Weakly Supervised Instance Segmentation using Class Peak Response

Yanzhao Zhou†[1], Yi Zhu[1], Qixiang Ye[1], Qiang Qiu[2] and Jianbin Jiao†[1]

[1]University of Chinese Academy of Sciences
[2]Duke University
{zhouyanzhao215, zhuyi215}@mails.ucas.ac.cn, {qxye, jiaojb}@ucas.ac.cn, qiang.qiu@duke.edu

Abstract

Weakly supervised instance segmentation with image-level labels, instead of expensive pixel-level masks, remains unexplored. In this paper, we tackle this challenging problem by exploiting class peak responses to enable a classification network for instance mask extraction. With image labels supervision only, CNN classifiers in a fully convolutional manner can produce class response maps, which specify classification confidence at each image location. We observed that local maximums, i.e.*, peaks, in a class response map typically correspond to strong visual cues residing inside each instance. Motivated by this, we first design a process to stimulate peaks to emerge from a class response map. The emerged peaks are then back-propagated and effectively mapped to highly informative regions of each object instance, such as instance boundaries. We refer to the above maps generated from class peak responses as Peak Response Maps (PRMs). PRMs provide a fine-detailed instance-level representation, which allows instance masks to be extracted even with some off-the-shelf methods. To the best of our knowledge, we for the first time report results for the challenging image-level supervised instance segmentation task. Extensive experiments show that our method also boosts weakly supervised pointwise localization as well as semantic segmentation performance, and reports state-of-the-art results on popular benchmarks, including PASCAL VOC 2012 and MS COCO.* [1]

1. Introduction

Most contemporary methods of semantic segmentation rely on large-scale dense annotations for training deep models; however, annotating pixel-level masks is expensive and labor-intensive [18]. In contrast, image-level annotations, *i.e.*, presence or absence of object categories in an image, are much cheaper and easier to define. This motivates the development of weakly supervised semantic segmentation methods, which use image labels to learn convolutional neural networks (CNNs) for class-aware segmentation.

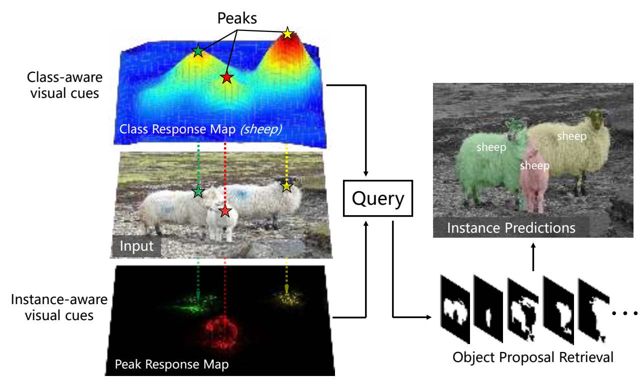

Figure 1: Class peak responses correspond to strong visual cues residing inside each respective instance. Those peaks can be back-propagated and effectively mapped to highly informative regions of each object, which allow instance masks to be extracted. Best viewed in color.

Most existing weakly supervised semantic segmentation methods consider convolutional filters in CNN as object detectors and aggregate the deep feature maps to extract class-aware visual evidence [47, 43]. Typically, pre-trained classification networks are first converted to fully convolutional networks (FCNs) to produce class response maps in a single forward pass. Such class response maps indicate essential image regions used by the network to identify an image class; however, cannot distinguish different object instances from the same category. Therefore, existing weakly supervised semantic segmentation methods cannot be simply generalized to instance-level semantic segmentation [16, 12], which aims to detect all objects in an image as well as predicting precise masks for each instance.

In this paper, we explore the challenging problem of training CNNs with image-level weak supervision for instance-level semantic segmentation (instance segmentation for short). Specifically, we propose to exploit peaks in a class response map to enable a classification network,

†Corresponding Authors
[1]Source code is publicly available at yzhou.work/PRM

e.g., VGGNet, ResNet, for instance mask extraction.

Local maximums, *i.e.*, peaks, in a class response map typically correspond to strong visual cues residing inside an instance, Fig. 1. Motivated by such observation, we first design a process to stimulate, during the training stage, peaks to emerge from a class response map. At the inference stage, the emerged peaks are back-propagated and effectively mapped to highly informative regions of each object instance, such as instance boundaries. The above maps generated from class peak responses are referred to as Peak Response Maps (PRMs). As shown in Fig. 1, PRMs serve as an instance-level representation, which specifies both spatial layouts and fine-detailed boundaries of each object; thus allows instance masks to be extracted even with some off-the-shelf methods [3, 38, 27].

Compared with many fully supervised approaches that typically use complex frameworks including conditional random fields (CRF) [46, 45], recurrent neural networks (RNN) [30, 32], or template matching [37], to handle instance extraction; our approach is simple yet effective. It is compatible with any modern network architectures and can be trained using standard classification settings, *e.g.*, image class labels and cross entropy loss, with negligible computational overhead. Thanks to its training efficiency, our method is well suited for application to large-scale data.

To summarize, the main contributions of this paper are:

- We observe that peaks in class response maps typically correspond to strong visual cues residing inside each respective instance, and such simple observation leads to an effective weakly supervised instance segmentation technique.
- We propose to exploit class peak responses to enable a classification network for instance mask extraction. We first stimulate peaks to emerge from a class response map and then back-propagate them to map to highly informative regions of each object instance, such as instance boundaries.
- We implement the proposed method in popular CNNs, *e.g.*, VGG16 and ResNet50, and show top performance on multiple benchmarks. To the best of our knowledge, we for the first time report results for the challenging image-level supervised instance segmentation task.

2. Related Work

Weakly supervised semantic segmentation. Semantic segmentation approaches typically require dense annotations in the training phase. Given the inefficiency of pixel-level annotating, previous efforts have explored various alternative weak annotations, *e.g.*, points on instances [1], object bounding boxes [5, 22], scribbles [17, 42], and human selected foreground [34]. Although effective, these approaches require significant more human efforts than image-level supervised methods [24, 25, 41, 14, 33].

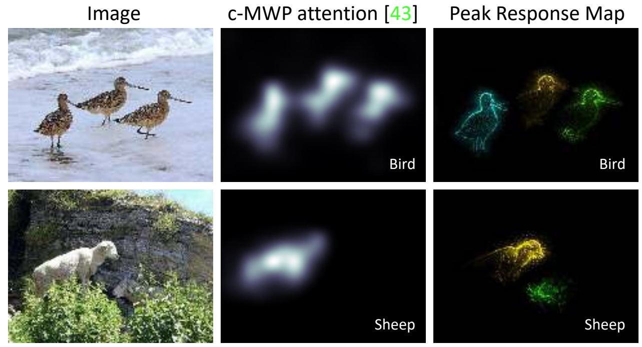

Figure 2: Compared to existing weakly supervised methods which aim to obtain a saliency map (middle) for each **class**, the proposed approach extracts fine-detailed representation (right), including both explicit layouts and boundaries, for each **instance** (visualized with different colors).

Some works leverage object cues in an unsupervised manner. For examples, graphical models have been used to infer labels for segments [44, 15], yet their object localization capacity remains limited. External localization network is therefore used to initialize object locations [26, 14, 23], and refining low-resolution CNN planes with pre-generated object segment proposal priors. Previous works usually involve time-consuming training strategies, *e.g.*, repeatedly model learning [40] or online proposal selection [29, 39]. In this work, instead, we use the standard classification networks to produce class-aware and instance-aware visual cues born with convolutional responses.

Instance segmentation. Compared with semantic segmentation that seeks to produce class-aware masks, instance segmentation requires to produce, at the same time, instance-aware region labels and fine-detailed segmentation masks and thus is much more challenging. Even with supervision from accurate pixel-level annotations, many instance segmentation approaches resort to additional constraints from precise object bounding boxes. The FCIS approach [16] combines a segment proposal module [6] and an object detection system [7]. Mask R-CNN [12] fully leverages the precise object bounding boxes generated with a proposal network [31] to aid the prediction of object masks.

With strong supervision from pixel-level GT masks, the above approaches have greatly boosted the performance of instance segmentation. However, the problem that how to perform instance segmentation under weak supervision remains open. Khoreva *et al.* [13] propose to obtain pseudo ground truth masks from bounding box supervision to alleviate labeling cost. In contrast, we leverage instance-aware visual cues naturally learned with classification networks; thus only image-level annotations are required for training.

Object prior information. When accurate annota-

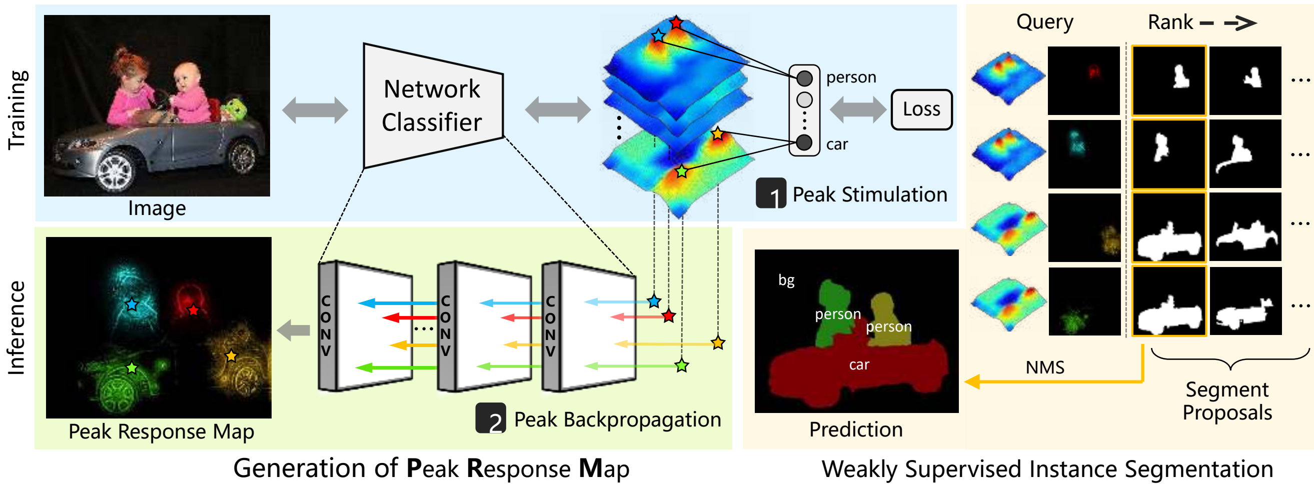

Figure 3: The generation and utilization of Peak Response Maps (PRMs). A stimulation procedure selectively activates strong visual cues residing inside each object into class peak responses. A back-propagation process further extracts fine details of each instance from the resulting peaks. Finally, class-aware cues, instance-aware cues, and object priors from proposals are considered together to predict instance masks. Best viewed in color.

tions are unavailable, visual recognition approaches leverage prior information typically to obtain additional visual cues. Object proposal methods that hypothesize object locations and extent are often used in weakly supervised object detection and segmentation to provide object priors. Selective Search [38] and Edge Boxes [49] use low-level features like color and edges as cues to produce object candidate windows. Multi-scale Combinatorial Grouping (MCG) [27] uses low-level contour information, *e.g.*, Structured Edge [8] or Ultrametric Contour Map [20], to extract object proposals, which contain fine-detailed object boundaries that is valuable to instance segmentation. In this paper, we perform instance mask extraction with the help of object priors from MCG proposals.

Image-level supervised deep activation. With image-level supervision only, it is required to aggregate deep responses, *i.e.*, feature maps, of CNNs into global class confidences so that image labels can be used for training. Global max pooling (GMP) [21] chooses the most discriminative response for each class to generate classification confidence scores, but many other informative regions are discarded. Global average pooling (GAP) [47] assigns equal importance to all responses, which makes it hard to differentiate foreground and background. The log-sum-exponential (LSE) [35] provides a smooth combination of GMP and GAP to constrain class-aware object regions. Global rank max-min pooling (GRP) [9] selects a portion of high-scored pixels as positives and low-scored pixels as negatives to enhance discrimination capacity.

Existing approaches usually activate deep responses from a global perspective without considering local spatial relevance, which makes it hard to discriminate object instances in an image. Peaks in the convolution response imply a maximal local match between the learned filters and the informative receptive field. In our method, the peak stimulation process aggregates responses from local maximums to enhance the network's localization ability.

Based on the deep responses, top-down attention methods are proposed to generate refined class saliency maps by exploring visual attention evidence [4, 43]. These class-aware and instance-agnostic cues can be used in semantic segmentation [14, 33] yet is insufficient for instance segmentation, Fig. 2. In contrast, our methods provide fine-detailed instance-aware cues that are suitable for weakly supervised instance-level problems.

3. Method

In this section, we present an image-level supervised instance segmentation technique that utilizes class peak response. CNN classifiers in the fully convolutional manner can produce class response maps, which specify classification confidence at each image location [21]. Based on our observation that local maximums, *i.e.*, peaks, of class response maps typically correspond to strong visual cues residing inside an instance, we first design an process to stimulate peaks to emerge from a class response map in the network training phase. During the inference phase, emerged peaks are back-propagated to generate maps that highlight informative regions for each object, referred to as Peak Response Maps (PRMs). PRMs provide a fine-detailed separate representation for each instance, which are further exploited to retrieve instance masks from object segment proposals off-the-shelf, Fig. 3.

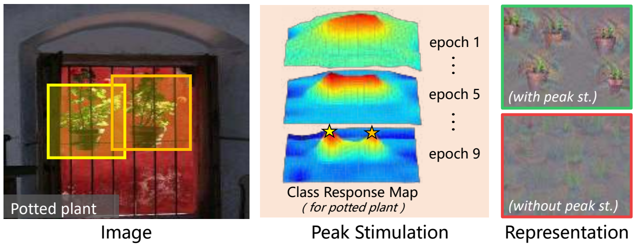

Figure 4: With peak stimulation, multiple instances can be better distinguished on the class response map (middle). The learned representations (right) are visualized by **activation maximization** [10]. Best viewed in color.

3.1. Fully Convolutional Architecture

By simply removing the global pooling layer and adapting fully connected layers to 1x1 convolution layers, modern CNN classifiers can be seamlessly converted to fully convolutional networks (FCNs) [19] that naturally preserve spatial information throughout the forwarding. The converted network outputs class response maps with a single forward pass; therefore are suitable for spatial predictions. In this work, networks are converted to FCN first.

3.2. Peak Stimulation

To stimulate peaks to emerge from class response maps, we construct a peak stimulation layer, to be inserted after the top layer, Fig. 3. Consider a standard network, let $M \in \mathbb{R}^{C \times H \times W}$ denotes the class response maps of the top convolutional layer, where C is the number of classes, and $H \times W$ denotes the spatial size of the response maps. Therefore, the input of the peak stimulation layer is M and the output is class-wise confidence scores $s \in \mathbb{R}^C$. Peaks of the c-th response map M^c are defined to be the local maximums within a window size of r [2], and the location of peaks are denoted as $P^c = \{(i_1, j_1), (i_2, j_2), ..., (i_{N^c}, j_{N^c})\}$, where N^c is the number of valid peaks for the c-th class. During the forwarding pass, a sampling kernel $G^c \in \mathbb{R}^{H \times W}$ is generated for computing the classification confidence score of the c-th object category. Each kernel element at the location (x, y) can be accessed with $G^c_{x,y}$. Without loss of generality, the kernel is formed as

$$G^c_{x,y} = \sum_{k=1}^{N^c} f(x - i_k, y - j_k), \tag{1}$$

where $0 \leq x < H, 0 \leq y < W$, (i_k, j_k) is the coordinate of the k-th peak, and f is a sampling function. In our settings, f is a Dirac delta function for aggregating features from the peaks only; therefore the confidence score of the c-th category s^c is then computed by the convolution between the class response map M^c and sampling kernel G^c, as

[2] The region radius r for peak finding is set to 3 in all our experiments.

$$s^c = M^c * G^c = \frac{1}{N^c} \sum_{k=1}^{N^c} M^c_{i_k, j_k}. \tag{2}$$

It can be seen from Eq. 2 that the network uses peaks only to make the final decision; naturally, during the backward pass, the gradient is apportioned by G^c to all the peak locations, as

$$\delta^c = \frac{1}{N^c} \cdot \frac{\partial L}{\partial s^c} \cdot G^c, \tag{3}$$

where δ^c is the gradient for the c-th channel of the top convolutional layer and L is the classification loss.

From the perspective of model learning, the class response maps are computed by the dense sampling of all receptive fields (RFs), in which most of RFs are negative samples that do not contain valid instances. Eq. 3 indicates that in contrast to conventional networks which unconditionally learn from the extreme foreground-background imbalance set, peak stimulation forces the learning on a sparse set of informative RFs (potential positives and hard negatives) estimated via class peak responses, thus prevents the vast number of easy negatives from overwhelming the learned representation during training, Fig. 4 (right).

3.3. Peak Back-propagation

We propose a probability back-propagation process for peaks to further generate the fine-detailed and instance-aware representation, *i.e.*, Peak Response Map. In contrast to previous top-down attention models [43, 36], which seek the most relevant neurons of an output category to generate class-aware attention maps, our formulation explicitly considers the receptive field and can extract instance-aware visual cues from the specific spatial locations, *i.e.*, class peak responses. The peak back-propagation can be interpreted as a procedure that a walker starts from the peak (top layer) and walk randomly to the bottom layer. The top-down relevance of each location in the bottom layer is then formulated as its probability of being visited by the walker.

Consider a convolution layer that has a single filter $W \in \mathbb{R}^{kH \times kW}$ for mathematical simplification, the input and output feature maps are denoted as U and V, where each spatial locations can be accessed by U_{ij} and V_{pq} respectively. The visiting probability $P(U_{ij})$ can be obtained by $P(V_{pq})$ and the transition probability between two maps, as

$$P(U_{ij}) = \sum_{p=i-\frac{kH}{2}}^{i+\frac{kH}{2}} \sum_{q=j-\frac{kW}{2}}^{j+\frac{kW}{2}} P(U_{ij}|V_{pq}) \times P(V_{pq}), \tag{4}$$

where the transition probability is defined as

$$P(U_{ij}|V_{pq}) = Z_{pq} \times \hat{U}_{ij} W^+_{(i-p)(j-q)}. \tag{5}$$

$\hat{U}_{ij}$ is the bottom-up activation (computed in the forward pass) at the location (i, j) of U, $W^+ = ReLU(W)$, which

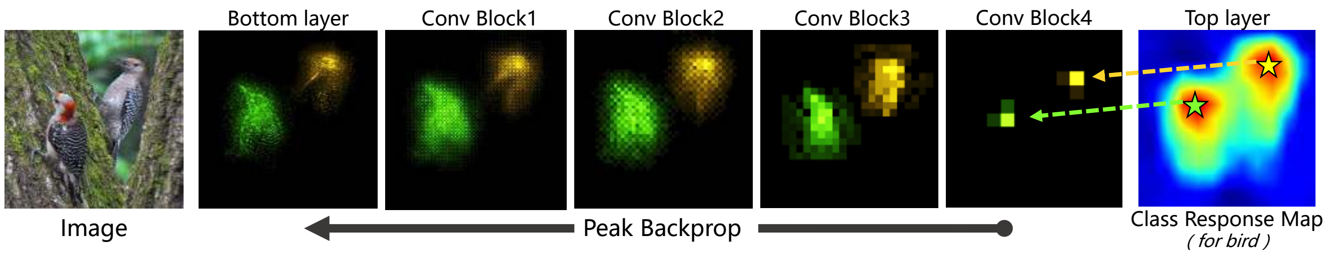

Figure 5: Peak back-propagation process maps class peak responses to fine detailed visual cues residing inside each object, *i.e*., Peak Response Maps (PRMs), enabling the instance-level masks to be extracted. Best viewed in color.

discards negative connections, and Z_{pq} is a normalization factor to guarantee $\sum_{p,q} P(U_{ij}|V_{pq}) = 1$. Note in most modern CNNs that adopt ReLU as transfer function, negative weights have no positive effects in enhancing the output response, thus are excluded from propagation.

Other commonly used intermediate layers, *e.g*., the average pooling and max-pooling layers, are regarded as the same type of layers that perform an affine transform of the input [43]; thus the corresponding back-propagation can be modeled in the same way of convolution layers.

With the probability propagation defined by Eq. 4 and Eq. 5, we can localize most relevant spatial locations for each class peak response in a top-down fashion, to generate fine-detailed instance-aware visual cues, referred to as Peak Response Map, Fig. 5.

3.4. Weakly Supervised Instance Segmentation

We further leverage the instance-aware cues of PRMs to perform challenging instance segmentation tasks. Specifically, we propose a simple yet effective strategy to predict mask for each object instance by combining instance-aware cues from PRMs, class-aware cues from class response maps, and spatial continuity priors from object proposals off-the-shelf [38, 27, 20].

We retrieve instance segmentation masks from a proposal gallery, Fig. 3, with the metric,

$$Score = \underbrace{\alpha \cdot R * S}_{\text{instance-aware}} + \underbrace{R * \hat{S}}_{\text{boundary-aware}} \underbrace{- \beta \cdot Q * S}_{\text{class-aware}}, \quad (6)$$

where R is the PRM corresponds to a class peak response, $\hat{S}$ is the contour mask of the proposal S computed by morphological gradient, and Q is the background mask obtained by the class response map and a bias (based on the mean value of the map). The class independent free parameters α and β are selected on the validation set.

In Eq. 6, the instance-aware term encourages proposal to maximize the overlap with PRM, while the boundary-aware term leverages the fine-detailed boundary information within the PRM to select proposal with a similar shape. Furthermore, the class-aware term uses class response map to suppress class-irrelevant regions. The effects of three terms are ablation studied in Sec. 4.3.

Algorithm 1 Segment Instances via Class Peak Response

Input: A test image I, segment proposals $\mathcal{S}$, and a network trained with peak stimulation.
Output: Instance segmentation prediction set $\mathcal{A}$

```
Initialize instance prediction set A = ∅;
Forward I to get class response maps M;
for map M^k of the k-th class in M do
    Detect peaks P_i^k and add to P, Sec. 3.2;
end for
for peak P_i^k in P do
    Peak backprop at P_i^k to get PRM R, Sec. 3.3;
    for proposal S_j in S do
        Compute score using R and M^k, Eq. 6;
    end for
    Add top-ranked proposal and label (S_*, k) to A;
end for
Do Non-Maximum Suppression (NMS) over A.
```

The overall algorithm for weakly supervised instance segmentation is specified in Alg. 1.

4. Experiment

We implement the proposed method using state-of-the-art CNN architectures, including VGG16 and ResNet50, and evaluate it on several benchmarks. In Sec. 4.1, we perform a detailed analysis of the peak stimulation and back-propagation process, to show that the proposed technique can generate accurate object localization and high-quality instance-aware cues. In Sec. 4.2, on weakly supervised semantic segmentation, the ability of PRMs to extract class-aware masks with the help of segment proposals is shown. In Sec. 4.3, we for the first time report results for challenging image-level supervised instance segmentation. Ablation study and upper bound analysis are further performed to demonstrate the effectiveness and potential of our method.

4.1. Peak Response Analysis

Pointwise localization. A pointwise object localization metric [21] is used to evaluate the localization ability of class peak responses and effectiveness of peak stimulation. We first upsample the class response maps to the size of the image via bilinear interpolation. For each predicted class,

Method	VOC 2012	MS COCO
DeepMIL [21]	74.5	41.2
WSLoc [2]	79.7	49.2
WILDCAT [9]	82.9	53.5
SPN [48]	82.9	55.3
Ours (w/o Peak Stimulation)	81.5	53.1
Ours (full approach)	**85.5**	**57.5**

Table 1: Mean Average Precision (mAP%) of pointwise localization on VOC2012 and COCO2014 val. set.

if the coordinate of the maximum class peak response falls into a ground truth bounding box of the same category, we count a true positive.

We fine-tune ResNet50 equipped with/without peak stimulation on the training set of PASCAL VOC 2012 [11] as well as MS COCO 2014 [18], and report performances on the validation set, Tab. 1. The results show that class peak responses correspond to visual cues of objects and can be used to localize objects. Our full approach shows top performance against state-of-the-arts and outperforms the baseline (w/o stimulation) by a large margin, which indicates the stimulation process can lead the network to discover better visual cues correspond to valid instances.

Quality of peak response maps. To evaluate the quality of extracted instance-aware cues, we measure the correlation between a Peak Response Map (PRM) R and a GT mask G with $\frac{\sum R \odot G}{\sum R}$, which indicates the ability of the PRM to discover visual cues residing inside the instance. For each PRM, we define its score to be the largest correlation with GT masks of the same class. Thus, a score of 0 indicates that the corresponding PRM does not locate any valid object region, while a score of 1 implies the PRM perfectly distinguishes the visual cues of an instance from the background. PRMs with a score higher than 0.5 are considered as true positives. On VOC 2012, we use classification data to train ResNet50 equipped with response aggregation strategies from different methods, and evaluate the quality of resulting PRMs on the validation set of the segmentation data in terms of mAP, Tab. 2. Peak stimulation forces networks to learn an explicit representation from informative receptive fields; thus obtaining higher quality of PRMs.

We perform statistical analysis on the relationship between the PRM quality and the crowding level of images, Fig. 6 (left). On average, the energy of PRMs that falls into an instance reaches 78% for images with a single object, and 67% for images with 2-5 objects. Surprisingly, even for crowded scenes with more than six objects, the instances collect more energy than the background on average, which shows that the instance-aware visual cues from PRMs are of high quality. We further analyze the impact of object size, Fig. 6 (right), and results show that PRMs can localize fine-detailed evidence from common size objects.

Method	Response Aggregation Strategy	mAP
CAM [47]	Global Average Pooling	55.7
DeepMIL [21]	Global Max Pooling	60.9
WILDCAT [9]	Global Max-Min Pooling	62.4
PRM (Ours)	Peak Stimulation	**64.0**

Table 2: Comparison of the effect of different response aggregation strategies on the quality of Peak Response Maps.

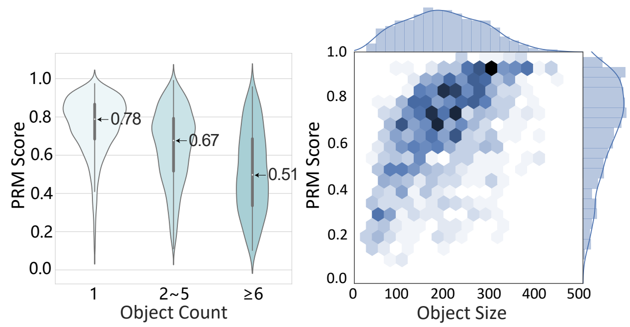

Figure 6: Statistical analysis of the effect of the number and size of objects on the quality of Peak Response Maps.

Method	mIoU	Comments
MIL+ILP+SP-seg † [26]	42.0	Object segment proposals
WILDCAT † [9]	43.7	CRF post-processing
SEC [14]	50.7	CRF as boundary loss
Check mask [34]	51.5	CRF & Human in the loop
Combining [33]	52.8	CRF as RNN
PRM (Ours) †	**53.4**	Object segment proposals

Table 3: Weakly supervised semantic segmentation results on VOC 2012 val. set in terms of the mean IoU (%). Mark † indicates methods that introduce negligible training costs.

4.2. Weakly Supervised Semantic Segmentation

Experiments above shows that the PRMs correspond to accurate instance "seeds" while another challenging thing is to expand each seed into full object segmentation. We evaluate the ResNet50 model equipped with peak stimulation on the weakly supervised semantic segmentation task, which requires assigning objects from the same categories as the same segmentation labels. On the validation set of VOC 2012 segmentation data, We merge the instance segmentation masks of the same class to produce semantic segmentation predictions. The performance is measured regarding pixel intersection-over-union averaged across 20+1 classes (20 object categories and background).

Instead of using time-consuming training strategies [33], or additional supervisions [1, 34], our method trains models using image-level labels and standard classification settings, and reports competitive results, on weakly supervised semantic segmentation without CRF post-processing, Tab. 3. Fig. 7 shows examples of predictions in different scenarios.

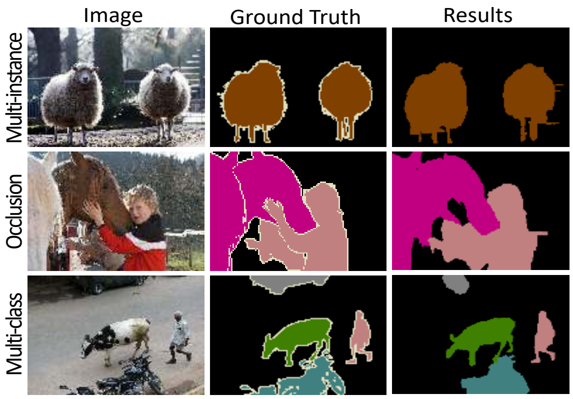

Figure 7: Examples of predicted semantic segmentations. Different colors indicate different classes.

4.3. Weakly Supervised Instance Segmentation

With the proposed technique, we perform instance segmentation on the PASCAL VOC 2012 segmentation set with ResNet50 and VGG16 models trained on the classification set. To the best of our knowledge, this is the first work reporting results for image-level supervised instance segmentation. We construct several baselines based on object bounding boxes obtained from ground truth and weakly supervised localization methods [39, 47, 48], Tab. 4. With the localized bounding boxes, we set three reasonable mask extraction strategies: (1) Rect. Simply filling in the object boxes with instance labels, (2) Ellipse. Fitting a maximum ellipse inside each box, and (3) MCG. Retrieving an MCG segment proposal of maximum IoU with the bounding box.

Numerical results. The instance segmentation is evaluated with the mAP^r at IoU threshold 0.25, 0.5 and 0.75, and the Average Best Overlap (ABO) [28] metric is also employed for evaluation to give a different perspective. Tab. 4 shows that our approach significantly outperforms weakly supervised localization techniques that use the same setting, *i.e.*, using image-level labels only for model training. The performance improvement at lower IoU thresholds, *e.g.*, 0.25 and 0.5, shows the effectiveness of peak stimulation for object location, while the improvement at higher IoU threshold, *e.g.*, 0.75, indicates the validity of peak back-propagation for capturing fine-detailed instance cues.

Compare with the latest state-of-the-art MELM [39], which is trained with multi-scale augmentation, online proposal selection, and a specially designed loss, our method is simple yet effective and shows a competitive performance.

Ablation study. To investigate the contribution of peak stimulation as well as each term in our proposal retrieval metric, we perform instance segmentation based on different backbones in which different factors were omitted. The results are presented in Tab. 5. From the ablation study, we can draw the following conclusions: 1). Peak stimulation process, which stimulates peaks during network training, is crucial to the instance segmentation performance of our method. 2). The $mAP^r_{0.5}$ dramatically drops from 26.8% to 13.3% when omitting the instance-aware term, which demonstrates the effectiveness of the well-isolated instance-aware representation generated by our method. 3). Boundary-aware term significantly improves the performance by 2.5% shows our method does extract fine-detailed boundary information of instances. 4). Class-aware cues depress class-irrelevant regions; thus substantially improve the instance segmentation performance of our method.

Method		$mAP^r_{0.25}$	$mAP^r_{0.5}$	$mAP^r_{0.75}$	ABO
Ground Truth	Rect.	78.3	30.2	4.5	47.4
	Ellipse	81.6	41.1	6.6	51.9
	MCG	69.7	38.0	12.3	53.3
Training requires image-level labels and object proposals					
MELM [39]	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
Training requires only image-level labels					
CAM [47]	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 [48]	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
PRM (Ours)		**44.3**	**26.8**	**9.0**	**37.6**

Table 4: Weakly supervised instance segmentation results on the PASCAL VOC 2012 val. set in terms of mean average precision (mAP%) and Average Best Overlap (ABO).

	ResNet50					VGG16	
Peak Stimulation		✓	✓	✓	✓		✓
Instance-aware term	✓		✓	✓	✓	✓	✓
Class-aware term	✓	✓		✓	✓	✓	✓
Boundary-aware term	✓	✓	✓		✓	✓	✓
$mAP^r_{0.5}$	22.8	13.3	16.5	24.3	**26.8**	11.9	22.0

Table 5: Ablation study on the PASCAL VOC2012 val. set based on different network backbones.

Qualitative results. In Fig. 8, we illustrate some instance segmentation examples including successful cases and typical failure cases. It can be seen that our approach can produce high quality visual cues and obtain decent instance segmentation results in many challenging scenarios. In the first and second columns, it can distinguish instances when they are closed or occluded with each other. Examples in the third and fourth columns show that it performs well with objects from different scales. In the fifth column, objects from different class are well segmented, which shows that the proposed method can extract both class-discriminative and instance-aware visual cues from classification networks. As is typical for weakly-supervised systems, PRMs can be misled by noisy co-occurrence patterns and sometimes have problems telling the difference between object parts and multiple objects. We address this

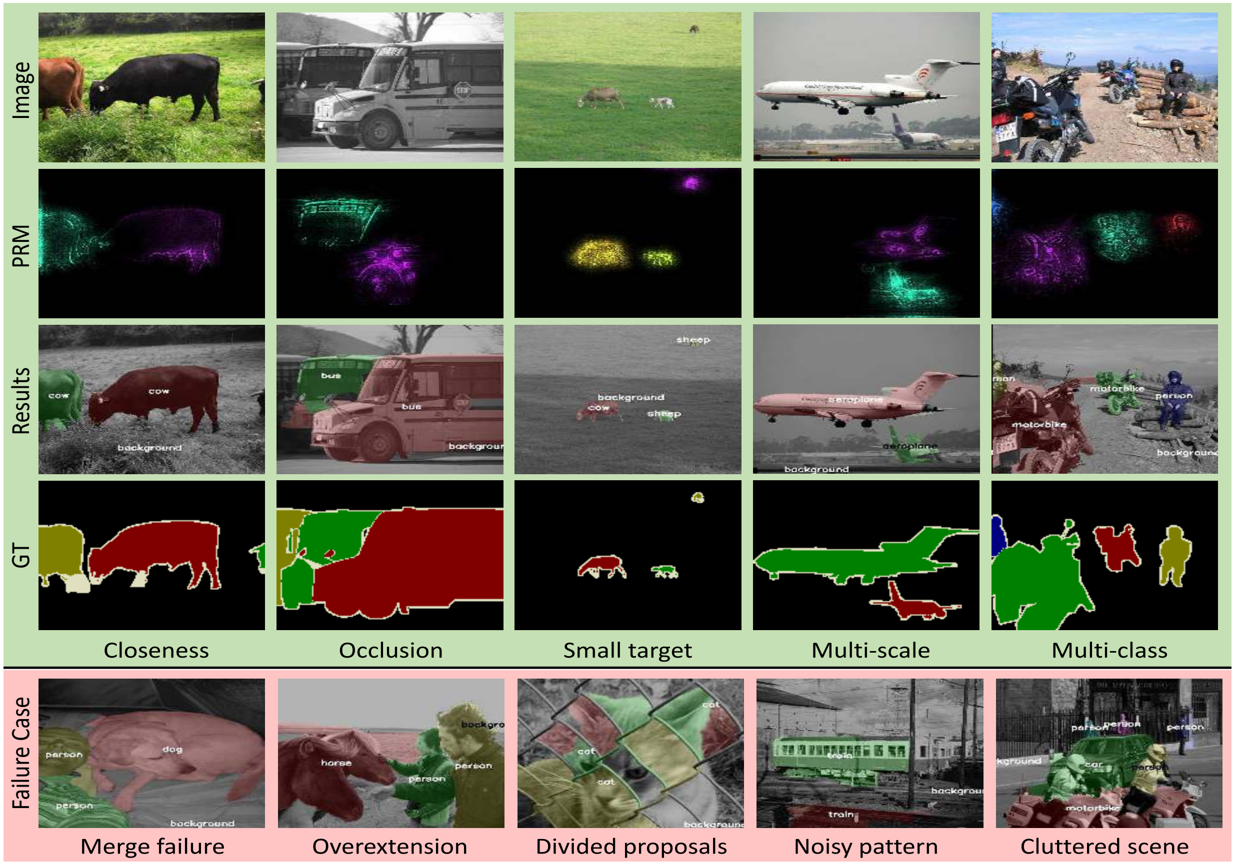

Figure 8: Instance segmentation examples on the PASCAL VOC 2012 val. set. It can be seen that the Peak Response Maps (second row) incorporate fine-detailed instance-aware information, which can be exploited to produce instance-level masks (third row). The last row shows typical failure cases. Best viewed in color.

Proposal gallery	GT mask	GT bbox	SPN [48]	PRM (Ours)
MCG	26.0	12.3	4.4	9.0
MCG + GT mask	100.0	29.2	10.4	26.9
GT mask	100.0	93.0	50.0	73.3

Table 6: Comparison of instance segmentation results ($mAP^r_{0.75}$) on the PASCAL VOC 2012 val. set.

problem with a proposal retrieval step; nevertheless, the performance remains limited by proposal quality.

Upper bound analysis. To explore the upper bound of our method, we construct different proposal galleries, Tab. 6. First, we mix GT masks into MCG proposals to get a gallery with 100% recall, and the results show that the capability of our method (image-level supervised) to retrieve proposals is comparable to GT bbox (26.9% vs. 29.2%). Next, we use GT masks as a perfect proposal gallery (note that GT bbox still fails in highly occlusion cases) to evaluate the instance localization ability of PRMs. Our result further boosts to 73.3% and outperforms SPN by a large margin, demonstrating the potential of the proposed technique on video/RGB-D applications where rich information can be exploited to generate proposals of high quality.

5. Conclusions

In this paper, we propose a simple yet effective technique to enable classification networks for instance mask extraction. Based on class peak responses, the peak stimulation shows effective to reinforce object localization, while the peak back-propagation extracts fine-detailed visual cues for each instance. We show top results for pointwise localization as well as weakly supervised semantic segmentation and, to the best of our knowledge, for the first time report results for image-level supervised instance segmentation. The underlying fact is that instance-aware cues are naturally learned by convolutional filters and encoded in hierarchical response maps. To discover these cues provides fresh insights for weakly supervised instance-level problems.

Acknowledgements

The authors are very grateful for support by the NSFC grant 61771447 / 61671427, BMSTC, and NSF.

References

[1] A. Bearman, O. Russakovsky, V. Ferrari, and L. FeiFei. Whats the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision (ECCV)*, pages 543–559, 2016. 2, 6

[2] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. S. Manjunath. Weakly supervised localization using deep feature maps. In *European Conference on Computer Vision (ECCV)*, pages 714–731, 2016. 6

[3] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 105–112. IEEE, 2001. 2

[4] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2956–2964, 2015. 3

[5] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643, 2015. 2

[6] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3150–3158, 2016. 2

[7] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 379–387, 2016. 2

[8] P. Dollár and C. L. Zitnick. Fast edge detection using structured forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(8):1558–1570, 2015. 3

[9] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 6

[10] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. 4

[11] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015. 6

[12] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 1, 2

[13] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[14] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 695–711. Springer, 2016. 2, 3, 6

[15] B. Lai and X. Gong. Saliency guided dictionary learning for weakly-supervised image parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3630–3639, 2016. 2

[16] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[17] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016. 2

[18] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 1, 6

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 4

[20] K. Maninis, J. Pont-Tuset, P. A. Arbeláez, and L. V. Gool. Convolutional oriented boundaries: From image segmentation to high-level tasks. *CoRR*, abs/1701.04658, 2017. 3, 5

[21] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 685–694, 2015. 3, 5, 6

[22] G. Papandreou, L. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. *IEEE International Conference on Computer Vision (ICCV)*, 2,4,6,5,7,11,14, 2015. 2

[23] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1742–1750, 2015. 2

[24] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. *ICLR Workshop*, 2, 2015. 2

[25] P. H. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1713–1721, 2015. 2

[26] P. O. Pinheiro and R. Collobert. Weakly supervised semantic segmentation with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6, 2015. 2, 6

[27] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marqués, and J. Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(1):128–140, 2017. 2, 3, 5

[28] J. Pont-Tuset and L. Van Gool. Boosting object proposals: From pascal to coco. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1546–1554, 2015. 7

[29] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *European Conference on Computer Vision (ECCV)*, pages 90–105, 2016. 2

[30] M. Ren and R. S. Zemel. End-to-end instance segmentation and counting with recurrent attention. *arXiv preprint arXiv:1605.09410*, 2016. 2

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 2

[32] B. Romera-Paredes and P. H. S. Torr. Recurrent instance segmentation. In *European Conference on Computer Vision*, pages 312–329. Springer, 2016. 2

[33] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3529–3538, 2017. 2, 3, 6

[34] 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 *European Conference on Computer Vision (ECCV)*, pages 413–432. Springer, 2016. 2, 6

[35] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. Bourdev. Pronet: Learning to propose object-specific boxes for cascaded neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3493, 2016. 3

[36] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1-2):507–545, 1995. 4

[37] J. Uhrig, M. Cordts, U. Franke, and T. Brox. Pixel-level encoding and depth layering for instance-level semantic labeling. In *German Conference on Pattern Recognition*, pages 14–25. Springer, 2016. 2

[38] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013. 2, 3, 5

[39] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye. Min-entropy latent model for weakly supervised object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7

[40] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[41] Y. Wei, X. Liang, Y. Chen, X. Shen, M. Cheng, J. Feng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(11):2314–2320, 2017. 2

[42] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3781–3790, 2015. 2

[43] J. Zhang, Z. L. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision (ECCV)*, pages 543–559, 2016. 1, 3, 4, 5

[44] W. Zhang, S. Zeng, D. Wang, and X. Xue. Weakly supervised semantic segmentation for social images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2718–2726, 2015. 2

[45] Z. Zhang, S. Fidler, and R. Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 669–677, 2016. 2

[46] Z. Zhang, A. G. Schwing, S. Fidler, and R. Urtasun. Monocular object instance segmentation and depth ordering with cnns. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2614–2622, 2015. 2

[47] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1, 3, 6, 7

[48] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft proposal networks for weakly supervised object localization. *arXiv preprint arXiv:1709.01829*, 2017. 6, 7, 8

[49] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, pages 391–405, 2014. 3