ContrastMask: Contrastive Learning to Segment Every Thing

Xuehui Wang\textsuperscript{1†}, Kai Zhao\textsuperscript{2}, Ruixin Zhang\textsuperscript{2}, Shouhong Ding\textsuperscript{2}, Yan Wang\textsuperscript{3}, Wei Shen\textsuperscript{1(\textcopyright)}

\textsuperscript{1}MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
\textsuperscript{2}Youtu Lab, Tencent
\textsuperscript{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.

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-

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Visualization results of Mask R-CNN [17], OPMask [2] and the proposed ContrastMask on novel categories.}
\end{figure}

\textsuperscript{†}Work done during an internship at Youtu Lab, Tencent.
\textsuperscript{\textcopyright}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 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, MaskX 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.
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 $3 \times 3$ 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 $X \in \mathbb{R}^{H \times W \times C}$ 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 $Y = f(X) \in \mathbb{R}^{H \times W \times C}$ which will be incorporated into the mask head for mask segmentation. Next, $Y$ is first flattened and then fed into the projector, which maps representations to the space where the pixel-level contrastive loss is applied: $Z = g(Y) \in \mathbb{R}^{HW \times C}$. Here, after projection, the feature map $Z$ is reshaped to the same size as $Y$.

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 $q^+$ and the other for background $q^-$, 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:

$$L_{\mathcal{K}^+, \mathcal{K}^-} = -\frac{1}{|\mathcal{K}^+|} \sum_{k^+ \in \mathcal{K}^+} \left[ \phi(q^+, k^+)/\tau - \log \left( \exp(\phi(q^+, k^+)/\tau) + \sum_{k^- \in \mathcal{K}^-} \exp(\phi(q^+, k^-)/\tau) \right) \right],$$

where $\tau$ 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 $\delta$, then perform partition and randomly sample easy and hard keys based on partitions. The foreground query $q^+$ and background query $q^-$ 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 $L_{K^+,K^+}$. Next, we describe the details of how to obtain the foreground and background queries $q^+, q^-$, as well as the foreground and background key sets $K^+, K^-$. We illustrate these details in Fig. 4.

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

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

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

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

**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 a RoI proposal
According to Eq. (1), our query-sharing pixel-level contrastive loss is given an RoI proposal \(Z\). Now, instantiation of the proposed contrastive loss is as follows.

A key set \(K\) contains four types of key sets (Fig. 4) from it: 1) foreground-easy and easy key sets from it. Then consequently, we can obtain background key sets from it as well as how to mine hard categories, we have described how to obtain foreground and background query \(Z\) and foreground query \(A\).

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

\[
L = L_{\text{box}} + L_{\text{mask}} + \lambda L_{\text{con}},
\]

where \(\lambda\) 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 → 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 \(\sigma\) is set as \(\sigma = 0.3\), and the temperature hyper-parameter \(\tau\) (Eq. 1) for easy and hard keys are set as \(\tau = 0.7\) and \(\tau’ = 1 - \tau = 0.3\), respectively. We linearly warmup the \(\lambda\) of \(L_{\text{con}}\) (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-
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 \(1\times\) represents for 12 epochs and \(130k\) is a customized schedule only used in OPMask [2]. We use two conventional schedules, i.e., \(1\times\) and \(3\times\), 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 do not report results on VOC.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Sche.</th>
<th>Layers.</th>
<th>mAP</th>
<th>AP_{50}</th>
<th>AP_{75}</th>
<th>AP_{S}</th>
<th>AP_{M}</th>
<th>AP_{L}</th>
<th>AP_{50} __50</th>
<th>AP_{75} __50</th>
<th>AP_{S} __50</th>
<th>AP_{M} __50</th>
<th>AP_{L} __50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN (Baseline) [17]</td>
<td>ResNet-50</td>
<td>1x__</td>
<td>4</td>
<td>39.2</td>
<td>25.4</td>
<td>35.3</td>
<td>31.8</td>
<td>21.1</td>
<td>20.9</td>
<td>12.9</td>
<td>13.0</td>
<td>17.0</td>
<td>16.3</td>
<td>13.0</td>
</tr>
<tr>
<td>MaskX R-CNN [18]</td>
<td>ResNet-50</td>
<td>1x__</td>
<td>4</td>
<td>39.2</td>
<td>25.4</td>
<td>35.3</td>
<td>31.8</td>
<td>21.1</td>
<td>20.9</td>
<td>12.9</td>
<td>13.0</td>
<td>17.0</td>
<td>16.3</td>
<td>13.0</td>
</tr>
<tr>
<td>Mask GrabCut [21]</td>
<td>ResNet-50</td>
<td>1x__</td>
<td>4</td>
<td>39.2</td>
<td>25.4</td>
<td>35.3</td>
<td>31.8</td>
<td>21.1</td>
<td>20.9</td>
<td>12.9</td>
<td>13.0</td>
<td>17.0</td>
<td>16.3</td>
<td>13.0</td>
</tr>
<tr>
<td>CPMask [15]</td>
<td>ResNet-50</td>
<td>1x__</td>
<td>4</td>
<td>39.2</td>
<td>25.4</td>
<td>35.3</td>
<td>31.8</td>
<td>21.1</td>
<td>20.9</td>
<td>12.9</td>
<td>13.0</td>
<td>17.0</td>
<td>16.3</td>
<td>13.0</td>
</tr>
<tr>
<td>ShapeProp [45]</td>
<td>ResNet-50</td>
<td>1x__</td>
<td>4</td>
<td>39.2</td>
<td>25.4</td>
<td>35.3</td>
<td>31.8</td>
<td>21.1</td>
<td>20.9</td>
<td>12.9</td>
<td>13.0</td>
<td>17.0</td>
<td>16.3</td>
<td>13.0</td>
</tr>
<tr>
<td>ContrastMask (Ours)</td>
<td>ResNet-50</td>
<td>1x__</td>
<td>4</td>
<td>39.2</td>
<td>25.4</td>
<td>35.3</td>
<td>31.8</td>
<td>21.1</td>
<td>20.9</td>
<td>12.9</td>
<td>13.0</td>
<td>17.0</td>
<td>16.3</td>
<td>13.0</td>
</tr>
<tr>
<td>OPMask [2]</td>
<td>ResNet-50</td>
<td>130k</td>
<td>7</td>
<td>39.2</td>
<td>25.4</td>
<td>35.3</td>
<td>31.8</td>
<td>21.1</td>
<td>20.9</td>
<td>12.9</td>
<td>13.0</td>
<td>17.0</td>
<td>16.3</td>
<td>13.0</td>
</tr>
</tbody>
</table>

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 \(1\times\) 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 \(3\times\) 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 \(3\times\) schedule, ContrastMask (ResNet-101) achieves new SOTA performance of 38.4/34.5 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 \(3\times\) schedule). Even OPMask adopts a heavier mask head, our ContrastMask still outperforms it. In addition, we also provide stronger results by using ResNetXt [40] backbones under the \(3\times\) 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 X from the backbone, feature map Y from the CL head and class activation map 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 \textit{(base)}. Each group consists of two results, one is obtained by ContrastMask without CL Head \textit{(Ours w/o CL)} and the other is obtained by ContrastMask \textit{(Ours)}. The results show that our ContrastMask performs more precisely segmentation on both \textit{base} and \textit{novel} objects benefited from the unified pixel-level contrastive learning framework conducted on all training data.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AP$_{S}$</th>
<th>AP$_{M}$</th>
<th>AP$_{L}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>23.9</td>
<td>42.9</td>
<td>23.5</td>
<td>11.6</td>
<td>24.3</td>
<td>33.7</td>
</tr>
<tr>
<td>Baseline + CM</td>
<td>32.3</td>
<td>57.6</td>
<td>31.9</td>
<td>15.2</td>
<td>31.6</td>
<td>44.6</td>
</tr>
<tr>
<td>Baseline + CM + CL</td>
<td>35.1</td>
<td>60.8</td>
<td>35.7</td>
<td>17.2</td>
<td>34.7</td>
<td>47.7</td>
</tr>
</tbody>
</table>

Table 2. \textbf{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.

CNN \cite{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.

\textbf{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.

\textbf{Robustness of Sampling Ratio}. A proportion ratio $\sigma$ 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 $\sigma$ is too small or too large, \textit{i.e.}, $\sigma = 0.1$ and $\sigma = 0.6$, performance is degraded. The reason is that a small $\sigma$ means only a few keys can be sampled and a small number of keys can not realize an accurate
Table 4. Discussion on the sample ratio σ.

<table>
<thead>
<tr>
<th>Sampling ratio σ</th>
<th>mAP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AP$_S$</th>
<th>AP$_M$</th>
<th>AP$_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>34.4</td>
<td>60.2</td>
<td>34.3</td>
<td>16.8</td>
<td>34.4</td>
<td>46.9</td>
</tr>
<tr>
<td>0.2</td>
<td>34.7</td>
<td>60.3</td>
<td>35.2</td>
<td>17.1</td>
<td>34.5</td>
<td>46.9</td>
</tr>
<tr>
<td>0.3</td>
<td>35.1</td>
<td>60.8</td>
<td>35.7</td>
<td>17.2</td>
<td>34.7</td>
<td>47.7</td>
</tr>
<tr>
<td>0.6</td>
<td>34.3</td>
<td>60.0</td>
<td>34.2</td>
<td>16.9</td>
<td>34.2</td>
<td>46.4</td>
</tr>
</tbody>
</table>

Table 5. Discussion on the temperature hyper-parameter. We apply τ and τ' = 1 − τ for easy and hard keys, respectively.

<table>
<thead>
<tr>
<th>Temperature τ</th>
<th>mAP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AP$_S$</th>
<th>AP$_M$</th>
<th>AP$_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>34.4</td>
<td>60.4</td>
<td>35.0</td>
<td>16.7</td>
<td>34.1</td>
<td>46.9</td>
</tr>
<tr>
<td>0.7</td>
<td>35.1</td>
<td>60.8</td>
<td>35.7</td>
<td>17.2</td>
<td>34.7</td>
<td>47.7</td>
</tr>
<tr>
<td>0.9</td>
<td>34.0</td>
<td>60.2</td>
<td>33.7</td>
<td>16.8</td>
<td>33.4</td>
<td>46.4</td>
</tr>
</tbody>
</table>

Table 6. Ablation on the necessity of query-sharing.

<table>
<thead>
<tr>
<th>Supervision</th>
<th>mAP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AP$_S$</th>
<th>AP$_M$</th>
<th>AP$_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>33.5</td>
<td>58.4</td>
<td>33.9</td>
<td>15.9</td>
<td>33.3</td>
<td>45.3</td>
</tr>
<tr>
<td>novel</td>
<td>33.4</td>
<td>58.0</td>
<td>34.2</td>
<td>15.8</td>
<td>33.1</td>
<td>45.8</td>
</tr>
<tr>
<td>all</td>
<td>35.1</td>
<td>60.8</td>
<td>35.7</td>
<td>17.2</td>
<td>34.7</td>
<td>47.7</td>
</tr>
</tbody>
</table>

Table 5. Discussion on the sample ratio σ.

Table 7. Discussion on the temperature hyper-parameter. We apply τ and τ' = 1 − τ for easy and hard keys, respectively.

The 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 τ' = 1 − τ 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., τ = 0.1. However, we expect more negative keys can be pushed away. Thus, we set τ = 0.7 for easy keys and τ' = 1 − τ = 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 to 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 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.

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 → 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 2021SHZDX0102 and the Fundamental Research Funds for the Central Universities.
References


