Exemplar-FreeSOLO: Enhancing Unsupervised Instance Segmentation with Exemplars

Taoseef Ishtiak,[1] Qing En,[1] Yuhong Guo[1,2]
[1]Carleton University, Ottawa, Canada [2]CIFAR AI Chair, Amii, Canada
taoseefishtiak@cmail.carleton.ca, qingen@cunet.carleton.ca, yuhong.guo@carleton.ca

Abstract

Instance segmentation seeks to identify and segment each object from images, which often relies on a large number of dense annotations for model training. To alleviate this burden, unsupervised instance segmentation methods have been developed to train class-agnostic instance segmentation models without any annotation. In this paper, we propose a novel unsupervised instance segmentation approach, Exemplar-FreeSOLO, to enhance unsupervised instance segmentation by exploiting a limited number of unannotated and unsegmented exemplars. The proposed framework offers a new perspective on directly perceiving top-down information without annotations. Specifically, Exemplar-FreeSOLO introduces a novel exemplar-knowledge abstraction module to acquire beneficial top-down guidance knowledge for instances using unsupervised exemplar object extraction. Moreover, a new exemplar embedding contrastive module is designed to enhance the discriminative capability of the segmentation model by exploiting the contrastive exemplar-based guidance knowledge in the embedding space. To evaluate the proposed Exemplar-FreeSOLO, we conduct comprehensive experiments and perform in-depth analyses on three image instance segmentation datasets. The experimental results demonstrate that the proposed approach is effective and outperforms the state-of-the-art methods.

1. Introduction

Instance segmentation is among the most fundamental and challenging tasks in computer vision, aiming to recognize and segment each object in an image. By utilizing a significant amount of densely annotated data to train segmentation models, existing techniques have achieved desirable results [4, 5, 10, 16, 26, 41, 47]. However, acquiring numerous pixel-level labels requires substantial labour and financial resources, limiting the developments and practical applications in the field. To reduce the costly annotation requirement, some solutions have been put forth to investigate ways of using less expensive training labels to complete complex tasks, such as weakly-supervised [18, 22, 29, 42], partially-supervised [20, 51] and semi-supervised instance segmentation [3, 52, 55].

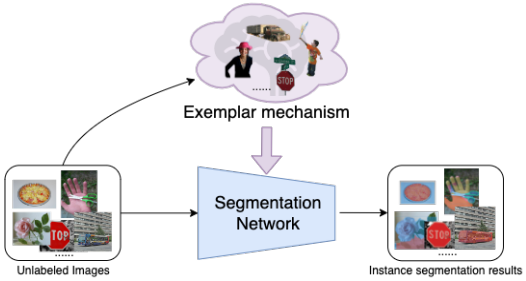

Figure 1. An illustration of the proposed idea. The proposed Exemplar-FreeSOLO framework addresses the unsupervised instance segmentation problem by excavating information from unlabeled data through an exemplar mechanism, which produces top-down knowledge guidance and enhances the discriminability of the segmentation model.

Although some significant advances have been achieved, the labelling conundrum obstacle remains as these methods still require nontrivial dense labels and precise position information. By contrast, unsupervised instance segmentation methods benefit from not requiring any annotated data; they can directly exploit many existing unannotated images while being able to continuously upgrade the effectiveness of the segmentation models with incoming data. Therefore it is important to investigate unsupervised instance segmentation, which enables learning class-agnostic instance segmentation models without any data annotation. Recently, an unsupervised instance segmentation framework, FreeSOLO [48], has been proposed to extract coarse object masks as pseudo-labels and train an instance segmentation model using a self-supervised method. Although FreeSOLO attempts to improve the quality of pseudo labels and prediction masks, it can hardly overcome the detrimental effects of the considerable noise in pseudo labels without any guid-

ance information in the training process.

The overall fundamental challenges for unsupervised instance segmentation lie in the following two aspects: (1) Unsupervised segmentation models are heavily influenced by the noisy pseudo-labels. When the objects of an image are part of the background relative to the features of interest, models are prone to generating a large number of false positive regions. (2) The unsupervised nature makes it difficult to learn discriminative information. Constructing comparison relations directly between different instances for the same type of target tends to lead to fragmentation problems. Meanwhile, as suggested by the generalized context model [36], humans can capture different categories of information through exemplars in their memory. Inspired by this idea, we aim to overcome the unsupervised instance segmentation challenges by developing new segmentation models to integrate an exemplar learning mechanism, which is desirable from both biological and practical perspectives.

In this paper, we propose a novel approach, *Exemplar-FreeSOLO*, for performing instance segmentation without any annotation. The core of the framework is an exemplar mechanism that aims to extract and utilize pertinent information from unlabeled data in order to obtain useful knowledge that can guide model training, as shown in Figure 1. Exemplar-FreeSOLO obtains beneficial top-down guidance for objects through exemplar knowledge extraction, while consequently enhancing the discriminability of instance segmentation models by exploiting the exemplar guidance information in a contrastive manner. Specifically, given randomly selected exemplar images, we design an exemplar knowledge abstraction module (EKA) to acquire top-down guidance knowledge for objects. The exemplar images are roughly cropped and fed into an unsupervised model to extract masked-out images for constructing a pool of exemplar objects, which are then used to produce the exemplar guidance knowledge. Next, an exemplar embedding contrastive module (EEC) is devised to capture homogeneous components of the same type of instances through a contrastive learning paradigm. This is achieved by considering similarities between the embeddings of unlabeled images and the exemplar embeddings and constructing contrastive relationships among them. This module is expected to enhance the discriminability of the instance segmentation model. Finally, we incorporate the two modules into the FreeSOLO framework to effectively train instance segmentation models. The main contributions of our paper are summarized as follows:

- We propose a novel Exemplar-FreeSOLO approach to tackle the unsupervised instance segmentation problem by leveraging useful information from the unlabelled data through an exemplar mechanism.
- We design an exemplar knowledge abstraction module to acquire beneficial top-down guidance knowledge by extracting exemplar objects in an unsupervised way.
- We devise an exemplar embedding contrastive module to enhance the discriminative capability of instance segmentation models by exploiting contrasting exemplar guidance knowledge in the embedding space.
- Experimental results on three datasets show that the proposed Exemplar-FreeSOLO can substantially outperform the state-of-the-art unsupervised instance segmentation and object detection methods.

2. Related Work

Instance Segmentation Instance segmentation is an important task in computer vision. Methods for instance segmentation can be divided into two categories: two-stage approaches and one-stage approaches. Mask-RCNN [16] is a representative two-stage instance segmentation approach. This approach first creates candidate ROIs, which are then segmented in the second stage. Other approaches like FPN [28] try to improve the performance of the two-stage models by addressing the incompatibility issue between a mask's confidence score and localization accuracy. By contrast, one-stage approaches map the final masks with position-sensitive pooling [8, 26]. SOLO [47] is a one-stage approach that addresses the trade-off between the domain's speed and accuracy. Based on the SOLO model, the recent work of FreeSOLO [48] has been by far the first and only work that proposes to perform unsupervised instance segmentation without any labels.

Cluster-based and Self-supervised Approaches HAIS is a clustering-based instance segmentation framework [6] that uses the spatial relationships between points and point sets. The hierarchical aggregation presented in this method generates instance proposals progressively. PointGroup [21] is a bottom-up segmentation network that focuses on better point-grouping. The network predicts semantic labels and offsets, and the clustering component is deployed for proper utilization of both the original and offset-shifted coordinates. Targeting real-time computation for autonomous vehicles, the work in [34] proposes a clustering-based loss function to achieve proposal-free instance segmentation. In [19], the graph colouring theorem is combined with FCN to show that deep FCN can cluster image pixels in an end-to-end manner. Self-supervised learning is another approach that has gained much attention in computer vision. Methods like MoCo [15], SimCLR [7], jigsaw puzzles [35], colorization [54], orientation discrimination [13], and inpainting [37] are pioneer works of self-supervised learning. Some self-supervised learning strategies, such as contrastive learning, have been used in fully supervised segmentation models, yielding good empirical results [51]. An

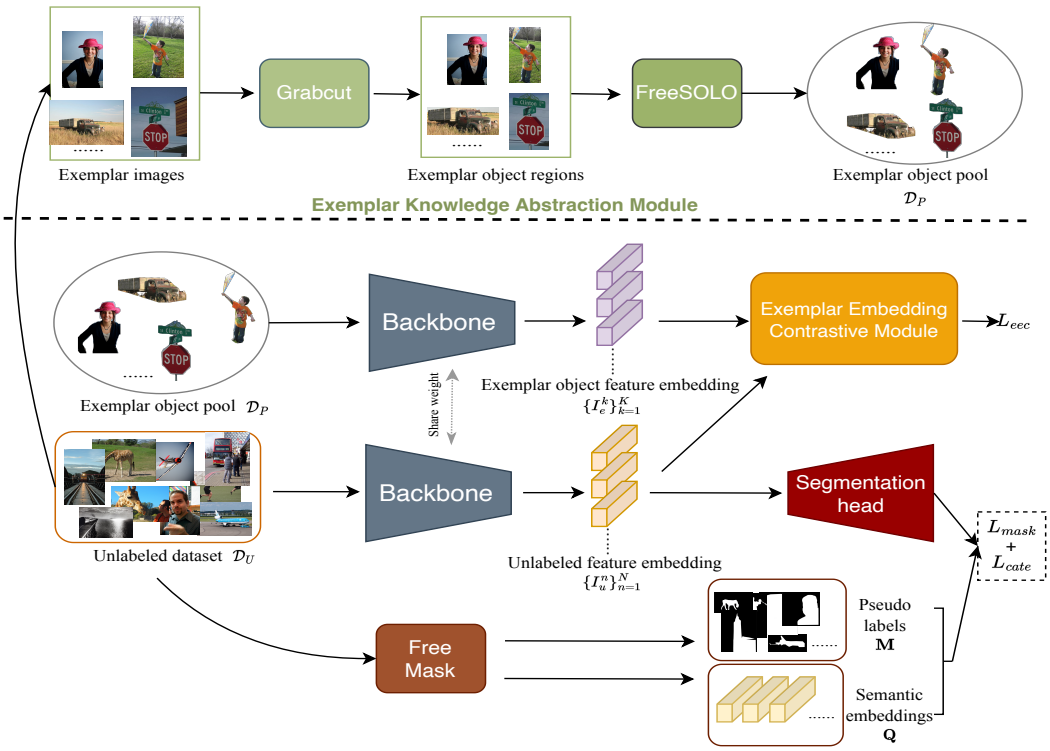

Figure 2. An overview of the proposed Exemplar-FreeSOLO. We first select exemplar images from the unlabeled dataset and build the exemplar object pool $\mathcal{D}_P$ through the exemplar knowledge abstraction module. Next, we extract the feature embeddings $\{I_u^n\}_{n=1}^N$ of the unlabeled images and the feature embedding pool $\mathcal{I}_P = \{I_e^k\}_{k=1}^K$ of the exemplar objects in $\mathcal{D}_P$ by using the backbone. Finally, we separately input $\{I_u^n\}_{n=1}^N$ into a segmentation head to obtain the instance segmentation results and input both $\{I_u^n\}_{n=1}^N$ and $\mathcal{I}_P = \{I_e^k\}_{k=1}^K$ into an exemplar embedding contrastive module to compute the exemplar embedding contrastive loss.

exemplar-based approach [30] has been proposed to address the object detection problem, but it requires supervised information to identify positive and negative samples, whereas our approach is unsupervised.

Weakly and Partially Supervised Instance Segmentation Obtaining pixel-level annotation in images is challenging and induces tremendous costs. To reduce this annotation burden, some recent works have used weak annotations or incomplete annotations for instance segmentation. For example, some weakly supervised methods employ box-level annotations [18, 25, 42] or image-level labels [12, 39] to perform instance segmentation. However, as these methods do not use pixel-level annotations, the results obtained in such works are much less satisfactory than the ones from a fully-supervised setup. By contrast, some other works adopt a partial-supervised setup [20, 24, 55], where a small number of categories are pixel annotated, and the rest of the categories have only box-level annotations. Such partial-supervised methods rely on the power of semi-supervised learning to produce good instance segmentation models without comprehensive image annotations. Different from the above-mentioned efforts to reduce the annotation cost, our proposed approach aims to learn a good instance segmentation model from unlabeled images without any annotations.

3. Method

In this section, we present the proposed Exemplar-FreeSOLO for unsupervised instance segmentation. We first briefly introduce the unsupervised instance segmentation framework FreeSOLO [48] in Sec. 3.1. Next, we depict the architecture of the Exemplar-FreeSOLO in Sec. 3.2. The proposed exemplar knowledge abstraction module (EKA) and exemplar embedding contrastive module (EEC) are then presented in Sec. 3.3 and Sec. 3.4, respectively. Finally, we present the overall loss function used to train the proposed Exemplar-FreeSOLO in Sec. 3.5.

3.1. Revisiting FreeSOLO

Built on top of the SOLO architecture [47], FreeSOLO [48] has achieved successful instance segmentation without any image annotations. The main idea is to train an instance segmentation model (*i.e.* SOLO) by generating

coarse masks $\mathbf{M}$ and semantic embeddings $\mathbf{Q}$ in an unsupervised manner. Specifically, the input for FreeSOLO is a set of N unlabeled images, $\mathcal{D}_U = \{(X_u^n)\}_{n=1}^N$, where each image is represented as $X_u^n \in \mathbb{R}^{H*W*3}$, with H and W denoting the height and width of the image, respectively. For each input image X_u^n, its feature embedding $I_u^n \in \mathbb{R}^{h\times w\times c}$ can be extracted from a self-supervised backbone [50], *e.g.*, ResNet [17]. I_u^n is then bilinearly downsampled to obtain queries $\mathbf{Q}_u^n \in \mathbb{R}^{h'\times w'\times c}$, while taking itself as keys $\mathbf{K}_u^n$. Next, both of them go through ℓ_2 normalization to generate new queries $\bar{\mathbf{Q}}_u^n$ and new keys $\bar{\mathbf{K}}_u^n$. Each query in $\bar{\mathbf{Q}}_u^n$ can be treated as a 1×1 convolutional kernel to perform a convolution operation on the keys in $\bar{\mathbf{K}}_u^n$ and generate score maps $\mathbf{S}_u^n = \bar{\mathbf{Q}}_u^n \circledast \bar{\mathbf{K}}_u^n$, which are then used to produce the coarse mask $\mathbf{M}_u^n$ as follows:

$$\mathbf{M}_u^n = \mathrm{NMS}(\mathrm{Maskness}(\mathrm{Norm}(\mathbf{S}_u^n))), \tag{1}$$

where $\circledast$ denotes the convolution operation; $\mathrm{Norm}(\cdot)$ is a normalization function that shifts the scores to the range of [0,1]; $\mathrm{Maskness}(\cdot)$ indicates a confidence score function [47]; and $\mathrm{NMS}(\cdot)$ represents the mask non-maximum-suppression operation. The coarse masks $\{\mathbf{M}_u^n\}_{n=1}^N$ and the query feature vectors $\{\mathbf{Q}_u^n\}_{n=1}^N$ from all the unlabeled images are then used as initial pseudo-labels and semantic embeddings to train a SOLO-based instance segmentation model [47] via self-training.

3.2. Overview of Exemplar-FreeSOLO

Fully unsupervised instance segmentation, nevertheless, is substantially more challenging than its supervised counterpart. Although FreeSOLO [48] has achieved surprising results, it still struggles to overcome the negative impact of the considerable pseudo-label noise. Moreover, it also does not optimize the potential of exploiting the discriminative contrastive information that naturally exists in unlabeled data. In view of these drawbacks, we propose a novel unsupervised instance segmentation method dubbed as Exemplar-FreeSOLO to enhance unsupervised instance segmentation through exemplar knowledge extraction and contrastive embedding learning.

The overall architecture of the Exemplar-FreeSOLO is illustrated in Figure 2, which is built on top of the FreeSOLO architecture with two extra modules, an exemplar knowledge abstraction (EKA) module and an exemplar embedding contrastive (EEC) module. The base instance segmentation network consists of an embedding backbone, a segmentation head and the proposed EEC module, where the first two parts are the same as the segmentation model in FreeSOLO. As a result, the embedding features extracted by the embedding backbone are fed into two branches, the segmentation head and the EEC module. In particular, given the set of unlabeled images $\mathcal{D}_U$, Exemplar-FreeSOLO aims to train a good instance segmentation network $\mathcal{N}_s$ without any annotation by enhancing the self-training process, starting with the initial pseudo-labels generated by the Free Mask of FreeSOLO. First, the EKA module is used to build an exemplar object pool $\mathcal{D}_P$ from randomly selected images from the unlabeled dataset $\mathcal{D}_U$, which can provide useful top-down guidance for embedding learning and the unsupervised segmentation model training process. Next, the EEC module is used to construct contrastive relationships between the exemplar embeddings and the unlabeled image embeddings and boost the discriminative capability of the instance segmentation network with an additional contrastive embedding loss during the self-training process. We elaborate on these modules of Exemplar-FreeSOLO and the training loss below.

3.3. Exemplar Knowledge Abstraction Module

We randomly select exemplar images from the unlabeled training set and devise an exemplar knowledge abstraction module (EKA) to extract exemplar objects from the selected exemplar images and build an exemplar pool of objects, which are later leveraged as pivots for self-supervised contrastive embedding learning.

For each selected exemplar image X_u, we employ an unsupervised object segmentation method (*e.g.*, Grabcut [38]) to extract the coarse foreground mask, and then crop out the object region from the image based on the surrounding boundaries of the mask. In this way, an exemplar object with a high probability of appearing in the region can be identified. Nevertheless, the coarse foreground mask may have a higher recall but a lower precision. We, therefore, further deploy a trained FreeSOLO to extract a refined object mask from the exemplar object region and apply the mask to the exemplar image to obtain an exemplar object X_e. Using this procedure, we can extract a set of K exemplar objects and create an exemplar object pool $\mathcal{D}_P$:

$$\mathcal{D}_P = \{X_e^k\}_{k=1}^K. \tag{2}$$

This exemplar object pool can be treated as representatives or pivots for the corresponding hidden categories of the extracted objects. This EKA module is specifically designed for unsupervised instance segmentation, and the exemplar object pool will be exploited as top-down guidance knowledge during the self-training process of the segmentation network through the subsequent EEC module.

3.4. Exemplar Embedding Contrastive Module

In an unsupervised scenario, selecting positive and negative samples to construct discriminative contrastive relationships is challenging as there is no explicit labelling information. Leveraging the exemplar object pool produced by the EKA module, we propose an exemplar embedding contrastive (EEC) module that exploits the exemplar objects as guidance knowledge to construct contrastive embedding

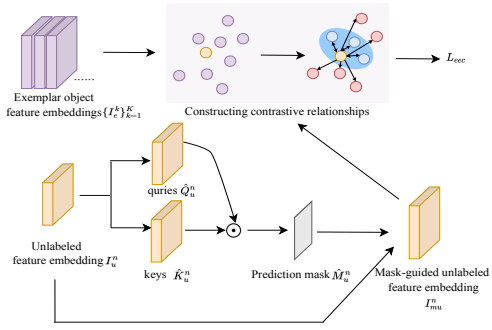

Figure 3. The exemplar embedding contrastive module

losses for the unlabeled images, aiming to improve image embedding and enhance the discriminative capacity of the instance segmentation network.

The structure of the proposed EEC module is shown in Figure 3. This module takes the feature embeddings $\{I_u^n\}_{n=1}^N$ of the unlabeled images and the feature embedding pool $\mathcal{I}_P = \{I_e^k\}_{k=1}^K$ of the exemplar objects in $\mathcal{D}_P$ as inputs. For each unlabeled image, we further compute a mask-guided feature embedding I_{mu}^n for it as follows. First, we apply ℓ_2 normalization and up-sampling on I_u^n to generate queries $\hat{\mathbf{Q}}_u^n$. Similarly, we generate keys $\hat{\mathbf{K}}_u^n$ by applying ℓ_2 normalization, up-sampling, and an extra 1×1 convolution operation on I_u^n. Next, following FreeSOLO, a prediction mask $\hat{M}_u^n$ can be produced by applying the series of operations in Eq.(1) to the element-wise multiplication of $\hat{\mathbf{Q}}_u^n$ and $\hat{\mathbf{K}}_u^n$. Finally, we produce the mask-guided feature embedding $I_{mu}^n \in \mathbb{R}^{h\times w\times c}$ by multiplying the prediction mask $\hat{M}_u^n$ with a transformed I_u^n.

Given the mask-guided feature embeddings, we construct a contrastive loss for each unlabeled input image by choosing positive and negative samples for it from the exemplar embedding pool $\mathcal{I}_P$ based on embedding similarities. For I_{mu}^n, we randomly select an exemplar embedding whose similarity with I_{mu}^n is larger than a threshold α as a positive sample, such as

$$I_{e,pos}^n \in \{I_e^k \in \mathcal{I}_P : \text{sim}(I_{mu}^n, I_e^k) > \alpha\}, \tag{3}$$

while using all the exemplars whose similarity with I_{mu}^n is smaller than a threshold β as negative samples:

$$\mathcal{I}_{e,neg}^n = \{I_e^k \in \mathcal{I}_P : \text{sim}(I_{mu}^n, I_e^k) < \beta\}. \tag{4}$$

In particular, we use the cosine similarity in our implementation. Consequently, the exemplar embedding contrastive loss is defined as follows:

$$L_{eec} = -\sum_n \log \frac{pos(n)}{pos(n) + neg(n)}, \tag{5}$$

where

$$pos(n) = \exp(\langle I_{mu}^n, I_{e,pos}^n \rangle / \tau) \tag{6}$$

$$neg(n) = \frac{1}{|\mathcal{I}_{e,neg}^n|} \sum\nolimits_{I_e \in \mathcal{I}_{e,neg}^n} \exp(\langle I_{mu}^n, I_e \rangle / \tau). \tag{7}$$

Here $\tau > 0$ is a hyperparameter, $|\cdot|$ denotes the size of the given set, and $\langle \cdot, \cdot \rangle$ denotes the inner product. If no positive or negative samples can be found for an unlabeled image, its contrastive loss will be set to zero.

This exemplar embedding contrastive module (EEC) is designed to possess the following advantages. First, by calculating similarities between the mask-guided unlabeled feature embeddings and the exemplar object embeddings, we can determine positive and negative samples in an unsupervised manner. Second, with the proposed exemplar embedding contrastive loss, by maximizing the similarities between the positive pairs in contrast to the negative pairs, we can enforce discriminative information learning in the embedding space by using diverse exemplar objects as pivots. As such, EEC can enhance the discriminative capability of unsupervised instance segmentation models.

3.5. Loss Function

The overall loss function for the proposed Exemplar-FreeSOLO contains three terms: a mask segmentation loss L_{mask}, a category loss L_{cate}, and an exemplar embedding contrastive loss L_{eec}:

$$L_{total} = L_{mask} + L_{cate} + \lambda_{eec} L_{eec}, \tag{8}$$

where λ_{eec} is a trade-off hyperparameter. The first two terms of Eq.(8) are derived from FreeSOLO together with its generated coarse masks and semantic embeddings (Sec. 3.1), and the last term is derived from the proposed modules (Sec. 3.3 and 3.4). In particular, L_{mask} is defined to constrain the predicted mask of the segmentation head:

$$\begin{aligned} L_{\text{mask}} =& \gamma L_{avg\_proj}(\mathbf{M}^*, \mathbf{M}) + \\ & L_{max\_proj}(\mathbf{M}^*, \mathbf{M}) + L_{pairwise}(\mathbf{M}^*), \end{aligned} \tag{9}$$

where $\mathbf{M}^*$ and $\mathbf{M}$ are the predicted mask from the segmentation head and the coarse mask generated from the Free Mask of FreeSOLO [48]; γ is a trade-off hyperparameter; L_{avg_proj} and L_{max_proj} are average projection loss [42] and max projection loss [48]; and $L_{pairwise}$ is a pairwise affinity loss [42]. Besides, L_{cate} is defined for foreground/background binary classification and semantic embedding learning:

$$L_{cate} = L_{focal}(\mathbf{M}^*, \mathbf{M}) + \mu L_{sem}(\mathbf{Q}^*, \mathbf{Q}), \tag{10}$$

where L_{focal} is the focal loss and L_{sem} is a negative cosine similarity function; μ is a trade-off hyperparameter; and $\mathbf{Q}^*$ and $\mathbf{Q}$ are the predicted embeddings from the segmentation head and the embeddings generated together with pseudo labels from the Free Mask of FreeSOLO [48].

Model	AP_{50}	AP_{75}	AP	AR_1	AR_{10}	AR_{100}
w/anns:						
MCG [1]	4.6	0.8	1.6	1.9	7.4	18.2
COB [31]	8.8	1.9	3.3	2.9	10.1	22.7
w/o anns:						
FreeSOLO [48]	9.8	2.9	4.0	4.1	10.5	12.7
Exemplar-FreeSOLO	13.2	6.3	8.4	7.3	15.8	15.5

Table 1. Class-agnostic instance segmentation results on MS COCO val2017. "w/anns" indicates the results obtained in a supervised scenario. "w/o anns" indicates the results obtained in an unsupervised scenario.

4. Experiments

4.1. Experimental Setting

Implementation details Following FreeSOLO, we set the input image size to 800 pixels and the object confidence threshold to 0.5 in the Free Mask approach. Besides, we employ the same network structure as FreeSOLO for the instance segmentation network $\mathcal{N}_s$. In particular, we adopt a ResNet-50-based DenceCL [50] model as our backbone, which is trained on the ImageNet with 1.28 million unlabeled images [48]. We use the FastNMS [4] and a mask confidence threshold of 0.7 to filter out the low-quality masks. For the loss terms in Eq.(8), Eq.(9) and Eq.(10), we set the values of the hyperparameters (λ_{eec}, γ and μ), to 1.3, 0.1 and 0.4, respectively. We set τ to 0.02, α to 0.8, and β to 0.3. During training, the batch size is set to 8, and the learning rate is set to 0.001 with SGD.

Datasets and evaluation metrics The Exemplar-FreeSOLO is trained on the MS COCO unlabeled2017 and train2017 datasets [27], and is tested on MS COCO val2017, UVO val set [46], and PASCAL VOC trainval07 [11]. We use average precision (AP) and average recall (AR) as performance assessment metrics. The AP scores are averaged over 10 results by varying the IoU threshold from 0.5 to 0.95. AP_{50} and AP_{75} indicate the AP scores by fixing the IoU threshold as 0.5 and 0.75, respectively. AP_s, AP_m and AP_l are reported for small, medium and large objects with areas less than 64^2, within [64^2, 192^2] and greater than 192^2, respectively, using an IoU threshold of 0.5. Similarly, AR_1, AR_{10} and AR_{100} are the recall values computed with different numbers of fixed detections (*i.e.*, 1, 10, 100) for each image [48]. As a byproduct of the masks, Exemplar-FreeSOLO also generates bounding boxes to address the unsupervised object detection tasks on COCO val2017, COCO 20k, and VOC trainval07.

4.2. Quantitative Evaluation Results

Comparison results on MS COCO val2017 The performance of our proposed Exemplar-FreeSOLO is compared with that of the state-of-the-art unsupervised methods on MS COCO val2017 in Table 1 and Table 2. We can see from Table 1 that the proposed Exemplar-FreeSOLO largely outperforms the state-of-the-art class-agnostic instance segmentation techniques. MCG [1] and COB [31] are trained using the BSD500 dataset [32] and the PASCAL Context dataset [33], respectively. Yet, the instance segmentation results of the proposed Exemplar-FreeSOLO are still noticeably better than those of MCG and COB. Moreover, the proposed Exemplar-FreeSOLO outperforms FreeSOLO in terms of AP value by more than 4%. Additionally, from the unsupervised class-agnostic object detection results in Table 2, we can see that our proposed method substantially outperforms all the comparison methods, achieving an AP value of 12.6%. These experimental results illustrate the effectiveness of our proposed Exemplar-FreeSOLO. Moreover, the remarkable performance gain over FreeSOLO suggests the effectiveness of Exemplar-FreeSOLO can be attributed to the novel exemplar mechanism that provides useful top-down guidance and discriminative information.

Model	AP_{50}	AP_{75}	AP	AR_1	AR_{10}	AR_{100}
UP-DETR [9]	0.0	0.0	0.0	0.0	0.0	0.4
SS [43]	0.5	0.1	0.2	0.2	1.5	10.9
DETReg [2]	3.1	0.6	1.0	0.6	3.6	12.7
FreeSOLO [48]	12.2	4.2	5.5	4.6	11.4	15.3
Exemplar-FreeSOLO	17.9	8.6	12.6	8.2	13.0	17.9

Table 2. Unsupervised class-agnostic object detection results on MS COCO val2017.

Model	AP_{50}	AP_{75}	AP
w/anns:			
SOLOv2 [49] w/COCO	38.0	20.9	21.4
Mask R-CNN [16] w/COCO	31.0	14.2	15.9
SOLOv2 [49] w/LVIS	14.8	5.9	7.1
Mask R-CNN [16] w/LVIS	18.1	4.1	6.8
w/o anns:			
FreeSOLO [48]	12.7	3.0	4.8
Exemplar-FreeSOLO	14.2	7.3	9.2

Table 3. Unsupervised instance segmentation results on UVO val in terms of average precision scores. "w/anns" indicates the results obtained in a fully supervised scenario. "w/o anns" indicates the results obtained in an unsupervised scenario.

Comparison results on UVO val We also evaluate our proposed framework on the UVO val dataset in terms of AP values, as shown in Table 3. UVO is a video dataset with more difficult characteristics like camera shake, a dynamic background, and motion blur. Even so, the proposed method still achieves impressive results, especially in narrowing the performance gap between unsupervised and fully supervised instance segmentation models. We can see

Model	AP_{50}	AP_{75}	AP
Kim et al. [23]	9.5	-	2.5
DDT+ [53]	8.7	-	3.0
rOSD [44]	13.1	-	4.3
LOD [45]	13.9	-	4.5
LOST [40]	19.8	-	6.7
FreeSOLO [48]	24.5	7.2	10.2
Exemplar-FreeSOLO	26.8	8.2	12.6

Table 4. Multi-object discovery results on PASCAL VOC trainval07 in terms of average precision scores.

that our proposed framework outperforms FreeSOLO by 4.4% in terms of the AP value. SOLOv2 [49] and Mask R-CNN [16] are trained on the fully supervised COCO dataset, and LVIS dataset [14] separately. Surprisingly, our unsupervised framework outperforms the SOLOv2 and Mask R-CNN methods trained on the LVIS dataset in terms of the AP value. Again, we attribute this to the proposed exemplar mechanism.

Comparison results on PASCAL VOC The multi-object discovery results for all the comparison methods on the PASCAL VOC dataset are reported in Table 4. The goal of this task is to find the location of multiple salient objects without any annotations. We can see that the proposed framework significantly outperforms the existing state-of-the-art methods and outperforms the second-best model (*i.e.* FreeSOLO) in terms of the AP value by 2.4%. Besides, LOST [40] is a transformer-based method that leverages the activation features to generate seeds for generating the location of objects. By contrast, our proposed framework effectively deploys a simple exemplar mechanism to exploit unsupervised discriminative information.

4.3. Ablation Study

We conducted an ablation study on the proposed approach by comparing the full Exemplar-FreeSOLO with three variant methods: (1) "vanilla FreeSOLO" denotes the standard FreeSOLO. (2) "Semi-super-box" denotes a variant based on FreeSOLO, in which the bounding box ground-truths of the exemplar images are added directly to model training as foreground supervision information. The variant can be viewed as a weakly semi-supervised model for instance segmentation and object detection. (3) "Semi-super-mask" denotes another variant based on FreeSOLO, in which the mask ground-truths of the exemplar images are added directly to model training as foreground supervision information. The comparison results of unsupervised instance segmentation and object detection on MS COCO val2017 are reported in Table 5. First, we can see that with additional supervision information, the two variants outperform FreeSOLO in most cases. However, our proposed unsupervised Exemplar-FreeSOLO consistently outperforms the two variants across all evaluations. In terms of the AP value, Exemplar-FreeSOLO outperforms the best variant by 2.2% for segmentation and 4.4% for detection. These experimental results demonstrate the effectiveness of the proposed EKA and EEC modules.

Model	AP_{50}	AP_{75}	AP	AP_s	AP_m	AP_l
Segmentation:						
vanilla FreeSOLO	9.8	2.9	4.0	3.6	13.5	10.8
Semi-super-box	11.2	3.5	5.8	3.0	7.7	19.2
Semi-super-mask	12.5	3.9	6.2	3.5	8.2	19.7
Exemplar-FreeSOLO	13.2	6.3	8.4	5.5	16.6	22.2
Detection:						
vanilla FreeSOLO	12.2	4.2	5.5	5.1	13.8	16.8
Semi-super-box	13.4	3.9	7.1	4.7	12.1	15.9
Semi-super-mask	14.3	4.7	8.2	4.8	12.6	16.7
Exemplar-FreeSOLO	17.9	8.6	12.6	6.8	15.9	19.9

Table 5. Ablation study for the proposed Exemplar-FreeSOLO on MS COCO val2017 with unsupervised instance segmentation and unsupervised object detection.

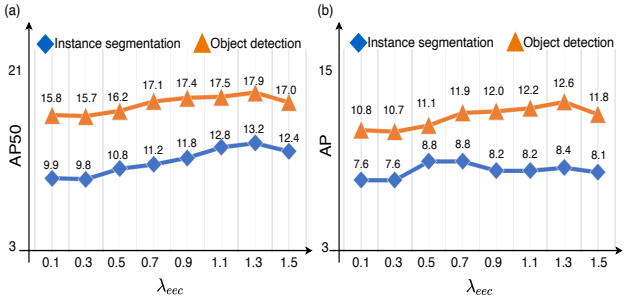

Figure 4. Impact of the weight of the proposed loss function in terms of AP_{50} (a) and AP (b). We report the results for unsupervised class-agnostic instance segmentation and unsupervised class-agnostic object detection on MS COCO val2017.

4.4. Further Analysis

Impact of the weight of the proposed loss function We summarize the experimental results about the impact of the weight of the exemplar embedding contrastive loss, λ_{eec}, in Figure 4. The experiments are performed by fixing the other hyperparameters and only changing the value of λ_{eec}. The best results in terms of both AP_{50} and AP values for the object detection task are obtained with λ_{eec}=1.3. For instance segmentation, the same λ_{eec} value produces the best result in terms of AP_{50} and produces a good AP value of 8.4%, which is only slightly smaller than the best result in terms of the AP metric, 8.8%, which is achieved with λ_{eec}= 0.5 or 0.7. Considering these results on both tasks, we used λ_{eec} = 1.3 in all the other experiments.

	Num	AP_{50}	AP_{75}	AP	AP_s	AP_m	AP_l
Segmentation	1	13.2	6.3	8.4	5.5	16.6	22.2
	3	13.5	6.4	8.7	5.5	16.8	22.4
	5	13.6	6.6	8.8	5.7	16.8	22.5
	7	13.6	6.7	8.8	5.6	16.8	22.8
	9	13.8	6.9	8.9	5.8	16.8	23.2
Detection	1	17.9	8.6	12.6	6.8	15.9	19.9
	3	18.5	8.8	12.7	6.8	16.2	20.2
	5	18.8	9.1	12.8	7.2	16.5	20.6
	7	18.8	9.2	12.8	7.4	16.5	20.8
	9	18.8	9.5	13.0	7.7	16.7	20.9

Table 6. Impact of the number of exemplars on the performance of Exemplar-FreeSOLO for unsupervised instance segmentation and unsupervised object detection on MS COCO val2017.

Impact of the number of exemplars In Table 6, we demonstrate an in-depth performance analysis by using different numbers of exemplar objects for each category. We can observe that the performance of the proposed model for both unsupervised segmentation and unsupervised detection improves slightly as the number of exemplars increases. But it is worth noting that even with only one exemplar in each category, very impressive results can still be obtained.

Impact of the exemplar distribution We summarize the impact of the exemplar distribution across different categories in Figure 5. In each case, we randomly select a given number of categories and choose one exemplar from each category. We can see that the performance of the proposed framework gradually improves as the number of classes increases. The largest improvement occurs when the number of classes increases from 50 to 60. The proposed framework achieves the best results when the exemplars are distributed among all categories. This is reasonable since as the exemplars come from more categories, there are not only more exemplars but also a high probability of providing more diverse representations. These experimental results also validate that randomly chosen images are suitable as unsupervised representative exemplars.

4.5. Qualitative Evaluation Results

For qualitative analyses of the results produced by Exemplar-FreeSOLO, we present some visualized examples of the unsupervised instance segmentation and object detection results in Figure 6 and Figure 7. Background clutter or edge ambiguity can exist in different images, which is very challenging for unsupervised learning scenarios. From the figures, we can observe that Exemplar-FreeSOLO can still segment and detect the corresponding targets in such scenes more accurately than FreeSOLO.

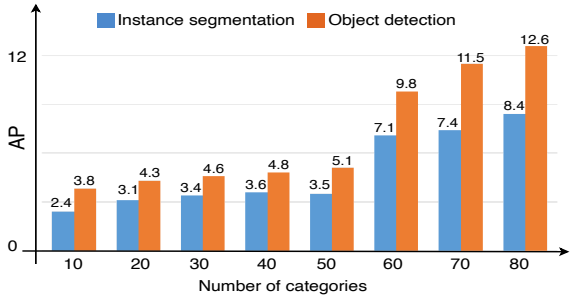

Figure 5. Impact of the exemplar distribution in terms of AP for unsupervised class-agnostic instance segmentation and unsupervised class-agnostic object detection on MS COCO val2017.

Figure 6. Visualized examples of the unsupervised instance segmentation and object detection results by Exemplar-FreeSOLO.

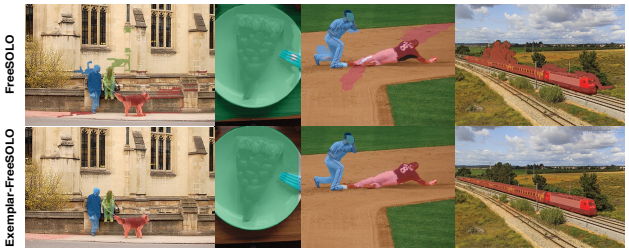

Figure 7. Examples of segmentation using FreeSOLO and the proposed Exemplar-FreeSOLO.

5. Conclusion

This paper proposes a novel framework, Exemplar-FreeSOLO, to address unsupervised instance segmentation by developing an effective exemplar mechanism through two consecutive modular functions. The Exemplar-FreeSOLO uses an exemplar knowledge abstraction module (EKA) to acquire beneficial top-down guidance knowledge by extracting exemplar objects in unsupervised ways, and uses an exemplar embedding contrastive module (EEC) to enhance the discriminative capability of the instance segmentation network by exploiting the exemplar guidance knowledge in a contrastive manner. Experimental results demonstrate that the proposed framework outperforms the existing state-of-the-art methods.

References

[1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *IEEE conference on computer vision and pattern recognition*, 2014. 6

[2] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[3] Míriam Bellver Bueno, Amaia Salvador Aguilera, Jordi Torres Viñals, and Xavier Giró Nieto. Budget-aware semi-supervised semantic and instance segmentation. In *CVPR Workshops*, 2019. 1

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *IEEE/CVF international conference on computer vision*, 2019. 1, 6

[5] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[6] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020. 2

[8] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *European conference on computer vision*, 2016. 2

[9] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *IEEE/CVF conference on computer vision and pattern recognition*, 2021. 6

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

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[12] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *European conference on computer vision (ECCV)*, 2018. 3

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[14] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE/CVF conference on computer vision and pattern recognition*, 2019. 7

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE international conference on computer vision*, 2017. 1, 2, 6, 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 2016. 4

[18] 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. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3

[19] Yen-Chang Hsu, Zheng Xu, Zsolt Kira, and Jiawei Huang. Learning to cluster for proposal-free instance segmentation. In *International Joint Conference on Neural Networks (IJCNN)*, 2018. 2

[20] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *IEEE conference on computer vision and pattern recognition*, 2018. 1, 3

[21] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *IEEE/CVF conference on computer vision and Pattern recognition*, 2020. 2

[22] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, 2017. 1

[23] Gunhee Kim and Antonio Torralba. Unsupervised detection of regions of interest using iterative link analysis. *Advances in neural information processing systems*, 22, 2009. 7

[24] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *IEEE/CVF international conference on computer vision*, 2019. 3

[25] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *IEEE/CVF conference on computer vision and pattern recognition*, 2021. 3

[26] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, 2017. 1, 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014. 6

[28] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE conference on computer vision and pattern recognition*, 2018. 2

[29] Yun Liu, Yu-Huan Wu, Pei-Song Wen, Yu-Jun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image- and dataset-level information for weakly supervised instance

segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[30] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *2011 International conference on computer vision*, 2011. 3

[31] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Convolutional oriented boundaries: From image segmentation to high-level tasks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):819–833, 2017. 6

[32] David Martin, Charless 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 *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, 2001. 6

[33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE conference on computer vision and pattern recognition*, 2014. 6

[34] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[35] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, 2016. 2

[36] Robert M Nosofsky. The generalized context model: An exemplar model of classification. *Formal approaches in categorization*, pages 18–39, 2011. 2

[37] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE conference on computer vision and pattern recognition*, 2016. 2

[38] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 4

[39] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[40] 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 preprint arXiv:2109.14279*, 2021. 7

[41] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, 2020. 1

[42] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 3, 5

[43] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 6

[44] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision*, 2020. 7

[45] Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. In *Advances in Neural Information Processing Systems*, 2021. 7

[46] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2021. 6

[47] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 4

[48] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 4, 5, 6, 7

[49] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: dynamic and fast instance segmentation. In *Advances in Neural information processing systems*, 2020. 6, 7

[50] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 4, 6

[51] Xuehui Wang, Kai Zhao, Ruixin Zhang, Shouhong Ding, Yan Wang, and Wei Shen. Contrastmask: Contrastive learning to segment every thing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[52] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[53] 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*, 88:113–126, 2019. 7

[54] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, 2016. 2

[55] Yanzhao Zhou, Xin Wang, Jianbin Jiao, Trevor Darrell, and Fisher Yu. Learning saliency propagation for semi-supervised instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3