ContrastMask: Contrastive Learning to Segment Every Thing

Xuehui Wang[1†], Kai Zhao[2], Ruixin Zhang[2], Shouhong Ding[2], Yan Wang[3], Wei Shen[1(✉)]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Youtu Lab, Tencent [3]Shanghai Key Lab of Multidimensional Information Processing, ECNU
{wangxuehui, wei.shen}@sjtu.edu.cn; {ruixinzhang, ericshding}@tencent.com;
kz@kaizhao.net; ywang@cee.ecnu.edu.cn

Abstract

Partially-supervised instance segmentation is a task which requests segmenting objects from novel categories via learning on limited base categories with annotated masks thus eliminating demands of heavy annotation burden. The key to addressing this task is to build an effective class-agnostic mask segmentation model. Unlike previous methods that learn such models only on base categories, in this paper, we propose a new method, named ContrastMask, which learns a mask segmentation model on both base and novel categories under a unified pixel-level contrastive learning framework. In this framework, annotated masks of base categories and pseudo masks of novel categories serve as a prior for contrastive learning, where features from the mask regions (foreground) are pulled together, and are contrasted against those from the background, and vice versa. Through this framework, feature discrimination between foreground and background is largely improved, facilitating learning of the class-agnostic mask segmentation model. Exhaustive experiments on the COCO dataset demonstrate the superiority of our method, which outperforms previous state-of-the-arts.

1. Introduction

Instance segmentation is one of the most fundamental tasks in computer vision, which requests pixel-level prediction on holistic images and identifies each individual object. Many works [8, 13, 17, 19, 26, 31, 39, 42] have boosted instance segmentation performance by relying on a large amount of available pixel-level annotated data. However, performing pixel-level annotation (mask annotation) is significantly burdensome, which hinders the further development of instance segmentation on massive novel categories.

Since box-level annotations are much cheaper and easier to obtain than mask annotations [12], a common way to address the aforementioned issue is to perform *partially-supervised instance segmentation* [15, 18, 22, 45]. This instance segmentation task was first proposed in the paper "Learning to Segment Every Thing" [18], where object categories are divided into two splits: *base* and *novel*. Both of them have box-level annotations, while only *base* categories have additional mask annotations. Then the goal of this task is by taking advantage of the data of *base* categories with mask annotations to generalize instance segmentation models to *novel* categories. The main obstacle to achieve favorable instance segmentation performance under the partially-supervised setting is how to distinguish foreground and background within each box for an arbitrary category via learning on the data with limited annotations.

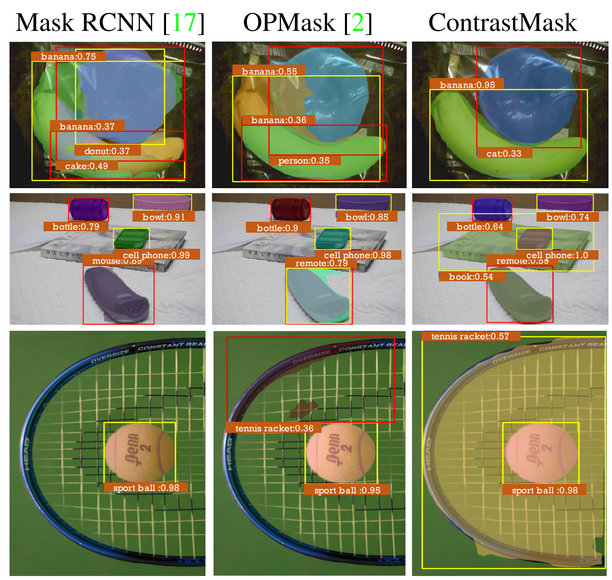

Figure 1. Visualization results of Mask R-CNN [17], OPMask [2] and the proposed ContrastMask on *novel* categories.

Previous methods [2, 15, 18, 22, 29, 30, 45] addressed this task via learning a class-agnostic mask segmentation model to separate foreground and background, by capturing class-agnostic cues, such as shape bases [22] and appear-

†Work done during an internship at Youtu Lab, Tencent.
✉Corresponding Author.

ance commonalities [15]. However, these methods learn the class-agnostic mask segmentation model only on *base* categories, ignoring a large amount of training data from *novel* categories, and consequently lack a bridge to transfer the segmentation capability of the mask segmentation model on *base* categories to *novel* categories.

To build this bridge, in this paper, we propose ContrastMask, a new partially-supervised instance segmentation method, which learns a class-agnostic mask segmentation model on both *base* and *novel* categories under a unified pixel-level contrastive learning framework. In this framework, we design a new query-sharing pixel-level contrastive loss to fully exploit data from all categories. To this end, annotated masks of *base* categories or pseudo masks of *novel* categories computed by Class Activation Map (CAM) [2, 44] serve as a region prior, which indicates not only the foreground and background separation, but also shared queries, positive keys and negative keys. Concretely, given a training image batch containing both *base* categories and *novel* categories, we establish two shared queries: a foreground query and a background query, which are obtained by averaging features within and outside the mask regions, including both the annotated and the pseudo masks, respectively. Then, we perform a special sampling strategy to select proper keys. By introducing the proposed loss, we expect to pull keys within/outside the mask regions towards the foreground/background shared query and contrast it against keys outside/within the mask regions. Finally, features learned by our pixel-level contrastive learning framework are fused into a class-agnostic mask head to perform mask segmentation.

Compared with previous methods, ContrastMask enjoys several benefits: 1) It fully exploits training data, making those from *novel* categories also contribute to the optimization process of the segmentation model; 2) More importantly, it builds a bridge to transfer the segmentation capability on *base* categories to *novel* categories by the unified pixel-level contrastive learning framework, especially the shared queries for both *base* and *novel* categories, which consistently improves feature discrimination between foreground and background for both *base* and *novel* categories. A visualization result of comparison with other methods is shown in Fig. 1.

Without bells and whistles, ContrastMask surpasses all previous state-of-the-art partially-supervised instances segmentation methods on the COCO dataset [25], by large margins. Notably, with the ResNeXt-101-FPN [24, 40] as the backbone, our method achieves 39.8 mAP for mask segmentation on *novel* categories.

2. Related Work

Instance Segmentation. Instance segmentation is a task that combines both object detection and semantic segmentation, *i.e.*, each pixel is assigned to a specific category and an individual instance simultaneously. Mask R-CNN [17] produced a mask for each detected bounding box by extending Faster R-CNN with a mask head. PANet [26] improved segmentation performance by building bottom-up path augmentations and lateral connections across features of multiple levels. HTC [8] presented interleaved execution and mask information flow and achieved considerable performance. DSC [13] formed a bi-directional relationship between detection and segmentation tasks, and achieved state-of-the-art performance. BMask [11] established a parallel head to predict the boundary of objects, which can be fused into the mask head to refine segmentation results. BCNet [20] adopted bilayer GCN and self-attention to regress the object contour and instance masks, respectively. In addition to these two-stage methods, one-stage methods, such as CondInst [31], BlendMask [7], SOLO [35], SOLOv2 [36], YOLACT [5], and FCIS [23], obtained comparable performance with favorable inference speed.

Pixel-level contrastive learning. Very recently, several works [1,6,28,37,38,41,43] have been proposed to perform pixel-level contrastive learning to remedy the misalignment between the classification task and the dense prediction task. However, these methods and ours **differ in both objective and design philosophy**: Their objective is to learn general dense representations for per-pixel multi-class categorization, so they perform pixel-level instance discrimination by sampling keys from two different augmented views; While ours is to improve the foreground and background discrimination, so we perform pixel-level instance discrimination by sampling keys from foreground and background areas within one image.

Partially-supervised instance segmentation. As the pioneer method of partially-supervised instance segmentation, Mask^X R-CNN [18] designed a parameterized transformation function between the bounding box head and the mask head in Mask R-CNN [17], which enables segmenting *novel* categories based on the assumption that the bounding box head encodes the embeddings of all categories. ShapeMask [22] learned shape priors from limited data with mask annotations, and expected these shape priors can generalize to *novel* objects. ShapeProp [45] exploited box supervision to learn salient regions and propagated these regions to the whole box proposals via an efficient message passing module which can generate a more accurate shape prior. CPMask [15] achieved promising performance by parsing shape commonality and appearance commonality among different categories. It claimed that sharing these commonalities can promote the generalization ability for mask prediction in a class-agnostic manner. OPMask [2] employed object mask prior (OMP) to provide general concepts of foreground for mask head learning, and thus the network can concentrate on the primary objects in a region.

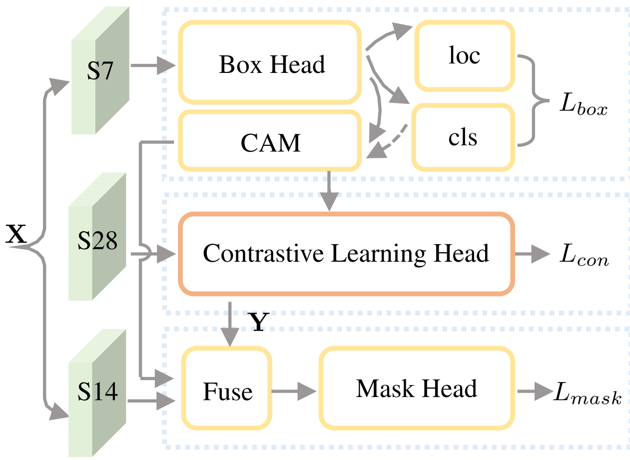

Figure 2. The whole architecture of ContrastMask, which is built on the Mask R-CNN [17], with an extra contrastive learning head. "Sn" denotes that size of the feature map is $n \times n$. $\mathbf{X}$ and $\mathbf{Y}$ are an intput RoI feature map and its enhanced feature map, respectively.

Very recently, Deep-MAC [4] explored the impact of mask head architectures to segmentation performance on *novel* categories. It adopted much heavier architectures, such as Hourglass-52 [27], for mask heads, and achieved outstanding performance. However, a lightweight mask head is always more popular in practice.

All these methods optimize their mask segmentation models only on *base* categories, ignoring a large amount of data from *novel* categories, and thus lack a bridge to transfer the segmentation capability on *base* categories to *novel* categories. We address this issue by introducing a unified contrastive learning framework for dense mask prediction, in which both *base* and *novel* categories contribute to mask segmentation model learning.

3. ContrastMask

We first depict the whole flowchart of the proposed ContrastMask. Then, we show how the unified pixel-level contrastive learning framework is instantiated to enhance feature discrimination between foreground and background on both *base* and *novel* categories. Finally, we introduce the loss functions to learn our partially-supervised instance segmentation model.

3.1. Overview

As shown in Fig. 2, our method, ContrastMask, is built on the classic two-stage Mask R-CNN [17] architecture with an extra "contrastive learning" head, termed as CL Head, which performs unified pixel-level contrastive learning on both *base* and *novel* categories. The CL Head takes an RoI feature map and a CAM generated by the box head as input. It is supervised by our proposed pixel-level contrastive loss (Sec. 3.3) and outputs an enhanced feature map for the mask head. Finally, the mask head predicts a class-agnostic segmentation map by taking a fused feature map as input. Next, we describe the details of each component of our method.

3.2. Contrastive Learning Head (CL Head)

The goal of the CL Head is to increase feature discrimination between foreground and background and decrease feature dissimilarity within each region (background or foreground) for both *base* and *novel* categories, so that it can facilitate mask head learning. We achieve this by learning it with a new pixel-level contrastive loss.

As illustrated in Fig. 3, the architecture of the contrastive learning head (CL Head) is inspired by SimCLR [10], which is composed of a lightweight encoder $f(\cdot)$ and a projector $g(\cdot)$ for contrastive learning. The encoder $f(\cdot)$ contains eight 3×3 Conv-ReLU layers and the projector $g(\cdot)$ is a three-layer MLP, where the last layer is not followed by a ReLU function.

Given an RoI feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ extracted by RoIAlign [17], where C, H and W represent channel dimension, height and width of the RoI, respectively, the CL Head feeds them into the encoder to get an enhanced feature map $\mathbf{Y} = f(\mathbf{X}) \in \mathbb{R}^{H \times W \times C}$ which will be incorporated into the mask head for mask segmentation. Next, $\mathbf{Y}$ is first flattened and then fed into the projector, which maps representations to the space where the pixel-level contrastive loss is applied: $\mathbf{Z} = g(\mathbf{Y}) \in \mathbb{R}^{HW \times C}$. Here, after projection, the feature map $\mathbf{Z}$ is reshaped to the same size as $\mathbf{Y}$.

3.3. Query-sharing Pixel-level Contrastive Loss

Now, we introduce our new pixel-level loss, which enables learning the mask segmentation model on both *base* and *novel* categories under a unified contrastive learning framework. A core design philosophy for this loss function is *base* and *novel* categories share two class-agnostic queries, one for foreground $\mathbf{q}^+$ and the other for background $\mathbf{q}^-$, so that a bridge is formed to transfer the segmentation capability on *base* categories to *novel* categories. For this reason, we name our loss function **query-sharing pixel-level contrastive loss**.

The query-sharing pixel-level contrastive loss consists of two symmetrical formulations for foreground and background, respectively. Taking foreground as an example, let $\mathcal{K}^+$ and $\mathcal{K}^-$ denote a set of foreground keys and a set of background keys, respectively. Then the query-sharing pixel-level contrastive loss for foreground is defined as:

$$L^{\mathbf{q}^+}_{\mathcal{K}^+,\mathcal{K}^-} = -\frac{1}{|\mathcal{K}^+|} \sum_{\mathbf{k}^+ \in \mathcal{K}^+} \Big[\phi(\mathbf{q}^+, \mathbf{k}^+)/\tau - \log\big(\exp(\phi(\mathbf{q}^+, \mathbf{k}^+)/\tau) + \sum_{\mathbf{k}^- \in \mathcal{K}^-} \exp(\phi(\mathbf{q}^+, \mathbf{k}^-)/\tau)\big) \Big], \tag{1}$$

where τ is a temperature hyper-parameter and $\phi(\cdot, \cdot)$ de-

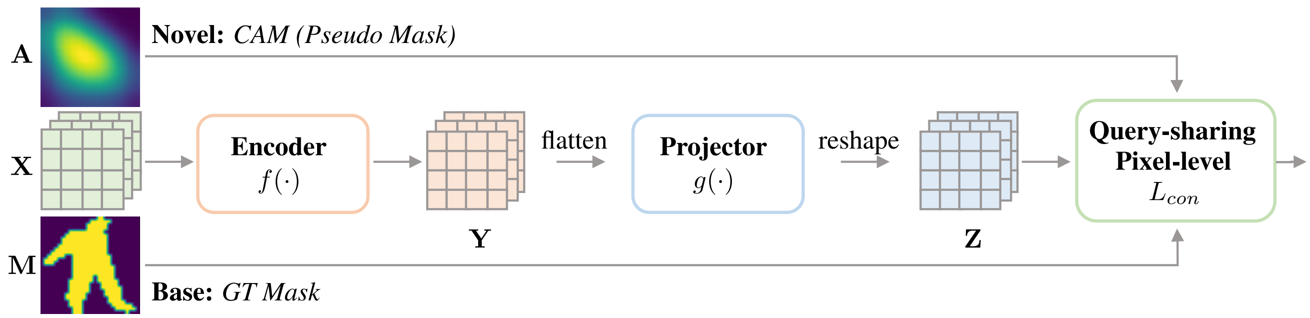

Figure 3. The flowchart of our contrastive learning head (CL Head) which consists of an encoder and a projector, supervised by a pixel-level contrastive loss. Ground-truth masks (if *base*) or pseudo masks converted from CAMs (if *novel*) are used to calculate the contrastive loss.

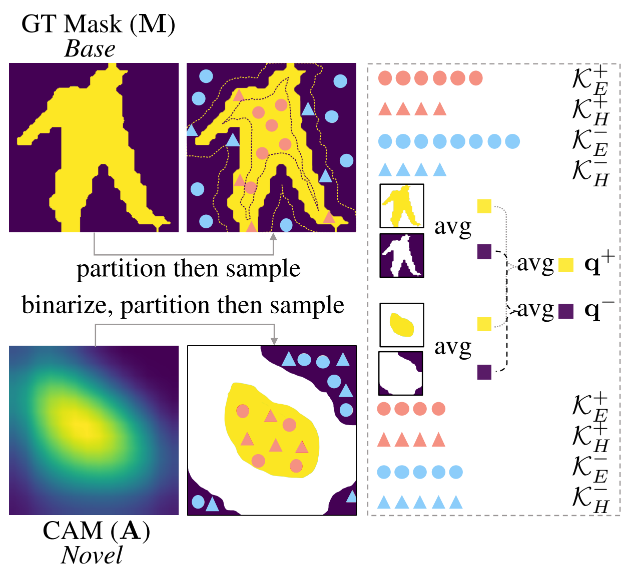

Figure 4. A schematic diagram to illustrate how to obtain queries and sample keys. For *base* categories, we use ground-truth masks to do partition and extract edges to guide sampling hard keys. For *novel* categories, we firstly binarize CAMs by a threshold δ, then perform partition and randomly sample easy and hard keys based on partitions. The foreground query $\mathbf{q}^+$ and background query $\mathbf{q}^-$ are obtained by averaging features from corresponding partitions of a batch of object proposals.

notes the cosine similarity. Similarly, we can obtain the query-sharing pixel-level contrastive loss for background $L^{q^-}_{\mathcal{K}^-,\mathcal{K}^+}$. Next, we describe the details of how to obtain the foreground and background queries $\mathbf{q}^+, \mathbf{q}^-$, as well as the foreground and background key sets $\mathcal{K}^+, \mathcal{K}^-$. We illustrate these details in Fig. 4.

Foreground and background partition. Given a projected feature map $\mathbf{Z} \in \mathbb{R}^{H\times W\times C}$, let $\mathbf{M} \in \{0,1\}^{H\times W}$ and $\mathbf{A} \in [0,1]^{H\times W}$ be the ground-truth mask and the class-activation map (CAM) aligned with $\mathbf{Z}$, repectively. Let $\mathcal{I}$ denote the $H \times W$ spatial location lattice of feature map $\mathbf{Z}$, then given a location $i \in \mathcal{I}$, we can obtain a feature vector $\mathbf{z}_i$ at location i from feature map $\mathbf{Z}$, and similarly the mask label m_i and the CAM value a_i at the location i from ground-truth mask $\mathbf{M}$ and CAM $\mathbf{A}$, respectively. The whole spatial location lattice can be partitioned into two disjoint lattices: foreground location lattice $\mathcal{I}^+$ and background location lattice $\mathcal{I}^-$. For *base* categories, we can achieve this partition by using the ground-truth mask: $\mathcal{I}^+ = \{i \in \mathcal{I}|m_i = 1\}$ and $\mathcal{I}^- = \{i \in \mathcal{I}|m_i = 0\}$; While for *novel* categories, as the ground-truth mask is not available, we alternatively use the CAM $\mathbf{A}$ to guide the foreground and background partition: $\mathcal{I}^+ = \{i \in \mathcal{I}|a_i \geq 1-\delta\}$ and $\mathcal{I}^- = \{i \in \mathcal{I}|a_i \leq \delta\}$, where $\delta = 0.1$ is a small threshold and is fixed in our method.

Query and key set generation. Let $\mathcal{I}^+_{(n)}$ and $\mathcal{I}^-_{(n)}$ be the foreground and background partitions of n^{th} RoI proposal in a batch consisting of N RoI proposals $\{\mathbf{Z}^{(n)}\}_{n=1}^N$, from both *base* and *novel* categories, respectively. The foreground and background queries $\mathbf{q}^+, \mathbf{q}^-$ are obtained by averaging features within foreground and background partitions across all proposals. Taking the foreground query $\mathbf{q}^+$ as an example, we obtain it by:

$$\mathbf{q}^+ = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|\mathcal{I}^+_{(n)}|}\sum_{i\in\mathcal{I}^+_{(n)}}\mathbf{z}_i^{(n)}. \tag{2}$$

The foreground and background key sets for a RoI proposal $\mathbf{Z}$ (here we omit the index n for notation simplicity) are obtained by $\mathcal{K}^+ = \{\mathbf{z_i}|i \in \S(\mathcal{I}^+,\sigma)\}$ and $\mathcal{K}^- = \{\mathbf{z_i}|i \in \S(\mathcal{I}^-,\sigma)\}$, respectively, where $\S(\cdot,\sigma)$ is a random sampling operator which samples a subset from a set randomly with a proportion ratio σ.

Hard and easy key mining. Previous studies reveal that mining hard and easy keys is helpful to learn discriminative features by contrastive learning [33, 34].

For *base* categories, we specify keys near an object boundary as hard keys and those far away from the boundary as easy keys. The ground-truth boundary can be obtained from the ground-truth mask easily. Let b_i be the nearest boundary location to location i on an RoI proposal

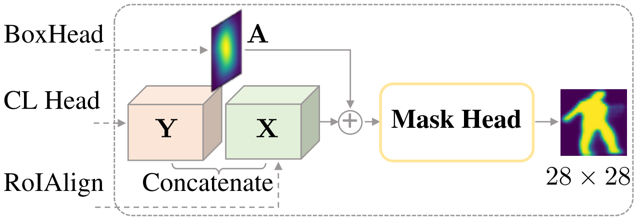

Figure 5. The input of the class-agnostic mask head consists of enhanced featuer map $\mathbf{Y}$, RoI feature map $\mathbf{X}$ and CAM $\mathbf{A}$. $\oplus$ represents an element-wise addition operation.

$\mathbf{Z}$. Then, we have the sets of hard keys and easy keys:

$$\begin{aligned}\mathcal{K}_H &= \{\mathbf{z}_i \mid i \in \S(\mathcal{I},\sigma), \|i - b_i\|_2^2 \le 2\} \\ \mathcal{K}_E &= \{\mathbf{z}_i \mid i \in \S(\mathcal{I},\sigma), \|i - b_i\|_2^2 > 2\}.\end{aligned} \tag{3}$$

For *novel* categories, since the ground-truth boundary is unavailable, we simply mine the hard and easy key sets by random sampling, *i.e.*, $\mathcal{K}_H = \{\mathbf{z}_i \mid i \in \S(\mathcal{I},\sigma)\}$ and $\mathcal{K}_E = \{\mathbf{z}_i \mid i \in \S(\mathcal{I},\sigma)\}$.

Instantiation of the proposed contrastive loss. Now, given an RoI proposal $\mathbf{Z}$, no matter from *base* or *novel* categories, we have described how to obtain foreground and background key sets from it as well as how to mine hard and easy key sets from it. Then consequently, we can obtain four types of key sets (Fig. 4) from it: 1) foreground-easy key set $\mathcal{K}_E^+$, 2) foreground-hard key set $\mathcal{K}_H^+$, 3) background-easy key set $\mathcal{K}_E^-$, and 4) background-hard key set $\mathcal{K}_H^-$. Referring to Eq. (1), our query-sharing pixel-level contrastive loss is defined as:

$$L_{con} = L^{\mathbf{q}^+}_{\mathcal{K}_E^+,\mathcal{K}_E^-} + L^{q^+}_{\mathcal{K}_H^+,\mathcal{K}_H^-} + L^{\mathbf{q}^-}_{\mathcal{K}_E^-,\mathcal{K}_E^+} + L^{\mathbf{q}^-}_{\mathcal{K}_H^-,\mathcal{K}_H^+}, \tag{4}$$

which contains four terms for the four key sets, respectively, and foreground query $\mathbf{q}^+$ and background query $\mathbf{q}^-$ are shared with keys from both *base* and *novel* categories.

3.4. Class-agnostic mask head

As shown in Fig. 5, the architecture and the corresponding loss function of the mask head is the same as those in Mask R-CNN [17] except for three modifications: 1) Following [2, 15, 45], we change the output channels of the last convolutional layer from 80 to 1, resulting in a class-agnostic mask head. 2) We concatenate the output feature map of the CL Head with the input feature map of the mask head, which makes the input features of the mask head more distinctive and facilitates its learning. 3) We utilize the CAM [2] to tell the mask head which region should focus on. This can be easily implemented by adding the CAM to the input feature maps.

3.5. Loss Function

The overall loss function for our ContrastMask contains three terms: a box detection loss L_{box}, a mask segmentation loss L_{mask}, and a contrastive learning loss L_{con}. The formulations of L_{box} and L_{mask} are the same as those defined in Mask R-CNN [17]:

$$L = L_{box} + L_{mask} + \lambda L_{con}, \tag{5}$$

where λ is a weight parameter.

4. Experiments

In this section, we first describe the experimental setup and implementation details. Then, we compare ContrastMask with state-of-the-art partially-supervised instance segmentation methods. Finally, we conduct ablation studies to show the contribution of each component in our method.

4.1. Experimental Setup

Our experiments are conducted on the challenging COCO dataset[1] [25]. To simulate *base* and *novel* categories, the training set is split into two subsets. Typically, categories presented in PASCAL VOC dataset [14] is termed as "*voc*" and remaining categories are "*nonvoc*". We mainly conduct experiments on these two subsets, and "*nonvoc* $\to$ *voc*" indicates that "*nonvoc*" categories are regarded as *base* and "*voc*" as *novel*, and vice versa. We use images in COCO-*train2017* for training and those in COCO-*val2017* for evaluation. Typical metrics for instance segmentation, *i.e.*, mask AP, including mAP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L, are used for evaluation. These metrics are calculated on the *novel* categories.

Implementation details. We implement our approach based on MMDetection[2] [9]. We adopt ResNet-50-FPN as the backbone for most ablation experiments and ResNet-101-FPN as the backbone for fair comparison with other methods. Typical training schedules, *i.e.*, $1\times$ and $3\times$, are both employed for a fair comparison, and all ablation experiments are conducted by $1\times$ schedule for efficiency. During training, we employ SGD with momentum for optimization, and the initial learning rate is 0.02. All experiments are conducted on 8 Tesla V100 GPUs and the batch size is 16, *i.e.*, 2 images per GPU. Each input image is resized to keep the rule that the long side of the image is less than 1,333 and the short side less than 800. The sampling ratio σ is set as $\sigma = 0.3$, and the temperature hyper-parameter τ (Eq. 1) for easy and hard keys are set as $\tau = 0.7$ and $\tau' = 1 - \tau = 0.3$, respectively. We linearly warmup the λ of L_{con} (Eq. (5)) from 0.25 to 1.0. Besides, commonly-used augmentations such as *random-flip* and multi-scale training are adopted.

4.2. Experimental Results

We compare our method ContrastMask with recent partially-supervised instance segmentation methods, in-

[1]It is released under the CC-BY 4.0 license.
[2]It is released under the Apache 2.0 license.

Method	Backbone	Sche.	Layers.	*nonvoc*→*voc*						*voc*→*nonvoc*					
				mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN (Baseline) [17]	ResNet-50	1×	4	23.9	42.9	23.5	11.6	24.3	33.7	19.2	36.4	18.4	11.5	23.3	24.4
Mask^X R-CNN [18]	ResNet-50	1×	4	28.9	52.2	28.6	12.1	29.0	40.6	23.7	43.1	23.5	12.4	27.6	32.9
Mask GrabCut [21]	ResNet-50	1×	-	19.5	46.2	14.2	4.7	15.9	32.0	19.5	39.2	17.0	6.5	20.9	34.3
CPMask [15]	ResNet-50	1×	4	-	-	-	-	-	-	28.8	46.1	30.6	12.4	33.1	43.4
ShapeProp [45]	ResNet-50	1×	4	34.4	59.6	35.2	13.5	32.9	48.6	30.4	51.2	31.8	14.3	34.2	44.7
ContrastMask (Ours)	ResNet-50	1×	4	**35.1**	**60.8**	**35.7**	**17.2**	**34.7**	**47.7**	**30.9**	**50.3**	**32.9**	**15.2**	**34.6**	**44.3**
OPMask [2]	ResNet-50	130k	7	36.5	62.5	37.4	17.3	34.8	49.8	31.9	52.2	33.7	16.3	35.2	46.5
ContrastMask (Ours)	ResNet-50	3×	4	**37.0**	**63.0**	**38.6**	**18.3**	**36.4**	**50.2**	**32.9**	**52.5**	**35.4**	**16.6**	**37.1**	**47.3**
ContrastMask (Ours)	ResNeXt-50	3×	4	**37.6**	**63.8**	**38.9**	**18.1**	**36.6**	**51.3**	**33.4**	**54.2**	**35.8**	**17.7**	**37.4**	**48.5**
Mask GrabCut [21]	ResNet-101	1×	-	19.6	46.1	14.3	5.1	16.0	32.4	19.7	39.7	17.0	6.4	21.2	35.8
Mask^X R-CNN [18]	ResNet-101	1×	4	29.5	52.4	29.7	13.4	30.2	41.0	23.8	42.9	23.5	12.7	28.1	33.5
ShapeMask [22]	ResNet-101	1×	8	33.3	56.9	34.3	17.1	38.1	45.4	30.2	49.3	31.5	16.1	38.2	28.4
ShapeProp [45]	ResNet-101	1×	4	35.5	60.5	36.7	15.6	33.8	50.3	31.9	52.1	33.7	14.2	35.9	46.5
ContrastMask (Ours)	ResNet-101	1×	4	**36.6**	**62.2**	**37.7**	**17.5**	**36.5**	**50.1**	**32.4**	**52.1**	**34.8**	**15.2**	**36.7**	**47.3**
ShapeMask (NAS-FPN) [22]	ResNet-101	3×	8	35.7	60.3	36.6	18.3	40.5	47.3	33.2	53.1	35.0	18.3	40.2	43.3
CPMask [15]	ResNet-101	3×	4	36.8	60.5	38.6	17.6	37.1	51.5	34.0	53.7	36.5	18.5	38.9	47.4
OPMask [2]	ResNet-101	130k	7	37.1	62.5	38.4	16.9	36.0	50.5	33.2	53.5	35.2	17.2	37.1	46.9
ContrastMask (Ours)	ResNet-101	3×	4	**38.4**	**64.5**	**39.8**	**18.4**	**38.1**	**52.6**	**34.3**	**54.7**	**36.6**	**17.5**	**38.4**	**50.0**
ContrastMask (Ours)	ResNeXt-101	3×	4	**39.8**	**66.2**	**42.3**	**19.2**	**39.3**	**53.6**	**35.0**	**56.4**	**36.9**	**18.6**	**38.9**	**50.5**

Table 1. Quantitative comparisons on the challenging COCO dataset. "*nonvoc*→*voc*" denotes that categories in *nonvoc* have the mask annotation and methods are required to be tested on *voc* categories that only have box annotations, and vice versa. "Sche." denotes the training schedule, where 1× represents for 12 epochs and *130k* is a customized schedule only used in OPMask [2]. We use two conventional schedules, *i.e.*, 1× and 3×, for fair comparison. "Layers." indicates the number of Conv blocks adopted in the mask head to perform mask prediction. Generally, a heavier mask head leads to better performance, which has been demonstrated in [4]. ResNeXt-50 and ResNeXt-101 indicate "ResNeXt-50-32x4d" [40] and "ResNeXt-101-64x4d" [40], respectively.

cluding Mask^X R-CNN [18], Mask GrabCut [21], ShapeMask [22], CPMask [15], ShapeProp [45] and OPMask [2].

Quantitative results. The quantitative results for *nonvoc* → *voc* and *voc* → *nonvoc* are shown in Tab. 1. When adopting ResNet-50 as the backbone and using the 1× schedule, our method surpasses the state-of-the-art method ShapeProp [45] by 0.7/0.5 mAP in *nonvoc* → *voc* and *voc* → *nonvoc* settings, respectively. We also outperforms CPMask [15] that uses a stronger detector, *i.e.*, FCOS [32], by a large margin (2.1 mAP). In addition, we provide comparison results under the 3× schedule. Our ContrastMask (ResNet-50) achieves 37.0 mAP which even outperforms the CPMask [15] (36.8 mAP) that uses ResNet-101 backbone by 0.2 mAP. This indicates that our method fully exploits all training data and builds a bridge to transfer the segmentation capability from *base* to *novel*.

Our method also offers superior performance using the ResNet-101 as the backbone, *e.g.*, outperforms the SOTA ShapeProp [45] by 1.1 mAP in the *nonvoc* → *voc* setting. By using the 3× schedule, ContrastMask (ResNet-101) achieves new SOTA performance of 38.4/34.3 mAP in the *nonvoc* → *voc* and *voc* → *nonvoc* settings. It outperforms CPMask [15] and ShapeMask [22] by 1.6/2.7 mAP, respectively, in the *nonvoc* → *voc* setting. Note that ShapeMask [22] adopts enhanced NAS-FPN [16] as the feature enhancement module to utilize multi-scale features.

We notice that the results of OPMask [2] are reported by adopting a heavier mask head, *i.e.*, 7 Conv layers, and a different training schedule, *i.e.*, 130k training iterations. We kindly refer readers to its arXiv version [3] (v1) for more comparison (They reported their result under the 3× schedule). Even OPMask adopts a heavier mask head, our ContrastMask still outperforms it. In addition, we also provide stronger results by using ResNeXt [40] backbones under the 3× schedule to show the potential of our method.

Qualitative results. Here, we visualize some example segmentation results of our method under two situations: with and without CL Head. We employ mask annotations from the *voc* subset to train our model. In Fig. 6, we show some samples from COCO-*val2017* dataset, including *voc* (*base*) and *nonvoc* (*novel*) categories. Our ContrastMask represents great capability to segment both of *base* and *novel* objects accurately. Even if objects are small and the background is clutter, our method still performs well. More visualization results are shown in the supplementary material.

4.3. Ablation Study

We conduct ablation studies to verify different designs of the components in our ContrastMask. Unless otherwise specified, we do ablations in the *nonvoc* → *voc* setting. All results are reported on *novel* (*voc*) categories.

Effectiveness of CL Head. Referring to Fig. 5, the input of the mask head in our ContrastMask is composed of three signals: feature map $\mathbf{X}$ from the backbone, feature map $\mathbf{Y}$ from the CL head and class activation map $\mathbf{A}$ from the CAM module. Here, we do an ablation study to show the benefit brought by each of the inputs. Since Mask R-

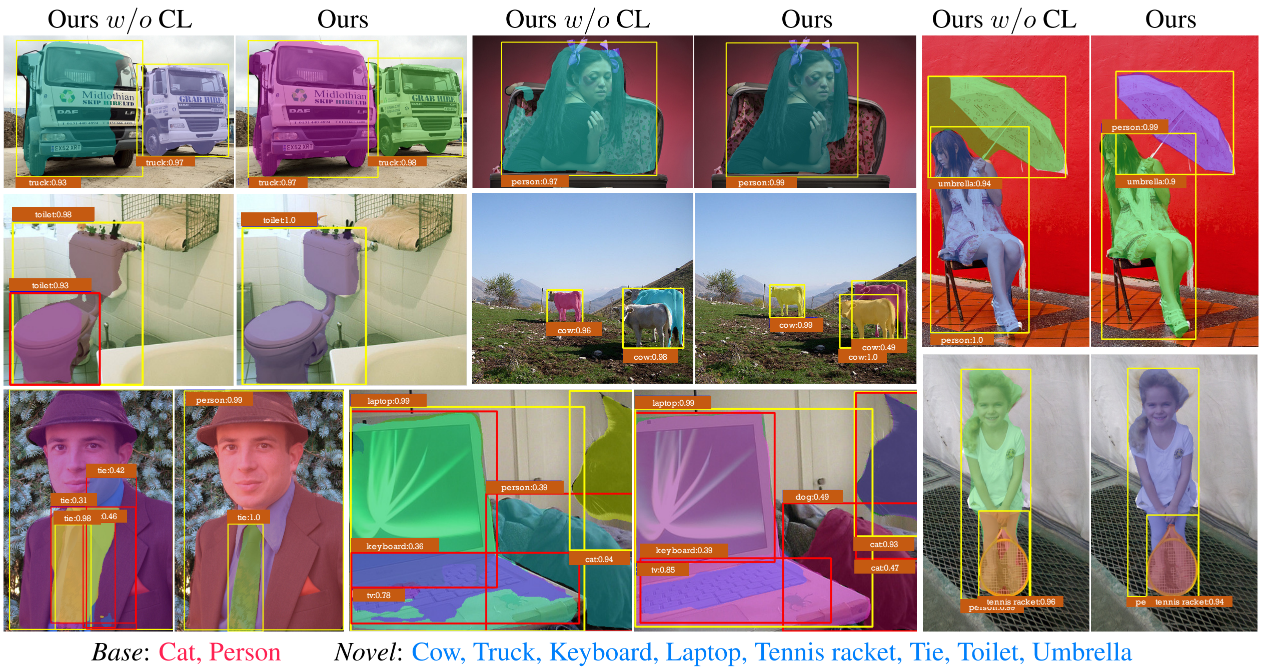

Figure 6. Qualitative results on COCO dataset when using *voc* as training data (*base*). Each group consists of two results, one is obtained by ContrastMask without CL Head (Ours w/o CL) and the other is obtained by ContrastMask (Ours). The results show that our ContrastMask performs more precisely segmentation on both *base* and *novel* objects benefited from the unified pixel-level contrastive learning framework conducted on all training data.

Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	23.9	42.9	23.5	11.6	24.3	33.7
Baseline + CM	32.3	57.6	31.9	15.2	31.6	44.6
Baseline + CM +CL	35.1	60.8	35.7	17.2	34.7	47.7

Table 2. **Ablation on the impact of each component.** The baseline is Mask R-CNN we built on. "CM" denotes CAM and "CL" represents for the CL head.

Architecture	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
C4F2	34.2	59.8	34.6	16.5	33.7	46.4
C8F3	35.1	60.8	35.7	17.2	34.7	47.7
C12F4	35.0	61.1	35.0	17.3	34.8	47.5

Table 3. **Ablation on the architecture of the CL head.** "CnFm" indicates n Conv-ReLU blocks in the encoder and m-layer MLP in the projector.

CNN [17] is our baseline, We first train it in a partially-supervised manner. The result is shown in Tab. 2. Then by involving the CAM module (CM) into the mask head, "Baseline + CM" obtains a much better result, 32.3 mAP, since CAM brings a latent cue for class-agnostic mask head to clearly point out which region is the foreground area. Furthermore, performance is boosted to 35.1 mAP after integrating the CL Head, termed as "CL", with the baseline model plus the CAM module. This evidences that the CL Head largely improves feature discrimination between foreground and background, and thus facilitates the learning of the class-agnostic mask segmentation model.

Architecture of CL Head. Since the input to our CLHead is ROI features from the backbone, unlike other contrastive learning methods, our encoder is relatively simpler and consists of several convolutional and linear layers. Here we ablate the architecture of the encoder. Tab. 3 illustrates different settings we explored. The base setting employs 4 Conv-ReLU blocks as the encoder and a two-layer MLP as the projector. After adding additional 4 Conv-ReLU blocks to the encoder and a one-layer MLP to the projector, an increase of 0.9 mAP (from 34.2 mAP to 35.1 mAP) is achieved, which explains that only 4 Conv-ReLU blocks are insufficient. When increasing the number of Conv-ReLU blocks to 12, the performance gain is limited. This indicates that adopting 8 Conv-ReLU blocks is adequate for an encoder, and more Conv-ReLU blocks bring limited benefits. Thus, we use "C8F3" as the architecture of CL Head, considering the trade-off between efficiency and accuracy.

Robustness of Sampling Ratio. A proportion ratio σ is applied to determine the number of sampled keys for each key set. Tab. 4 shows the performance change by varying the proportion ratio. When σ is too small or too large, *i.e.*, $\sigma = 0.1$ and $\sigma = 0.6$, performance is degraded. The reason is that a small σ means only a few keys can be sampled and a small number of keys can not realize an accurate

Sampling ratio σ	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.1	34.4	60.2	34.3	16.8	34.4	46.9
0.2	34.7	60.3	35.2	17.1	34.5	46.9
0.3	35.1	60.8	35.7	17.2	34.7	47.7
0.6	34.3	60.0	34.2	16.9	34.2	46.4

Table 4. **Discussion on the sample ratio σ.**

Temperature τ	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.1	34.4	60.4	35.0	16.7	34.1	46.9
0.7	35.1	60.8	35.7	17.2	34.7	47.7
0.9	34.0	60.2	33.7	16.8	33.4	46.4

Table 5. **Discussion on the temperature hyper-parameter.** we apply τ and $\tau' = 1 - \tau$ for easy and hard keys, respectively.

representation of foreground and background. A large σ encounters a dilemma that the rate of error keys will increase because the foreground and background partition for *novel* categories are produced by a predicted and coarse CAM. In general, a minor discrepancy arises among different σ, which demonstrates the robustness of our method to this hyper-parameter. We attribute this characteristic to the fact that only two classes, *i.e.*, foreground and background, are considered in our method, which requires a small number of keys to optimize the model.

Temperature hyper-parameter. We apply τ to easy keys and $\tau' = 1-\tau$ to hard keys when computing our contrastive loss. From Tab. 5, we notice that a very small τ is unsuitable for easy or hard keys, which leads to performance degradation. This can be explained from a perspective [33] that only a few negative keys near the query are focused when using a small τ, *i.e.*, $\tau = 0.1$. However, we expect more negative keys can be pushed away. Thus, we set $\tau = 0.7$ for easy keys and $\tau' = 1 - \tau = 0.3$ for hard keys.

Supervisions for our contrastive learning. In this study, we guide our query-sharing pixel-level contrastive learning by three different types of supervisions, *i.e.*, only *base*, only *novel* and *all*. As shown in Tab. 6, both only using *base* categories and only using *novel* categories to contribute in loss calculation lead to obvious performance drops, 1.6 mAP and 1.7 mAP respectively, compared with adopting *all* categories. This demonstrates that involving training data from all categories is important to learn a segmentation model with good generalization capability between *base* and *novel* categories.

Necessity of query-sharing. We ablate this experiment to validate the influence of the query-sharing strategy. In Tab. 7, "✗" means that we obtain different query $\mathbf{q}$ for different proposal, and thus the pixel-level contrastive loss is calculated for each proposal individually. It achieves worse performance compared with "✓", which indicates that the query-sharing strategy is essential for the proposed unified pixel-level contrastive learning framework.

Supervision	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
base	33.5	58.4	33.9	15.9	33.3	45.3
novel	33.4	58.0	34.2	15.8	33.1	45.8
all	35.1	60.8	35.7	17.2	34.7	47.7

Table 6. **Ablation on different supervision for our contrastive learning head.** "*base*", "*novel*" and "*all*" denote that only *base* categories, only "*novel*" categories and all categories are considered when calculating our contrastive loss, respectively.

Query-Sharing	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
✗	32.7	56.9	33.1	15.7	32.0	44.7
✓	35.1	60.8	35.7	17.2	34.7	47.7

Table 7. **Ablation on the necessity of query-sharing**.

5. Discussions

Since pseudo masks converted from CAMs are not accurate, the foreground and background partitions for *novel* categories are not guaranteed to be correct, which inevitably damages segmentation performance. If ground-truth masks for *novel* categories are available for sampling keys, an improvement about 1.4 mAP can be further achieved on the *voc* $\rightarrow$ *nonvoc* setting. There are two ways to approach this upper bound: 1) Utilizing stronger techniques to produce more precise pseudo masks. 2) Providing scribble or point annotations for *novel* categories, which are also cheaper than mask annotations. Besides, we also provide more discussions in the supplementary material, *e.g.*, possible application scenarios, relation to a teacher-student model, etc.

6. Conclusion

We developed an effective method for partially-supervised instance segmentation, named as ContrastMask, which introduces a unified pixel-level contrastive learning framework to learn a segmentation model on both *base* and *novel* categories. ContrastMask utilized a query-sharing pixel-level contrastive loss to make data from *novel* categories also contribute to the optimization process, and thus largely improved the feature discrimination between foreground and background areas for all categories. These enhanced features further facilitated the learning of the class-agnostic segmentation model, resulting in a better mask segmentor. Extensive results on the COCO dataset showed that ContrastMask consistently outperformed other methods by a large margin, achieving states-of-the-art under the partially-supervised setting.

Acknowledgments. This work was supported by NSFC 62176159, Natural Science Foundation of Shanghai 21ZR1432200, Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102 and the Fundamental Research Funds for the Central Universities.

References

[1] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C. Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8219–8228, October 2021. 2

[2] David Biertimpel, Sindi Shkodrani, Anil S. Baslamisli, and Nóra Baka. Prior to segment: Foreground cues for weakly annotated classes in partially supervised instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2824–2833, October 2021. 1, 2, 5, 6

[3] David Biertimpel, Sindi Shkodrani, Anil S. Baslamisli, and Nóra Baka. Prior to segment: Foreground cues for weakly annotated classes in partially supervised instance segmentation. *arXiv preprint arXiv:2011.11787v1*, 2021. 6

[4] Vighnesh Birodkar, Zhichao Lu, Siyang Li, Vivek Rathod, and Jonathan Huang. The surprising impact of mask-head architecture on novel class segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7015–7025, October 2021. 3, 6

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[6] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020. 2

[7] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4974–4983, 2019. 1, 2

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Internationa Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. 3

[11] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *European Conference on Computer Vision (ECCV)*, pages 660–676. Springer, 2020. 2

[12] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1

[13] Hao Ding, Siyuan Qiao, Alan Yuille, and Wei Shen. Deeply shape-guided cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8278–8288, June 2021. 1, 2

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, June 2010. 5

[15] Qi Fan, Lei Ke, Wenjie Pei, Chi-Keung Tang, and Yu-Wing Tai. Commonality-parsing network across shape and appearance for partially supervised instance segmentation. In *European Conference on Computer Vision (ECCV)*, pages 379–396. Springer, 2020. 1, 2, 5, 6

[16] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[17] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 3, 5, 6, 7

[18] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 6

[19] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6409–6418, 2019. 1

[20] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4019–4028, June 2021. 2

[21] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[22] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 6

[23] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. 2

[25] 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 (ECCV)*, pages 740–755. Springer, 2014. 2, 5

[26] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8759–8768, 2018. 1, 2

[27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016. 3

[28] Pedro O O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 4489–4500. Curran Associates, Inc., 2020. 2

[29] Pedro O O. Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to segment object candidates. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28. Curran Associates, Inc., 2015. 1

[30] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European Conference on Computer Vision (ECCV)*, pages 75–91. Springer, 2016. 1

[31] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020. 1, 2

[32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 9627–9636, 2019. 6

[33] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, June 2021. 4, 8

[34] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, pages 9929–9939. PMLR, 2020. 4

[35] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision (ECCV)*, pages 649–665. Springer, 2020. 2

[36] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 17721–17732. Curran Associates, Inc., 2020. 2

[37] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[38] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8392–8401, 2021. 2

[39] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 12193–12202, 2020. 1

[40] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 6

[41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16684–16693, 2021. 2

[42] Kai Zhao, Qi Han, Chang bin Zhang, Jun Xu, and Ming ming Cheng. Deep hough transform for semantic line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 1

[43] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7273–7282, October 2021. 2

[44] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 2

[45] 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 (CVPR)*, June 2020. 1, 2, 5, 6