FreeSOLO: Learning to Segment Objects without Annotations*

Xinlong Wang[1], Zhiding Yu[2], Shalini De Mello[2], Jan Kautz[2],
Anima Anandkumar[2,3], Chunhua Shen[4], Jose M. Alvarez[2]

[1] The University of Adelaide [2] NVIDIA [3] Caltech [4] Zhejiang University

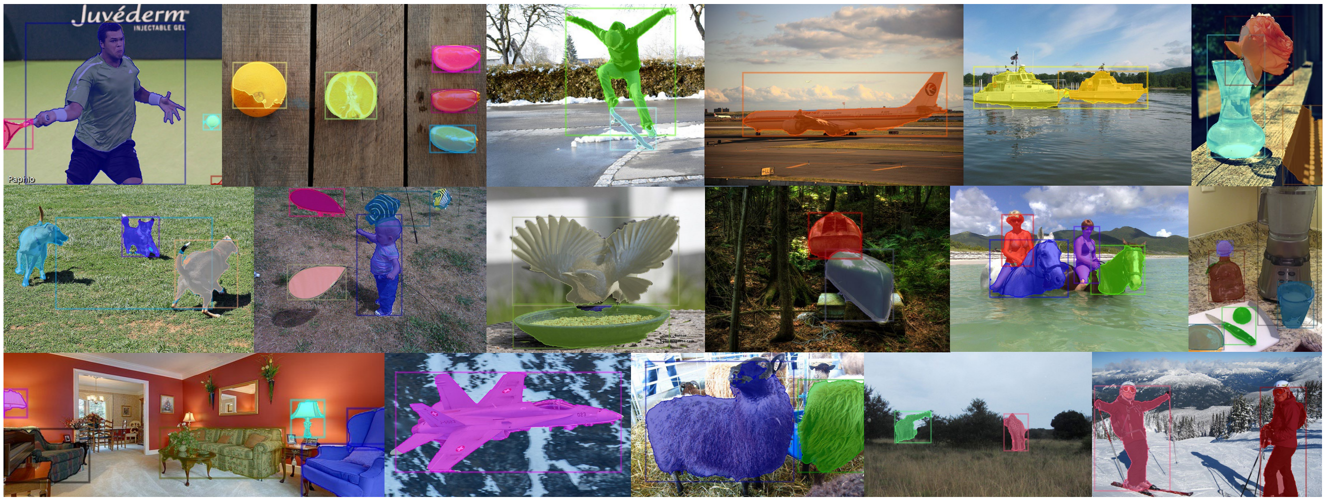

Figure 1. Qualitative results of FreeSOLO for the task of class-agnostic instance segmentation. The model is trained *without any kind of manual annotations* and can infer at 16 FPS on a V100 GPU. Best viewed on screen.

Abstract

Instance segmentation is a fundamental vision task that aims to recognize and segment each object in an image. However, it requires costly annotations such as bounding boxes and segmentation masks for learning. In this work, we propose a fully unsupervised learning method that learns class-agnostic instance segmentation without any annotations. We present FreeSOLO, a self-supervised instance segmentation framework built on top of the simple instance segmentation method SOLO. Our method also presents a novel localization-aware pre-training framework, where objects can be discovered from complicated scenes in an unsupervised manner. FreeSOLO achieves 9.8% AP$_{50}$ on the challenging COCO dataset, which even outperforms several segmentation proposal methods that use manual annotations. For the first time, we demonstrate unsupervised class-agnostic instance segmentation successfully. FreeSOLO's box localization significantly outperforms state-of-the-art unsupervised object detection/discovery methods, with about 100% relative improvements in COCO AP. FreeSOLO further demonstrates superiority as a strong pre-training method, outperforming state-of-the-art self-supervised pre-training methods by +9.8% AP when fine-tuning instance segmentation with only 5% COCO masks.

Code is available at: `github.com/NVlabs/FreeSOLO`

1. Introduction

Instance segmentation is a fundamental computer vision task that requires recognizing the objects in an image and segmenting each of them at the pixel level. Instance segmentation subsumes object detection, as bounding box can be thought of as a coarse parametric representation of a segmentation mask. Therefore, it is a more demanding and challenging task than object detection by requiring both instance-level and pixel-level predictions. Recently, significant progress [1–7] has been made to address the instance segmentation task. However, the dense prediction nature of the task requires rich and expensive annotations during training. Weakly-supervised instance segmentation methods are thus proposed to relax the annotation requirements [8–13]. Latest methods such as BoxInst [11] and Dis-

*Part of this work was done when XW was an intern at NVIDIA, and CS was with The Univerity of Adelaide. CS is the corresponding author.

coBox [13] have significantly closed the gap to fully supervised methods. However, their competitive result still relies on box or point annotations that contain strong localization information.

In this work, we explore *learning class-agnostic instance segmentation without any annotations.* The work here is built upon our recent work of SOLO [7], a simple yet strong instance segmentation framework, and the self-supervised dense feature learning method of DenseCL [14]. SOLO adopts a one-stage design, which contains a category branch and a mask branch to encode the object category information and segmentation proposals, respectively. Our main intuition is that this "top-down meets bottom-up" design allows us to unify pixel grouping, object localization and feature pre-training in a fully self-supervised manner.

Our proposed framework, **FreeSOLO**, contains two major pillars: Free Mask and Self-supervised SOLO, as shown in Figure 2. Specifically, Free Mask contains self-supervised design elements that promote objectness in network attention. It contains a "query-key" attention design, where the queries and keys are constructed from self-supervised features. The method takes the cosine similarity between each query with all the keys, thus obtaining a set of query-conditioned (seeded) attention maps as coarse masks. The coarse masks are ranked and filtered by their maskness scores, followed by non-maximum suppression (NMS) to further remove the redundant masks. Self-Supervised SOLO then takes the coarse masks as pseudo-labels to train a SOLO model. Since the coarse masks can be inaccurate, Self-Supervised SOLO contains a weakly-supervised design to better accommodate the label noise. This is followed by a self-training strategy to further refine mask quality and to improve accuracy. Our network design is almost the same as SOLO with minimal modifications, thus leading to simple and fast inference process.

FreeSOLO provides an effective solution to the challenging problem of self-supervised instance segmentation. With the bounding boxes obtained from the predicted masks, FreeSOLO also shows significant advantage as an unsupervised object discovery method. In addition to the above roles, we further consider FreeSOLO as a strong self-supervised pretext task for instance segmentation by jointly learning object-level and pixel-level representations. Compared to pre-training for image classification [15–17], object detection [18, 19] and semantic segmentation [20, 21], pre-training for instance segmentation is still under-studied. General instance segmentation requires not only localizing objects at the pixel level, but also recognizing their semantic categories. Interestingly, the design of FreeSOLO allows us to directly learn object-level semantic representations in an unsupervised manner. Upon completing the pre-training, all the learned parameters except for the last classification layer can be used to initialize the supervised instance segmentation models to improve accuracy.

Our contributions can be summarized as follows.

- We propose the Free Mask approach, which leverages the specific design of SOLO to effectively extract coarse object masks and semantic embeddings in an unsupervised manner.
- We further propose Self-Supervised SOLO, which takes the coarse masks and semantic embeddings from Free Mask and trains the SOLO instance segmentation model, with several novel design elements to overcome label noise in the coarse masks.
- With the above methods, FreeSOLO presents a simple and effective framework that demonstrates unsupervised instance segmentation successfully for the first time. Notably, it outperforms some proposal generation methods that use manual annotations. FreeSOLO also outperforms state-of-the-art methods for unsupervised object detection/discovery by a significant margin (relative +100% in COCO AP).
- In addition, FreeSOLO serves as a strong self-supervised pretext task for representation learning for instance segmentation. For example, when fine-tuning on COCO dataset with 5% labeled masks, FreeSOLO outperforms DenseCL [14] by +9.8% AP.

2. Related Work

Instance segmentation. Instance segmentation has attracted much attention in recent years. Most existing works focus on learning instance segmentation with full annotations. Top-down methods [1, 2, 4, 22] solve the problem from the perspective of object detection, *i.e.*, detecting the bounding box of objects first and then segmenting the object in the box. Bottom-up methods [3, 23–25] view the task as a label-then-cluster problem, *e.g.*, by learning per-pixel embeddings first and then clustering them into groups. Some recent methods [5, 6, 26–28] seek a combination of top-down and bottom-up approaches to perform faster inference and better segmentation. Among these methods, SOLO has shown a promising speed/accuracy trade-off with a very simple architecture. A few works explore learning instance segmentation with weak annotations, *e.g.*, image-level and box-level labels [8, 9, 11, 29]. To the best of our knowledge, none have additionally explored learning instance segmentation without any labels at all.

In particular, BoxInst [11] attains strong instance segmentation results using box annotations only, demonstrating that instance segmentation may not necessarily be more difficult to solve than box-level object detection. We move one-step forward by reporting strong instance segmentation results in an unsupervised setting, without any annotations.

Self-supervised learning. To learn a good visual representation from unlabeled data, a wide range of pretext tasks have been explored, *e.g.*, colorization [30], inpaint-

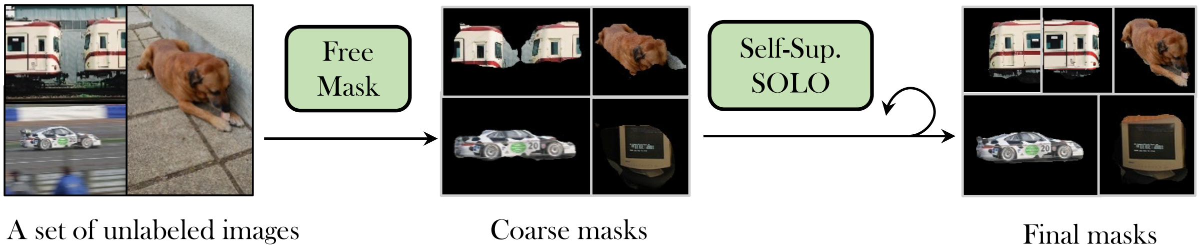

Figure 2. Overview of FreeSOLO. Unlabeled images are first input to Free Mask to generate coarse object masks. The segmentation masks as well as their associated semantic embeddings are used to train a SOLO-based instance segmentation model via weak supervision. We use self-training to improve object mask segmentation.

ing [31], jigsaw puzzles [32] and orientation discrimination [33]. The breakthroughs came from the contrastive learning methods, *e.g.*, SimCLR [16] and MoCo [15] that perform an instance discrimination pretext task [34]. Besides pre-training for image classification [17,35,36], some recent works [14, 19, 37–39] design self-supervised pre-training methods for dense prediction tasks, *e.g.*, object detection and semantic segmentation. Different from them, our method can not only learn intermediate representations, but also train instance segmenters, which can segment objects in the wild. Our FreeSOLO naturally serves as a strong pretext task for learning representations for instance segmentation. The pre-trained model can be seamlessly transferred to supervised fine-tuning and can achieve significant gains compared to existing pre-training methods.

Unsupervised object discovery. A wide range of approaches have been proposed for unsupervised object discovery, including statistical topic discovery models [40, 41], link analysis technique [42], clustering by composition [43], and part-based matching [44]. Some recent works [45, 46] formulate object discovery as an optimization problem. LOD [47] further proposes to formulate unsupervised object discovery as a ranking problem. Yet, the existing methods have achieved limited success in challenging and complicated scenes. Furthermore, most of these methods can only find coarse bounding boxes of objects. By contrast, our method discovers and localizes objects in the wild with pixel-wise segmentation masks. With bounding boxes obtained from predicted masks, FreeSOLO outperforms the state-of-the-art unsupervised object discovery methods by a large margin.

Unsupervised segmentation. To remove the dependency on manual supervision, some object co-segmentation works [48–50] make a strong assumption about the image collection, *i.e.*, to segment common repeated objects in a collection of images. Besides, there are a few works [51–53] that explore unsupervised semantic segmentation. Some [51] only deal with simple scenarios, and some [52, 53] still require a salient object estimator or boundary annotations. In addition, the key difference lies in the task. Instead of semantic segmentation, our method solves the harder problem of instance segmentation, *i.e.*, to segment each object individually.

3. Method

Background. We briefly introduce the supervised instance segmentation method SOLO [7]. SOLO shows that instance segmentation can be solved by directly mapping an input image to the desired object categories and instance masks using fully convolutional networks (FCNs), eliminating the need for bounding box detection or grouping via post-processing. Its main idea is to formulate instance segmentation into two simultaneous category-aware pixel-level prediction problems. It conceptually divides the input image into $S \times S$ grids. A grid cell is responsible for predicting the semantic category as well as the segmentation mask for an object whose center falls into that grid cell. The model consists of two branches, *i.e.*, a category branch and a mask branch. The category branch predicts the semantic categories. The mask branch generates S^2 sized masks, one corresponding to each grid cell. Specifically, the dynamic SOLO variant employs dynamic convolutions to separately predict the mask kernels and mask features. The mask features are then convolved with the predicted mask kernels to generate the masks. This operation can be written as:

$$\mathbf{S} = \mathbf{G} \circledast \mathbf{F}, \tag{1}$$

where $\mathbf{G}$ is the convolution kernel, and $\mathbf{S}$ denotes the score maps for all the S^2 masks. $\mathbf{S}$ is then normalized via a `sigmoid` operation, and input to mask NMS to form the final object masks.

3.1. Overview of FreeSOLO

We propose a novel framework for self-supervised instance segmentation, termed FreeSOLO. FreeSOLO does not require any type of annotations, neither pixel-level nor image-level labels, and simply uses a collection of unlabeled images for training. Its overall pipeline is illustrated in Figure 2. We first propose the Free Mask approach to generate segmentation masks from a self-supervised pre-trained model. For each unlabeled image, the coarse object masks can be generated fast with simple operations, *e.g.*, at 21 FPS on a V100 GPU with a ResNet-50-based backbone. We further propose Self-Supervised SOLO, which trains the SOLO-based instance segmenter using the coarse masks and semantic embeddings from Free Mask, with several novel design elements including weaky-supervised design, self-training, and semantic embedding learning.

With FreeSOLO, we obtain an instance segmentation model given only unlabeled images. In addition to unsupervised instance segmentation itself, the well-trained model serves as a strong pre-trained model for downstream fine-tuning. All its parameters except the last classification layer can be transferred to supervised instance segmentation as a strong initialization.

3.2. Free Mask

Free Mask generates object masks from unlabeled images. As shown in Figure 3, given an input image, dense feature maps $\mathbf{I} \in \mathbb{R}^{H \times W \times E}$ are extracted by a backbone model trained via self-supervision, *e.g.*, ResNet [54] or any other convolutional neural network. This pre-trained model can be from supervised or unsupervised pre-training, as discussed below. We first construct queries $\mathbf{Q}$ and keys $\mathbf{K}$ from the features $\mathbf{I}$, which work together to generate the coarse segmentation masks. We bilinearly downsample $\mathbf{I}$ to form the queries $\mathbf{Q} \in \mathbb{R}^{H' \times W' \times E}$, where H' and W' denote the downsampled spatial size. $\mathbf{I}$ itself is used as the set of keys $\mathbf{K}$. For each query in $\mathbf{Q}$, we compute its cosine similarity with every key in $\mathbf{K}$, thus obtaining the score maps $\mathbf{S} \in \mathbb{R}^{H \times W \times N}$, where $N = H' \times W'$ is the total number of queries. This operation can be written as:

$$\mathbf{S}_{i,j,q} = \mathtt{sim}(\mathbf{Q}_q, \mathbf{K}_{i,j}), \tag{2}$$

where $\mathbf{Q}_q \in \mathbb{R}^E$ is the q^{th} query, and $\mathbf{K}_{i,j} \in \mathbb{R}^E$ is the key at spatial location (i, j). $\mathtt{sim}(\boldsymbol{u}, \boldsymbol{v})$ denotes the cosine similarity, calculated by the dot product between ℓ_2-normalized $\boldsymbol{u}$ and $\boldsymbol{v}$, *i.e.*, $\mathtt{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$. The process can also be viewed as a convolution where the ℓ_2 normalized queries $\mathbf{Q}'$ and keys $\mathbf{K}'$ are respectively the convolutional kernels and the features to be convolved together. Each of the normalized queries is treated as a 1×1 convolutional kernel. Thus the operation can also be written as:

$$\mathbf{S} = \mathbf{Q}' \circledast \mathbf{K}'. \tag{3}$$

The score maps are then normalized as soft masks by shifting the scores to the range $[0, 1]$. We compute the '$\mathtt{maskness}$' score defined further below for each of the N soft masks, which serves as a confidence score of each extracted mask. The soft masks are converted to binary masks using a threshold τ. We then sort the binary masks by their maskness scores and remove the redundant masks via mask non-maximum-suppression (NMS). The overall process can be formulated as:

$$\mathbf{M} = \mathtt{NMS}\big(\mathtt{Maskness}(\mathtt{Norm}(\mathbf{Q}' \circledast \mathbf{K}'))\big), \tag{4}$$

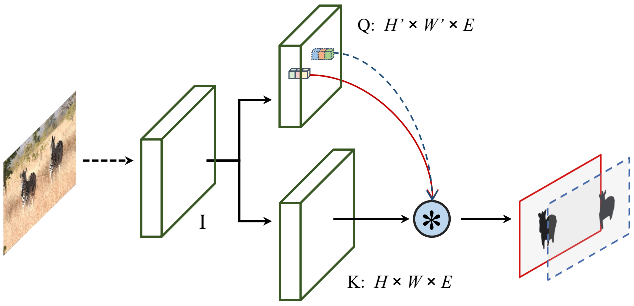

Figure 3. The Free Mask approach. Given queries and keys from the backbone feature $\mathbf{I}$, the keys are convolved by the queries to generate segmentation masks. The masks go through NMS to form the object mask outputs.

where $\mathbf{M}$ denotes the object masks that Free Mask outputs.

Self-supervised pre-training. Free Mask uses a pre-trained backbone via self-supervision as the starting point. We propose to leverage the self-supervised model pre-trained with dense correspondence. Specifically, we find that dense contrastive learning [14] achieves considerably better results with our Free Mask approach, compared to the conventional self-supervised learning by global image-level contrasting. This can be attributed to the similar objective of Free Mask and dense contrastive learning. Here we briefly introduce how the dense contrastive learning is performed. It optimizes a pairwise (dis)similarity loss at the level of local features between two views of the input image. A local feature vector, *i.e.*, a query vector, should be similar to the corresponding positive key in the other view while being dissimilar to other negative keys. Observe that this is also aligned with Equation (2) where the cosine similarity between a query and the keys is evaluated. This also explains why Free Mask extracts reasonable masks. We believe that there could be even better pre-training methods for Free Mask, *e.g.*, those which tackle how to learn fine-grained representations at higher resolutions to generate better masks. We leave this for future research.

Pyramid queries. When constructing the queries $\mathbf{Q}$ from $\mathbf{I}$, we design a pyramid queries method to generate masks for instances at different scales. Specifically, we set a list of scale factors, *e.g.*, $[1.0, 0.5, 0.25]$, when downsampling $\mathbf{I}$, thus leading to a list of $\mathbf{Q}$ at different scales from large to small. All pyramid queries are flattened and concatenated together as the final $\mathbf{Q}$.

Maskness score. A scoring function is required for evaluating the quality of each generated coarse mask, which cannot be learned from annotations. We use the non-parametric maskness method [27], *i.e.*, $\mathtt{maskness} = \frac{1}{N_f} \sum_i^{N_f} \mathbf{p}_i$, to obtain the confidence score of an extracted mask. Here N_f denotes the number of foreground pixels of the soft mask $\mathbf{p}$, *i.e.*, the pixels that have values greater than threshold

τ. Intuitively, this score weighs more heavily on masks that have high confidence on foreground pixels and down weights masks with uncertain foreground pixels.

Unified with SOLO. We can see that the pipeline in Equation (4) is unified with that of SOLO, as introduced in the above background section. They both go through FCN, dynamic convolution, normalization and NMS operations to generate object masks. However, the two are proposed to solve different problems. The latter aims to learn instance segmentation with rich annotated data, while the former is for segmenting objects in unlabeled images. This provides a unifying perspective on segmenting objects in images.

3.3. Self-Supervised SOLO

We aim to train the SOLO-based instance segmenter using the segmentation masks and semantic embeddings, *i.e.*, feature embeddings with high-level semantics, from Free Mask. We separately introduce the methods for learning with coarse masks, self-training, and the semantic representation learning.

Learning with coarse masks. In SOLO, the Dice loss [55] is used to supervise the predicted masks with their ground truth labels. However, this is not ideally suited for our case of learning with noisy masks. As the masks are coarse, directly using them as ground-truth masks can lead to unsatisfactory results. We propose to use the coarse masks as a type of weak annotation and perform weakly supervised instance segmentation with them.

Inspired by the latest weakly-supervised method of BoxInst [11], we project the predicted masks and the coarse masks on to the x-axis and the y-axis via a `max` operation along each axis. The model is supervised to minimize the discrepancy between the projections of predicted masks and the coarse masks. The loss term can be defined as:

$$\begin{aligned}\mathcal{L}_{max\_proj} &= \mathcal{L}(\max_x(\boldsymbol{m}), \max_x(\boldsymbol{m}^*)) \\ &+ \mathcal{L}(\max_y(\boldsymbol{m}), \max_y(\boldsymbol{m}^*)),\end{aligned} \tag{5}$$

where $\mathcal{L}(\cdot,\cdot)$ is the Dice loss, $\boldsymbol{m}$ and $\boldsymbol{m}^*$ are the predicted mask and the coarse mask. $\max_x$ and $\max_y$ denote the `max` operations along each axis.

We further propose to project the predicted and coarse masks onto the x and y axes via an `average` operation along each axis. The motivation is that the `max` operation may emphasize outlier segmentations in coarse masks, while the `average` operation de-emphasize the outliers. In addition, `average` operation preserves solid shape of the object mask, which can benefit the training. The loss term can be written as:

$$\begin{aligned}\mathcal{L}_{avg\_proj} &= \mathcal{L}(\text{avg}_x(\boldsymbol{m}), \text{avg}_x(\boldsymbol{m}^*)) \\ &+ \mathcal{L}(\text{avg}_y(\boldsymbol{m}), \text{avg}_y(\boldsymbol{m}^*)),\end{aligned} \tag{6}$$

where avg_x and avg_y denote the `average` operation along each axis. We also employ a pairwise affinity loss $\mathcal{L}_{pairwise}$ [11] to leverage the prior that the proximal pixels are likely to be in the same class, *i.e.*, foreground or background, if they have similar colors in the raw image.

Overall, the total loss for mask prediction can be formulated as:

$$\mathcal{L}_{mask} = \alpha\mathcal{L}_{avg\_proj} + \mathcal{L}_{max\_proj} + \mathcal{L}_{pairwise}, \tag{7}$$

where α acts as the weight to balance the various loss terms.

Self-training. With our carefully-designed loss function, we are able to train a SOLO-based instance segmenter with the free and noisy coarse masks. As shown in Figure 2, the object masks predicted by the instance segmenter are considerably better than the original coarse masks from Free Mask, which is also validated by the boosted accuracy in Table 7c. As such, we propose to perform self-training with the initially trained instance segmenter to further improve accuracy. We input unlabeled images into the instance segmenter and collect their predicted object masks. The low-confidence predictions are removed and the remaining ones are treated as a new set of coarse masks. We again train an instance segmenter with the unlabeled images and the new masks, using the loss function in Equation (7). Performing self-training once already brings clear improvements and more iterations do not provide additional gains.

Semantic representation learning. General instance segmentation requires not only localizing objects at the pixel level, but also recognizing their semantic categories. In SOLO, the category branch predicts the semantic categories (including background) for each of the objects. In our case without annotations, we propose to decouple the category branch to perform two sub-tasks: foreground/background binary classification, and semantic embedding learning. The former task is trained with the conventional Focal loss [56], termed $\mathcal{L}_{focal}$. For the latter task, we propose a simple approach for learning object-level semantic representations. From Free Mask (introduced in Section 3.2), in addition to the segmentation masks, we can also directly obtain the semantic embedding of the discovered objects. As shown in Figure 3, each mask is associated with a query feature vector $\mathbf{Q}_q \in \mathbb{R}^E$. When training the instance segmenter, we add a branch in parallel to the last layer of the original category branch, which consists of a single convolution layer to predict the semantic embedding of each object. Given the predicted and extracted embeddings $\boldsymbol{q}$ and $\boldsymbol{q}^*$, we train the model by minimizing their negative cosine similarity:

$$L_{sem} = 1 - \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2} \cdot \frac{\boldsymbol{q}^*}{\|\boldsymbol{q}^*\|_2}. \tag{8}$$

The total loss for the category branch can be written as:

$$\mathcal{L}_{cate} = \mathcal{L}_{focal} + \beta\mathcal{L}_{sem}, \tag{9}$$

where β acts as the weight to balance the two terms. Overall, we train the instance segmenter with a combination of $\mathcal{L}_{mask}$ and $\mathcal{L}_{cate}$, corresponding to the losses for the mask branch and category branch, respectively.

4. Experiments

4.1. Experimental Settings

Technical details. For Free Mask, the shorter side of the input image is set to 800 pixels. Threshold τ is set to 0.5. DenseCL [14] with a pre-trained ResNet-50 [54] architecture is adopted as the backbone unless specified. Matrix NMS [27] is used for mask NMS. After NMS, we filter out the low-quality masks with a maskness threshold of 0.7. When training the SOLO model, we initialize the backbone with the pre-trained model used in Free Mask. We set the α and β parameters to 0.1 and 4.0, respectively. We employ the simple copy-paste strategy [57] for data augmentation. During self-training, we set the confidence threshold for removing the low-confidence predictions to 0.3.

Datasets. For FreeSOLO, we use the images in COCO `train2017` and COCO `unlabeled2017` [58] as the set of unlabeled images, containing a total of ~241k images. These unlabeled images are input to Free Mask and are used to train the instance segmenter. The self-supervised backbone in Free Mask is pre-trained on ImageNet with ~1.28 million unlabeled images. We further employ COCO `val2017`, UVO `val` [59], and PASCAL VOC `trainval07` [60] datasets for evaluation.

Evaluation protocol. We evaluate self-supervised instance segmentation with the standard COCO protocol. We report class-agnostic COCO mask average precision (AP) and average recall (AR) on 5k `val2017` split, which is averaged over 10 intersection-over-union (IoU) thresholds evenly-spaced between 0.5 and 0.95. AP considers recall and precision simultaneously, which computes the average precision value for recall values over 0 to 1. AR allows redundant or random detection results, as it computes the maximum recall given a fixed number of detections per image.

To compare with unsupervised object detection methods, we convert the masks to boxes and report the box AP on both the COCO `val2017`, COCO `20k`, and VOC `trainval07`. We further evaluate the pre-trained model by fine-tuning with annotations. Specifically, we fine-tune the instance segmenter on COCO `train2017` and evaluate on COCO `val2017`. We provide two settings, *i.e.*, limited fully annotated images, and limited segmentation masks (see Appendix A.2). Mask AP averaged across all 10 IoU thresholds and all 80 categories is reported.

4.2. Main Results

Self-supervised instance segmentation. For evaluating the self-supervised instance segmenter, we first provide qualitative results to show how FreeSOLO performs at the task of class-agnostic instance segmentation. As shown in Figure 1, without any annotations, FreeSOLO is able to segment object instances of many different categories. To provide a quantitative comparison with previous methods, we report the results of unsupervised class-agnostic instance segmentation in Table 1 and Table 2. As there is no reported result for this new problem, we evaluate a few popular segmentation proposal methods on this benchmark. Among the compared methods, MCG [61] uses the annotated BSDS500 dataset [63] for training a boundary detector, and COB [62] trains its hierarchies and combinatorial grouping on PASCAL Context dataset [64]. By contrast, our FreeSOLO method achieves better results without any annotations. We further compare against the supervised methods trained with full annotations. It is worth noting that FreeSOLO even performs closely to the fully supervised Mask R-CNN [2] trained on the LVIS dataset [65], *e.g.*, 4.8% vs 6.8% AP on the UVO dataset.

method	AP_{50}	AP_{75}	AP	AR_1	AR_{10}	AR_{100}
w/ anns:						
MCG [61]	4.6	0.8	1.6	1.9	7.4	18.2
COB [62]	8.8	1.9	3.3	2.9	10.1	22.7
w/o anns:						
FreeSOLO	9.8	2.9	4.0	4.1	10.5	12.7

Table 1. Class-agnostic instance segmentation on MS COCO `val2017`. Both MCG and COB require annotations more or less.

method	AP_{50}	AP_{75}	AP
w/ full anns:			
SOLOv2 w/ COCO	38.0	20.9	21.4
Mask R-CNN w/ COCO	31.0	14.2	15.9
SOLOv2 w/ LVIS	14.8	5.9	7.1
Mask R-CNN w/ LVIS	18.1	4.1	6.8
w/o anns:			
FreeSOLO	12.7	3.0	4.8

Table 2. Class-agnostic instance segmentation on UVO `val` split. Results of Mask R-CNN are from the paper of UVO [59].

Self-supervised object detection. By converting the masks into boxes, our self-supervised instance segmenter naturally serves as a self-supervised object detector as well. We report the results of class-agnostic object detection on COCO `val2017` benchmark in Table 3. Our method shows significantly superior performance. To compare with existing object discovery methods, we also evaluate FreeSOLO on VOC `trainval07` and COCO `20k` for multi-object discovery. As shown in Table 4, our method largely outperforms the state-of-the-art object discovery methods, including a concurrent work [66]. Its relative improvements are up to 100% on the COCO dataset.

Supervised fine-tuning. In addition to evaluating the self-

method	AP$_{50}$	AP$_{75}$	AP	AR$_1$	AR$_{10}$	AR$_{100}$
UP-DETR [18]	0.0	0.0	0.0	0.0	0.0	0.4
Selective Search [67]	0.5	0.1	0.2	0.2	1.5	10.9
DETReg [68]	3.1	0.6	1.0	0.6	3.6	12.7
FreeSOLO	12.2	4.2	5.5	4.6	11.4	15.3

Table 3. Unsupervised class-agnostic object detection on MS COCO val2017. Compared results are directly from DETReg.

method	VOC			COCO		
	AP$_{50}$	AP$_{75}$	AP	AP$_{50}$	AP$_{75}$	AP
Kim et al. [42]	9.5	-	2.5	3.9	-	1.0
DDT+ [69]	8.7	-	3.0	2.4	-	0.7
rOSD [46]	13.1	-	4.3	5.2	-	1.6
LOD [47]	13.9	-	4.5	6.6	-	2.0
LOST* [66]	19.8	-	6.7	7.9	-	2.5
FreeSOLO	24.5	7.2	10.2	12.4	4.4	5.6

Table 4. Multi-object discovery on PASCAL VOC trainval07 and MS COCO 20k. LOST* is a concurrent work.

	pre-train	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
5% images	sup.	18.0	32.2	17.6	5.5	18.9	27.8
	MoCo-v2 [70]	19.0	32.7	19.2	5.4	19.9	28.9
	DenseCL [14]	20.0	33.7	20.5	5.5	21.5	30.1
	FreeSOLO	22.0	36.0	22.9	6.5	23.2	33.8
10% images	sup.	22.3	38.0	22.9	6.3	24.0	34.8
	MoCo-v2 [70]	23.2	39.0	23.9	6.7	24.6	36.2
	DenseCL [14]	23.7	39.3	24.5	7.3	25.2	37.1
	FreeSOLO	25.6	41.6	26.7	8.3	27.5	40.3

Table 5. Supervised instance segmentation with limited fully annotated images.

	pre-train	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
5% masks	sup.	17.8	36.1	15.9	6.3	19.5	27.4
	MoCo-v2 [70]	17.2	34.9	14.9	5.8	19.0	26.2
	DenseCL [14]	20.1	39.0	18.3	7.6	21.4	31.2
	FreeSOLO	29.9	50.5	30.5	10.7	32.5	46.7
10% masks	sup.	25.4	45.6	25.1	8.8	26.9	40.7
	MoCo-v2 [70]	25.6	45.1	25.5	8.7	27.2	40.4
	DenseCL [14]	26.1	45.2	26.3	9.1	28.0	40.8
	FreeSOLO	31.1	51.4	32.0	11.2	34.1	48.4

Table 6. Supervised instance segmentation with limited segmentation masks.

supervised instance segmenter directly, we also evaluate the performance of our approach in a supervised setting by fine-tuning the self-supervised instance segmenter with annotations. As shown in Table 5, FreeSOLO pre-training outperforms ImageNet supervised pre-training by 4.0% AP when using 5% COCO training images. The gains over the state-of-the-art self-supervised pre-training methods are also clear, *e.g.*, 2.0% AP better than DenseCL [14].

To further compare the pre-training methods with different amount of mask annotations, in Table 6, we conduct fine-tuning experiments with only limited masks available. When fine-tuning with 5% masks, FreeSOLO achieves significant gains of 9.8% AP over supervised pre-training. These fine-tuning experiments demonstrate that FreeSOLO serves as a strong instance segmentation pre-training method, outperforming both the supervised and state-of-the-art self-supervised pre-training methods.

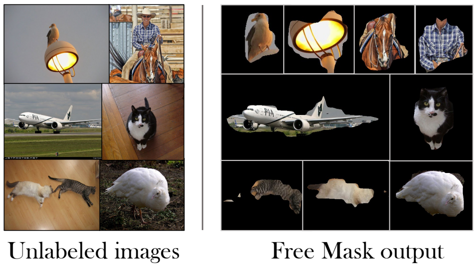

Figure 4. Qualitative results of the Free Mask. Free Mask extracts coarse masks of the common objects in unlabeled images.

4.3. Ablation Study

We conduct ablation experiments to show how each component contributes to FreeSOLO. The ablation studies are performed on the COCO val2017 split.

Free Mask with different pre-trained backbones. In Table 7a, we show how Free Mask performs with different pre-trained backbones. The conventional self-supervised learning methods that contrast the global representations of image pairs, *e.g.*, SimCLR and MoCo-v2, show worse results compared to supervised ImageNet pre-training. The self-supervised learning methods that consider dense correspondence, *e.g.*, EsViT and DenesCL, yield better results than those that do not. DenseCL shows the best results compared to both supervised and other self-supervised methods. This aligns with our hypothesis in Section 3.2 that DenseCL's objective is consistent with Free Mask's. We provide some visualizations of Free Mask in Figure 4.

Pyramid queries. We compare different scales of the queries $\mathbf{Q}$ used in Free Mask in Table 7b. A smaller scale is better for large objects but worse for medium and small objects. A large scale is just the opposite. Pyramid queries with scales $[1.0, 0.5, 0.25]$ yield the best results.

Loss functions. In Table 7d, we compare our weakly-supervised design against the full mask supervision, *i.e.*, the original Dice loss used in SOLO computed with the full masks. Directly using the coarse masks to provide full supervision to the instance segmenter leads to unsatisfactory results. Our weakly-supervised loss outperforms the original full mask loss by a large margin. In Table 7e, we study the mask loss terms in Equation (7). The performance drops sharply when learning without $\mathcal{L}_{avg\_proj}$, *i.e.*, with only the projection loss from max operation and pairwise loss as in [11]. The model even collapses to only segmenting the contours when trained longer (Figure 5). Our method tack-

pre-train	AR	AR$_S$	AR$_M$	AR$_L$
sup.	7.8	0.1	11.3	16.4
SimCLR [16]	6.1	1.0	12.1	6.7
MoCo-v2 [70]	4.7	1.6	8.1	5.4
DINO [71]	3.2	2.8	5.2	0.9
EsViT [72]	6.3	0.0	6.0	17.8
DenseCL [14]	11.5	0.1	6.0	39.5

(a) **Different pre-training methods** with Free Mask. DenseCL works the best.

scale	AR	AR$_S$	AR$_M$	AR$_L$
0.25	10.1	0.0	1.9	39.5
1.0	11.3	0.1	6.0	38.6
pyramid	11.5	0.1	6.0	39.5

(b) **Pyramid queries** in Free Mask. Pyramid queries improve over single scale queries.

iters	AP$_{50}$	AP$_{75}$	AP
-1	2.3	0.2	0.7
0	7.9	2.5	3.3
1	8.3	2.8	3.7
2	7.7	2.9	3.5

(c) **Self-training iterations.** '-1' refers to coarse masks. '0' means learning without self-training.

mask loss	AP$_{50}$	AP$_{75}$	AP
full	6.2	1.6	2.4
weak	7.9	2.5	3.3

(d) **Full vs. weak supervision.** Weakly-supervised design is effective.

mask loss	AP$_{50}$	AP$_{75}$	AP
combination	7.9	2.5	3.3
- w/o $\mathcal{L}_{avg\_proj}$	3.8	1.6	2.0
- w/o $\mathcal{L}_{max\_proj}$	7.1	1.6	2.6
- w/o $\mathcal{L}_{pairwise}$	6.1	0.9	2.1

(e) **Mask loss terms.** Each loss component contributes to the final results.

$\mathcal{L}_{sem}$?	AP	AP$_{50}$	AP$_{75}$
	24.9	40.5	26.1
✓	25.6	41.6	26.7

(f) **Semantic embedding.** Semantic embedding learning improves the fine-tuning results.

Table 7. FreeSOLO ablation experiments. All the experiments are with a ResNet-50 backbone. We report class-agnostic instance segmentation results (a-e) and supervised fine-tuning results (f) on the COCO `val2017` split.

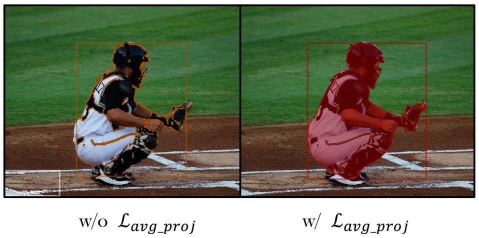
w/o $\mathcal{L}_{avg\_proj}$ w/ $\mathcal{L}_{avg\_proj}$

Figure 5. Qualitative comparison of with and without $\mathcal{L}_{avg\_proj}$ when learning from coarse masks. The model trained without $\mathcal{L}_{avg\_proj}$ tends to only segment the contours when trained longer.

les this problem by leveraging the projection from `average` operation, which not only preserves the shape but is also less sensitive to outlier pixels.

Self-training. Our method performs self-training by selecting high-confidence predictions of the self-supervised instance segmenter and training the instance segmenter again with them. We compare the results of performing different iterations of self-training in Table 7c. '-1' refers to the initial coarse masks. Zero iteration refers to learning from the coarse masks without self-training. We show that performing self-training once already brings clear improvements, but additional iterations do not provide additional gains.

Semantic embedding. To validate the effectiveness of the semantic embedding learning, in Table 7f we compare the models trained with or without the semantic embedding loss defined in Equation (8). The models are fine-tuned with 10% of fully annotated COCO images. It shows that the semantic embedding loss yields clear improvements when fine-tuning instance segmentation with annotations.

5. Discussion and Conclusion

In this work, we have developed a simple and effective self-supervised instance segmentation framework FreeSOLO. FreeSOLO enables learning to segment objects without any annotations, neither pixel-level nor image-level labels. We hope that its novel design elements provide insights for future works on unsupervised visual learning, *e.g.*, unsupervised panoptic segmentation, and beyond.

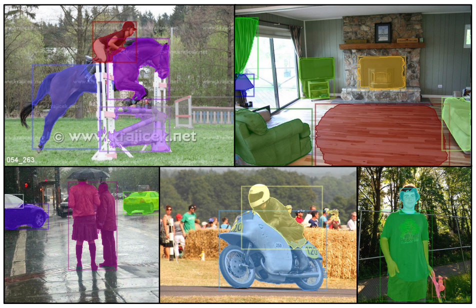

Figure 6. Failure cases of FreeSOLO. Our method could fail to localize objects that are truncated, crowded or small.

Limitations. Without category labels, our self-supervised instance segmenter cannot predict the categories of the detected objects, but generate class-agnostic object masks. There is still a large gap between our self-supervised model and the supervised one trained with rich annotations. Our method could fail in some scenarios (Figure 6). We believe there is plenty of room to improve based on our method.

Broader impacts. This work shows that one can learn a class-agnostic instance segmenter without any annotations. In the future, there is a chance for self-supervised segmenter to reach or even outperform the supervised model trained with manual annotations, which may eliminate the need for annotating masks or boxes for common objects. We expect that the proposed technique can be used to largely reduce data annotation effort for a few instance-level recognition tasks in computer vision.

References

[1] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Int. Conf. Comput. Vis.*, 2017. 1, 2, 6

[3] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv:1708.02551*, 2017. 1, 2

[4] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Int. Conf. Comput. Vis.*, 2019. 1, 2

[6] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2

[7] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. SOLO: A simple framework for instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 1, 2, 3

[8] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 2

[9] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *Adv. Neural Inform. Process. Syst.*, 2019. 1, 2

[10] Yun Liu, Yu-Huan Wu, Peisong Wen, Yujun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 1

[11] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. BoxInst: High-performance instance segmentation with box annotations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 5, 7

[12] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. *arXiv preprint arXiv:2104.06404*, 2021. 1

[13] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Int. Conf. Comput. Vis.*, 2021. 1, 2

[14] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 3, 4, 6, 7, 8

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. Mach. Learn.*, 2020. 2, 3, 8

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Adv. Neural Inform. Process. Syst.*, 2020. 2, 3

[18] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 7

[19] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aäron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Int. Conf. Comput. Vis.*, 2021. 2, 3

[20] Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron C. Courville. Unsupervised learning of dense visual representations. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[21] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[22] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[23] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Adv. Neural Inform. Process. Syst.*, 2017. 2

[24] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sequential grouping networks for instance segmentation. In *Int. Conf. Comput. Vis.*, 2017. 2

[25] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. In *Int. Conf. Comput. Vis.*, 2019. 2

[26] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-down meets bottom-up for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[27] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting objects by locations. In *Eur. Conf. Comput. Vis.*, 2020. 2, 4, 6

[28] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Adv. Neural Inform. Process. Syst.*, 2020. 2

[29] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[30] Richard Zhang, Phillip Isola, and Alexei Efros. Colorful image colorization. In *Eur. Conf. Comput. Vis.*, 2016. 2

[31] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[32] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Eur. Conf. Comput. Vis.*, 2016. 3

[33] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Int. Conf. Learn. Represent.*, 2018. 3

[34] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[35] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[36] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[37] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[38] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Int. Conf. Comput. Vis.*, 2021. 3

[39] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Int. Conf. Comput. Vis.*, 2021. 3

[40] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering object categories in image collections. In *Int. Conf. Comput. Vis.*, 2005. 3

[41] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2006. 3

[42] Gunhee Kim and Antonio Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *Adv. Neural Inform. Process. Syst.*, 2009. 3, 7

[43] Alon Faktor and Michal Irani. "clustering by composition" - unsupervised discovery of image categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014. 3

[44] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 3

[45] Huy V. Vo, Francis R. Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[46] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *Eur. Conf. Comput. Vis.*, 2020. 3, 7

[47] Huy V. Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. *arXiv: Comp. Res. Repository*, 2021. 3, 7

[48] Armand Joulin, Francis R. Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2010. 3

[49] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *IJCAI*, 2018. 3

[50] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 3

[51] Xu Ji, Andrea Vedaldi, and João F. Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *Int. Conf. Comput. Vis.*, 2019. 3

[52] Jyh-Jing Hwang, Stella X. Yu, Jianbo Shi, Maxwell D. Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Int. Conf. Comput. Vis.*, 2019. 3

[53] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *arXiv: Comp. Res. Repository*, 2021. 3

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 4, 6

[55] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proc. Int. Conf. 3D Vision*, 2016. 5

[56] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, 2017. 5

[57] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 6

[58] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. 6

[59] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. *arXiv: Comp. Res. Repository*, 2021. 6

[60] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.*, 2010. 6

[61] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 6, 12

[62] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Convolutional oriented boundaries: From image segmentation to high-level tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 6, 12

[63] David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Int. Conf. Comput. Vis.*, 2001. 6

[64] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 6

[65] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 6

[66] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv: Comp. Res. Repository*, 2021. 6, 7

[67] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *Int. J. Comput. Vis.*, 2013. 7

[68] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. *arXiv: Comp. Res. Repository*, 2021. 7

[69] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 2019. 7

[70] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv: Comp. Res. Repository*, 2020. 7, 8

[71] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, 2021. 8

[72] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv: Comp. Res. Repository*, 2021. 8

[73] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 12