Exemplar-FreeSOLO: Enhancing Unsupervised Instance Segmentation with Exemplars

Taoseef Ishtiak,\textsuperscript{1} Qing En,\textsuperscript{1} Yuhong Guo\textsuperscript{1,2}
\textsuperscript{1}Carleton University, Ottawa, Canada \textsuperscript{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 \cite{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 \cite{18,22,29,42}, partially-supervised \cite{20,51} and semi-supervised instance segmentation \cite{3,52,55}.

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 \cite{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-
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 unlabeled 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
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 $M$ and semantic embeddings $Q$ in an unsupervised manner. Specifically, the input for FreeSOLO is a set of $N$ unlabeled images, $D_U = \{X^n_u\}_{n=1}^N$, where each image is represented as $X^n_u \in \mathbb{R}^{H \times W \times 3}$, with $H$ and $W$ denoting the height and width of the image, respectively. For each input image $X^n_u$, its feature embedding $I^n_u \in \mathbb{R}^{H \times W \times c}$ can be extracted from a self-supervised backbone [50], e.g., ResNet [17]. $I^n_u$ is then bilinearly downsampled to obtain queries $Q^n_u \in \mathbb{R}^{H' \times W' \times c}$, while taking itself as keys $K^n_u$. Next, both of them go through $\ell_2$ normalization to generate new queries $\bar{Q}^n_u$ and new keys $\bar{K}^n_u$. Each query in $\bar{Q}^n_u$ can be treated as a $1 \times 1$ convolutional kernel to perform a convolution operation on the keys in $\bar{K}^n_u$ and generate score maps $S^n_u = \bar{Q}^n_u \odot \bar{K}^n_u$, which are then used to produce the coarse mask $M^n_u$ as follows:

$$M^n_u = \text{NMS}(\text{Maskness}(\text{Norm}(S^n_u))),$$

where $\odot$ denotes the convolution operation; $\text{Norm}(\cdot)$ is a normalization function that shifts the scores to the range of $[0,1]$; $\text{Maskness}(\cdot)$ indicates a confidence score function [47]; and $\text{NMS}(\cdot)$ represents the mask non-maximum-suppression operation. The coarse masks $\{M^n_u\}_{n=1}^N$ and the query feature vectors $\{Q^n_u\}_{n=1}^N$ from all the unlabeled images are then used as initial pseudo-labels and semantic embeddings to train a SOLO-based instance segmentation network 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 $D_U$, Exemplar-FreeSOLO aims to train a good instance segmentation network $N^e$ 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 $D_P$ from randomly selected images from the unlabeled dataset $D_U$, 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 $X_e$, 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 $D_P$:

$$D_P = \{X^k_e\}_{k=1}^K.$$

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
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^k_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 \hat{I}^n_{mu} for it as follows. First, we apply \ell_2 normalization and up-sampling on \hat{I}^n_{u} to generate queries \hat{Q}^n_{u}. Similarly, we generate keys \hat{K}^n_{u} by applying \ell_2 normalization, up-sampling, and an extra \times 1 convolution operation on \hat{I}^n_{u}. Next, following FreeSOLO, a prediction mask \hat{M}^n_{u} can be produced by applying the series of operations in Eq.(1) to the element-wise multiplication of \hat{Q}^n_{u} and \hat{K}^n_{u}. 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 \hat{I}^n_{mu}, we randomly select an exemplar embedding whose similarity with \hat{I}^n_{mu} is larger than a threshold \alpha as a positive sample, such as

\[ I^n_{e,pos} \in \{I^k_k \in \mathcal{I}_P : \text{sim}(\hat{I}^n_{mu}, I^k_k) > \alpha\}, \]  

while using all the exemplars whose similarity with \hat{I}^n_{mu} is smaller than a threshold \beta as negative samples:

\[ I^n_{e,neg} \in \{I^k_k \in \mathcal{I}_P : \text{sim}(\hat{I}^n_{mu}, I^k_k) < \beta\}. \]

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

\[ L_{eec} = -\sum_n \log \frac{\text{pos}(n)}{\text{pos}(n) + \text{neg}(n)}, \]

where

\[ \text{pos}(n) = \exp(\langle \hat{I}^n_{mu}, I^n_{pos}\rangle / \tau) \]

\[ \text{neg}(n) = \frac{1}{|\mathcal{I}_{e,neg}|} \sum_{I_e \in \mathcal{I}_{e,neg}} \exp(\langle \hat{I}^n_{mu}, I_e\rangle / \tau). \]

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 \( L_{ece} \):

\[ L_{total} = L_{mask} + L_{cate} + \lambda L_{ece}, \]  

where \( \lambda \) 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:

\[ L_{mask} = \gamma L_{avg-proj}(M^*, M) + L_{max-proj}(M^*, M) + L_{pairwise}(M^*), \]

where \( M^* \) and \( M \) are the predicted mask from the segmentation head and the coarse mask generated from the Free Mask of FreeSOLO [48]; \( \gamma \) 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:

\[ L_{cate} = L_{focal}(M^*, M) + \mu L_{sem}(Q^*, Q), \]

where \( L_{focal} \) is the focal loss and \( L_{sem} \) is a negative cosine similarity function; \( \mu \) is a trade-off hyperparameter; and \( Q^* \) and \( Q \) are the predicted embeddings from the segmentation head and the embeddings generated together with pseudo labels from the Free Mask of FreeSOLO [48].
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 $N_x$. 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 [3] 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 ($\lambda_{rec}, \gamma$ and $\mu$), to 1.3, 0.1 and 0.4, respectively. We set $\tau$ to 0.02, $\alpha$ to 0.8, and $\beta$ 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. “w/anns” indicates the results obtained in a supervised scenario. “w/o anns” indicates the results obtained in an unsupervised scenario.

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

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.

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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>$AP_{50}$</th>
<th>$AP_{75}$</th>
<th>AP</th>
<th>AR</th>
<th>$AR_{10}$</th>
<th>$AR_{100}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/anns:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOLOv2 [49] w/COCO</td>
<td>38.0</td>
<td>20.9</td>
<td>21.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mask R-CNN [16] w/COCO</td>
<td>31.0</td>
<td>14.2</td>
<td>15.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOLOv2 [49] w/LVIS</td>
<td>14.8</td>
<td>5.9</td>
<td>7.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mask R-CNN [16] w/LVIS</td>
<td>18.1</td>
<td>4.1</td>
<td>6.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o anns:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FreeSOLO [48]</td>
<td>12.7</td>
<td>3.0</td>
<td>4.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exemplar-FreeSOLO</td>
<td>14.2</td>
<td>7.3</td>
<td>9.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>$AP_{50}$</th>
<th>$AP_{75}$</th>
<th>AP</th>
<th>AR</th>
<th>$AR_{10}$</th>
<th>$AR_{100}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UP-DETR [9]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>SS [43]</td>
<td>0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>1.5</td>
<td>10.9</td>
</tr>
<tr>
<td>DETReg [2]</td>
<td>3.1</td>
<td>0.6</td>
<td>1.0</td>
<td>0.6</td>
<td>3.6</td>
<td>12.7</td>
</tr>
<tr>
<td>FreeSOLO [48]</td>
<td>12.2</td>
<td>4.2</td>
<td>5.5</td>
<td>4.6</td>
<td>11.4</td>
<td>15.3</td>
</tr>
<tr>
<td>Exemplar-FreeSOLO</td>
<td>17.9</td>
<td>8.6</td>
<td>12.6</td>
<td>8.2</td>
<td>13.0</td>
<td>17.9</td>
</tr>
</tbody>
</table>

Table 3. Unsupervised instance segmentation results on UVO val2017.
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.

### 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, $\lambda_{ee}$, in Figure 4. The experiments are performed by fixing the other hyperparameters and only changing the value of $\lambda_{ee}$. The best results in terms of both $AP_{50}$ and AP values for the object detection task are obtained with $\lambda_{ee} = 1.3$. For instance segmentation, the same $\lambda_{ee}$ 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 $\lambda_{ee} = 0.5$ or 0.7. Considering these results on both tasks, we used $\lambda_{ee} = 1.3$ in all the other experiments.

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

**Table 4.** Multi-object discovery results on PASCAL VOC train-val07 in terms of average precision scores.

<table>
<thead>
<tr>
<th>Model</th>
<th>$AP_{50}$</th>
<th>$AP_{75}$</th>
<th>$AP$</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanilla FreeSOLO</td>
<td>9.8</td>
<td>2.9</td>
<td>4.0</td>
</tr>
<tr>
<td>Semi-super-box</td>
<td>11.2</td>
<td>3.5</td>
<td>5.8</td>
</tr>
<tr>
<td>Semi-super-mask</td>
<td>12.5</td>
<td>3.9</td>
<td>6.2</td>
</tr>
<tr>
<td>Exemplar-FreeSOLO</td>
<td>13.2</td>
<td>6.3</td>
<td>8.4</td>
</tr>
</tbody>
</table>

**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.
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.

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.

<table>
<thead>
<tr>
<th>Num</th>
<th>$AP_{50}$</th>
<th>$AP_{75}$</th>
<th>$AP$</th>
<th>$AP_s$</th>
<th>$AP_m$</th>
<th>$AP_l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.2</td>
<td>6.3</td>
<td>8.4</td>
<td>5.5</td>
<td>16.6</td>
<td>22.2</td>
</tr>
<tr>
<td>3</td>
<td>13.5</td>
<td>6.4</td>
<td>8.7</td>
<td>5.5</td>
<td>16.8</td>
<td>22.4</td>
</tr>
<tr>
<td>5</td>
<td>13.6</td>
<td>6.6</td>
<td>8.8</td>
<td>5.7</td>
<td>16.8</td>
<td>22.5</td>
</tr>
<tr>
<td>7</td>
<td>13.6</td>
<td>6.7</td>
<td>8.8</td>
<td>5.6</td>
<td>16.8</td>
<td>22.8</td>
</tr>
<tr>
<td>9</td>
<td>13.8</td>
<td>6.9</td>
<td>8.9</td>
<td>5.8</td>
<td>16.8</td>
<td>23.2</td>
</tr>
</tbody>
</table>

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


[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.


[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


[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


[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


[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


