. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020)
13. 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)
14. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9640–9649 (2021)
15. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* **26** (2013)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *North American Chapter of the Association for Computational Linguistics* (2019), <https://api.semanticscholar.org/CorpusID:52967399>
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.:

- An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
18. 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)
  19. Ermolov, A., Siarohin, A., Sangineto, E., Sebe, N.: Whitening for self-supervised representation learning. In: International Conference on Machine Learning. pp. 3015–3024. PMLR (2021)
  20. Feng, C., Patras, I.: Maskcon: Masked contrastive learning for coarse-labelled dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19913–19922 (2023)
  21. Gao, Z., Feng, C., Patras, I.: Self-supervised representation learning with cross-context learning between global and hypercolumn features. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1773–1783 (2024)
  22. Gidaris, S., Bursuc, A., Puy, G., Komodakis, N., Cord, M., Pérez, 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)
  23. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018), <https://openreview.net/forum?id=S1v4N210->
  24. 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)
  25. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
  26. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
  27. Hénaff, O.J., Koppula, S., Alayrac, J.B., Van den Oord, A., Vinyals, O., Carreira, J.: Efficient visual pretraining with contrastive detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10086–10096 (2021)
  28. Hénaff, O.J., Koppula, S., Shelhamer, E., Zoran, D., Jaegle, A., Zisserman, A., Carreira, J., Arandjelović, R.: Object discovery and representation networks. In: European Conference on Computer Vision. pp. 123–143. Springer (2022)
  29. Huang, C., Goh, H., Gu, J., Susskind, J.M.: MAST: masked augmentation subspace training for generalizable self-supervised priors. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), <https://openreview.net/pdf?id=5KUPKjHYD-1>
  30. Kostas, D., Aroca-Ouellette, S., Rudzicz, F.: Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience* **15**, 653659 (2021)
  31. Lee, D.H., Choi, S., Kim, H.J., Chung, S.Y.: Unsupervised visual representation learning via mutual information regularized assignment. *Advances in Neural Information Processing Systems* **35**, 29610–29623 (2022)
  32. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. In: ICLR (2021)



33. Li, Y., Pogodin, R., Sutherland, D.J., Gretton, A.: Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems* **34**, 15543–15556 (2021)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: *ECCV* (2014)
35. 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)
36. Mitrovic, J., McWilliams, B., Walker, J.C., Buesing, L.H., Blundell, C.: Representation learning via invariant causal mechanisms. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net (2021), <https://openreview.net/forum?id=9p2ekP904Rs>
37. Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., Kashino, K.: Byol for audio: Self-supervised learning for general-purpose audio representation. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–8. IEEE (2021)
38. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: *European conference on computer vision*. pp. 69–84. Springer (2016)
39. Pang, B., Zhang, Y., Li, Y., Cai, J., Lu, C.: Unsupervised visual representation learning by synchronous momentum grouping. In: *European Conference on Computer Vision*. pp. 265–282. Springer (2022)
40. Qian, Q., Xu, Y., Hu, J., Li, H., Jin, R.: Unsupervised visual representation learning by online constrained k-means. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16640–16649 (2022)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
42. Stegmüller, T., Lebailly, T., Bozorgtabar, B., Tuytelaars, T., Thiran, J.P.: Croc: Cross-view online clustering for dense visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7000–7009 (2023)
43. Wang, F., Kong, T., Zhang, R., Liu, H., Li, H.: Self-supervised learning by estimating twin class distribution. *IEEE Transactions on Image Processing* (2023)
44. Wang, G., Wang, K., Wang, G., Torr, P.H., Lin, L.: Solving inefficiency of self-supervised representation learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9505–9515 (2021)
45. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3024–3033 (2021)
46. Wen, X., Zhao, B., Zheng, A., Zhang, X., Qi, X.: Self-supervised visual representation learning with semantic grouping. *Advances in Neural Information Processing Systems* **35**, 16423–16438 (2022)
47. Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., Hu, H.: Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16684–16693 (2021)
48. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9653–9663 (2022)

49. Yeh, C.H., Hong, C.Y., Hsu, Y.C., Liu, T.L., Chen, Y., LeCun, Y.: Decoupled contrastive learning. In: European Conference on Computer Vision. pp. 668–684. Springer (2022)
50. 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)
51. Zhao, Y., Karypis, G.: Criterion functions for document clustering: Experiments and analysis (2001)
52. Zhu, J., Moraes, R.M., Karakulak, S., Sobol, V., Canziani, A., LeCun, Y.: Tico: Transformation invariance and covariance contrast for self-supervised visual representation learning. arXiv preprint arXiv:2206.10698 (2022)