Commonality-Parsing Network across Shape and Appearance for Partially Supervised Instance Segmentation

Qi Fan[*1], Lei Ke[*1], Wenjie Pei[†2], Chi-Keung Tang[1], and Yu-Wing Tai[1,3]

[1] Hong Kong University of Science and Technology
{qfanaa, lkeab, cktang, yuwing}@cse.ust.hk
[2] Harbin Institute of Technology, Shenzhen
{wenjiecoder}@gmail.com
[3] Kwai Inc.

Abstract. Partially supervised instance segmentation aims to perform learning on limited mask-annotated categories of data thus eliminating expensive and exhaustive mask annotation. The learned models are expected to be generalizable to novel categories. Existing methods either learn a transfer function from detection to segmentation, or cluster shape priors for segmenting novel categories. We propose to learn the underlying class-agnostic commonalities that can be generalized from mask-annotated categories to novel categories. Specifically, we parse two types of commonalities: 1) shape commonalities which are learned by performing supervised learning on instance boundary prediction; and 2) appearance commonalities which are captured by modeling pairwise affinities among pixels of feature maps to optimize the separability between instance and the background. Incorporating both the shape and appearance commonalities, our model significantly outperforms the state-of-the-art methods on both partially supervised setting and few-shot setting for instance segmentation on COCO dataset. The code is available at https://github.com/fanq15/CPMask.

Keywords: Partially supervised; Few-shot; Instance segmentation

1 Introduction

Instance segmentation is a fundamental research topic in computer vision due to its extensive applications ranging from object selection [31], image editing [44,46] to scene understanding [30]. Typical methods [8,21,24,32,36] for instance segmentation have achieved remarkable progress, relying on the fully supervised learning on the precise mask-annotated data. However, this kind of pixel-level mask annotation is extremely labor-consuming and thus expensive to be performed on large amount of data which is typically required for deep learning methods. On the other hand, it is less expensive and more feasible to perform annotation of

---

[*] Equal contribution
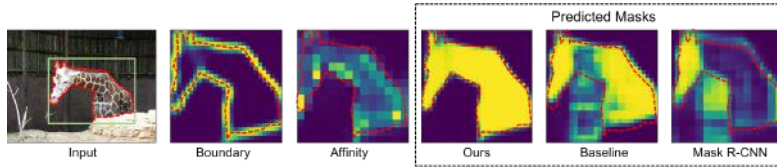[†] Corresponding author: Wenjie Pei

Fig. 1. Given an input image, our model captures shape commonalities by predicting instance boundaries and learns the appearance commonalities by modeling pairwise affinities among all pixels. The learned class-agnostic commonalities in both shape and appearance enable our model to segment more accurate mask than other models. Note that the similar background (wall color) misguides other methods. Herein, the affinity heatmap encodes the mean of the affinity maps for each instance pixel. The baseline model refers to basis framework of our model without commonality-parsing modules. The red dash line indicates the ground-truth of annotated mask.

bounding box for instances, which motivates the newly proposed task: *partially supervised instance segmentation* [23,28]. It aims to learn instance segmentation models on limited mask-annotated categories of data, which can be generalized to new (novel) categories with only bounding-box annotations available. The partially supervised instance segmentation is much more challenging than the typical instance segmentation in full supervision. The major difficulty lies in how to learn the class-agnostic features for instance segmentation that can be generalized from the mask-annotated categories to novel categories.

A straightforward way for partially supervised instance segmentation is to directly extend existing fully supervised algorithms to segmentation of novel categories by class-agnostic training [39,40], which treats all mask-annotated categories of instances involved in training as one foreground category and forces the model to learn to distinguish between foreground and background regions for segmentation. This brute-force way of class-agnostic training expects the model to learn all the generalized features between annotated and novel categories by itself, which is hardly achieved. As the initiator of the partially supervised instance segmentation, Mask^X R-CNN [23] transfers the visual information from the modeling of bounding box to the mask head through a parameterized transfer function. Subsequently, ShapeMask [28] seeks to extract the generic class-agnostic shape features across different categories by summarizing a collection of shape priors as reference for segmenting new categories.

Whilst both Mask^X R-CNN and ShapeMask have distinctly advanced the performance of partially supervised instance segmentation, there are two important features have not been fully exploited. First, the generalized **appearance** features that shared across different categories, e.g., similar hairy body surface between dogs and cats or similar textures on the furniture surface, are not explicitly explored. These class-agnostic appearance features can be potentially generalized from mask-annotated categories of data to novel categories for segmentation. Second, the common **shape** features that can be generalized across different categories are not explicitly learned in a supervised way, though ShapeMask refines the shape priors by simply clustering the annotated masks and adapts them to a given novel object. In this work we intend to tackle the partially supervised instance segmentation by fully exploiting these two features.

We propose to capture the underlying commonalities which can be generalized across different categories by supervised learning for partially supervised instance segmentation. In particular, we aim to learn two types of generalized commonalities: 1) the shape commonalities that can be generalized between different categories like similar instance contour or similar instance boundary features; 2) the appearance commonalities that shared among categories of instances owning similar appearance features such as similar texture or similar color distribution. The resulting model, Commonality-Parsing Network (denoted as CPMask), can be trained in an end-to-end manner. Consider the example in Fig. 1, to segment the giraffe in the red bounding box, our model extracts its shape information by predicting the boundaries of giraffe and captures the appearance information by modeling the pairwise affinities among pixels. Taking into account both the shape and appearance information, our model is able to predict more accurate segmentation mask than other models. It is worth noting that although giraffe is a novel category whose mask-annotation is not provided in the training data, our model is able to accurately predict its boundary and affinity due to the learned class-agnostic commonalities w.r.t. both shape and appearance information.

We evaluate our model on two settings on COCO dataset: 1) partially-supervised instance segmentation, in which partial categories are provided with the ground-truth for both bounding boxes and segmentation masks while the other (novel) categories are only provided with the annotated bounding boxes during training; 2) few-shot instance segmentation, in which each of the novel categories only contain a small number of training samples (with both annotated bounding boxes and masks). Our model outperforms the state-of-the-art performance significantly on both settings. We further qualitatively demonstrate the generalization ability of our model by directly applying our trained model on COCO dataset to other 9 datasets with various scenes. It is worth mentioning that our model is more effective given fewer mask-annotated categories of training data compared to methods for fully supervised (routine) instance segmentation. To conclude, our contributions includes:

- We design a supervised learning mechanism for predicting instance boundaries to learn the class-agnostic shape commonalities that can be generalized from mask-annotated categories to novel categories.
- We propose to model the affinities among pixels of feature maps in a supervised way to optimize the separability between the instance region and the background and learn the class-agnostic appearance commonalities that can be generalized to novel objects.
- Incorporating both learned shape and appearance commonalities, our model substantially outperforms state-of-the-art methods on COCO dataset for instance segmentation in both partially supervised and few-shot setting.

2   Related Works

Conventional Instance Segmentation. is fully supervised by numerous high-quality pixel-level annotations [10,11,17,18,19,20,39,40]. Lots of methods have

made great progress on this task by embracing the classical "detect then segment" paradigm, which first generates detection results using the powerful two-stage detector and then segments each object in the bounding box. Mask R-CNN [21] attaches one simple mask predictor on Faster R-CNN [43] to segment each object in the box. PANet [36] merges multi-level features to enhance the performance. FCIS [32] and MaskLab [8] use position-sensitive score maps to encode the segmentation information. Kong and Fowlkes [27] propose to use pairwise pixel affinity for instance segmentation. Mask Scoring R-CNN [24] introduces a mask IoU branch to predict the mask quality and then selects good mask results accordingly. HTC [6] fully leverages the relationship between detection and segmentation to build a successful instance segmentation cascade network. Most recently, some works attempt to build instance segmentation network on the one-stage detector [34,45] for its simplicity and efficiency. In YOLACT [5], a set of prototype masks and coefficients are used to assemble masks for each instance. CenterMask [29] builds an attention-based mask branch on FCOS [45] for fast mask prediction. Compared to these previous works, our model mainly targets for novel objects segmentation by learning shape and appearance commonalities, although it also achieves superior performance in the fully supervised task.

Instance Segmentation for Novel Objects. Generalizing instance segmentation model to novel categories with limited annotations is meaningful and challenging, which mainly has three different settings: **Weakly supervised** instance segmentation methods are developed to use weak labels to segment novel categories where the training samples are only annotated with bounding boxes [26,42] or image-level labels [1,57] without pixel-level annotations. **Few-shot supervised** instance segmentation [52] is proposed to solve this problem by imitating the human visual systems to learn new visual concepts with only a few well-annotated samples. **Partially supervised** instance segmentation is formulated in a mixture of strongly and weakly annotated scenario where only a small subset of base categories are well-annotated with both box and mask annotations while the novel categories only have box annotations. In Mask^X R-CNN [23], a parameterized weight transfer function is designed to transfer the visual information from detection to segmentation while ShapeMask [28] learns the intermediate concept of object shape as the prior knowledge. Different from the above two works, which solve the partially supervised segmentation task either from transfer learning perspective or utilizing additional shape priors, our model focuses on learning class-agnostic features with great generalization ability by parsing the shape and appearance commonalities and clearly outperforms the existing methods by a large margin.

3   Commonality-Parsing Network

The crux of performing novel instance segmentation is to learn the underlying commonalities that can be generalized from the mask-annotated categories to novel categories. To surmount this crux, our Commonality-Parsing Network performs class-agnostic learning for partially supervised instance segmentation by
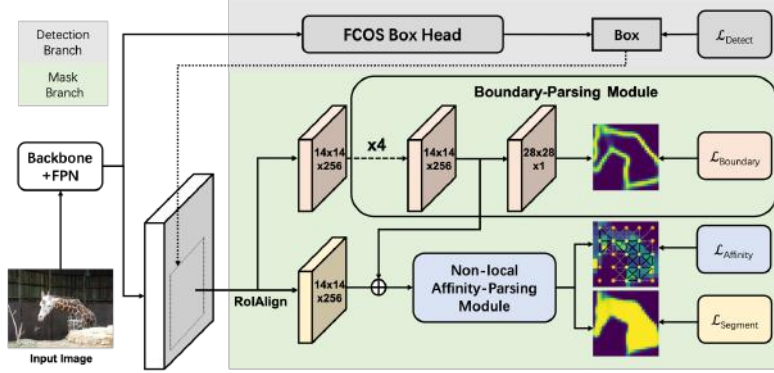
Fig. 2. Architecture of our Commonality-Parsing Network which consists of a detection branch and a mask branch. The cropped RoI feature based on predicted bounding boxes is first processed by Boundary-Parsing Module of the mask branch for predicting instance boundaries to guide the learning of shape commonalities in intermediate feature maps. Then the feature maps are fed into Non-local Affinity-Parsing Module (presented in Fig. 3) to learn the appearance commonalities by modeling the pairwise affinities among all pixels in feature maps. Finally, the feature maps incorporating both shape and appearance commonalities are used for mask prediction.

two proposed modules: 1) Boundary-Parsing Module for learning shape commonalities and 2) Non-local Affinity-Parsing Module for learning appearance commonalities. We will first present the overall framework of the proposed Commonality-Parsing Network, then we will elaborate on the aforementioned two modules specifically designed for class-agnostic learning.

3.1 Class-Agnostic Learning Framework

Fig. 2 presents the architecture of our Commonality-Parsing Network. Following typical models [8,21,36] for instance segmentation, our model contains two branches: 1) the object detection branch in charge of predicting bounding boxes as instance proposals, and 2) the mask branch for predicting segmented masks for the instance proposals obtained from the object detection branch.

We adopt FCOS [45], which is an excellent one-stage detection model, as our object detection backbone. As illustrated in Fig. 2, a backbone network equipped with FPN [33] is first employed to extract intermediate convolutional features for downstream processing. The object detection branch is then utilized to predict bounding boxes with positions as well as categories for potential instances. In the training phrase, supervision on both the position prediction and the category classification is performed to guide the optimization of the backbone network and FPN as in [45]:

$$\mathcal{L}_{\text{Detect}} = \mathcal{L}_{\text{regression}} + \mathcal{L}_{\text{centerness}} + \mathcal{L}_{\text{classification}}. \tag{1}$$

The mask branch is responsible for segmenting each of target instances predicted by the object detection branch. It is composed of two core modules designed specifically for class-agnostic learning by parsing the commonalities across both the shape and appearance features: Boundary-Parsing Module and Non-local Affinity-Parsing Module. These two modules are trained on a small set of

mask-annotated categories of data (termed as base categories) and the learned inter-category commonality of both shape and appearance information enables our model to perform instance segmentation on novel categories of image data.

3.2   Boundary-Parsing Module for Learning Shape Commonality

Boundary-Parsing Module is designed to learn the underlying commonalities with respect to the shape information that can be generalized from the mask-annotated categories to mask-unseen novel categories of data. Specifically, the Boundary-Parsing Module focuses on learning to predict the boundaries between the instance (foreground) and the background. The rationale behind this design is that there are common shape features shared among different categories on discrimination of the instance-background boundaries, which can be leveraged during class-agnostic learning for instance segmentation of novel categories. Besides, accurate boundary localization is able to explicitly contribute to the mask prediction for segmentation, which has been proved by many works [2,3,4,7,12,37,41,47,54,56]. Hence, we perform supervised learning for the prediction of boundaries to learn the shape commonalities among different categories.

There are several ways to design the structure of Boundary-Parsing Module and we just investigate a straightforward yet effective way: four 3×3 convolutional layers with ReLU as the activation functions, followed by one upsampling layer and one 1×1 convolutional layer to output one channel of feature map as boundary predictions. The Boundary-Parsing Module is trained with the boundary loss:

$$\mathcal{L}_{\mathrm{boundary}} = \mathcal{L}_{\mathrm{BCE}}(\mathcal{F}_B(\mathbf{X}), \mathcal{GT}_B), \tag{2}$$

where $\mathcal{L}_{\mathrm{BCE}}$ denotes the binary cross-entropy loss, $\mathcal{F}_B$ denotes the nonlinear transformation functions by Boundary-Parsing Module, $\mathbf{X}$ is the RoI feature cropped by the *RoIAlign* operation corresponding to a target instance predicted by the object detection branch and $\mathcal{GT}_B$ is the off-the-shelf boundary ground-truth that can be readily obtained from mask annotations.

3.3   Non-local Affinity-Parsing Module for Learning Appearance Commonality

Similar categories tend to share similar appearance commonality, e.g., similar hairy body surface between dogs and cats, or similar texture on the furniture surface. This kind of appearance commonalities can be leveraged for class-agnostic learning to generalize the instance segmentation to novel categories. Therefore, we propose Non-local Affinity-Parsing Module to learn the appearance commonalities across different categories by parsing the affinities among pixels of feature maps in a non-local way. The pixels belonging to an instance (in the foreground region) are expected to have much closer affinities than the affinities between foreground and background pixels.

Formally, given the RoI feature $\mathbf{X}$ after *RoIAlign* operation for an instance proposal, we first fuse it with the output feature maps $\mathcal{F}_B(\mathbf{X})$ from Boundary-Parsing Module by a simple attention module which incorporates the shape
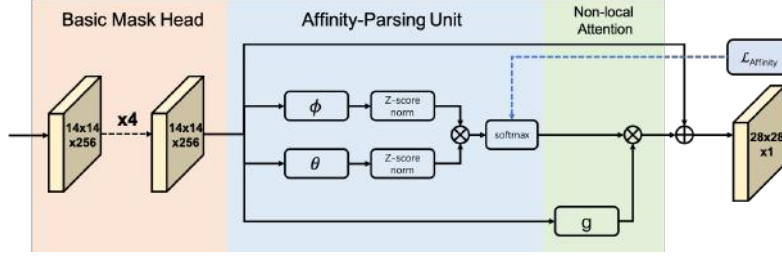
Fig. 3. Architecture of our Non-local Affinity-Parsing Module, which is composed of three units. The Basic Mask Head processes the input feature with four convolutional layers with 3×3 kernel and ReLU. Subsequently, the Affinity-Parsing unit performs supervised learning to model the pairwise affinities among pixels in feature maps. Finally, the non-local attention is employed to coordinate feature maps based on affinities to enable our model perceive more context information and increase the appearance separation between the instance and the background. Herein, "$\otimes$" denotes the matrix multiplication and "$\oplus$" represents element-wise addition.

commonality information by weighted element-wise additions. Then the non-linear transformation $\mathcal{G}$ by four convolutional layers is performed on the fused features as a basic mask head of operations:

$$\mathbf{C} = \mathcal{G}(\mathbf{X} \oplus \mathcal{F}_B(\mathbf{X}))). \tag{3}$$

The obtained feature maps $\mathbf{C} \in \mathrm{R}^{c \times h \times w}$, with c feature maps of size $h \times w$, is then fed into the non-local affinity-parsing unit for modeling affinity. Specifically, we model the affinity between the pixel at (i, j) and the pixel at (m, n) in a latent embedding space by:

$$\mathbf{A}_{(<i,j>,<m,n>)} = f\left[\frac{(\theta(\mathbf{C}_{i,j}) - \mu_{i,j})}{\sigma_{i,j}}, \frac{(\phi(\mathbf{C}_{m,n}) - \mu_{m,n})}{\sigma_{m,n}}\right], \tag{4}$$

where $\mathbf{C}_{i,j} \in \mathrm{R}^c$ corresponds to the vectorial representation (in channel dimension) for the pixel at (i, j) and the same goes for $\mathbf{C}_{m,n}$. Herein, θ, ϕ are embedding functions and f is a kernel function for encoding affinity. In practice, we opt for the dot-product operator for f, which is a typical way of modeling similarity. μ and σ are the mean value and the standard deviation respectively. Note that here we apply the *z-score* normalization for both $\theta(\mathbf{C}_{i,j})$ and $\phi(\mathbf{C}_{m,n})$ to ease the convergence during optimization.

Larger affinity value indicates closer relationship while smaller affinity value implies larger difference. We expect that the affinities between pixels belonging to an instance (foreground) region are much higher than that between foreground and background pixels. To this end, we introduce a supervision signal to guide the optimization to achieve the desired affinity distribution. In particular, we impose an affinity constraint to maximize the affinities among pixels in the foreground region Fg and minimize the affinities between foreground Fg and background Bg pixels:

$$\mathbf{A} = \mathrm{softmax}(\mathbf{A}),$$
$$\mathcal{L}_{\mathrm{Affinity}} = \mathcal{L}_1(1, \sum_{\substack{<i,j>\in Fg \\ <m,n>\in Fg}} \mathbf{A}_{<i,j>,<m,n>}) + \mathcal{L}_1(0, \sum_{\substack{<i,j>\in Fg \\ <m,n>\in Bg}} \mathbf{A}_{<i,j>,<m,n>}). \tag{5}$$

Here we first normalize $\mathbf{A}$ using a *softmax* operator and then impose the loss function that encourages the sum of affinities among foreground pixels to be close to 1 for more appearance affinities while pushing the affinities between foreground and background pixels to be 0 for larger appearance separation.

The supervised learning on the affinity distribution enables our model to perceive the appearance separability between the foreground (instance) and background regions. To further increase this appearance separation, we propose to coordinate feature maps by explicitly incorporating the learned affinities in a non-local attention manner [48,55]:

$$\widetilde{\mathbf{C}}_{i,j} = \sum_{\forall <m,n>} \mathbf{A}_{<i,j>,<m,n>} \cdot g(\mathbf{C}_{m,n}), \tag{6}$$

where g is a embedding function. Here we coordinate the vectorial representation for the pixel at (i,j) in the feature maps by attending each pixel with the corresponding affinity. Such coordination on feature maps enables our model to perceive the context of whole image region with affinity-based attention, thus resulting in more separation of appearance between foreground and background and closer affinities among pixels in foreground (instance) region, which is beneficial for learning appearance commonalities and instance segmentation.

Together with original feature maps $\mathbf{C}$, the output coordinated feature maps $\tilde{\mathbf{C}}$ from the Non-local Affinity-Parsing Module is subsequently fed into one upsampling layer and one 1×1 convolutional layer for the final prediction of segmented mask:

$$\mathcal{L}_{\text{Segment}} = \mathcal{L}_{\text{BCE}}(\mathcal{F}_{1 \times 1 conv}(\tilde{\mathbf{C}} \oplus \mathbf{C}), \mathcal{GT}_S), \tag{7}$$

where $\mathcal{F}_{1 \times 1 conv}$ denotes the nonlinear transformation functions by 1×1 convolutional layer and $\mathcal{GT}_S$ is the ground-truth mask annotations.

3.4   End-to-End Parameter Learning

The whole model of our Commonality-Parsing Network can be trained in an end-to-end manner on two different types of training data:

- For the mask-annotated training data in base categories, the model is optimized by integrating all the aforementioned loss functions:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{Detect}} + \lambda_2 \mathcal{L}_{\text{Boundary}} + \lambda_3 \mathcal{L}_{\text{Affinity}} + \lambda_4 \mathcal{L}_{\text{Segment}}, \tag{8}$$

  where λ_1, λ_2, λ_3, and λ_4 are hyper-parameter weights to balance the loss functions. In our implementation, they are tuned to be $\{1, 0.5, 0.5, 1\}$ respectively on a validation set.
- For the training data without mask-annotation in novel categories, we train the model with only detection loss, i.e., only the parameters in backbone network, FPN and detection branch are optimized:

$$\mathcal{L} = \mathcal{L}_{\text{Detect}}. \tag{9}$$

4   Experiments

We conduct experiments on MS COCO dataset [35] to evaluate our model. We first perform ablation study to investigate the effect of Boundary-Parsing Module and Non-local Affinity-Parsing Module, then we compare our model with state-of-the-art methods in three different settings for instance segmentation: 1) partially supervised setting, 2) few-shot setting and 3) fully supervised setting.

model		voc → non-voc				
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	20.7	37.9	20.4	10.6	24.7	27.3
Baseline + BM w/o FF	21.6	38.8	21.1	11.6	26.5	28.8
Baseline + BM	27.4	45.1	28.7	12.4	32.3	39.5
Baseline + AM w/o AL	26.9	45.0	27.8	11.6	31.3	39.2
Baseline + AM	27.2	45.2	28.3	11.7	31.5	40.3
Baseline + BM + AM	28.8	46.1	30.6	12.4	33.1	43.4

Table 1. Experimental results of ablation studies on the COCO *val* set. The models are trained on the *voc* base categories and evaluated on the *non-voc* novel categories. The "BM" denotes the Boundary-Parsing Module, the "AM" denotes the Non-local Affinity-Parsing Module, the "FF" denotes fusing boundary feature to the mask head and the "AL" denotes the affinity loss.

4.1 Experimental Setup

Evaluation Protocol. We follow the typical data split on COCO in our experiment: *train2017* for training and *val2017* for test. In both of our experiments on partially supervised setting and few-shot setting, we split the 80 COCO categories into "*voc*" and "*non-voc*" category subsets where the *voc* categories are those in PASCAL VOC [13] dataset while the remaining categories are included in the *non-voc* categories. Each time we select classes in one subset as base categories with annotations of both bounding boxes and masks, and those in the other subset as novel categories. Note that the training samples of novel categories have only bounding box annotation (no mask annotation) for partially supervised setting. For few-shot setting, each novel category in the training data only contains a small amount of samples with annotations of both bounding boxes and masks. We adopt the typical evaluation metrics for instance segmentation in our experiments, i.e., AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L.

Implementation Details. SGD with Momentum is employed for training our model, starting with 1 K constant warm-up iterations. The batch size is set to 16 and initial learning rate is set to 0.01. For efficiency, ResNet-50 [22] is used as backbone network for ablation study and the input images are resized in such a way that the short side and long side are no more than 600 and 1000 pixels respectively (denoted as (600, 1000)). For other experiments on comparison with other methods, ResNet-101 [22] backbone with multi-scale training is employed.

4.2 Ablation Study

We investigate the effectiveness of our Boundary-Parsing Module and Non-local Affinity-Parsing Module by carrying out ablation experiments for partially supervised instance segmentation in this section. The *voc* classes is used as base categories and the *non-voc* as novel categories. We refer to the variant of our model without Boundary-Parsing Module and Non-local Affinity-Parsing Module as *Baseline* model. The class-agnostic version of Mask R-CNN [23] is compared for reference in this section.

Quantitative Evaluation. Table 1 presents the experimental results. The baseline model obtains 20.7 AP on the novel categories. Boundary-Parsing Module improves the performance by 6.7 AP and explicitly adopting the boundary feature to guide the mask prediction is crucial for the overall performance. Non-local
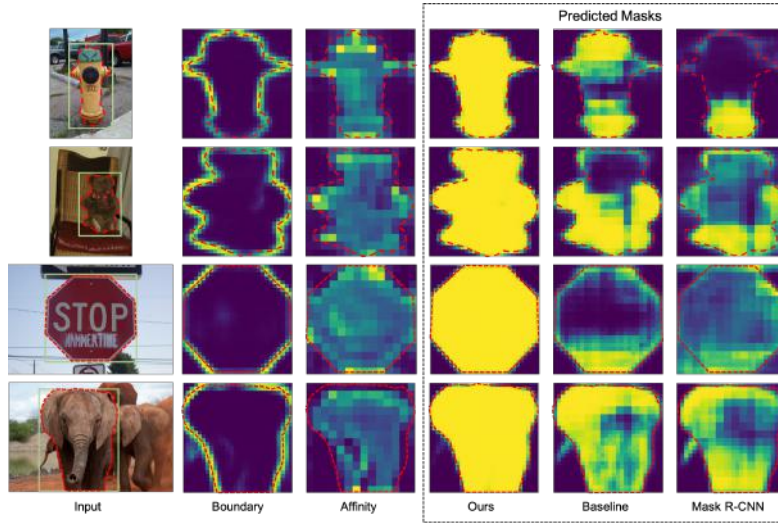
Fig. 4. Visualization of boundary heatmaps and affinity heatmaps learned by our model for four novel categories of cases. The red dash lines indicate the ground-truth mask. The affinity heatmap is obtained by calculating the mean of the affinity maps for each instance pixel. The yellow color in the heatmap indicates higher response value and the blue color indicates lower response value.

Affinity-Parsing Module promotes the performance by 6.2 AP and the better pixel relationship introduced by the affinity loss further boosts the performance to 27.2 AP. Both the shape and appearance commonalities learned by these two modules from the base categories generalize well to the novel categories. After integrating both modules, our model achieves 28.8 AP which is distinctly better than the performance by each individual module. It implies that the learned shape and appearance commonalities contribute in their own way for instance segmentation.

Qualitative Evaluation. To further reveal the mechanism of these two modules, we visualize boundary and affinity heatmaps on novel categories in Fig. 4. The affinity heatmap is obtained by calculating the mean of the affinity maps for each instance pixel in the Non-local Affinity-Parsing Module. We observe that our model is able to accurately estimate instance boundaries to capture the shape commonalities. Meanwhile, the affinities between instance pixels are evidently higher (closer) than affinities between instance and background pixels, which indicates that appearance commonalities are well learned via affinity modeling for these novel categories. Both of the shape and appearance commonalities can help our model to segment novel instances from background, because the commonalities learned from these modules are successfully generalized from base categories to novel categories. By contrast, the baseline model without these two modules and the Mask R-CNN for fully supervised instance segmentation performs quite poorly on these cases.

Complementary advantages of Boundary (Shape) and Affinity (Appearance). We present two challenging examples in Fig. 5. Our model cannot
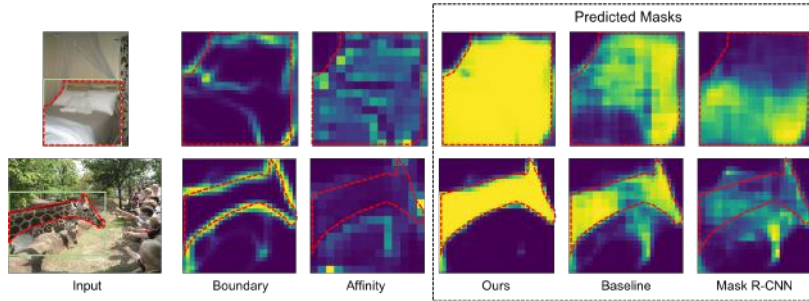
Fig. 5. Two Challenging examples that indicate the complementation between Boundary-Parsing Module and Non-local Affinity-Parsing Module.
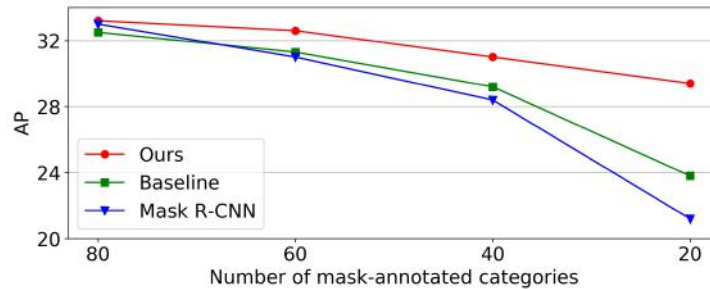

Fig. 6. The segmentation performance of different models on a fixed set of novel categories as a function of number of mask-annotated (base) categories. The novel categories are randomly selected from COCO dataset.

precisely estimate the boundary for the bed due to the confusing color differences. On the other hand, our Non-local Affinity-Parsing Module can tackle this problem very well based on appearance commonalities, which lead to the accurate mask prediction for bed. On the other hand, the affinity heat map is not precise for the giraffe due to the similar appearance of another overlapping giraffe behind. In such scenario with multiple instances of same category within a box, our Boundary-Parsing Module can still predict the boundaries very accurately.

Evaluation of Generalization. To further evaluate the ability of generalization from base (mask-annotated) categories to novel categories for our model, we conduct experiments to investigate the effect of varying the number of mask-annotated categories in Fig. 6. The performances of the both baseline model and Mask R-CNN decay much faster than our model as the number of base categories for training decreases, which indicates that our method is particularly more effective given fewer annotated categories of training data compared to fully supervised methods and benefits from the class-agnostic learning of our model by Boundary-Parsing Module and Non-local Affinity-Parsing Module.

4.3   Partially Supervised Instance Segmentation

In this section we compare our model to other state-of-the-art methods for partially supervised instance segmentation.

Table 2 presents the quantitative results on COCO dataset with two sets of experiments: use *voc* or *non-voc* classes as the base categories and treat the

method	voc → non-voc						non-voc → voc					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN [21]	18.5	34.8	18.1	11.3	23.4	21.7	24.7	43.5	24.9	11.4	25.7	35.1
Mask GrabCut [23]	19.7	39.7	17.0	6.4	21.2	35.8	19.6	46.1	14.3	5.1	16.0	32.4
Mask^X R-CNN [23]	23.8	42.9	23.5	12.7	28.1	33.5	29.5	52.4	29.7	13.4	30.2	41.0
ShapeMask [28]	30.2	49.3	31.5	16.1	38.2	38.4	33.3	56.9	34.3	17.1	38.1	45.4
ShapeMask (NAS-FPN) [28]	33.2	53.1	35.0	18.3	**40.2**	43.3	35.7	60.3	36.6	**18.3**	**40.5**	47.3
CPMask (Ours)	**34.0**	**53.7**	**36.5**	**18.5**	38.9	**47.4**	**36.8**	**60.5**	**38.6**	17.6	37.1	**51.5**
Oracle Mask^X R-CNN [23]	34.4	55.2	36.3	15.5	39.0	52.6	39.1	64.5	41.4	16.3	38.1	55.1
Oracle ShapeMask [28]	35.0	53.9	37.5	17.3	41.0	49.0	40.9	65.1	43.4	18.5	41.9	56.6
Oracle ShapeMask (NAS-FPN) [28]	37.6	57.7	40.2	20.1	44.4	51.1	43.1	67.9	45.8	20.1	44.3	57.8
Oracle CPMask (Ours)	37.6	58.2	40.2	19.9	42.6	54.2	42.9	67.6	46.6	21.6	42.1	58.9

Table 2. Experimental results of partially supervised instance segmentation on the COCO *val* set. The "voc → non-voc" means that we use the *voc* classes as base categories and the *non-voc* as novel categories, and vice versa. The oracle models indicates the upper-bound performance for reference which are trained on masks from all categories (in full supervision).

Fig. 7. Qualitative results on novel COCO categories. We use *voc* classes as the base (mask-annotated) categories for training.

remaining classes as novel categories. Our model outperforms the state-of-the-art ShapeMask by a large margin: 3.8 AP on the *non-voc* novel categories and 3.5 AP on the *voc* novel categories respectively. Even compared to its stronger version equipped with NAS-FPN [16] backbone which boosts the performance of both detection and segmentation, our model still performs better than Shape-Mask. Besides, we also provide the *oracle* performance which corresponds to the performance under full supervision and can be considered as the performance upper bound for partially supervised learning. We observe that the performance gap between our model and its oracle version is narrowed to 3.6/6.1 AP compared to 4.8/7.6 (4.4/7.4) AP by ShapeMask (ShapeMask with NAS-FPN) and 10.6/9.6 AP by Mask^X R-CNN, indicating the advantages of agnostic learning by our specifically designed modules.

Fig. 7 shows qualitative results on multiple samples that randomly selected from COCO dataset including various scenes, which shows that our model is able to segment all different kinds of objects precisely, even for quite small ones.
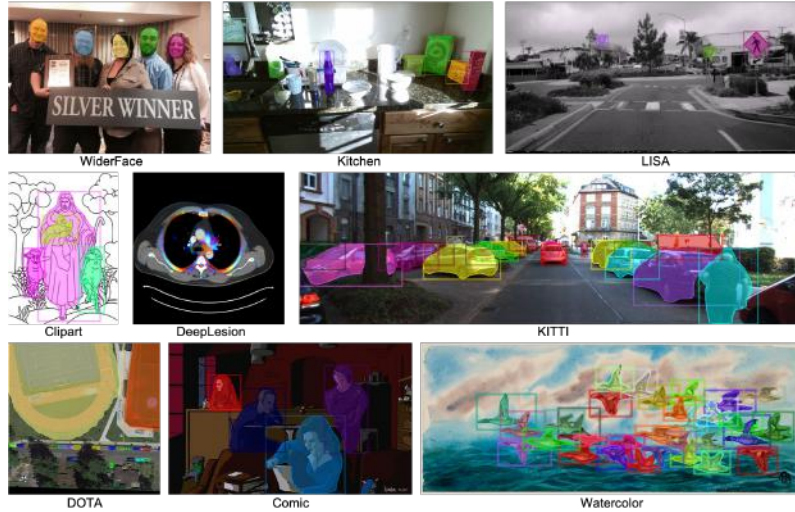
Fig. 8. Qualitative results of generalization by our model to 9 different datasets. The model is only trained on COCO and directly applied on these datasets.

Application on other datasets. We further qualitatively demonstrate our model on other 9 datasets across various styles and domains [49]: Clipart [25], Comic [25], Watercolor [25], DeepLesions [51], DOTA [50], KITTI [14], LISA [38], Kitchen [15], and WiderFace [53]. It is worth noticing that this is a much harder task due to the cross-dataset generalization. Specifically, we train our model on COCO dataset and feed it ground-truth boxes to obtain the segmentation results on these datasets. As shown in Fig. 8, our model successfully segments novel objects from various domains.

4.4   Few-shot Instance Segmentation

In this section, we directly apply our model to the challenging few-shot instance segmentation without any network adaption. Few-shot instance segmentation is another challenging task for novel categories. In this task, the model is first trained on base categories with numerous training samples and then generalizes to novel categories with only a few (10 or 20 shots) training samples by direct fine-tuning. Following Meta R-CNN [52], the *non-voc* classes is used as base categories with full samples per category and the *voc* as the novel categories with only 10/20 training samples per category. For fair comparison, we follow Meta R-CNN [52] and use ResNet-50 as backbone and input image size is resized to (600, 1000). Note that the annotations of both bounding box and mask are provided for training samples in novel categories in the few-shot setting.

As shown in Table 3, our model outperforms Meta R-CNN (the state-of-the-art method) by 2.7/3.9 AP in the 10/20-shot settings. Even equipped with the Faster R-CNN detector like Meta R-CNN, our model still performs much better. Although not specifically designed for few-shot learning, our model still obtains the state-of-the-art performance, demonstrating that our proposed model is not limited to the partially supervised learning, and is general for other novel instance segmentation tasks.

method	10-shot						20-shot					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN-ft [52]	1.9	4.7	1.3	0.2	1.4	3.2	3.7	8.5	2.9	0.3	2.5	5.8
Meta R-CNN [52]	4.4	10.6	3.3	**0.5**	3.6	7.2	6.4	14.8	4.4	**0.7**	4.9	9.3
CPMask* (Ours)	6.5	11.6	6.3	0.3	4.1	11.9	9.3	16.0	9.4	0.3	5.8	17.2
CPMask (Ours)	**7.1**	**12.0**	**7.2**	0.3	**5.5**	**12.2**	**10.3**	**16.6**	**10.7**	0.7	**8.0**	**17.5**

Table 3. Experimental results of few-shot instance segmentation on COCO *val* set. The models are trained on the *voc* base categories and fine-tuned on the *non-voc* novel categories with 10/20 instances per category. The evaluation is performed on the held-out *non-voc* novel categories. * denotes using the Faster R-CNN detector.

	method	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	Mask R-CNN [21]	ResNet-101	35.7	58.0	37.8	15.5	38.1	52.4
	MaskLab [8]	ResNet-101	37.3	59.8	39.6	19.1	40.5	50.6
Two-stage	MS R-CNN [24]	ResNet-101	38.3	58.8	41.5	17.8	40.4	54.4
	HTC [6]	ResNet-101	39.7	61.8	43.1	21.0	42.2	53.5
	PANet [36]	ResNeXt-101	**42.0**	**65.1**	**45.7**	**22.4**	**44.7**	**58.1**
	YOLACT [5]	ResNet-101	31.2	50.6	32.8	12.1	33.3	47.1
	TensorMask [9]	ResNet-101	37.1	59.3	39.4	17.4	39.1	51.6
One-stage	ShapeMask [28]	ResNet-101	37.4	58.1	40.0	16.1	40.1	53.8
	CenterMask [29]	ResNet-101	38.3	-	-	17.7	40.8	**54.5**
	CPMask (Ours)	ResNet-101	**39.2**	**60.8**	**42.2**	**22.2**	**41.8**	50.1

Table 4. Experimental results of fully supervised instance segmentation on COCO *test-dev* set. The mask AP is reported and all entries are single-model results.

4.5   Fully Supervised Instance Segmentation

In this section we investigate the performance of our model for fully supervised instance segmentation, namely the routine task for instance segmentation. Table 4 compares our model with other methods on COCO using COCO *train2017* as train set and *test-dev2017* as test set. The experimental results indicate that our model achieves best performance among one-stage methods, although our method focuses on segmenting novel categories. Particularly, our model outperforms the best one-stage model CenterMask [29] by 0.9 AP which is also built on FCOS detection backbone like ours. These encouraging results proves the effectiveness of model on fully supervised instance segmentation.

5   Conclusion

In this paper we present a novel "Commonality-Parsing Network" for partially supervised instance segmentation. Our model learns the class-agnostic commonality knowledge that can be generalized from mask-annotated categories to novel categories without mask annotations. Specifically, we design Boundary-Parsing Module to capture shape commonalities by performing supervised learning on boundary estimation. Further, we propose Non-local Affinity-Parsing Module to model pairwise affinities among pixels in intermediate feature maps to learn appearance commonalities across different categories. Benefiting from these two modules, our model outperforms state-of-the-art methods significantly for instance segmentation in both partially-supervised setting and few-shot setting.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: CVPR (2019)
2. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From contours to regions: An empirical evaluation. In: CVPR (2009)
3. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence **33**(5), 898–916 (2010)
4. Bertasius, G., Shi, J., Torresani, L.: High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In: ICCV (2015)
5. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: real-time instance segmentation. In: ICCV (2019)
6. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: CVPR (2019)
7. Chen, L.C., Barron, J.T., Papandreou, G., Murphy, K., Yuille, A.L.: Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In: CVPR (2016)
8. Chen, L.C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., Adam, H.: Masklab: Instance segmentation by refining object detection with semantic and direction features. In: CVPR (2018)
9. Chen, X., Girshick, R., He, K., Dollár, P.: Tensormask: A foundation for dense object segmentation. In: ICCV (2019)
10. Dai, J., He, K., Sun, J.: Convolutional feature masking for joint object and stuff segmentation. In: CVPR (2015)
11. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: CVPR (2016)
12. Ding, H., Jiang, X., Liu, A.Q., Thalmann, N.M., Wang, G.: Boundary-aware feature propagation for scene segmentation. In: ICCV (2019)
13. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
15. Georgakis, G., Reza, M.A., Mousavian, A., Le, P.H., Košecká, J.: Multiview rgb-d dataset for object instance detection. In: International Conference on 3D Vision (2016)
16. Ghiasi, G., Lin, T.Y., Le, Q.V.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: CVPR (2019)
17. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
18. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: ECCV (2014)
19. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR (2015)
20. Hayder, Z., He, X., Salzmann, M.: Boundary-aware instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)

21. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
23. Hu, R., Dollár, P., He, K., Darrell, T., Girshick, R.: Learning to segment every thing. In: CVPR (2018)
24. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: CVPR (2019)
25. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: CVPR (2018)
26. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: CVPR (2017)
27. Kong, S., Fowlkes, C.C.: Recurrent pixel embedding for instance grouping. In: CVPR (2018)
28. Kuo, W., Angelova, A., Malik, J., Lin, T.Y.: Shapemask: Learning to segment novel objects by refining shape priors. In: ICCV (2019)
29. Lee, Y., Park, J.: Centermask: Real-time anchor-free instance segmentation. In: CVPR (2020)
30. Li, L., Huang, W., Gu, I.Y., Tian, Q.: Foreground object detection from videos containing complex background. In: ACM Multimedia (2003)
31. Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Statistical modeling of complex backgrounds for foreground object detection. IEEE Transactions on Image Processing **13**(11), 1459–1472 (2004)
32. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: CVPR (2017)
33. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
34. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
36. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: CVPR (2018)
37. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: CVPR (2017)
38. Mogelmose, A., Trivedi, M.M., Moeslund, T.B.: Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. IEEE Transactions on Intelligent Transportation Systems **13**(4), 1484–1497 (2012)
39. Pinheiro, P.O., Collobert, R., Dollár, P.: Learning to segment object candidates. In: NeurIPS (2015)
40. Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P.: Learning to refine object segments. In: ECCV (2016)
41. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: CVPR (2019)
42. Remez, T., Huang, J., Brown, M.: Learning to segment via cut-and-paste. In: ECCV (2018)
43. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
44. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. ACM transactions on graphics (TOG) **23**(3), 309–314 (2004)

45. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: ICCV (2019)
46. Vezhnevets, V., Konouchine, V.: Growcut: Interactive multi-label nd image segmentation by cellular automata. proc. of Graphicon **1**, 150–156 (2005)
47. Wang, W., Zhao, S., Shen, J., Hoi, S.C., Borji, A.: Salient object detection with pyramid attention and salient edges. In: CVPR (2019)
48. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
49. Wang, X., Cai, Z., Gao, D., Vasconcelos, N.: Towards universal object detection by domain attention. In: CVPR (2019)
50. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: CVPR (2018)
51. Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A.P., Bagheri, M., Summers, R.M.: Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In: CVPR (2018)
52. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn : Towards general solver for instance-level low-shot learning. In: ICCV (2019)
53. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: CVPR (2016)
54. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: CVPR (2018)
55. Zhang, S., Yan, S., He, X.: LatentGNN: Learning efficient non-local relations for visual recognition. In: ICML (2019)
56. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: ICCV (2019)
57. Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J.: Weakly supervised instance segmentation using class peak response. In: CVPR (2018)