Object discovery and representation networks

Olivier J. Hénaff, Skanda Koppula, Evan Shelhamer,
Daniel Zoran, Andrew Jaegle, Andrew Zisserman,
João Carreira, and Relja Arandjelović

DeepMind, London, UK

Abstract. The promise of self-supervised learning (SSL) is to leverage large amounts of unlabeled data to solve complex tasks. While there has been excellent progress with simple, image-level learning, recent methods have shown the advantage of including knowledge of image structure. However, by introducing hand-crafted image segmentations to define regions of interest, or specialized augmentation strategies, these methods sacrifice the simplicity and generality that makes SSL so powerful. Instead, we propose a self-supervised learning paradigm that discovers this image structure by itself. Our method, **Odin**, couples **o**bject **di**scovery and representation **n**etworks to discover meaningful image segmentations without any supervision. The resulting learning paradigm is simpler, less brittle, and more general, and achieves state-of-the-art transfer learning results for object detection and instance segmentation on COCO, and semantic segmentation on PASCAL and Cityscapes, while strongly surpassing supervised pre-training for video segmentation on DAVIS.

Keywords: self-supervised learning, object discovery, transfer learning

1 Introduction

Self-supervised learning proposes to leverage large amounts of unlabeled data to solve complex visual tasks. Early attempts hand-designed pretext tasks, which required some semantic understanding of images and their layout to solve [21, 60, 64, 89]. Contrastive learning departed from this tradition by radically simplifying the self-supervised protocol, in that the pretext task is specified by the data itself: representations must learn to distinguish a given example from the others in the dataset [25, 36, 62, 81]. Modern instances of the contrastive framework have proven to be very powerful, leading to strong performance on a variety of downstream tasks [13, 39, 43]. More recent self-supervised methods have simplified the framework further, removing the need for negative samples [35], bespoke architectural components [15], and learning dynamics [87], suggesting that increasingly domain-agnostic and data-driven methods might enable learning from ever-larger and more general sources of data.

However, a parallel line of work has asked whether the current self-supervised paradigm—which maximizes the similarity of the same data-point under different views—is too simple. By treating data-points as monolithic instances, these

methods overlook the complexity of real-world data: natural scenes are composed of many objects, natural speech of multiple speakers, and natural videos of many scenes. Ignoring this variability and encouraging models to represent different parts of an image in a similar manner risks dampening their selectivity for objects, their relationships, and layouts in real-world scenes. Indeed, several works have demonstrated the benefits of properly handling this variability when learning task-relevant representations [42, 70, 76, 77, 79]. While such object- and segmentation-aware approaches have yielded impressive empirical gains, they have relied on more domain-specific prior knowledge to expose the structure of the data—for example by using hand-crafted segmentation algorithms [42], or salience estimation networks trained on human annotations [77]—bounding how much they can learn from the data and what data they can be used on.

In this work we ask whether this knowledge can instead be derived from the data itself. To do so, we propose to couple two learning processes: object discovery and object representation. We use object discovery to uncover the structure of individual data-points, allowing the self-supervised task to focus on learning invariant representations of object-level instances. In turn, we use the resulting object representations as features for unsupervised object discovery, which feeds back into the representation learning process. These object discovery and representation networks thus engage in a virtuous cycle of representation and segmentation quality: better representations lead to better segmentations, and vice versa. Crucially, we derive the unsupervised segmentations with no prior knowledge about image structure or content, using a simple k-means clustering of local features to partition each image. We thus open the possibility of applying the algorithm to different domains, modalities, and their combination.

We make the following contributions: **1)** Our object discovery networks uncover, in an entirely self-supervised manner and without any prior knowledge of image structure or segmentation, meaningful decompositions of real-world scenes. **2)** Our object representation networks lead to state-of-the-art results in transfer learning to object detection and instance segmentation on COCO, and semantic segmentation on PASCAL and Cityscapes, surpassing prior works which exploit segmentation and saliency information, without requiring this prior knowledge. **3)** Our object representation networks seamlessly generalize to video understanding, surpassing supervised pre-training for video object segmentation on DAVIS. Finally, we test the resilience of our method by varying its essential components, and find it to be very robust, supporting further computational benefits. Together these results suggest that knowledge of scene structure, and the benefits it confers in representing objects, can—with the right learning paradigm—be extracted from the data itself.

2   Related Work

Pre-contrastive self-supervised learning: hand-designed tasks. Early self-supervised approaches focused on injecting human expertise and intuition into the design of proxy tasks for pretraining. For example, to stimulate the

network to learn object parts, [21] designed the task of predicting the spatial arrangement between local image patches. A rich collection of such intuitions and objectives was further developed, ranging from pixel-wise reconstruction-based approaches, such as denoising [78], inpainting [64], colorization [49, 89], and more [23, 90], to higher-level pretext tasks, such as predicting spatial layouts [21, 59, 60], orientation [30], egomotion [1], and temporal ordering [58].

Contrastive learning and its variants. Instance discrimination [25] has proven to be a very powerful pretext task which, we argue, owes its superior performance to being minimally hand-designed and maximally data-driven. By minimizing a contrastive loss [36, 62], the similarity of a representation across different 'views' of the same image is maximized, while minimizing their similarity with distracting negative samples. Multiple views of a single data-point can naturally be extracted from multimodal or multisensory data [2, 48, 56, 63, 68, 72] while for a single image-only modality they are typically constructed via local and global cropping [5, 43, 44, 62] or data-augmentation [13, 22, 25, 39, 81]. Positive pairs then correspond to views of the same data point, while negatives are sampled views of different data-points (typically from the same mini-batch), although the need for negative samples has recently been questioned [15, 35, 87].

Baking prior knowledge back into self-supervised learning. A growing body of research has brought hand-designed supervisory signals back into the self-supervised paradigm. For example, [42, 77, 79, 82, 88] decompose input images into their constituent objects and regions of interest using supervised segmentation algorithms, or hand-crafted heuristics. Object-level features are then computed for each region, and optimized using a contrastive objective. Other approaches use object-agnostic learning objectives, but integrate knowledge from segmentation heuristics or models in their augmentation strategies [57, 76, 92].

This trend is reflected in the broader research in self-supervised learning for other modalities. For example, [50] uses domain-specific knowledge to improve the masking strategies of BERT and other masked-language models. [37, 85] leverage motion and flow information to improve learning from video. And similar to previously described works in vision, [61] uses a segmentation step prior to applying SSL on point clouds. In all cases, we aim to remove the dependency on such prior knowledge while retaining its benefits for representation learning.

Clustering and representation learning. In parallel to the advent of contrastive methods, clustering-based representation learning methods have seen similar success, particularly in harnessing large amounts of uncurated images for transfer learning [4, 9, 10, 11, 33, 45]. Although they differ in their formulation of the self-supervised objective, these works also treat entire images as monolithic entities.

In contrast, IIC [45] performs within-image clustering using similarity losses counterbalanced by information maximization, obtaining compelling results in unsupervised segmentation. PiCIE [17] improves on this approach by imposing carefully-chosen, modality-specific geometric data augmentations and corresponding invariance and equivariance constraints. Neither of these works ex-

plicitly leverage their unsupervised segmentations for transfer learning across datasets and tasks however, which we investigate here.

Object discovery. Recent years have seen a growing interest in developing generative models that perform object discovery. By introducing different inductive biases such as mixture-model likelihoods [8, 34], attention [54, 93] and specific forms of factorization [46, 47], such models are able to discover objects and their interactions [20, 32, 71]. While much progress has been made, models from this family have yet to be demonstrated to work on natural images [34] and their application has been limited to synthetic data and simple highly structured environments typically used in robotics. Here we investigate object discovery on natural images in the wild, leveraging contrastive representation learning to enable this with simple k-means clustering.

3   Method

3.1   Self-supervised learning with Odin

Our method learns two sets of networks which work in collaboration. The *object discovery* network produces feature maps from high-resolution images. These feature maps are then spatially clustered to produce a segmentation of the image. The *object representation* networks learns better features via a contrastive loss which uses the masks proposed by the object discovery network. The resulting improved features are then used by the object discovery network to create better segmentations, and this process is continuously repeated. Figure 1 illustrates the full method, which we detail below.

Object discovery network: from representations to segmentations.
Given an image $\boldsymbol{x}$, we compute a spanning view $\boldsymbol{v}^0$ which encompasses most of the area of the image (Figure 1, *spanning view*, defined below) and which is simply cropped and resized. We use a feature extractor f_τ to encode this view into a spatial map of hidden vectors $\boldsymbol{h}^0 = f_\tau(\boldsymbol{v}^0)$ and projections $\boldsymbol{z}^0 = g_\tau(\boldsymbol{h}^0)$, where g_τ is a two-layer MLP which is applied to each vector independently, and τ are the parameters of the *object discovery network*. We apply K-means clustering to the spatial map of features $\boldsymbol{h}^0$ or $\boldsymbol{z}^0$, segmenting it (independently across images) into K non-overlapping binary masks $\boldsymbol{m}^{k,0}$ (Figure 1, top row).

Object representation networks: from segmentations to representations. We produce two views $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ of the image by augmenting $\boldsymbol{x}$ twice, using the random preprocessing pipeline of BYOL [35], which includes random cropping, flipping, blurring, and point-wise color transformations (Figure 1, *augmented views* and appendix).

The *spanning view* $\boldsymbol{v}^0$ is chosen as the smallest crop which spans the spatial extent of the augmented views $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$. We can therefore obtain two sets of masks $\boldsymbol{m}^{k,1}, \boldsymbol{m}^{k,2}$ which are consistent with each other and aligned with the underlying image content, by simply cropping, flipping, and resizing each
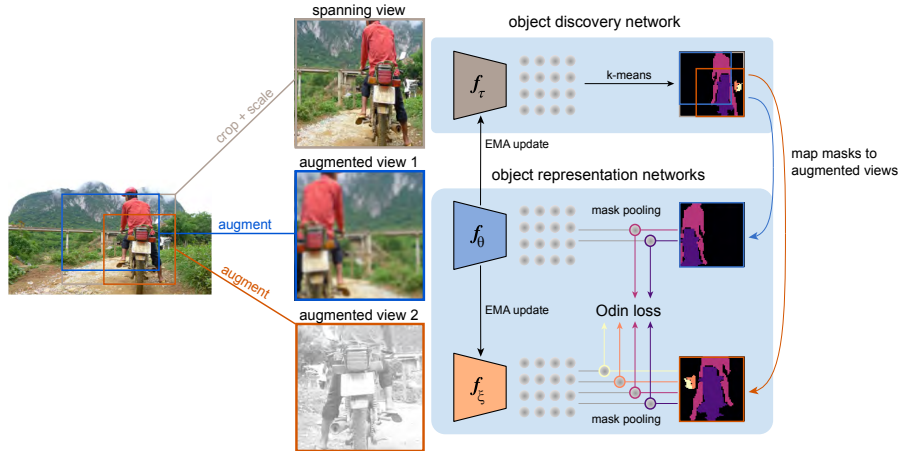
Fig. 1. Object discovery and representation networks. The object discovery network takes as input a cropped but otherwise un-augmented view of the image, and parses it using k-means clustering on its representation of it. The resulting segmentation is mapped into two augmented views of the same image, such that the masks are aligned across views and with the underlying image. The object representation networks take as input the augmented views of the image, and are trained using a self-supervised objective based on features pooled within each mask. The object discovery network is regularly updated with the parameters of the object representation network.

mask $\boldsymbol{m}^{k,0}$ as necessary (Figure 1, right). Despite the significant differences in appearance across views, these masks contain the same underlying image content (up to differences in cropping), which we leverage in our objective.

Each augmented view $\boldsymbol{v}^l \in \{\boldsymbol{v}^1, \boldsymbol{v}^2\}$ is encoded with a feature extractor f_θ into a spatial map of hidden vectors: $\boldsymbol{h}_\theta^l = f_\theta(\boldsymbol{v}^l)$ where θ are the parameters of the *object representation network* being optimized. For every mask $\boldsymbol{m}^{k,l}$ in the image, we compute a mask-pooled hidden vector

$$\boldsymbol{h}_\theta^{k,l} = \frac{1}{\sum_{i,j} \boldsymbol{m}^{k,l}[i,j]} \sum_{i,j} \boldsymbol{m}^{k,l}[i,j] \, \boldsymbol{h}_\theta^l[i,j], \tag{1}$$

discarding masks that are empty (due to cropping). Our goal is to ensure that these object-level features are roughly invariant across views. Specifically, we wish for an object-level feature in one view to be predictive of the same image content in the other view. To that end we transform the object-level hidden vectors $\boldsymbol{h}_\theta^{k,l}$ with two-layer MLPs g_θ and q_θ, yielding non-linear *projections* $\boldsymbol{z}_\theta^{k,l} = g_\theta(\boldsymbol{h}_\theta^{k,l})$ and *predictions* $q_\theta(\boldsymbol{z}_\theta^{k,l})$. In theory, we could regress the prediction $q_\theta(\boldsymbol{z}_\theta^{k,1})$ directly onto its target $\boldsymbol{z}_\theta^{k,2}$, however it is helpful to stabilize these targets by encoding them instead with specific *target networks* g_ξ and f_ξ, where the parameters ξ vary more slowly [16, 35, 42, 73]. We therefore instead use the projections $\boldsymbol{z}_\xi^{k,l} = g_\xi(\boldsymbol{h}_\xi^{k,l})$ as targets for the online prediction networks.

Jointly learning to discover and represent objects. Given a set of masks which approximately segment an image into objects, we wish to learn representations which distinguish these objects, while being invariant to identity-preserving transformations. Contrastive learning provides a straightforward objective for achieving this. Specifically, *contrastive detection* [42] trains a network to recognize object-level features across views, in the presence of many distracting "negative" features from other objects. The resulting objective maximizes the similarity between different views of the same object, while minimizing the similarity between different objects. We define the similarity between object-level features across views as

$$s_k^{1 \to 2} = \frac{1}{\alpha} \frac{\langle q_\theta(\boldsymbol{z}_\theta^{k,1}), \boldsymbol{z}_\xi^{k,2} \rangle}{\|q_\theta(\boldsymbol{z}_\theta^{k,1})\| \|\boldsymbol{z}_\xi^{k,2}\|} \tag{2}$$

where α is temperature hyper-parameter. We define the similarity between an object-level feature and a distracting negative sample $s_k^{1 \to n}$ analogously, by replacing the paired feature $\boldsymbol{z}_\xi^{k,2}$ with one from a different mask in the same image, or a different image altogether. The contrastive loss function for an individual feature is then

$$\ell_k^{1 \to 2}(\theta; \xi, \tau) = -\log \frac{\exp(s_k^{1 \to 2})}{\exp(s_k^{1 \to 2}) + \sum_n \exp(s_k^{1 \to n})}, \tag{3}$$

which we sum across objects, views, and images in the mini-batch (summation across images not shown for clarity)

$$\mathcal{L}(\theta; \xi, \tau) = \frac{1}{K} \sum_{k=1}^{K} \ell_k^{1 \to 2}(\theta; \xi, \tau) + \ell_k^{2 \to 1}(\theta; \xi, \tau). \tag{4}$$

We optimize the object discovery and representation networks using a strategy inspired by BYOL [35]. One object representation network (the *online network* with parameters θ) is updated with gradients from the contrastive loss. The second object representation network (the *target network* with parameters ξ) and the object discovery network (with parameters τ) are updated using an exponential moving average of the online network:

$$\theta \leftarrow \text{optimizer}(\theta, \nabla_\theta \mathcal{L}(\theta; \xi, \tau), \lambda_\theta) \tag{5}$$
$$\xi \leftarrow (1 - \lambda_\xi)\xi + \lambda_\xi \theta \tag{6}$$
$$\tau \leftarrow (1 - \lambda_\tau)\tau + \lambda_\tau \theta, \tag{7}$$

where the optimizer is LARS [86], and λ_θ, λ_ξ, λ_τ are learning rates for the online, target, and discovery networks respectively. We adopt the learning rates for online and targets networks from BYOL without modification. For the object discovery network, we consider two schedules: a constant learning rate which continuously updates the object discovery network with the online one (e.g. $\lambda_\tau = 10^{-3}$), and a small number of discrete updates which copy the online

representation network into the object discovery network (e.g. $\lambda_\tau = 1$ every 100 epochs and $\lambda_\tau = 0$ otherwise). The advantage of the second scheme is computational: if the object discovery network does not change between updates, the segments for training the object representation networks can be cached, removing the need to evaluate the object discovery network at every iteration.

Pretraining details. We train object discovery and representation networks (Odin) on ImageNet [69] for 1000 epochs, using a ResNet-50 [41] or Swin Transformer [53] backbone equipped with Feature Pyramid Networks (FPN; [51]) as the feature extractor f. The FPN takes as input the hierarchy of latent vectors output by the ResNet or Swin backbone, and progressively upsamples them while adding information from intermediate feature arrays, yielding high-level and high-resolution representations of the input image. We use the highest-resolution output of the FPN (subsampling the image by a factor of 4) as the array of hidden vectors $\boldsymbol{h}$.

After pretraining, we discard the target and object discovery networks, and use only the online object representation network for evaluation, facilitating the comparison to other methods which have their own means of learning the model.

3.2   Evaluating object discovery and representation

Having trained object representation networks using the Odin framework, we evaluate the quality of their representation by fine-tuning them for object detection and instance segmentation on COCO, and segmentatic segmentation on PASCAL and Cityscapes. For consistency with prior work we retain only the pretrained backbone (ResNet or Swin transformer) for transfer learning, discarding feature pyramid networks and projection heads.

Object detection and instance segmentation. For instance segmentation we use Mask-RCNN [40] while for object detection we report results for Mask-RCNN and FCOS*. Both methods are equipped with feature pyramid networks [51] and cross-replica batch-norm [65]. For Mask-RCNN we adopt the Cloud TPU implementation [31] and use it without modification. FCOS* is our implementation of a single-stage detector based on FCOS [75], and improved with IoU prediction [80], ATSS [91] and T-Head [29]; full details are available in the appendix. We follow the common transfer setup and evaluate on COCO [52] – the pretrained network is used to initialize the backbone of a Mask-RCNN or FCOS* model, which is then fine-tuned on the `train2017` set, and report bounding-box AP (AP$^{\mathrm{bb}}$) and mask AP (AP$^{\mathrm{mk}}$) on the `val2017` set. We use two standard training schedules: 12 epochs and 24 epochs [39].

Semantic segmentation with FCN. Following [39] we initialize the backbone of a fully-convolutional network (FCN, [55]) with our model. For PASCAL [27], we fine-tune on the `train_aug2012` set for 45 epochs and report the mean intersection over union (mIoU) on the `val2012` set. For Cityscapes [19], we fine-tune on the `train_fine` set for 160 epochs and evaluate on the `val_fine` set.

Object discovery on COCO. We wish to assess whether our representations uncover the structure of real-world scenes during self-supervised pretraining. Simply visualizing saliency maps induced by the model only weakly tests this ability however [26], hence we use the COCO dataset comprised of complex scenes and human-annotated object segments. Specifically, we evaluate models on COCO images, cluster their features, and measure the overlap between the resulting segments and human-annotated ones. Given the diversity of object scales in COCO, we run multiple K-means segmentations (for K in $[1, 2, \ldots, 128]$) on the same set of latents, resulting in 255 object proposals which we resize to the input image resolution.

For each ground-truth segment $\boldsymbol{g}_t$ we compute the overlap with all proposals $\boldsymbol{m}_k$ using their intersection-over-union (IoU), and record the "best overlap" by taking the maximum across proposals. Averaging this metric across ground-truth segments, we obtain the "average best overlap" (ABO) metric [3], and computing the fraction of "best overlaps" greater than 50% yields the "object recovery" metric [18]. We then average each of these metrics across images.

Video object segmentation on DAVIS. As a further test of scene understanding, we assess whether learned representations can continue to recognize parts of an object as they evolve over time. Video object segmentation, specifically in its semi-supervised setting, captures this ability, which we evaluate on the DAVIS'17 benchmark [66]. Having evaluated a learned representation on a video independently across frames, we segment these features with nearest neighbor matching from frame to frame, given a segmentation of the first frame. In this way, the segmentation is propagated according to the similarity of the representation across space and time.

4   Experiments

4.1   Transfer learning

Our first goal is to assess whether strong transfer learning performance can be obtained without resorting to prior knowledge of scene segmentations. To that end we train a ResNet-50 on ImageNet for 1000 epochs using the proposed Odin framework, and transfer it to object detection and instance segmentation on COCO, and semantic segmentation on PASCAL and Cityscapes.

Object detection and instance segmentation on COCO. Self-supervised learning has made steady gains on transfer learning from ImageNet to COCO, with a majority of methods surpassing supervised pretraining. The top-performing methods are ReLIC v2 and DetCon$_B$ which make heavy use of saliency or segmentation information in their learning paradigm. DetCon uses the same learning objective as Odin, but relies on a hand-crafted image segmentation algorithm [28] applied to the pixel lattice rather than a learned object discovery network. ReLIC v2 does not use segmentation information explicitly in its objective, but uses a hand-crafted saliency network to separate objects from their background

Table 1. Transfer to COCO object detection and instance segmentation with Mask-RCNN: all methods pretrain a ResNet-50 on ImageNet before fine-tuning on COCO with Mask-RCNN for 12 epochs ($1\times$ schedule) or 24 epochs ($2\times$ schedule). We report average precision on object detection (AP^{bb}) and instance segmentation (AP^{mk})

pretraining		fine-tune $1\times$		fine-tune $2\times$	
Method	Knows obj?	AP^{bb}	AP^{mk}	AP^{bb}	AP^{mk}
Supervised	no	39.6	35.6	41.6	37.6
VADeR [67]	no	39.2	35.6	-	-
MoCo [39]	no	39.4	35.6	41.7	37.5
SimCLR [13]	no	39.7	35.8	41.6	37.4
MoCo v2 [14]	no	40.1	36.3	41.7	37.6
InfoMin [74]	no	40.6	36.7	42.5	38.4
DeepCluster-v2 [11]	no	41.1	37.1	-	-
DINO [12]	no	41.2	37.1	42.3	38.1
PixPro [84]	no	41.4	-	-	-
BYOL [35]	no	41.6	37.2	42.4	38.0
SwAV [11]	no	41.6	37.8	-	-
ReLIC v2 [76]	**yes**	42.5	38.0	43.3	38.6
DetCon$_B$ [42]	**yes**	42.7	38.2	43.4	38.7
Odin	no	**42.9**	**38.4**	**43.8**	**39.1**

in the data-augmentation pipeline. Both represent a step-change in performance relative to previous methods. Odin, which instead derives segmentations from its own learned representations, surpasses both of these methods (Table 1).

A recent self-supervised method, DINO [12], reports high-quality unsupervised segmentations, however it appears to do so at the cost of object representation. We fine-tune the publicly available ResNet checkpoint in our framework, and find it underperforms relative to simple methods such as BYOL. Other SSL methods such as SwAV and DeepCluster-v2 [11] which cluster representations *across* images rather than within also underperform in this setting.

Semantic segmentation on PASCAL and Cityscapes. We assess the generality of these results by transferring them to two separate datasets and tasks, semantic segmetation on PASCAL and Cityscapes. Similarly to when transferring to COCO, DetCon and ReLIC v2 substantially outperform supervised and BYOL pretraining, confirming the utility of prior knowledge about segmentation and saliency. In this case as well, Odin successfully recovers this knowledge and surpasses both methods in a fully learned manner (Table 2).

In this setting DINO performs better, surpassing BYOL, possibly because semantic segmentation on PASCAL, which contains only 20 classes compared with 80 in COCO, weights object discovery more than object representation—isolating objects from the background rather than distinguishing object classes from each other. Nevertheless, Odin surpasses it as well, indicating that it achieves a better trade-off between object representation and discovery.

Table 2. Transfer to PASCAL and Cityscapes semantic segmentation with fully convolutional networks: all methods pretrain a ResNet-50 on ImageNet before fine-tuning for semantic segmentation on PASCAL or Cityscapes, and report the mean intersection-over-union

Method	Knows obj?	PASCAL	Cityscapes
Supervised	no	74.4	74.9
BYOL [35]	no	75.7	74.6
DeepCluster-v2 [11]	no	75.8	76.8
SwAV [11]	no	76.0	76.2
DINO [12]	no	76.9	75.6
DetCon$_B$ [42]	**yes**	77.3	77.0
ReLIC v2 [76]	**yes**	77.9	75.2
Odin	no	**78.6**	**77.1**

Table 3. Transfer to COCO object detection with FCOS*: all methods pretrain on ImageNet before fine-tuning on COCO with FCOS* for 30 epochs, and report average precision on object detection (AP^{bb}).

Pretraining	Knows obj?	ResNet-50	Swin-T	Swin-S
Supervised	no	44.2	46.7	48.3
DINO [12]	no	44.3	-	-
MOBY [83]	no	-	47.6	-
DetCon$_B$ [42]	**yes**	45.4	48.4	**50.4**
Odin	no	**45.6**	**48.5**	**50.4**

Transfer learning with high-performance architectures. While Mask-RCNN has become a standard method for evaluating the quality of object-level representations, we asked whether the performance gains afforded by the Odin framework persisted with more sophisticated models. For this we turned to our FCOS* implementation, whose supervised baseline surpasses Mask-RCNN by 4.6% AP^{bb}. In this setting as well, Odin surpasses the supervised baseline and DINO (+1.3% AP^{bb}, Table 3, 1st column).

Swin transformers appear as a compelling candidate for general-purpose vision architectures, surpassing ResNet's in a variety of tasks [53]. Despite the almost universal success of self-supervised pretraining in improving the transfer learning performance of ResNet architectures, similar results have yet to become widespread for Swin transformers.

We therefore pretrain Swin-T and Swin-S transformers on ImageNet using Odin, and transfer them to COCO object detection using FCOS*. We also evaluate a pre-trained Moby checkpoint in the same setting. Moby pretraining marginally improves the performance of a Swin-T, whereas Odin furthers these gains (+1.8% AP^{bb}, Table 3, 2nd column). The benefits of Odin pretraining are emphasized when pretraining and transferring the larger Swin-S backbone (+2.1% AP^{bb}, Table 3, 3rd column).

We return to our original question of whether Odin has successfully recovered knowledge of scene structure by pretraining ResNet-50, Swin-T, and Swin-S with DetCon (which uses a hand-crafted segmentation algorithm instead of the object discovery network [28, 42]), and transferring them with FCOS⋆. We find Odin to match or slightly surpass their performance, confirming our previous results.

4.2 Object discovery in COCO

We have found thus far that Odin surpasses the transfer learning performance of state-of-the-art self-supervised methods which rely on prior knowledge of scene segmentations, suggesting it has derived this knowledge from the data itself. In this section, we directly evaluate the extent to which Odin has discovered objects in real-world scenes. We extract Odin features from COCO images, cluster them, and visualize the resulting segments (Figure 2). Comparing unsupervised object proposals (last column) to human-annotated segments (2nd column) we see that Odin recovers a reasonable decomposition of real-world scenes: figures are separated from their background, small objects are isolated, and even different instances of the same class (such as cars, last row) are roughly separated from one-another. Failure modes, such as grouping multiple fruit—or two shirts—together, reflect the ambiguity of unsupervised object discovery.

Comparing these proposals to those obtained from a randomly-initialized network (3rd column) and an ImageNet supervised one (4th column), we appreciate the benefits of learning with the Odin framework. Both of these networks make erroneous proposals, failing to delineate object boundaries, or lacking the coherence and locality of real-world objects. We quantify this difference by evaluating the average best overlap (ABO) and fraction of recovered objects (OR) of the segments derived from each network. Consistently with the qualitative results, Odin strongly surpasses both baselines in all metrics (Table 4, left).

We also evaluate the accuracy of a recently-proposed self-supervised method, DINO, which specializes in object discovery. In this challenging task of discovering multiple objects in an unsupervised setting, we find that it underperforms relative to Odin. We test Odin in two regimes, one using the ResNet and FPN used for pretraining, the other with the ResNet only. Although its performance degrades slightly with the lower-resolution ResNet, it continues to outperform all other methods in all metrics. In particular, Odin surpasses DetCon by a large margin (+7% ABO, +16% OR), indicating that it has discovered more relevant image structure than the hand-crafted segmentations used in DetCon.

Finally, we note that the DINO method was primarily designed for use with vision transformers [24]. We therefore train a ViT-B/8 (as in DINO) on ImageNet for 100 epochs using the Odin framework (all other parameters unchanged). In this setting we find Odin to achieve compelling results, surpassing the high-resolution ResNet-FPN, and a supervised and DINO-pretrained vision-transformer (Table 4, right). Figure A.1. illustrates that Odin seems particularly effective at discovering small objects and differentiating instances of the same class. In sum, Odin provides a powerful means of discovering objects in real-world scenes irrespective of the architecture used.
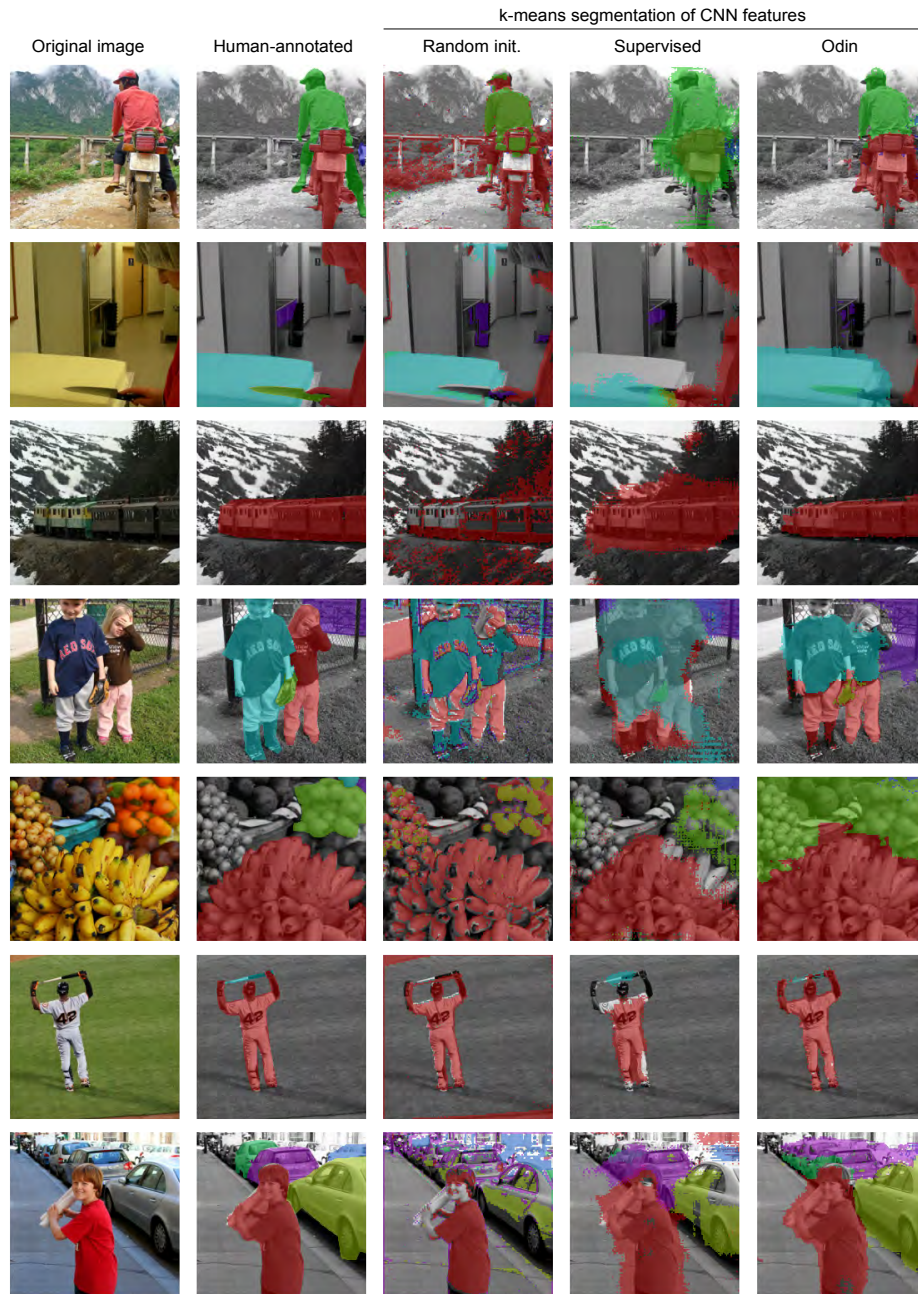
Fig. 2. Object discovery with Odin. 1st column: original image, 2nd: human-annotated COCO segmentations, 3rd, 4th, 5th: segmentations obtained from k-means clustering on randomly intialized, ImageNet-supervised, and Odin-trained features, respectively.

Table 4. Object discovery on COCO: all methods pretrain on ImageNet before evaluating object discovery on COCO in an unsupervised manner, reporting average best overlap of instance masks (ABO^i) and categorical masks (ABO^c), and average object recovery (OR). By default we retain only the pretrained ResNet-50 from Odin's feature extractor, such that all methods are matched in their architecture. **Odin**† denotes the model equipped with the FPN used during training. ResNet's use 1024×1024 images for evaluation with a stride of 32, yielding a 32×32 feature grid. ViT's use 448×448 resolution with a patch size of 8, yielding 56×56 feature grids.

| | ResNet-50 | | | ViT-B | | |
Pretraining	ABO^i	ABO^c	OR	ABO^i	ABO^c	OR
Random init	28.1	33.6	17.0	27.8	33.6	17.5
DetCon$_B$ [42]	34.1	40.0	20.4	-	-	-
Supervised	35.8	41.1	23.8	43.9	53.6	41.9
DINO [12]	38.3	46.5	30.8	42.7	51.7	39.7
Odin	**41.5**	**48.6**	**36.5**	**45.9**	**53.9**	**44.1**
Odin†	**43.0**	**53.0**	**42.3**			

Table 5. Video Object Segmentation on DAVIS'17: we evaluate representation quality for video segmentation by nearest neighbor inference. We report the standard region $\mathcal{J}$ and contour $\mathcal{F}$ metrics and their mean. All representations are trained on ImageNet then evaluated without fine-tuning on the 30 validation videos of DAVIS'17. Odin† includes a feature pyramid network to reduce the output stride from 32× to 8×.

Pretraining	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}$	$\mathcal{F}$
Random	15.2	15.9	14.6
Supervised	27.0	33.0	20.9
Odin	**35.6**	**41.3**	**29.9**
Odin†	**54.1**	**54.3**	**53.9**

4.3 Video Object Segmentation

We evaluate on the DAVIS'17 benchmark [66] following the experimental setup and nonparametric inference method of DINO [12]. Given a video and a segmentation of its first frame, we propagate the segmentation between consecutive frames by nearest neighbor matching of the extracted representation. In Table 5 we evaluate random, supervised, and our self-supervised representations with the ResNet-50 architecture. This evaluation does not fine-tune or train on the DAVIS benchmark, and so accuracy is a measure of object representation, as the fixed representation must support segmentation of the novel objects in these held-out videos. Consistently with our previous results, Odin strongly surpasses supervised pre-training in all metrics.

4.4 Ablations and analysis

What components are necessary for driving Odin's ability to represent and discover objects? We systematically vary the two hyper-parameters governing the

Table 6. Ablating the components of Odin: We use the variant of Odin equipped with FPN for object discovery. Transfer learning is performed with the ResNet backbone only. K denotes the number of segments obtained through K-means during pretraining

Odin pretraining		Object discovery		Mask-RCNN transfer	
K	update sched.	ABO^i	OR	AP^{bb}	AP^{mk}
8	every 100 ep.	38.3	34.6	42.6	38.1
16	every 100 ep.	43.0	**42.3**	42.6	38.1
32	every 100 ep.	**43.1**	42.0	42.5	38.0
16	$\lambda_\tau = 10^{-2}$	41.0	39.5	42.5	38.1
16	$\lambda_\tau = 10^{-3}$	41.3	39.9	**42.9**	**38.4**
16	$\lambda_\tau = 10^{-4}$	41.6	40.1	42.6	38.1

behavior of the object discovery network: the number of segments used for learning, and the schedule used for updating the network (Table 6). Starting with the number of segments K we find object discovery degrades substantially when using too coarse segmentations (e.g. $K = 8$). However, given a fine enough segmentations (K greater than 16) its performance is stable.

Regarding the rate at which the object discovery network is updated, we find both schemes to be viable: continuously updating the network leads to slightly better representations, whereas discrete updates lead to slightly better object discovery. The advantage of the later scheme is that the computational cost of the object discovery network becomes negligible, as it only needs to be evaluated every 100 epochs and resulting segmentations cached in-between.

5   Conclusions

We have presented Odin, a new approach to self-supervised training which couples object discovery and representation learning. The resulting framework benefits from the same representation quality as methods which utilize explicit priors about image segmentation [42, 76], while deriving this knowledge from the data itself. The result is a simpler and more generally-applicable learning paradigm, and leads to state-of-the-art performance on a range of transfer learning tasks.

In this work, we have shown the utility of coupling representation learning and object discovery, for transfer learning and unsupervised scene understanding. Nevertheless, we have presented a single instance of this coupling, and there remain several open questions around how best to tie them together. This may require greater integration of the learning procedure and architecture—for example our self-supervised algorithm learns mask-pooled features which are different from those used in downstream tasks. The learning dynamics of Odin also warrant further investigation, as well as the objective used for representation learning. Recent work has revived interest in masked-autoencoding [7, 24, 38] and masked-distillation [6] as viable alternatives to contrastive learning. Odin, by proposing to leverage learned representations in the design of iteratively refined self-supervised tasks, is well positioned to benefit them as well.

Bibliography

[1] Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: ICCV (2015)

[2] Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017)

[3] Arbeláez, P., Pont-Tuset, J., Barron, J.T., Marques, F., Malik, J.: Multi-scale combinatorial grouping. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 328–335 (2014)

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

[5] Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. Advances in Neural Information Processing Systems **32**, 15535–15545 (2019)

[6] 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. arXiv preprint arXiv:2202.03555 (2022)

[7] Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)

[8] Burgess, C.P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., Lerchner, A.: Monet: Unsupervised scene decomposition and representation. CoRR **abs/1901.11390** (2019), http://arxiv.org/abs/1901.11390

[9] Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: The European Conference on Computer Vision (ECCV) (September 2018)

[10] Caron, M., Bojanowski, P., Mairal, J., Joulin, A.: Unsupervised pre-training of image features on non-curated data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2959–2968 (2019)

[11] 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)

[12] 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)

[13] 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)

[14] Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)

[15] 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)

[16] 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)

[17] Cho, J.H., Mall, U., Bala, K., Hariharan, B.: Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: CVPR. pp. 16794–16804 (2021)

[18] Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1201–1210 (2015)

[19] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)

[20] Didolkar, A., Goyal, A., Ke, R., Blundell, C., Beaudoin, P., Heess, N.M.O., Mozer, M., Bengio, Y.: Neural production systems (2021)

[21] Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1422–1430 (2015)

[22] Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2051–2060 (2017)

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

[24] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[25] Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)

[26] 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)

[27] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)

[28] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International journal of computer vision **59**(2), 167–181 (2004)

[29] Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: TOOD: Task-aligned one-stage object detection. In: Int. Conf. Comput. Vis. (2021)

[30] Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. International Conference on Learning Representations (ICLR) (2018)

[31] GitHub: Tpu object detection and segmentation framework (2021),
https://github.com/tensorflow/tpu/tree/master/models/official/
detection

[32] Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio,
Y., Schölkopf, B.: Recurrent independent mechanisms. arXiv preprint
arXiv:1909.10893 (2019)

[33] Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M.,
Liptchinsky, V., Misra, I., Joulin, A., et al.: Self-supervised pretraining of
visual features in the wild. arXiv preprint arXiv:2103.01988 (2021)

[34] Greff, K., Kaufman, R.L., Kabra, R., Watters, N., Burgess, C., Zoran, D.,
Matthey, L., Botvinick, M., Lerchner, A.: Multi-object representation learn-
ing with iterative variational inference. In: International Conference on Ma-
chine Learning. pp. 2424–2433. PMLR (2019)

[35] 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** (2020)

[36] Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an
invariant mapping. In: 2006 IEEE Computer Society Conference on Com-
puter Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742.
IEEE (2006)

[37] Han, T., Xie, W., Zisserman, A.: Self-supervised co-training for video rep-
resentation learning. Advances in Neural Information Processing Systems
33, 5679–5690 (2020)

[38] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoen-
coders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)

[39] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsu-
pervised visual representation learning. In: Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition. pp. 9729–9738
(2020)

[40] 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)

[41] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recog-
nition. In: Proceedings of the IEEE conference on computer vision and
pattern recognition. pp. 770–778 (2016)

[42] Hénaff, O.J., Koppula, S., Alayrac, J.B., Oord, A.v.d., Vinyals, O., Carreira,
J.: Efficient visual pretraining with contrastive detection. In: ICCV (2021)

[43] Hénaff, O.J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S.,
Oord, A.v.d.: Data-efficient image recognition with contrastive predictive
coding. arXiv preprint arXiv:1905.09272 (2019)

[44] Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P.,
Trischler, A., Bengio, Y.: Learning deep representations by mutual informa-
tion estimation and maximization. In: International Conference on Learning
Representations (2018)

[45] Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9865–9874 (2019)

[46] Kabra, R., Zoran, D., Erdogan, G., Matthey, L., Creswell, A., Botvinick, M., Lerchner, A., Burgess, C.: Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. Advances in Neural Information Processing Systems **34** (2021)

[47] Kipf, T., Elsayed, G.F., Mahendran, A., Stone, A., Sabour, S., Heigold, G., Jonschkowski, R., Dosovitskiy, A., Greff, K.: Conditional object-centric learning from video. arXiv preprint arXiv:2111.12594 (2021)

[48] Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Adv. Neural Inform. Process. Syst. (2018)

[49] Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR. pp. 6874–6883 (2017)

[50] Lin, C., Miller, T., Dligach, D., Bethard, S., Savova, G.: Entitybert: Entity-centric masking strategy for model pretraining for the clinical domain. In: Proceedings of the 20th Workshop on Biomedical Language Processing. pp. 191–201 (2021)

[51] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)

[52] 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: Eur. Conf. Comput. Vis. pp. 740–755. Springer (2014)

[53] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)

[54] Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in Neural Information Processing Systems **33**, 11525–11538 (2020)

[55] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

[56] Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100M: Learning a text-video embedding by watching hundred million narrated video clips. In: Int. Conf. Comput. Vis. (2019)

[57] Mishra, S., Shah, A., Bansal, A., Jagannatha, A., Sharma, A., Jacobs, D., Krishnan, D.: Object-aware cropping for self-supervised learning. arXiv preprint arXiv:2112.00319 (2021)

[58] Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: ECCV (2016)

[59] Nathan Mundhenk, T., Ho, D., Chen, B.Y.: Improvements to context based self-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9339–9348 (2018)

[60] Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision. pp. 69–84. Springer (2016)

[61] Nunes, L., Marcuzzi, R., Chen, X., Behley, J., Stachniss, C.: Segcontrast: 3d point cloud feature representation learning through self-supervised segment discrimination. IEEE Robotics and Automation Letters (2022)

[62] van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

[63] Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Eur. Conf. Comput. Vis. (2018)

[64] 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. pp. 2536–2544 (2016)

[65] Peng, C., Xiao, T., Li, Z., Jiang, Y., Zhang, X., Jia, K., Yu, G., Sun, J.: Megdet: A large mini-batch object detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6181–6189 (2018)

[66] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)

[67] Pinheiro, P.O., Almahairi, A., Benmalek, R.Y., Golemo, F., Courville, A.C.: Unsupervised learning of dense visual representations. In: NeurIPS (2020)

[68] Recasens, A., Luc, P., Alayrac, J.B., Wang, L., Hemsley, R., Strub, F., Tallec, C., Malinowski, M., Patraucean, V., Altché, F., Valko, M., Grill, J.B., van den Oord, A., Zisserman, A.: Broaden your views for self-supervised video learning. In: Int. Conf. Comput. Vis. (2021)

[69] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)

[70] Ryali, C., Schwab, D.J., Morcos, A.S.: Learning background invariance improves generalization and robustness in self-supervised learning on imagenet and beyond. In: Adv. Neural Inform. Process. Syst. (2021)

[71] Shanahan, M., Nikiforou, K., Creswell, A., Kaplanis, C., Barrett, D., Garnelo, M.: An explicitly relational neural network architecture. In: International Conference on Machine Learning. pp. 8593–8603. PMLR (2020)

[72] Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: VideoBERT: A joint model for video and language representation learning. In: Int. Conf. Comput. Vis. (2019)

[73] Tian, Y., Henaff, O.J., van den Oord, A.: Divide and contrast: Self-supervised learning from uncurated data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10063–10074 (2021)

[74] Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning. NeurIPS (2020)

[75] Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: Int. Conf. Comput. Vis. (2019)

[76] Tomasev, N., Bica, I., McWilliams, B., Buesing, L., Pascanu, R., Blundell, C., Mitrovic, J.: Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? arXiv preprint arXiv:2201.05119 (2022)

[77] Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Van Gool, L.: Unsupervised semantic segmentation by contrasting object mask proposals. In: ICCV (2021)

[78] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103 (2008)

[79] Wei, F., Gao, Y., Wu, Z., Hu, H., Lin, S.: Aligning pretraining for detection via object-level contrastive learning. Advances in Neural Information Processing Systems **34** (2021)

[80] Wu, S., Li, X., Wang, X.: IoU-aware single-stage object detector for accurate localization. Image and Vision Computing (2020)

[81] Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via nonparametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018)

[82] Xie, J., Zhan, X., Liu, Z., Ong, Y., Loy, C.C.: Unsupervised object-level representation learning from scene images. Advances in Neural Information Processing Systems **34** (2021)

[83] Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., Hu, H.: Self-supervised learning with swin transformers. arXiv preprint arXiv:2105.04553 (2021)

[84] 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)

[85] Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: Int. Conf. Comput. Vis. (2021)

[86] You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888 (2017)

[87] 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)

[88] Zhang, F., Torr, P., Ranftl, R., Richter, S.: Looking beyond single images for contrastive semantic segmentation learning. Advances in Neural Information Processing Systems **34** (2021)

[89] Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016)

[90] 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. pp. 1058–1067 (2017)

[91] Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)

[92] Zhao, N., Wu, Z., Lau, R.W., Lin, S.: Distilling localization for self-supervised representation learning. arXiv preprint arXiv:2004.06638 (2020)

[93] Zoran, D., Kabra, R., Lerchner, A., Rezende, D.J.: Parts: Unsupervised segmentation with slots, attention and independence maximization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10439–10447 (2021)