4-2021

Weakly supervised segmentation via instance-aware propagation

Huang XIN

Qianshu ZHU

Yongtuo LIU

Shengfeng HE
Singapore Management University, shengfenghe@smu.edu.sg

Citation

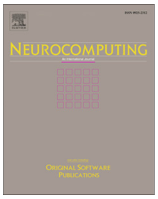
XIN, Huang; ZHU, Qianshu; LIU, Yongtuo; and HE, Shengfeng. Weakly supervised segmentation via instance-aware propagation. (2021). *Neurocomputing*. 447, 1-9.
Available at: https://ink.library.smu.edu.sg/sis_research/7843

Neurocomputing 447 (2021) 1–9

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Weakly supervised segmentation via instance-aware propagation

Xin Huang [1], Qianshu Zhu [1], Yongtuo Liu, Shengfeng He *

School of Computer Science and Engineering, South China University of Technology, China

ARTICLE INFO

Article history:
Received 29 January 2020
Revised 11 February 2021
Accepted 22 February 2021
Available online 2 April 2021
Communicated by Shijian Lu

Keywords:
Weakly supervised learning
Instance segmentation

ABSTRACT

Peak Response Map (PRM) highlighting the discriminative regions can be extracted from a pre-trained classification network. We can accurately localize instances of each class with the help of these response maps. However, these maps cannot provide reliable information for segmentation even with off-the-shelf object proposals. This is because neither PRM nor the proposals know which regions can be regarded as a complete instance. In this paper, we tackle this problem by proposing an Instance-aware Cue-propagation Network (ICN) with a new proposal-matching strategy. In particular, the ICN aims to filter out background distractions and cover the complete instance, while our proposed proposal-matching strategy adds a re-balancing constraint on the contributions of multi-scale object proposals. Extensive experiments conducted on the PASCAL VOC 2012 dataset show the superior performance of our method over weakly-supervised state-of-the-arts for both semantic and instance segmentation.

1. Introduction

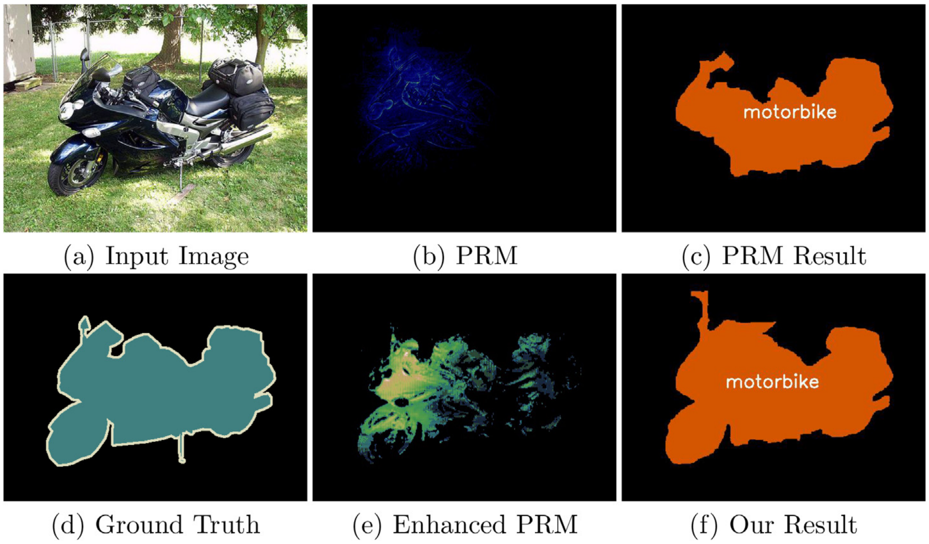
Convolutional neural networks have proved its huge feature representation ability for various tasks, such as image recognition [31], object detection [25], object segmentation [30,43,7], and so forth. While supervised learning using pixel-level annotation datasets [5,16] dramatically promotes the advancement of object segmentation, it requires cumbersome labelling efforts. Moreover, supervised learning tends to easily overfit the intrinsical distribution of training data, which leads to poor performance with many outliers as handling various objects in the real world. Weakly supervised learning, on the other hand, can relieves the burden of pixel-level labelling and alleviates the overfitting problem to some extent. Several prior works [1,35,42] have shown promising results in weakly-supervised object segmentation under the supervision of image-level labels.

Fully Convolutional Networks (FCNs) [30] first implement supervised semantic segmentation by fine-tuning a classification network, which implies that pixel-wise classification tasks can benefit from feature representations pre-trained on an image-level classification task. Grad-CAM [29] further explores the black box inside the convolutional neural network by visualizing the class activation maps. Inspired by Grad-CAM [29] and FCNs [30], Zhou et al. [42] realize instance segmentation by selecting off-the-shelf object proposals based on Peak Response Map (PRM), which depicts the most critical receptive fields in the input image that the neural network employs to ascertain object classes. However, we find that PRM tends to only highlight discriminative parts of objects for classification. As shown in Fig. 1, PRM of the input image only focuses on the front of the motorbike while overlooking the other parts. Analogous to the human visual system, the neural network optimized for classification recognizes an object class just from the prominent parts of instances rather than the whole regions. Since the final segmentation prediction relies on a matching process between PRM and off-the-shelf object proposals, the property of PRM that only focuses on local regions heavily limits the performance of weakly-supervised methods for object segmentation under the supervision of image-level labels.

Based on the above observation, we tailor an Instance-aware Cue-propagation Network (ICN) with a re-balancing constraint for proposal matching. In particular, the proposed ICN aims to propagate instance-aware appearance cues to render the entire shape of each instance. In the testing phase, we further design an iterative optimization process for progressive PRM enhancing. Specifically, the initially enhanced PRM is further input into ICN for better results in an iterative manner. Fig. 1 shows an example output of our method compared with the naive PRM method [42], and we can observe that the enhanced PRM can activate more regions belonging to the motorbike and as well suppress noises initially estimated by the naive PRM method, which finally results in better instance segmentation. Additionally, we propose a new proposal-matching strategy to balance the contributions of multi-scale proposals. By doing so, requiring only image-level labelled data, our method can perform object segmentation in

* Corresponding author.
E-mail address: hesfe@scut.edu.cn (S. He).
[1] Joint first authors.

https://doi.org/10.1016/j.neucom.2021.02.093
0925-2312/© 2021 Elsevier B.V. All rights reserved.



(a) Input Image (b) PRM (c) PRM Result

(d) Ground Truth (e) Enhanced PRM (f) Our Result

Fig. 1. Visualization of **P**eak **R**esponse **M**ap (PRM), the enhanced one and their corresponding segmentation results. The enhanced PRM comes from our Instance-aware Cue-propagation Network (ICN) which concatenates (*a*) and (*b*) as input. For a better view, (*b*) and (*c*) are colored according to the predicted probability values, e.g., warmer color means higher values.

many challenging scenarios. Extensive experiments demonstrate our approach surpasses previous weakly supervised approaches by a large margin on the PASCAL VOC 2012 dataset, even superior to many fully supervised methods.

The contributions of this paper are summarized as follows:

- We analyze the main limitation of the naive PRM method, and design an Instance-aware Cue-propagation Network (ICN) to render clear shape of objects. Furthermore, we iteratively enhance the initially generated PRM which hugely boosts segmentation accuracy.
- We present a new proposal-matching strategy to rebalance the contributions of multi-scale object proposals.
- Extensive experiments conducted on the PASCAL VOC 2012 dataset show a superior performance of our method over other counterparts by a large margin.

2. Related work

2.1. Weakly supervised semantic segmentation

Weakly supervised semantic segmentation aims to predict object masks for each class without the help of large-scale pixel-level annotations, which leaves it more challenging than its supervised counterparts. Recent weakly-supervised methods, however, get off to a promising start, supervised by various kinds of weak labels including points [2,17], scribbles [14,34,35], bounding boxes [3,19,32], and foreground [28]. Besides, the image-level class label is the most commonly used annotation as it is available on large-scale classification datasets, e.g, ImageNet. Most methods supervised by image-level class labels learn to leverage class activation cues which roughly aggregate pixels in the input image which prominently contribute to the final class estimation. However, these class activation cues overemphasize the most discriminative part of each object without crisp boundaries. To alleviate this problem, Some methods [11,40] utilize graphical models to generate coarse segmentation as a pseudo label, thus their performance is limited by the performance of object localization. As a consequence, other localization approaches provide initialized accurate object locations [10,20]. While Zhou et al. [42] perform object segmentation based on Peak Response Map, they are also restricted by the overemphasizing issue as aforementioned. In this paper, we propose to enhance the initially generated naive PRM to cover the entire shape of objects.

2.2. Weakly supervised instance segmentation

In comparison with semantic segmentation that predicts only class-level masks, instance segmentation is much more challenging since it further demands to distinguish various instances among the same object class. To address this problem, extra evidence, like bounding boxes, is explored to provide additional supervision. For example, inspired by adversarial learning, Tal et al. [24] learn an object shape generator that can produce a realistic fake image by cropping the predicted masks to a random background area. Further, GraphCut [9] obtains better instance shape by taking into account boundaries after introducing generic boundary detector [37]. Paired with the Conditional Random Field (CRF), Rajchl et al. [23] upgrade GraphCut to DeepCut. However, all of these methods require extra instance-aware information, either pseudo labels or bounding boxes. To deal with this issue, a few recent works center on challenging image-level class label supervision. Jiwoon et al. [1] present an AffinityNet-based model that generates object masks by randomly dilating the affinities centering at the local discriminative part. Zhou et al. [42] leverage Peak Response Map to locate object instances and couple it with off-the-shelf object proposals [21] to segment instances. Ge et al. [6] propose to refine the class activation maps via a multi-task learning in a coarse-to-fine manner. However, these methods are limited by the quality of PRM and off-the-shelf object proposals, which are usually attributed to low-quality. In contrast, our method introduces a simple yet effective architecture to enhance initially generated PRM and optimize the proposal-matching strategy with a new re-balance constraint.

2.3. Global class response activation

Deep networks tend to be immersed in the most distinguishing area if merely supervised by image-level labels, so that it is necessary to join shallow feature representations into global class responses. Global max pooling (GMP) [18] takes the maximal response score as result while abandoning the other values. Global average pooling (GAP) [18] samples all responses uniformly, leaving it hard to tell different objects. The log-sum-exponential (LSE) [33] reaches a trade-off between GMP and GAP. However, all of them ignore local spatial relevance which is critical to localize diverse object instances. Differently, We reveal the relevance by comparing each class response map with their filtered version, which is beneficial to promote the instance discriminative ability. Furthermore, we stimulate peaks extracted from class activation maps via gradient back-propagation and combine the output Peak Response Map with object proposals to predict the final segmentation.

2.4. Prediction in coarse-to-fine strategy

Learning in a coarse-to-fine fashion has been widely adopted in various tasks. Jing et al. [8] design a novel saliency detection network that recursively refines the previously generated mask, which is trained with only image-level class labels. Similarly, supervised by dense masks, RefineNet [15] utilizes sub-sampled feature representations that are updated progressively in a multi-path fashion to predict the final object segmentation. In contrast, our approach does not generate segmentation in this strategy directly but enhance the Peak Response Map iteratively which can result in better segmentation when matching with object proposals.

3. Proposed method

Our presented network is shown in Fig. 2, which consists of three parts, e.g., the PRM generation module, the PRM enhancing module, and the proposal-matching strategy. The PRM generation module (discussed in Section 3.1) takes an RGB image as input and aims to generate the Peak Response Map (PRM). The PRM enhancing module (discussed in Section 3.2) is designed to enhance the initially generated naive PRM to cover more parts of each instance instead of only prominent regions. The proposal-matching strategy (discussed in Section 3.3) predicts the final results of object segmentation via matching the enhanced PRM with off-the-shelf object proposals.

3.1. PRM generation module

The PRM generation module is introduced for localizing and stimulating the visual cues of each object instance inside the input image domain. Given a pre-trained classification network, the activation map of each class, which is called Class Activation Map (CAM) can be extracted from the last convolutional layer. Formally, we denote CAM $\in R^{H* W* C}$, where $H * W$ represents the spatial resolution of CAM and C denotes the number of channels. So far, we can directly generate class-aware attention maps through gradient

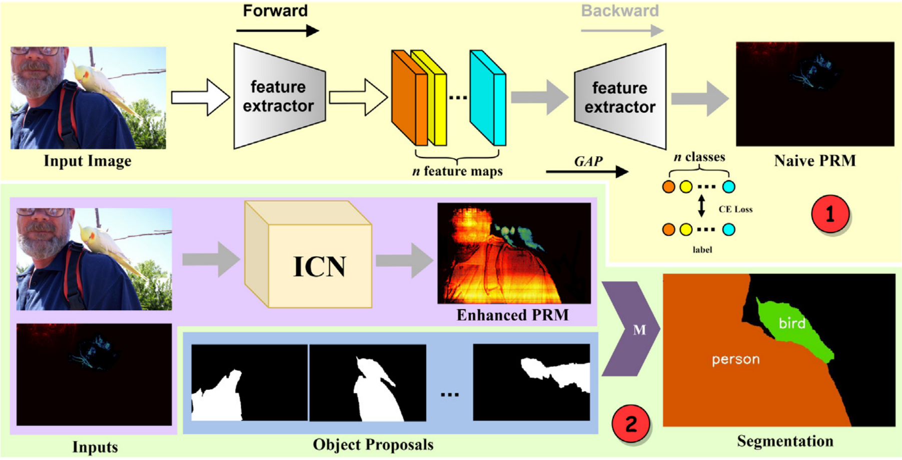


Fig. 2. An overview of our method for weakly supervised object segmentation. This pipeline follows a two-stage fashion. At stage 1, marked with the number 1 inside the red circle, we first train a classification network supervised by a **C**ross **E**ntropy (CE) loss and then use the pre-trained network to extract Class Activation Map (CAM) of each class (the orange, yellow and blue cubes), which is further utilized to launch gradient propagation backward through the feature extractor to generate the naive Peak Response Map (PRM). At stage 2, our proposed ICN (Instance-aware Cue-propagation Network) conducts iterative optimization to obtain the enhanced PRM, taking the naive PRM and input image as inputs. Finally, the enhanced PRM is utilized to match with off-the-shelf object proposals. Note that each color in the enhance PRM indicates an instance, and the proposal matching process marked as 'M' is performed independently for each instance.

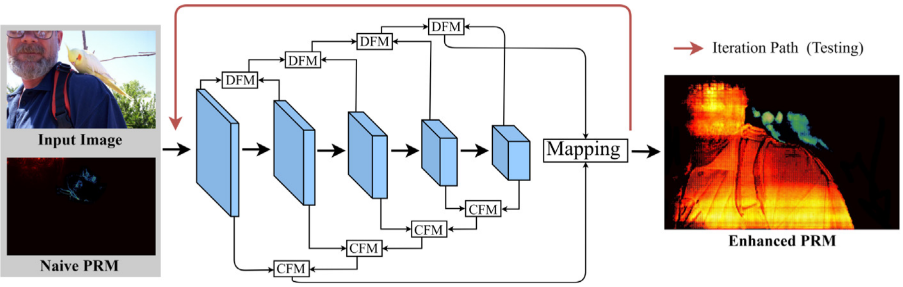


Fig. 3. Illustration of **I**nstance-aware **C**ue-propagation **N**etwork (ICN). ICN contains three components: 1). the **C**ontext-driven **F**usion **M**odule (CFM). 2). the **D**etail-driven **F**usion **M**odule (DFM). 3). the feature mapping module (Mapping). The orange arrow indicates the flow path that works for iterative optimization in the testing phase. Note that the PRM enhancing process is performed independently for each PRM (or instance).

back-propagation from this kind of CAM. However, the generated class-aware maps in this way is neither accurate nor instance-aware. To resolve these issues, we figure out local maximal responses from CAM which can be formulated as:

$$\mathscr{P}^c = max\{M_c - \mathscr{G}(M_c) - E_\eta, 0\}, \tag{1}$$

where $\mathscr{P}^c$ is the returned peak map of the object class c. M_c represents CAM of the c-th class (i.e., the c-th channel of the last convolutional layers), as is shown in Fig. 2. $\mathscr{G}$ denotes the gaussian filter operation. E_η is a threshold matrix with the same spatial resolution as M_c, aiming to suppress noises of class activations and accurately localize the real instances with higher activation values. In our experiments, we set E_η as the mean value of M_c.

To obtain instance-aware Peak Response Maps, we further launch gradient back-propagation for each individual peak in $\mathscr{P}^c$ instead of the original CAM M_c. Considering a two-layer convolutional network for simplification, where the weight matrixes are denoted as W_1 and W_2 respectively and let U denote the output of the first layer. Based on the chain rule, PRM of each peak in $\mathscr{P}^c$ can be formulated as:

$$\mathscr{Q}_k^c = \frac{\partial L}{\partial \mathscr{P}_{i,j}^c} \cdot \frac{\partial \mathscr{P}_{i,j}^c}{\partial U} \cdot \frac{\partial U}{\partial W_1}, \tag{2}$$

where $\mathscr{Q}_k^c$ means the Peak Response Map of the k-th peak of the c-th class. L denotes the standard classification loss. $\mathscr{P}_{i,j}^c$ represents each peak of the c-th class with the coordinates (i,j). It is worth noting that only peaks are enabled during gradient back-propagation.

With the peak activation and gradient back-propagation, we can assign each instance the most prominent area in the input image, which is also called Peak Response Map (PRM) as shown in Fig. 1 (b).

3.2. PRM enhancing module

The PRM enhancing module is utilized to enhance the naive PRM that only activates the prominent part of each object, ending in more complete object shapes for segmentation. To reconstruct the naive PRM, we design an iterative optimization network called ICN (Instance-aware Cue-propagation Network), as shown in Fig. 3. It is composed of three components: the Context-driven Fusion Module (CFM), the Detail-driven Fusion Module (DFM) and the feature mapping module. ICN takes as input the concatenation of the naive PRM and the corresponding input image, and utilizes the VGG-16 backbone as the feature extractor for hierarchically multi-level feature representations. As we know, deeper layers of the neural network contain more high-level context, while shallower layers are abundant with fine details. Then we design two parallel feature fusion processes from this perspective. The Context-driven Fusion Module (CFM) is first leveraged to fuse feature maps from the last and penultimate convolutional layers, and the fused features are then regarded as one input of another CFM to fuse feature maps from relatively shallower layers. This process is continued until the shallowest convolutional layers, which enables high-level features to progressively aggregate low-level details driven by global and local context. On the contrary, the Detail-driven Fusion Module (DFM) integrates multi-level features from the opposite direction, which encourages low-level details gradually matching their corresponding high-level context to suppress superfluous noises. Without loss of generality, we simply implement DFM and CFM as a concatenation operation followed by two consecutive convolutional layers. Formally, given two sets of feature maps to be fused, the output of DFM or CFM can be formulated as:

$$f_o = E(C(U(f_s), f_l), \theta), \tag{3}$$

where f_s and f_l represent the feature maps with small and large resolution respectively. $U(\cdot)$ denotes the up-sampling operation. $C(\cdot,\cdot)$ stands for the concatenation operation along the channel dimension. $E(\cdot,\theta)$ represents the two convolutional layers with parameters θ. Finally, to make better use of the above two different feature fusion strategies, which are complementary with each other, the feature mapping module takes as input the last aggregated features of the two paths and leverages concatenation and convolution operations to obtain the enhanced RPM with the same resolution as the input image.

Ideally, the enhanced PRM should be trained with pixel-level annotations. However, only image-level labels are available in our weakly-supervised setting. To train the PRM enhancing module, we leverage the matching results generated by matching the naive PRM to the object proposals (the matching process is discussed in Section 3.3). Specifically, we rank the proposals by their matching scores in descending order, and pick up k scores one by one from the top only if the current one is not fully contained by previously selected ones. At last, we take the union of the selected top k proposals and utilize the merged proposal as the pseudo label of the enhanced PRM. Although the proposal-based pseudo label is not accurate compared with GT segmentation annotations, it can provide complete visual cues of each instance than the naive PRM and benefit the proposal-matching process for better instance segmentation (the comparison results can be found later in Table 1). In our experiments, we empirically set k to 3 to balance the influence of positive and negative proposals.

In addition, to further validate the effectiveness of the PRM enhancing mechanism, we present another PRM enhancing module that governed by pixel-level salient object annotations. This variant is trained on the dataset proposed in [12], consists of 1000 images with annotations of salient instances. To adapt this dataset for the supervision of PRM enhancing module, we first input each training image into the PRM generation module to obtain its corresponding PRM, and then match each PRM with the labelled instance saliency map by computing IOU to generate pairs of an instance and its corresponding PRM. In this way, given each pair and the training image, we can train ICN to enhance the naive PRM with the instance saliency map as label. To optimize the weights of ICN, we take the pixel-wise cross entropy as the loss function, which is formulated as:

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\tilde{O}_{i,j}\log(O_{i,j}), \tag{4}$$

where $\tilde{O}$ and O denote the prediction and the ground truth respectively, which share the same spatial resolution with the height as H and width as W. N is the number of training samples. In the testing phase, we follow a multi-stage optimization manner where the previously enhanced PRM is iteratively input into ICN to progressively refine the enhanced prediction.

Table 1
Comparison results of different variants of our method on the PASCAL VOC 2012 dataset for weakly-supervised semantic segmentation. Note that the PRM enhancing module (or ICN) is trained with object proposals in all ablation experiments.

Method	*mIoU*(%)
w/o ICN	53.5
w/o CFM and DFM of ICN	53.8
w/o input PRM of ICN	36.7
w/o multi-scale rebalance matching	53.9
Full model (Ours-proposal)	**54.1**

3.3. Proposal Matching strategy

To realize instance segmentation, we design a re-balancing matching strategy to pick out appropriate off-the-shelf object proposals. Specifically, given an input image, we first leverage the object proposals generated by the traditional methods (e.g., MCG [21]) to construct a candidate collection of instance segmentation, and then compute the matching score between the enhanced PRM and each candidate, which can be formulated as:

$$\mathscr{S} = \frac{1}{sum(P)} \cdot \alpha \cdot V * P + \beta \cdot E(V) * P + \gamma \cdot B(V) * P, \qquad (5)$$

where P is a candidate object proposal. V represents the enhanced PRM of each instance. $E(\cdot)$ and $B(\cdot)$ denote a contour extractor and a background extractor respectively. $\frac{1}{sum(P)}$ is the multi-scale rebalance factor of the first term, where $sum(P)$ means the number of pixels belonging to P. α, β, γ are three Lagrange multipliers which are set as 0.95, 0.1, and 0.8, respectively via cross validation.

In Eq. (5), the first term enforces the enhanced PRM and the selected object proposals to share more overlapped regions. The second term coupled with the contour information extracted from the enhanced PRM, encourages it to pick out a proposal that shares a similar outline. Furthermore, the last term takes background as a template to suppress irrelevant regions.

The algorithm of our method for weakly-supervised instance segmentation is elaborated in Algorithm 1.

Algorithm 1: Instance segmentation with image-level labels

```
Input:
A testing image $\mathscr{I}$, corresponding object proposal gallery $\mathscr{S}$,
  our network equipped with the PRM generation module $\mathscr{G}$
  and the PRM enhancing module $\mathscr{E}$. Outpuy:
Instance segmentation set $\mathscr{O}$;
Initialize instance segmentation set $\mathscr{O} = \varnothing$, peak map set
  $\mathscr{P} = \varnothing$;
Input $\mathscr{I}$ into $\mathscr{G}$ to extract class activation maps $\mathscr{M}$;
for all $\mathscr{M}^c$ such that $\mathscr{M}^c$ is the $c$-th class activation map of
  $\mathscr{M}$ do
    Obtain peak map $\mathscr{P}^c$ and append it to $\mathscr{P}$;
end for
for all peak map $\mathscr{P}^c \in \mathscr{P}$ do
    Gradient back-propagation for $\mathscr{P}^c$ to obtain naive PRM;
    Iteratively enhance naive PRM to get enhanced PRM
  using $\mathscr{E}$;
    for all object proposal $\mathscr{S}_j \in \mathscr{S}$ do
        Calculate matching score between $\mathscr{S}_j$ and enhanced
  PRM;
    end for
    Append top-ranked proposal and label $(\mathscr{S}_t, c)$ to $\mathscr{O}$;
end for
Purge $\mathscr{O}$ via Non-Maximum Suppression (NMS);
return $\mathscr{O}$;
```

4. Experiments

4.1. Dataset

For a fair comparison, following previous methods [36,42], we use the PASCAL VOC 2012 dataset [5] to train the PRM generation module. It is made of 11,530 training and validation images containing 27,450 ROI-annotated objects, 6,929 dense segmentations and class annotations. Since this work explores image-level supervision, only the class annotations are used for training the PRM generation module. On the other hand, the dataset presented in [12] for salient instance segmentation is leveraged to enhance the naive PRM. This dataset contains 1,000 images that are mostly selected from existing datasets (e.g., DUT-OMRON [38], HKU-IS [13], and MSO [39]).

4.2. Training details

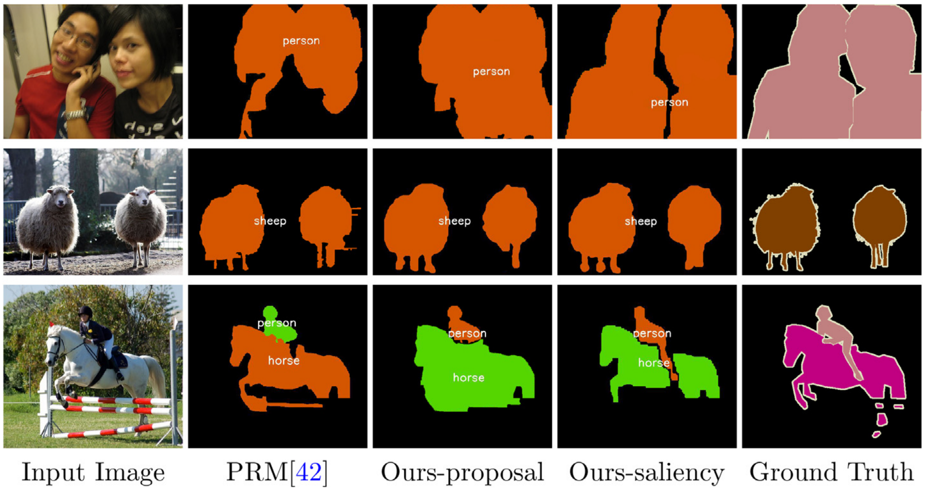
The training procedure of the entire architecture follows a two-stage fashion. In the first stage, we train the classification network in the PRM generation module for 20 epochs, and then freeze it to stimulate the naive PRM by gradient backpropagation from the last convolution layer. In the second stage, we train ICN in the PRM enhancing module for 120 epochs. In consideration of the scale of the training dataset in the second stage, we perform abundant data augmentation techniques, including random flipping and cropping. In addition, the two stages both employ Adam optimizer [27]. We set the initial learning rate of the first stage as 1e-4 and multiplied by 0.1 after 10 epochs. The learning rate of the second stage is 1e-3 and decay 0.5X every 20 epochs.

4.3. Ablation study

In this section, we perform ablation experiments on the PASCAL VOC 2012 dataset. Table 1 reports the comparison results of various variants of our method. Note that the PRM enhancing module (or ICN) is trained with object proposals in all ablation experiments (so as our full model). Specifically, to validate the effectiveness of the PRM enhancing mechanism, we drop ICN from our model, which is denoted as *w/o ICN*. We can see from Table 1 that compared with the full model, the mIoU of *w/o ICN* decreases from 54.1 to 53.5, which indicates that the enhanced PRM can benefit a lot when matching with object proposals. Besides, we train another variant by removing CFM, DFM and the input PRM from ICN (denoted as *w/o CFM and DFM of ICN*, *w/o input PRM of ICN*) to explore the optimal setting of the PRM enhancing module. As can be seen, without CFM and DFM, the PRM enhancing ability of ICN is limited and leads to worse segmentation accuracy than the full ICN. When we drop the input PRM and only input the testing image into ICN, the mIoU decreases dramatically by 32.2%. It is because that each input image may contain multiple classes and instances, and without the input PRM as a hint, ICN cannot know which instance to enhance. In addition, to evaluate the proposed multi-scale rebalance matching strategy, we show the comparison result of the vanilla matching strategy [42] (denoted as *w/o multi-scale rebalance matching*). As can be seen, the multi-scale rebalance constraint contributes to the segmentation performance by a considerable margin, which demonstrates the effectiveness of balancing the importance of multi-scale proposals. It is worth noting that compared with PRM [42], the method of *w/o ICN* has only one additional change (e.g., the multi-scale rebalance constraint), and still has a higher segmentation accuracy than PRM (see Table 2). This

Table 2
Comparison results of weakly-supervised semantic segmentation with state-of-the-art methods on the PASCAL VOC 2012 dataset. Top two methods are highlighted using bold fonts.

Method	*mIoU*(%)	Commentaries
WILDCAT [4]	43.7	CRF post-processing
SEC [10]	50.7	CRF as boundary loss
Check mask [28]	51.5	CRF & Human in the loop
Combining [26]	52.8	CRF as RNN
PRM [42]	53.4	Object proposals
Ours-proposal	**54.1**	Object proposals
Ours-saliency	**54.9**	Object proposals



Input Image PRM[42] Ours-proposal Ours-saliency Ground Truth

Fig. 4. Qualitative results of our method for weakly-supervised semantic segmentation on the PASCAL VOC 2012 dataset.

Table 3
Quantitative comparisons with weakly-supervised instance segmentation state-of-the-arts on the PASCAL VOC 2012 dataset w.r.t. mean Average Precision (mAP%) and Average Best Overlap (ABO). Top two methods are highlighted using bold fonts.

Method		$mAP^r_{0.25}$	$mAP^r_{0.5}$	$mAP^r_{0.75}$	ABO
CAM [41]	Rect.	18.7	2.5	0.1	18.9
	Ellipse	22.8	3.9	0.1	20.8
	MCG	20.4	7.8	2.5	23.0
SPN [45]	Rect.	29.2	5.2	0.3	23.0
	Ellipse	32.0	6.1	0.3	24.0
	MCG	26.4	12.7	4.4	27.1
MELM [36]	Rect.	36.0	14.6	1.9	26.4
	Ellipse	36.8	19.3	2.4	27.5
	MCG	36.9	22.9	8.4	32.9
PRM [42]		44.3	26.8	9.0	37.6
IAM-S5 [44]		45.9	28.8	11.9	41.9
Ours-proposal		**47.1**	**29.2**	**12.5**	**42.8**
Ours-saliency		**47.7**	**29.4**	**13.1**	**43.7**

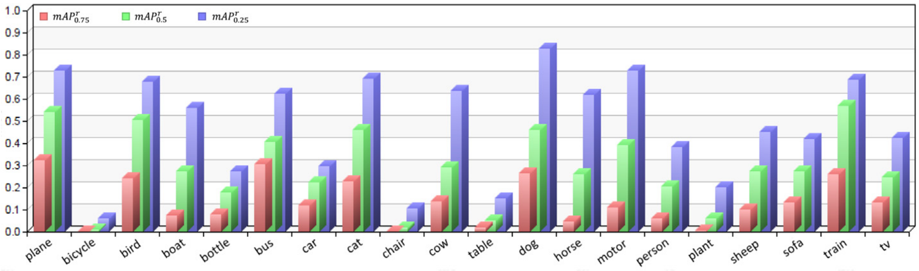
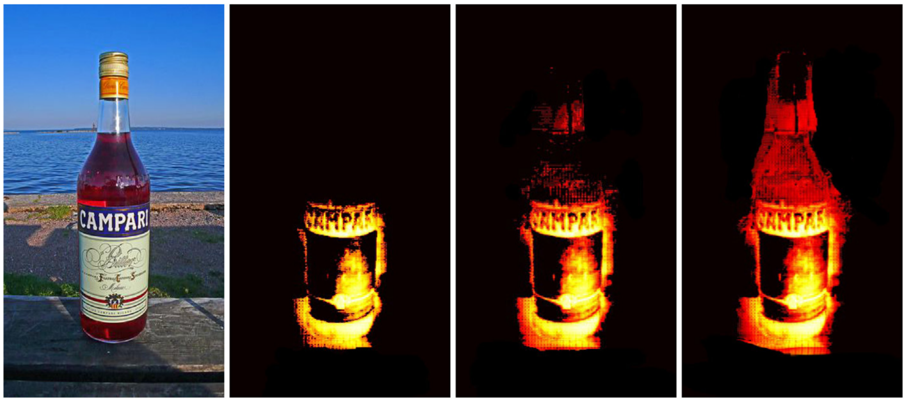


Fig. 5. Per-class mAP of our method at three different IoU threshold settings.



(a) Input Image (b) 1-th Step (c) 2-th Step (d) 3-th Step

Fig. 6. Visualization of the iterative optimization process of the enhanced PRM.

indicates that the multi-scale rebalance constraint is effective not only for proposal matching with the enhanced PRM, but also for the naive PRM.

Table 4
Comparison results of different time steps on the PASCAL VOC 2012 dataset.

Time step(s)	1	2	3	4
$mAP^r_{0.5}$	27.3	28.2	**29.4**	29.1

4.4. Weakly supervised semantic segmentation

We evaluate the effectiveness of our approach by comparing with weakly-supervised semantic segmentation state-of-the-arts on the PASCAL VOC 2012 dataset. As our framework directly produces an instance-level segmentation, we merge the instance masks of the same class for semantic segmentation. Table 2 summarizes the performance of weakly-supervised state-of-the-art methods using the metric of mean intersection-over-union (IoU) across 21 classes (20 object categories plus background). We can see that our method trained with object proposals (Ours-proposal) can outperform

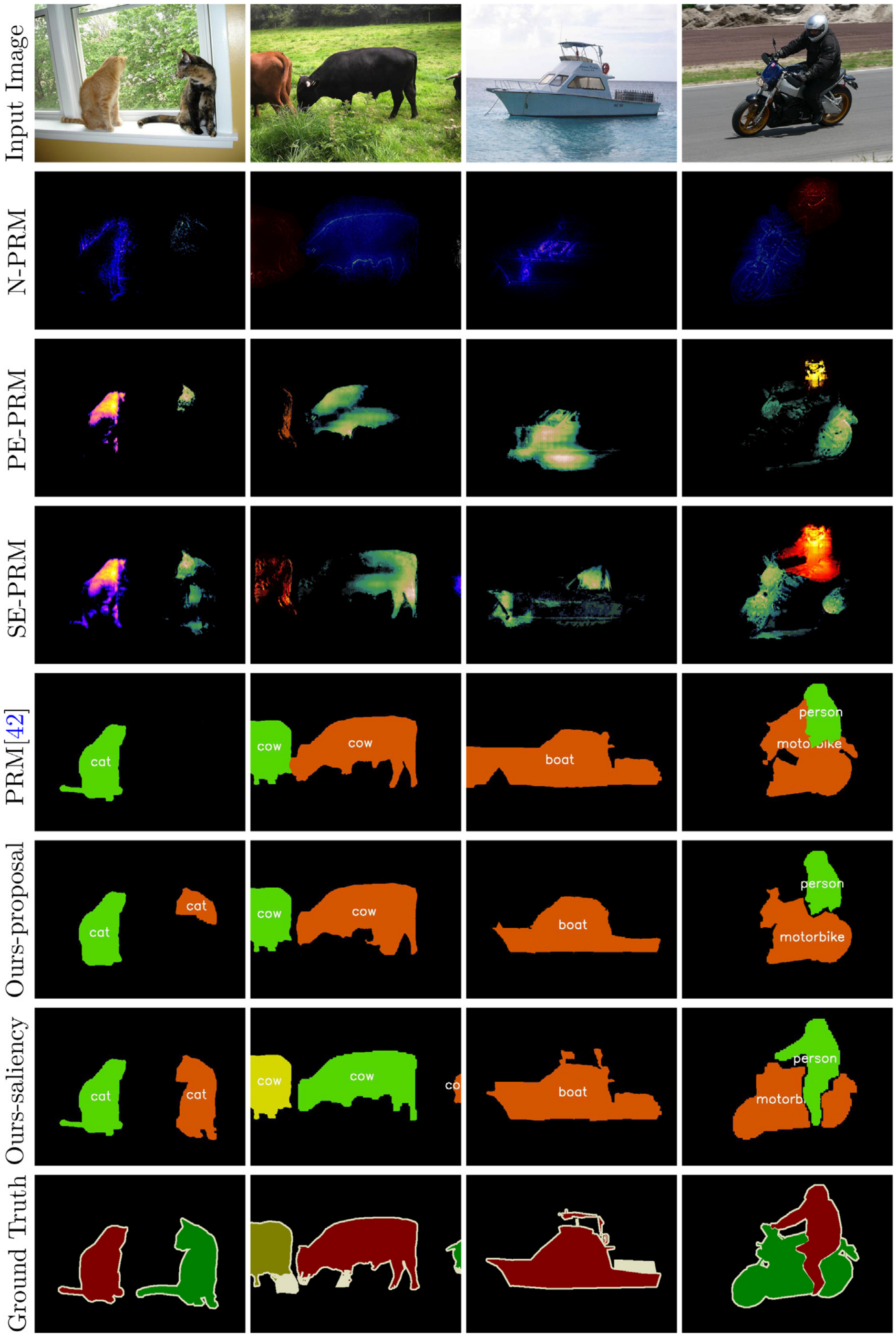


Fig. 7. Qualitative results of our method for weakly-supervised instance segmentation on the PASCAL VOC 2012 dataset. Due to limited space, "N-PRM", "PE-PRM" and "SE-PRM" represent "Naive PRM", "Proposal-enhanced PRM" and "Saliency-enhanced PRM" respectively.

previous works that even rely on additional post-processing operations (e.g., CRF), which demonstrates the superiority of the proposed PRM enhancing mechanism and multi-scale rebalance matching strategy. Additionally, our method trained with extra saliency dataset can further improve the segmentation accuracy. This indicates a more accurate enhanced PRM can benefit to the proposal matching process and match more appropriate object proposals for the final segmentation. Fig. 4 visualizes some examples of semantic segmentation achieved by our method.

4.5. Weakly supervised instance segmentation

In this section, we evaluate the performance of our method for weakly-supervised instance segmentation on the PASCAL VOC 2012 dataset. We compare our method with several previous works [41,42,44,45]. Unless otherwise explicitly stated, all the results shown in the Table 3 are computed based on the off-the-shelf object proposals generated by MCG [21].

4.5.1. Quantitative analysis

To assess the performance quantitatively, we compare our method with state-of-the-arts based on two widely used metrics: the mAP^{*} at three different IoU threshold settings and the Average Best Overlap (ABO) [22]. The comparison results are in Table 3. We can see that our method can surpass previous weakly supervised localization methods (e.g., CAM [41], SPN [45] and MELM [36]), especially CAM [41] that utilizes additional prior knowledge. In addition, by comparing with proposal-matching based approaches (e.g., PRM [42] and IAM-S5 [44]), our method can consistently improve the instance segmentation accuracy even without extra saliency annotations, which confirms the effectiveness of our PRM enhancing mechanism and multi-scale rebalance matching strategy. Fig. 5 illustrates the per-class mAP of our method at three different threshold settings.

4.5.2. Qualitative analysis

Here we aim at visual understanding to validate the effectiveness of our approach through qualitative analysis. As shown in the second and third rows of Fig. 7, the enhanced PRM can cover more parts of each instance compared with the naive PRM, which only activates the prominent small regions. High-quality predictions in the fourth row indicate that the enhanced PRM benefit a lot for the final segmentation based on the improved proposal matching strategy.

4.6. Iterative refinement

The mechanism of iterative refinement is a widely-used technique in the computer vision community due to its simplicity and efficiency. We adopt the refinement mechanism to iteratively enhance PRM for covering more regions of each instance. To explore the optimal number of refinement steps, we perform ablation experiments to test the PRM enhancing module with different time steps. Note that in this section, the PRM enhancing module is trained with only one time step on the salient instance segmentation dataset [12]. Fig. 6 shows an example of the iterative optimization process. Table 4 presents the accuracy of instance segmentation on the PASCAL VOC 2012 dataset at different time steps. We can see from Table 4 that the performance reaches the optimum after three times steps, so we use three refinement steps in the other experiments to balance the computational time and accuracy. Note that before processing the input PRM, we first normalize it using min–max normalization considering the wide range of values. In addition, we set the final enhanced PRM as a binary image so that we can treat each part of an object equally rather than focusing on the most discriminative regions.

5. Conclusion

In this paper, we present a novel framework for weakly-supervised object segmentation. In particular, we propose an Instance-aware Cue-propagation Network (ICN) to enhance the naive PRM, making it not only focusing on the prominent part but also covering the whole regions of each instance. Additionally, we propose a new proposal-matching strategy that imposes a constraint to balance the contributions of multi-scale object proposals. Extensive experiments conducted on the PASCAL VOC 2012 dataset show that our method can outperform previous counterparts by a large margin.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project is supported by the National Natural Science Foundation of China (No. 61972162, and No. 61702194) and the CCF-Tencent Open Research fund (CCF-Tencent RAGR20190112).

References

[1] J. Ahn and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In CVPR, pages 4981–4990, 2018..

[2] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, ECCV, pages 549–565, 2016..

[3] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In ICCV, pages 1635–1643, 2015..

[4] T. Durand, T. Mordan, N. Thome, and M. Cord. WILDCAT: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In CVPR, pages 5957–5966, 2017..

[5] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, IJCV 111 (1) (2015) 98–136.

[6] W. Ge, S. Guo, W. Huang, and M. R. Scott. Label-penet: Sequential label propagation and enhancement networks for weakly supervised instance segmentation. In ICCV, pages 3345–3354, 2019..

[7] J. Huang, S. Lu, D. Guan, and X. Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In ECCV, pages 705–722, 2020..

[8] L. Jing, Y. Chen, Y. Tian, Coarse-to-fine semantic segmentation from image-level labels, IEEE TIP 29 (2020) 225–236.

[9] A. Khoreva, R. Benenson, J. H. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In CVPR, pages 1665–1674, 2017..

[10] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In ECCV, pages 695–711, 2016..

[11] B. Lai and X. Gong. Saliency guided dictionary learning for weakly-supervised image parsing. In CVPR, pages 3630–3639, 2016..

[12] G. Li, Y. Xie, L. Lin, and Y. Yu. Instance-level salient object segmentation. In CVPR, pages 247–256, 2017..

[13] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In CVPR, June 2015..

[14] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In CVPR, pages 3159–3167, 2016..

[15] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In CVPR, pages 5168–5177, 2017..

[16] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr. Microsoft coco: Common objects in context, 2014..

[17] K. Maninis, S. Caelles, J. Pont-Tuset, and L. V. Gool. Deep extreme cut: From extreme points to object segmentation. In CVPR, pages 616–625, 2018..

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In CVPR, pages 685–694, 2015..

[19] G. Papandreou, L. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. In ICCV, 2015..

[20] P.O. Pinheiro, R. Collobert, Weakly supervised semantic segmentation with convolutional networks, CVPR 2 (2015) page 6.

[21] J. Pont-Tuset, P. Arbelaez, J.T. Barron, F. Marqués, J. Malik, Multiscale combinatorial grouping for image segmentation and object proposal generation, IEEE TPAMI 39 (1) (2017) 128–140.

[22] J. Pont-Tuset and L. Van Gool. Boosting object proposals: From pascal to coco. In ICCV, pages 1546–1554, 2015..

[23] M. Rajchl, M.C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M.A. Rutherford, J.V. Hajnal, B. Kainz, et al., Deepcut: Object segmentation from bounding box annotations using convolutional neural networks, IEEE Trans. Medical Imaging 36 (2) (2016) 674–683.

[24] T. Remez, J. Huang, and M. Brown. Learning to segment via cut-and-paste. In ECCV, pages 37–52, 2018..

[25] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE TPAMI 39 (6) (2017) 1137–1149.

[26] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In CVPR, July 2017..

[27] S. Ruder. An overview of gradient descent optimization algorithms. CoRR, abs/1609.04747, 2016..

[28] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In ECCV, pages 413–432, 2016..

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In ICCV, pages 618–626, 2017..

[30] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE TPAMI 39 (4) (2017) 640–651.

[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, ICLR (2015).

[32] C. Song, Y. Huang, W. Ouyang, and L. Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In CVPR, pages 3136–3145, 2019..

[33] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. D. Bourdev. Pronet: Learning to propose object-specific boxes for cascaded neural networks. In CVPR, pages 3485–3493, 2016..

[34] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers. Normalized cut loss for weakly-supervised CNN segmentation. In CVPR, pages 1818–1827, 2018..

[35] P. Vernaza, M. Chandraker, Learning random-walk label propagation for weakly-supervised semantic segmentation, CVPR (2017).

[36] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye. Min-entropy latent model for weakly supervised object detection. In CVPR, June 2018..

[37] S. Xie, Z. Tu, Holistically-nested edge detection, IJCV 125 (1–3) (2017) 3–18.

[38] C. Yang, L. Zhang, R. X. Lu, Huchuan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In CVPR, pages 3166–3173. IEEE, 2013..

[39] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. Unconstrained salient object detection via proposal subset optimization. In CVPR, June 2016..

[40] W. Zhang, S. Zeng, D. Wang, and X. Xue. Weakly supervised semantic segmentation for social images. In CVPR, pages 2718–2726, 2015..

[41] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In CVPR, pages 2921–2929, 2016..

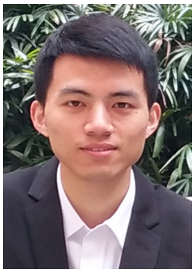
[42] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao. Weakly supervised instance segmentation using class peak response. In CVPR, pages 3791–3800, 2018..

[43] H. Zhu, F. Meng, J. Cai, S. Lu, Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation, J. Visual Commun. Image Representation 34 (2016) 12–27.

[44] Y. Zhu, Y. Zhou, H. Xu, Q. Ye, D. Doermann, J. Jiao, Learning instance activation maps for weakly supervised instance segmentation, CVPR (2019).

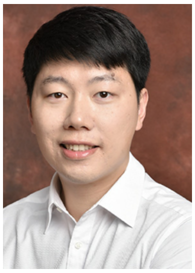
[45] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft proposal networks for weakly supervised object localization. In ICCV, Oct 2017..

Xin Huang is a master student in the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing and deep learning.



Qianshu Zhu is a master student in the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing and deep learning.

Yongtuo Liu is a master student in the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing and deep learning.



Shengfeng He is an Associate Professor in the School of Computer Science and Engineering, South China University of Technology. He was a Research Fellow at City University of Hong Kong. He obtained his B.Sc. degree and M.Sc. degree from Macau University of Science and Technology, and the Ph.D degree from City University of Hong Kong. His research interests include computer vision, image processing, computer graphics, and deep learning.