Learning Instance Activation Maps for Weakly Supervised Instance Segmentation

Yi Zhu[1], Yanzhao Zhou[1], Huijuan Xu[2], Qixiang Ye[1], David Doermann[3], Jianbin Jiao[1]

[1]University of Chinese Academy of Sciences [2]University of California, Berkeley [3]University at Buffalo

{zhuyi215,zhouyanzhao215}@mails.ucas.ac.cn huijuan@berkeley.edu
{jiaojb,qxye}@ucas.ac.cn doermann@buffalo.edu

Abstract

Discriminative region responses residing inside an object instance can be extracted from networks trained with image-level label supervision. However, learning the full extent of pixel-level instance response in a weakly supervised manner remains unexplored. In this work, we tackle this challenging problem by using a novel instance extent filling approach. We first design a process to selectively collect pseudo supervision from noisy segment proposals obtained with previously published techniques. The pseudo supervision is used to learn a differentiable filling module that predicts a class-agnostic activation map for each instance given the image and an incomplete region response. We refer to the above maps as Instance Activation Maps (IAMs), which provide a fine-grained instance-level representation and allow instance masks to be extracted by lightweight CRF. Extensive experiments on the PASCAL VOC12 dataset show that our approach beats the state-of-the-art weakly supervised instance segmentation methods by a significant margin and increases the inference speed by an order of magnitude. Our method also generalizes well across domains and to unseen object categories. Without fine-tuning for the specific tasks, our model trained on VOC12 dataset (20 classes) obtains top performance for weakly supervised object localization on the CUB dataset (200 classes) and achieves competitive results on three widely used salient object detection benchmarks.

1. Introduction

Powered by the recent advances of Deep Convolutional Neural Networks (DCNNs), instance segmentation has made remarkable progress [13, 6, 25]. Deep learning approaches, however, typically require large amounts of data for training and rely on detailed ground truth (GT) in the form of instance masks which often requires extensive human effort. For example, in the CityScapes dataset [9], fine-detailed pixel-level annotation typically requires more than 1.5 hours for a single image. In contrast, image-level labels, *i.e.*, the presence or absence of object categories in an image, are much easier to define and can even be automatically collected from the Internet.

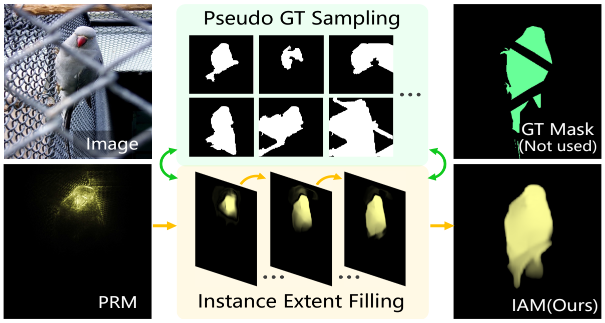


Figure 1: Peak Response Maps [43] from classification networks can only identify the discriminative parts of each instance. Our approach collects pseudo ground-truth masks from noisy segment proposals obtained with off-the-shelf techniques and learns an extent filling module. The resulting Instance Activation Maps (IAMs) effectively localize instance-level spatial extent.

Learning instance segmentation with image-level labels is a challenging task as the annotation does not inform the location or spatial extent of objects in an image. Zhou *et al*. [43] took the first step in addressing this task by extracting instance-aware visual cues from classification networks. They produce response maps for each object category, *i.e.*, Class Activation Maps (CAMs) [42], to indicate essential receptive fields used by the network when identifying the object class. The peaks, *i.e.*, local maximas, of CAMs are stimulated and back-propagated to generate Peak Response Maps (PRMs) that highlight informative regions residing inside each object instance. PRMs could identify discriminative parts of each object yet failed to localize other regions. As shown in Fig. 2, PRMs highlight the dog head while ignoring the dog's body. The reason may lie in the fact that the body region could be obscured when identifying "dog" but the head is essential for classification. Therefore, PRMs do not complete the instance's extent. This limits its performance and reduces its inference efficiency due to its dependence on a costly instance mask generation strategy, *i.e.*, retrieving segment object proposals obtained with low-level vision techniques [33, 29, 28].

In this paper, we address the problem of learning instance extent in a weakly supervised manner by developing a novel instance extent filling approach. We first leverage incomplete region responses obtained with the previously developed PRM method [43] to collect pseudo ground-truth (GT) masks from noisy object segment proposals. The pseudo GT masks are then used to learn a differentiable filling module that predicts a class-agnostic activation map for each instance conditioned on the image and an incomplete region response. The result is an Instance Activation Map (IAM) that specifies both spatial layout and fine-detailed instance boundaries, Fig. 1. This allows instance masks to be directly extracted with the lightweight and GPU-friendly dense CRF post-processing [19, 32]. As a result, we significantly improve the state-of-the-art weakly supervised instance segmentation performance as well as increase the inference speed by an order of magnitude.

Our approach learns instance extent knowledge from image-level labels and noisy segment proposals. We also show that the learned knowledge generalizes well across domains and to unseen object categories. This extends the application of the proposed approach to many other object extent related visual tasks. Our model obtains competitive performance on weakly supervised object localization and salient object detection benchmarks without fine-tuning the extent filling module for the specific tasks.

The main contributions of this paper include:

- The development of an instance extent filling approach to tackle the challenging problem of weakly supervised instance segmentation task by collecting pseudo GT masks from noisy segment proposals, and then train a differentiable filling module to learn common knowledge of class-agnostic object extent.

- An implementation of our approach with popular DCNNs, *e.g.*, ResNet50, that demonstrate substantial improvement over the state-of-the-art with respect to both performance and inference speed.

- A demonstration of the fact that the extent knowledge learned by the proposed approach generalizes well and achieves a performance that matches or exceeds state-of-the-art on object extent related tasks such as weakly supervised object localization and salient object detection without fine-tuning for the specific tasks.

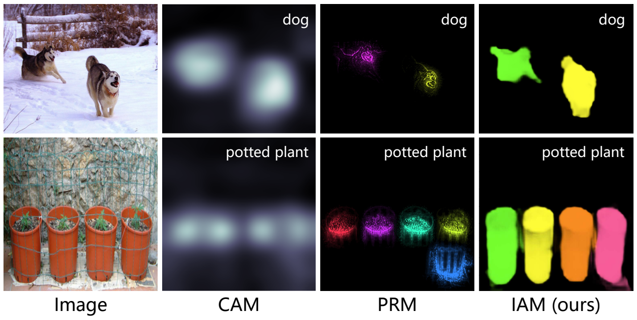


Figure 2: The activation maps from three methods, Class Activation Map (CAM) [42], Peak Response Map (PRM) [43] and our Instance Activation Map (IAM), shown in sequence for two example images. Our IAM covers full object extent while the other two methods only show coarse location or discriminative object parts.

2. Related Work

Weakly supervised instance segmentation: As one of the most challenging problems in computer vision, instance segmentation has been extensively investigated [22, 10, 2, 13, 6, 25]. Nevertheless, many of these works require strong supervision in the form of human annotated instance masks which limits their application on large-scale datasets with weaker forms of labeling. Weakly supervised instance segmentation tries to break this limitation. To perform instance segmentation with few annotations, partial supervision [16] performs instance segmentation on datasets where that a subset of classes have instance mask annotations during training. The remaining classes have only bounding box annotations. Weakly supervised instance segmentation with object bounding box supervision [18] uses object bounding boxes to construct pseudo GT masks to train instance segmentation models.

Although these methods have relaxed their reliance on accurate pixel-level masks, they still require instance-level labeling, which requires the location of each object. In [43], Zhou *et al*. for the first time proposed to tackle weakly supervised instance segmentation by exploiting the class peak response of classification networks to extract instance-aware visual cues. The cues were then used to retrieve proposals as instance masks. Nevertheless, as the filters learned for image classification typically corresponds to discriminative object parts, this approach failed to locate the full object extent and thus misled the instance mask generation. We aim to address this issue and extract complete instance-level representations by compensating for the missing extent information by learning knowledge of object extent from segment proposals off-the-shelf. As more areas of the instance are activated in the proposed Instance Activation Maps (IAMs), our method can better improve instance segmentation performance and increase the inference speed via lightweight post-processing strategy.

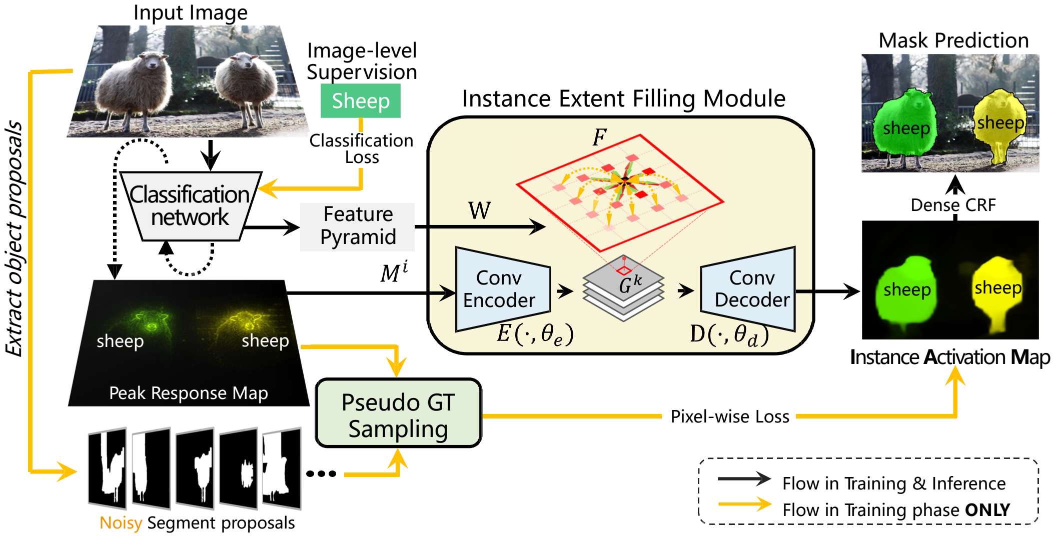


Figure 3: An overview of learning Instance Activation Maps for weakly supervised instance segmentation. The images are first fed to a classification network to generate deep feature pyramids and Peak Response Maps (PRMs) which highlight object parts. The deep features are used to construct the weights for an Instance Extent Filling module which recover instance extent from the PRMs. During training, the filling module collects common knowledge of object extent from pseudo GT masks.

Learning pixel-level affinity: Our work is also related to the approaches which leverage/learn pixel-level affinity for image segmentation [8, 3, 34, 1]. Some of these approaches use semantic segmentation labels to estimate a pixel-level affinity matrix of an image by training deconvolutional networks [3] or refinement modules, such as dense CRFs [8]. In contrast, our goal is to fill object regions by leveraging the sparse and partial visual cues under weak supervision. With solely image-level categories available, synthetic labels extracted from class response maps are employed to train a network which learns pairwise semantic affinity [1]. Our approach learns instance-aware affinity which extends beyond the semantic affinity. Our approach is also related to the Spatial Propagation Network [24] which learns semantically-aware affinity values for high-level vision tasks. The difference lies that [24] relies on GT masks while our approach uses image-level labels and inaccurate class-agnostic proposals off-the-shelf.

Region proposal: Due to the lack of object mask annotations, weakly supervised methods typically introduce object priors from region proposals. Classical region proposal methods [33, 46, 29, 11] hypothesize object candidates based on class-agnostic low-level features, *e.g.*, color, texture, edge, and contours. Therefore, proposal techniques typically generalize well and can be used off-the-shelf without introducing human labeling efforts for each specific task. The pre-computed proposals could be used to narrow the solution space in the pre-processing stage [4, 35] or to refine prediction boundaries during post-processing [31, 43]. We random sample noisy proposals to construct pseudo GT masks during training and statistically learn a differentiable instance extent filling module.

3. Method

In this section, we first revisit the previously published method [43] that we use to extract incomplete instance region responses from CNNs trained with image-level class labels. We then introduce the proposed instance extent filling approach, starting with the process of collecting pseudo GT masks and followed by the design of the extent filling module. Finally, we discuss the insights of the method and specify the implementation details. The overall architecture of our approach is illustrated in Fig. 3.

3.1. Revisiting Peak Response Mapping

We use the technique in [43] to extract Peak Response Maps (PRMs) from classification networks. The network is first converted to a fully convolutional network (FCN) by removing the global pooling layer and transforming the weights of the fully connected layers to 1x1 convolutional filters. The FCN outputs CAMs $M \in \mathbb{R}^{C \times N \times N}$ with a single forward pass, where C denotes the number of image classes, and $N \times N$ is the spatial size of the maps. The class peak responses (local maximums) of the c-th class response map M^c, are then detected and averaged to predict a confidence score for the c-th image class.

During training, the classification loss drives the network to learn multiple discriminative class peaks, and the learned peaks can be back-propagated to PRMs by leveraging the top-down relevance between spatial locations of adjacent layers as:

$$P(U^k_{pq}) = \sum_{k \in \mathcal{I}_c} \sum_{(p,q) \in \mathcal{M}^k_{ij}} P(U^k_{pq} | V^c_{ij}) P(V^c_{ij}), \qquad (1)$$

where $P(U^k_{pq} | V^c_{ij}) = \sum_{(i,j) \in \mathcal{N}^c_{pq}} Z_{pq} \times U^k_{ij} \hat{W}_{pq}$. U, V are

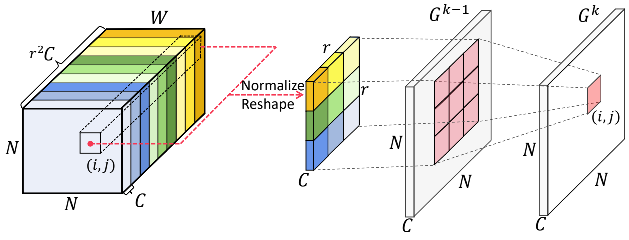


Figure 4: Illustration of the filling process. The value of each pixel is filled with the value from its neighbors. The process is performed in a convolutional manner to facilitate Cuda acceleration.

outputs of two adjacent layers. $\mathcal{I}_c$ is a set of feature maps connected to V^c. $\mathcal{H}^k_{ij}$ is the set of locations in U that connect to V^c_{ij} via non-negative convolutional weights. $\mathcal{N}^c_{pq}$ is the set of locations in V that connected to U^c_{pq} via non-negative convolutional weights $\hat{W}$. The negative weights are discarded as the commonly used ReLU activation layer prevents them from contributing to the final response. Z_{pq} is a normalization factor which guarantees the transition probabilities sum to one. With the probability propagation defined by Eq. 1 and an initial probability map that only a peak location is 1.0, we can identify which locations in the bottom layer (pixel space) contributes to the specific class peak response from the top layer (semantic space), and generate Peak Response Maps (PRMs), $\mathcal{M}$, to highlight discriminative instance regions, Fig. 3. Note that each PRM $M^i \in \mathcal{M}$ is a map with the same shape as the input image, and its predicted class label is the channel index of the corresponding class peak response. Before further processing, we compute mean across its channel dimension and normalize it by dividing the sum.

3.2. Learning Instance Activation Maps

Instance Activation Maps are generated by recovering the full object extent from incomplete PRMs using an extent filling process.

Collecting pseudo supervision: Low-level vision object proposal methods often use the hypothesis that an object has consistent color, texture, and/or closed boundary to estimate class-agnostic object masks, Fig. 3. Although incomplete and noisy, redundant segment proposals can statistically cover the object and are sufficient for learning to fill the objects extent in local areas.

Given an image, we first extract a set of segment proposals $\mathcal{S}$. We then calculate matching scores between each PRM $M^i \in \mathcal{M}$ and each segment proposal $S^j \in \mathcal{S}$ of the image as $f_{ij} = \alpha \cdot M^i * S^j + M^i * \hat{S}^j$, where $\hat{S}_j$ is the proposal contour mask computed by morphological gradient operation. α is a class independent balance factor. The score comprises both the extent matching and boundary matching between M^i and S^j. After ranking the match scores for each PRM, we retain the top k proposals to provide locally correct object extent. When k increases, more false proposals could be retained to cause large disagreement between the k proposals, which might affect the model performance. Therefore, we then compute overlaps between the $\mathcal{M}$ and the k proposals and discard proposals with overlap $\max_i M^i * S^j$ lower than a threshold, *i.e.*, 0.2 in our settings.

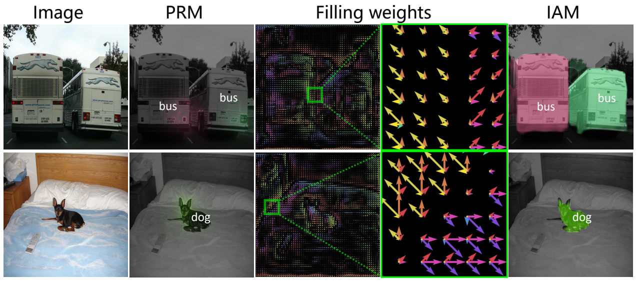


Figure 5: Visualization of the filling weights. The arrows indicate the filling from a pixel to its eight adjacent neighbors. Pixels in the flat area (*e.g.*, "bed") share a filling direction with their neighbors. On both sides of the edge, the arrows point in opposite directions; thus preserve the instance boundary. Best viewed zooming on screen.

During training, for each PRM, the approach randomly samples a proposal from the top k proposal candidates to construct the pseudo GT mask in each forward pass. Note that the difference between our approach with the proposal retrieval procedure in PRM [43] is two-fold: 1) the masks in our approach are used to learn the instance filling module while those of PRM is used to generate predictions and 2) we use a random strategy to sample multiple candidates while PRM only retrieves a single candidate with a maximum score. Therefore, the advantage is also two-fold: 1) our approach avoids proposals in the inference phase; thus improve the inference speed (order of magnitude). 2) we can statistically learn from multiple noisy proposals.

Instance Extent Filling: From each PRM-proposal pair produced above, we learn common knowledge of object extent from the segment proposals and image features to recover the instance extent conditioned on the PRM. To this end, we develop a differentiable extent filling module with an encoding-filling-decoding architecture (Fig. 3). The encoding process $E(\cdot, \theta_e)$ is the forward pass of a tiny convolutional network including two Conv-BatchNorm-ReLU-MaxPooling stacks. θ_e denotes the learnable parameters in the network. The contracting path of E squeezes the spatial size of the input PRM M^i; thus the filling process can better capture the long-range dependency between spatial locations in a computationally efficient way. Moreover, the encoded instance cues $E(M^i, \theta_e)$ are embedded in a feature space, making the filling process more stable to the response noises in PRMs. The decoding process $D(\cdot, \theta_d)$ is also a forward pass of a network that contains a symmetric expanding path and decodes the filling processed features as an Instance Activation Map (IAM). Note that this encoding-decoding process is built with standard CNN components; thus it can pass the gradients to each input. Compare to commonly used Auto-Encoder architectures, we design a filling process to effective model spatial relevance in the encoded feature space. The filling process (see Fig. 1) is an iterative process that consists of N filling steps F defined as:

$$\begin{aligned} G^k &= F(G^{k-1}, W) \\ G^0 &= E(M^i, \theta_e), \end{aligned} \tag{2}$$

where $G^k \in \mathbb{R}^{C\times N\times N}$ are the features after the k-th iteration, $0 < k \leq N$, and $W \in \mathbb{R}^{C\times(N\times N)\times(r\times r)}$ are the filling weights constructed from intermediate features maps M of the classification backbone as $W = R(M, \theta_r)$. R denotes a feature pyramid structure [23] which sequentially applies two 1×1 convolutions to adapt the feature maps at different levels. It then upsamples and fuses them from deep to shallow (Fig. 3). θ_r is the learnable parameters in the feature pyramid. As illustrated in Fig. 4, in each step of the filling process, locations of the c-th channel of G^k are filled according to its neighbors (and itself) and the predicted filling weights as

$$G^k_{ij} = \sum_{u,v\in\mathcal{N}_{ij}} Y_{ij} W_{c;i,j;u,v} E^{k-1}_{uv}(M^i, \theta_e), \tag{3}$$

where $\mathcal{N}_{ij}$ denotes the r^2 neighbors of coordinate (i, j), Y_{ij} is a normalizer to guarantee $\sum_{u,v\in\mathcal{N}_{ij}} W_{c;i,j;u,v} = 1$. The filling process stops when the maximum number of iterations N is reached. We set N to the size of the encoded maps to ensure the access to any location on the map.

The examples of learned filling weights W are shown in Fig. 5. The eight adjacent neighbors of a pixel are visualized in the form of vector fields, where the angle represents the corresponding neighbors and length represents the value. We compute the mean of W across channels and subtract the average from each map to suppress the "flat" regions that connected with all neighbors. It can be seen in the zoomed area that the filling weights clearly identify the instance boundaries. Interestingly, it can be seen from the example of the third row that our filling module successfully identifies the boundaries of "bed" even though it is not a valid object category in the dataset. This demonstrates that our approach can learn the common knowledge of object extent that generalizes to unseen object categories.

3.3. Implementation

Training details: Our proposed model is trained with image-level labels and class-agnostic segment proposals off-the-shelf. We implement our method based on standard ResNet50 [14] architecture. We first train the backbone network equipped with peak stimulation for image classification [43], using the Multi-label Soft Margin Loss and SGD optimizer, with a learning rate of 0.01. Then we optimize the filling module using Binary Cross Entropy loss. The initial learning rate of the SGD optimizer is set to 0.1. We use feature maps from res-block 2, 3, 4 of the backbone ResNet50 to form the feature pyramid. Following [43], we use the Multi-scale Combinatorial Grouping (MCG) framework [29] in conjunction with high-quality region hierarchies obtained with Convolutional Oriented Boundaries [28] to extract segment proposals. Note that our method does not constrain the choice of proposal technique.

Post-processing: Since our proposed IAMs covers instance extents, we choose the Convolutional Conditional Random Field (ConvCRF) [32] for further boundary refinement. This is in contrast to previous work [43] that has to employ computationally intensive proposal retrieval strategies to recover object extents. Experiments show that with ConvCRF, when the state-of-the-art fails, we maintain top performance while reducing the inference time by order of magnitude (0.3s vs. 3.0s per image).

3.4. Discussion

The proposed approach leverages instance-aware cues from classification networks, object prior from proposals, and instance extent filling operations to learn the full object extent. During the training phase, it actually implements a special kind of "semantic mosaicking". It collects redundant segments, absorbs broken semantic information into convolutional filters, and then fits complete semantics and full object extents. During the test phase, the peak response maps act as semantic anchors which correspond to the most discriminative parts, while the extent filling module produces instance activation maps (IAMs). The procedure is similar to the classical process of "flood-filling" [5]. The difference lies in the flood-filling is defined for grey-level stable image regions while IAMs are for semantically stable regions. The emergence of IAMs shows that the instance-level extent can be learned from redundant and noisy proposal segments, Fig. 1, which provide fresh insight for weakly supervised instance segmentation.

4. Experiments

We evaluate the proposed object extent filling approach on several popular benchmarks. In Sec. 4.1, we compare our Instance Activation Maps (IAMs) with state-of-the-art weakly supervised instance segmentation methods, demonstrating the effectiveness and efficiency of our approach. In Sec. 4.2, statistical analyses are performed to measure the quality of IAMs, which shows that our method can generate accurate instance-aware activations that cover object extent. In Sec. 4.3, we apply the trained filling module to fine-grained object localization and saliency detection without further fine-tuning of the instance extent filling module, validating the generalization ability of our method.

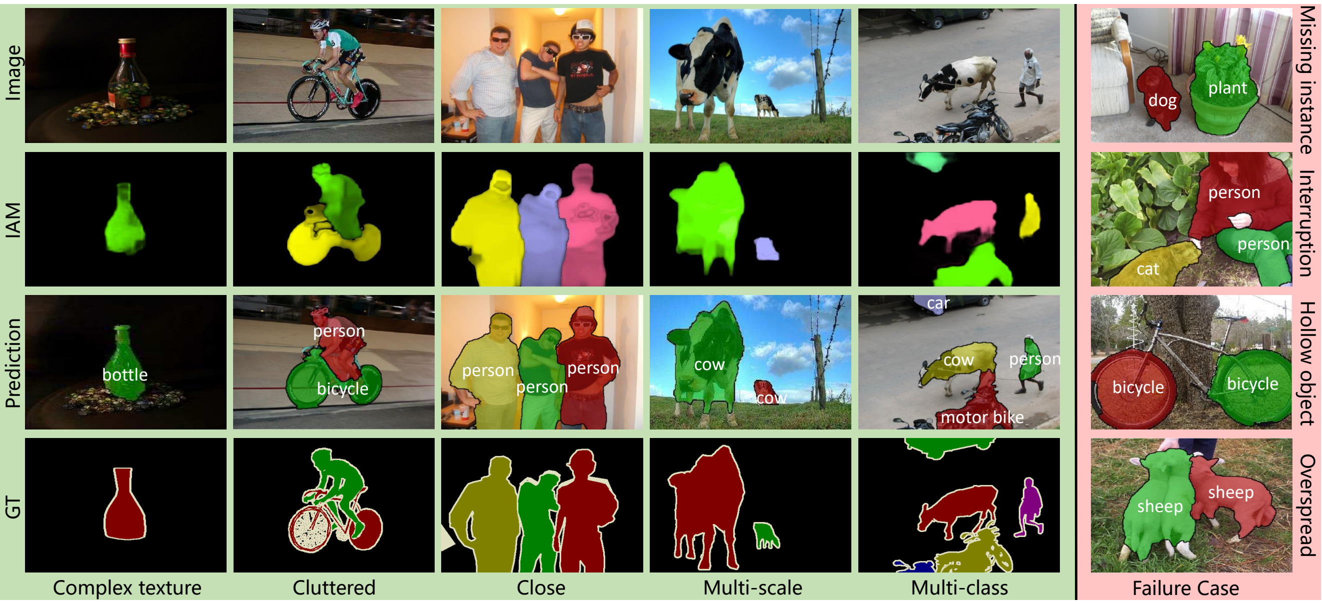


Figure 6: Weakly supervised instance segmentation examples. The Instance Activation Maps (2nd row) incorporate complete instance activation, which is exploited to produce instance-level masks (3rd row). The last column shows typical failure cases.

Method		$mAP^r_{0.25}$	$mAP^r_{0.5}$	$mAP^r_{0.75}$	ABO
Ground Truth Box	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
Baselines constructed from Weakly Supervised Object Localization					
CAM [42]	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 [35]	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
Weakly Supervised Instance Segmentation					
PRM [43]		44.3	26.8	9.0	37.6
IAM-S1		45.6	28.3	10.4	41.5
IAM-S5		**45.9**	**28.8**	**11.9**	**41.9**
IAM-S9		45.7	27.8	10.5	41.7

Table 1: Weakly supervised instance segmentation results - mean average precision (mAP%) and Average Best Overlap (ABO). Our method is evaluated with different random sampling numbers, *i.e.*, 1, 5, 9.

	Feed-Forward	Proposal Retrieval	CRF	Total
PRM [43]	0.05	3.0 (+8.1)	N/A	11.15
IAM (Ours)	0.07	N/A	0.3	0.37

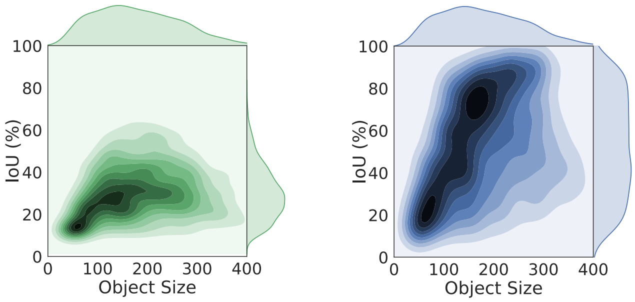
Table 2: Per-image inference time (seconds). The feed-forward and the CRF modules are tested with a Tesla P100 GPU while the proposal retrieval is tested with the official code on CPU. It takes 8.1s per image to extract proposals.

4.1. Weakly Supervised Instance Segmentation

We compare the performance of the proposed IAM with some baselines on the PASCAL VOC 2012 [12] segmentation benchmark.

Numerical results: In Tab. 1, the instance segmentation results are presented as the mean Average Precision (mAP) at IoU thresholds of 0.25, 0.5, and 0.75. Our IAM-S5 model outperforms the state-of-the-art by a margin 1.6%, 2.0%, and 2.9% respectively. The improvement at a higher IoU threshold of 0.75 is more significant than that 0.25 and 0.5, which indicates the effectiveness of our approach for generating high-quality instance activation and the capture of the fine-detailed object boundary. The Average Best Overlap (ABO) [30] score increased by a large margin of 4.3%, showing the ability of the IAMs to cover full object extents. Several baselines are constructed from weakly supervised object localization methods via three reasonable bbox-to-mask generation strategies [18].

The Effect of the random sampling number k: In the filling process, we learn IAMs from pseudo GT Masks obtained by random proposal sampling. The filling module summarizes the common knowledge of the object extent from the top k noisy masks. In Tab. 1, we evaluate the impact of the sampling number k. We first set k to 1 to verify if our model can explore the common object extent cross instances with only one noisy mask for each PRM. As a result, the IAM-S1 consistently improves the performance on the mAP^r at 0.25, 0.5, 0.75 as well as the ABO metric. As we increase k to 5, the performance of IAM-S5 is higher than IAM-S1, showing that our method can summarize the common knowledge of object extent from the noisy masks corresponding to the same PRM. Despite the ability to learn object extent from noisy masks, our model would be



(a) PRM [43]

(b) IAM (Ours)

Figure 7: The density map of samples from PRMs and IAMs. Darker area indicates more samples are of the corresponding IoU (%) value and object size.

affected when the sampling number increases to 9, as more and more false proposals could be sampled in the training procedure, which makes it hard for our model to summarize the knowledge of object extent.

Inference time: Tab. 2 shows the time cost for the inference period. PRM usually highlights object parts of instances and thus relies on the time-consuming proposal retrieval process (3.0s per image plus 8.1s for proposal generation) to get the instance masks. In contrast, our IAM can fill the object extent using CRF to refine object boundary, dramatically improves the inference speed (0.3 vs. 3.0s) while boosting instance segmentation performance.

Qualitative results: In Fig. 6, we illustrate instance segmentation examples including successful cases and typical failure cases. In the first column, our approach can distinguish instances with the complex texture. Examples in the second and third columns show that our approach performs well with cluttered or objects close to others. In the fourth and fifth columns, objects from different scales and different classes are well segmented. This shows that our method can extract both class-aware and instance-aware activation from classification networks. The last column shows failure cases of IAMs. It could miss an instance without proper instance-aware cues at first. Typically, IAMs can be misled by differences in color or texture in large areas and sometimes have problems connecting the parts of obscured or hollow objects. IAM may also fail to identify the boundary of huddled objects that are similar to each other.

4.2. Statistical Analysis for IAMs

A series of experiments are performed to analyze IAMs with respect to object size and object category, demonstrating that our approach outperforms the state-of-the-art approaches including Peak Response Map (PRM). IAMs are assigned to GT (ground truth) masks and judged to be overlapping or not by measuring the best matching IoU (Intersection over Union). To be considered a perfect IAM that completely coincides with a GT mask, the IoU between the predicted IAM M and GT masks $\mathcal{T}$ must be close to 100% as computed using the metric $\max_{\theta, T_i \in \mathcal{T}} \frac{area(f_b(M,\theta) \cap T_i)}{area(f_b(M,\theta) \cup T_i)}$, where the function $f_b(M,\theta) = M \geq \theta$ produces the best matching binary instance masks based on the probabilistic IAMs over a set of threshold values $\theta \in (0,1)$.

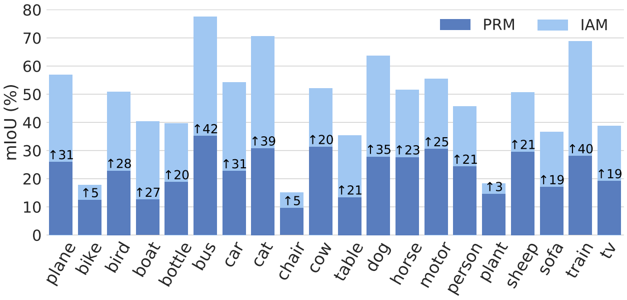


Figure 8: Per-class mean IoU (%) of PRMs and IAMs.

Sample distribution over object size: We first visualize the density of the IoU for PRMs and IAMs to see whether IAMs can cover objects of different sizes (Fig. 7). Fig. 7a shows that samples from PRMs are predominantly clustered in the area where the IoU value is less than 50% and failed to cover large objects. In contrast, in Fig. 7b most of the IAMs have high IoUs and perform well on large objects.

Sample distribution over object classes: We further calculate the mean IoU of each class to analyze the impact of different object categories, Fig. 8. Our method achieves consistent improvement across all categories. On “bus” and “cat”, IAM outperforms PRM by a large margin (∼40%). The reason is that PRMs can highlight the discriminative parts such as a tire for “bus” and head for “cat”, while IAMs cover complete object regions.

4.3. Generalization to Unseen Categories

The IAMs trained for weakly supervised instance segmentation are directly applied to localize the full extent of objects from a fine-grained species and unseen categories. Without any fine-tuning or re-training of the instance extent filling module, this procedure can be seen as unsupervised domain adaptation.

Localizing objects from fine-grained species: We use the pre-trained model to localize the bird species in the CUB-200-2011 dataset [36]. The dataset contains 11788 images over 200 categories of birds. There are 5994 images for training and 5794 images for testing. We chose this dataset to validate that the knowledge of object commonality learned from PASCAL VOC12 dataset can adapt to the fine-grained species which contain many unusual bird objects. Note that there are only 705 images defined for the bird category in the VOC12 training set. We first calculate IAMs for the images using the pre-trained IAM-S5 model, then extract bounding boxes using a mean value threshold. A bounding box is considered to have correct localization prediction if 1) the predicted class label is correct; 2) the overlap between the predicted box and ground-truth box is

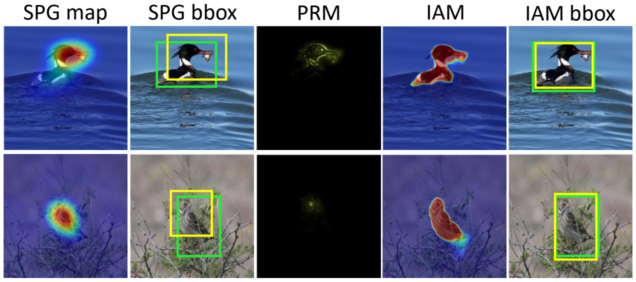


Figure 9: Visualization of object localization from fine-grained bird species. Green boxes are ground-truth and yellow boxes are predictions. Best viewed in color.

Methods	GoogLeNet-GAP [42]	ACol [40]	SPG [41]	IAM (Ours)
Loc. Err	59.00	54.08	53.36	**52.21**

Table 3: Localization error (%) on CUB-200-2011 test set for weakly supervised methods and our transferable IAMs. Note that our model is not trained on the target dataset.

higher than 0.5. To make a fair comparison, we use the scores predicted by GoogLeNet-GAP as the object extent predicted by our model is class agnostic. In Tab. 3, we compare the localization results with weakly supervised localization methods. We find that IAM performs well on the fine-grained bird species, achieving 52.21% error. This demonstrates that the model can generalize to objects from diversified sub-classes despite it being trained in another domain. Fig. 9 shows that IAMs can cover the full object extent and therefore benefit the bbox localization tasks.

Localizing salient objects from unseen categories: We humans can tell objects extent even when we don't know what the object is, motivating learning class-agnostic object commonality. To explore whether our approach can localize objects from unseen categories, we apply it to the salient object detection task. A Resnet50 classification network pre-trained on ImageNet is used to extract instance-aware visual cues which provide the coarse position of salient objects. These cues and the image are then fed to the model trained on VOC12. We obtain saliency detection results after performing ReLU on IAMs. We evaluated the performance on three popular saliency datasets including THUR [7], MSRA-B [26], and ECSSD [37] using the F-measure (F_β) as a performance metric. We compared our approach with state-of-the-art methods, including three supervised approaches based on deep learning frameworks, three unsupervised methods based on handcraft features and two unsupervised models based on deep learning. The results in Tab. 4 show that our model performs as well as a generic object extent localizer, despite the fact that it is not trained for the particular task. Fig. 10 presents some saliency detection results, which show that IAMs can fill object extent even though there is a large gap between the object appearance of the target categories and those of VOC12 object categories. This further validates the generalization capability of our approach.

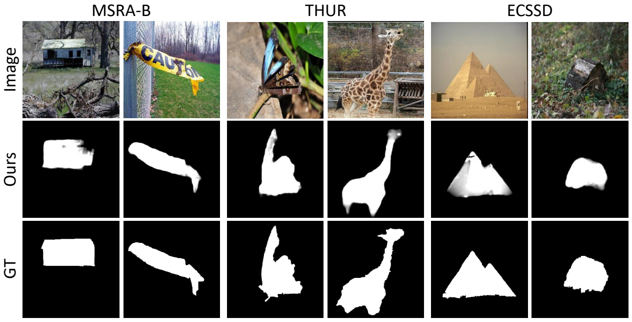


Figure 10: Salient object detection examples. Our method can fill salient object extent in a class-agnostic manner, even if the object appearance is unseen during training.

Finetune	Use GT	Methods	THUR	MSRA-B	ECSSD
✓	✓	DSS [15]	0.7081	0.8941	0.8796
✓	✓	NLDF [27]	-	0.8970	**0.8908**
✓	✓	DC [20]	0.6940	**0.8973**	0.8315
✓	✗	SBF [38]	-	-	0.7870
✓	✗	Multi-Noise [39]	0.7322	0.8770	0.8783
✗	✗	DRFI [17]	0.5613	0.7282	0.6440
✗	✗	RBD [44]	0.5221	0.7508	0.6518
✗	✗	DSR [21]	0.5498	0.7227	0.6387
✗	✗	IAM (Ours)	**0.7364**	0.8643	0.8613

Table 4: Mean F-measure (F_β) on salient object detection. The first column indicates if a finetuning model is used on saliency datasets while the second column indicates if the ground-truth masks are used in the training procedure.

5. Conclusions

We developed a framework which is trained to generate Instance Activation Maps (IAMs) driven by image-level supervision and prior knowledge from object segment proposals off-the-shelf. By extracting instance-aware cues from the classification network and iterative completing the cues according to the predicted extent filling weights, IAMs provide fine-detailed instance-level representation that highlight the spatial extent for each object. Our approach implements a special kind of "semantic mosaicking" that collects redundant noisy segments, absorbs broken semantic information into convolutional filters to learn class-agnostic object extent knowledge. On commonly used datasets, it significantly improves the state-of-the-art performance, increases inference speed by an order of magnitude, and can generalize to unseen categories, showing great potential on instance-level weakly supervised learning problems.

Acknowledgements

The authors are very grateful for the support by NSFC grant 61836012, 61771447, and 61671427.

References

[1] J. Ahn and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[2] A. Arnab and P. H. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, volume 1, page 5, 2017. 2

[3] G. Bertasius, L. Torresani, X. Y. Stella, and J. Shi. Convolutional random walk networks for semantic image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6137–6145, 2017. 3

[4] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016. 3

[5] S. V. Burtsev and Y. P. Kuzmin. An efficient flood-filling algorithm. *Computers & Graphics*, 17(5):549–561, 1993. 5

[6] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[7] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu. Salientshape: Group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014. 8

[8] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang. Localitysensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017. 3

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[10] 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

[11] 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

[12] 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

[13] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1, 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[15] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5300–5309. IEEE, 2017. 8

[16] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick. Learning to segment every thing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[17] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2083–2090, 2013. 8

[18] 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)*, pages 1665–1674, 2017. 2, 6

[19] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 2

[20] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 478–487, 2016. 8

[21] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2976–2983, 2013. 8

[22] 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)*, pages 4438–4446, 2017. 2

[23] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017. 5

[24] S. Liu, S. D. Mello, J. Gu, G. Zhong, M. Yang, and J. Kautz. Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1519–1529, 2017. 3

[25] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8759–8768, 2018. 1, 2

[26] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 33(2):353–367, 2011. 8

[27] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, and P.-M. Jodoin. Non-local deep features for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 7, 2017. 8

[28] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool. Convolutional oriented boundaries: From image segmentation to high-level tasks. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 40(4):819–833, 2018. 2, 5

[29] 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

[30] 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. 6

[31] 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. 3

[32] M. T. Teichmann and R. Cipolla. Convolutional crfs for semantic segmentation. *arXiv preprint arXiv:1805.04777*, 2018. 2, 5

[33] 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

[34] P. Vernaza and M. Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 3, 2017. 3

[35] 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. 3, 6

[36] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. 2010. 7

[37] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1162, 2013. 8

[38] D. Zhang, J. Han, and Y. Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, page 3, 2017. 8

[39] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9029–9038, 2018. 8

[40] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 8

[41] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang. Self-produced guidance for weakly-supervised object localization. In *European Conference on Computer Vision (ECCV)*, 2018. 8

[42] 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, 2, 6, 8

[43] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao. Weakly supervised instance segmentation using class peak response. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 4, 5, 6, 7

[44] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2814–2821, 2014. 8

[45] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft proposal networks for weakly supervised object localization. *IEEE International Conference on Computer Vision (ICCV)*, pages 1859–1868, 2017. 6

[46] 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