Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations

Jiwoon Ahn
DGIST, Kakao Corp.
jyun@dgist.ac.kr

Sunghyun Cho*
DGIST
scho@dgist.ac.kr

Suha Kwak*
POSTECH
suha.kwak@postech.ac.kr

Abstract

This paper presents a novel approach for learning instance segmentation with image-level class labels as supervision. Our approach generates pseudo instance segmentation labels of training images, which are used to train a fully supervised model. For generating the pseudo labels, we first identify confident seed areas of object classes from attention maps of an image classification model, and propagate them to discover the entire instance areas with accurate boundaries. To this end, we propose IRNet, which estimates rough areas of individual instances and detects boundaries between different object classes. It thus enables to assign instance labels to the seeds and to propagate them within the boundaries so that the entire areas of instances can be estimated accurately. Furthermroe, IRNet is trained with inter-pixel relations on the attention maps, thus no extra supervision is required. Our method with IRNet achieves an outstanding performance on the PASCAL VOC 2012 dataset, surpassing not only previous state-of-the-art trained with the same level of supervision, but also some of previous models relying on stronger supervision.

1. Introduction

Instance segmentation is a task that jointly estimates class labels and segmentation masks of individual objects. As in other visual recognition tasks, supervised learning of Convolutional Neural Networks (CNNs) has driven recent advances in instance segmentation [7, 9, 10, 16, 17, 23, 30, 35]. Due to the data-hungry nature of deep CNNs, this approach demands an enormous number of training images with groundtruth labels, which are given by hand in general. However, manual annotation of instance-wise segmentation masks is prohibitively time-consuming, which results in existing datasets limited in terms of both class diversity and the amount of annotated data. It is thus not straightforward to learn instance segmentation models that can handle diverse object classes in the real world.

One way to alleviate this issue is weakly supervised learning that adopts weaker and less expensive labels than instance-wise segmentation masks as supervision. Thanks to low annotation costs of weak labels, approaches in this category can utilize more training images of diverse objects, although they have to compensate for missing information in weak labels. For instance segmentation, bounding boxes have been widely used as weak labels since they provide every property of objects except shape [22, 41]. However, it is still costly to obtain box labels for a variety of classes in a large number of images as they are manually annotated.

To further reduce the annotation cost, one may utilize image-level class labels for learning instance segmentation since such labels are readily available in large-scale image classification datasets, *e.g.*, ImageNet [42]. Furthermore, although image-level class labels indicate only the existence of object classes, they can be used to derive strong cues for instance segmentation, called *Class Attention Maps* (CAMs) [37, 43, 45, 49]. A CAM roughly estimates areas of each class by investigating the contribution of local image regions to the classification score of the class. However, CAMs cannot be directly utilized as supervision for instance segmentation since they have limited resolution, often highlight only partial areas of objects, and most importantly, cannot distinguish different instances of the same class. To resolve this issue, a recent approach [50] incorporates CAMs with an off-the-shelf segmentation proposal technique [2], which however has to be trained separately on an external dataset with additional supervision.

In this paper, we present a novel approach for learning instance segmentation using image-level class labels, which outperforms the previous state-of-the-art trained with the same level of supervision [50] and even some of approaches relying on stronger supervision [16, 22]. Moreover, it requires neither additional supervision nor any segmentation proposals unlike the previous approaches [16, 50]. Our method generates pseudo instance segmentation labels of training images given their image-level labels and trains a known CNN model with the pseudo labels. For generating the pseudo labels, it utilizes CAMs, but as mentioned earlier, they can neither distinguish different instances nor find

*Co-corresponding authors.

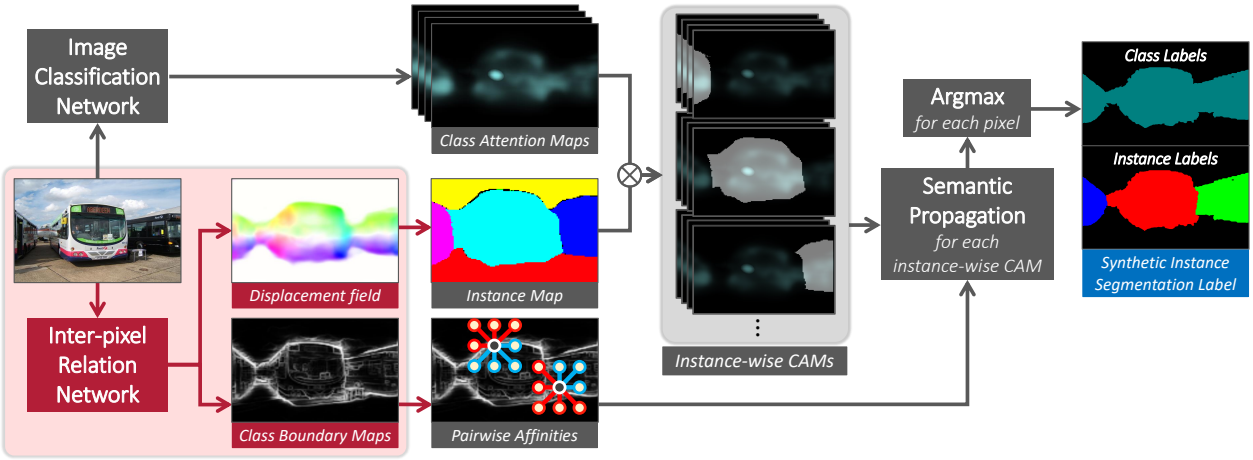

Figure 1. Overview of our framework for generating pseudo instance segmentation labels.

entire instance areas with accurate boundaries.

To overcome these limitations of CAMs, we introduce *Inter-pixel Relation Network* (IRNet) that is used to estimate two types of additional information complementary to CAMs: a class-agnostic instance map and pairwise semantic affinities. A class-agnostic instance map is a rough instance segmentation mask without class labels nor accurate boundaries. On the other hand, the semantic affinity between a pair of pixels is a confidence score for class equivalence between them. By incorporating instance-agnostic CAMs with a class-agnostic instance map, we obtain instance-wise CAMs, which are in turn enhanced by propagating their attention scores to relevant areas based on the semantic affinities between neighboring pixels. After the enhancement, a pseudo instance segmentation label is generated by selecting the instance label with the highest attention score in the instance-wise CAMS at each pixel. The entire procedure for label synthesis is illustrated in Fig. 1.

IRNet has two branches estimating an instance map and semantic affinities, respectively. The first branch predicts a displacement vector field where a 2D vector at each pixel indicates the centroid of the instance the pixel belongs to. The displacement field is converted to an instance map by assigning the same instance label to pixels whose displacement vectors point at the same location. The second branch detects boundaries between different object classes. Pairwise semantic affinities are then computed from the detected boundaries in such a way that two pixels separated by a strong boundary are considered as a pair with a low semantic affinity. Furthermore, we found that IRNet can be trained effectively with inter-pixel relations derived from CAMs. Specifically, we collect pixels with high attention scores and train IRNet with the displacements and class equivalence between the collected pixels. Thus, no supervision in addition to image-level class labels is required.

The contribution of this paper is three-fold:

- We propose a new approach to identify and localize instances with image-level supervision through class-agnostic instance maps. This enables instance segmentation without off-the-shelf segmentation proposals.
- We propose a new way to learn and predict semantic affinities between pixels with image-level supervision through class boundary detection, which is more effective and efficient than previous work [1].
- On the PASCAL VOC 2012 dataset [12], our model substantially outperforms the previous state-of-the-art trained with the same level of supervision [50]. Also, it even surpasses previous models based on stronger supervision like SDI [22] that uses bounding box labels and SDS [16], an early model that uses full supervision.

2. Related Work

This section reviews semantic and instance segmentation models closely related to our method. We first introduce weakly supervised approaches for the two tasks, and discuss models that are based on ideas similar with the displacement field and pairwise semantic affinity of our framework.

Weakly Supervised Semantic Segmentation: For weak supervision of semantic segmentation, various types of weak labels such as bounding boxes [8, 38], scribbles [27, 44], and points [3] have been utilized. In particular, image-level class labels have been widely used as weak labels since they require minimal or no effort for annotation [1, 11, 19, 20, 36, 39, 40, 45, 50]. Most approaches using the image-level supervision are based on CAMs [37, 43, 49] that roughly localize object areas by drawing attentions on discriminative parts of object classes. However, CAMs often fail to reveal the entire object areas with accurate boundaries. To address this issue, extra data or supervision have been exploited to obtain additional evidences like saliency [20, 36], motion in videos [19, 39] and class-agnostic object proposals [40]. Recent approaches tackle the issue without external information by mining comple-

mentary attentions iteratively [20, 45] or propagating CAMs based on semantic affinities between pixels [1].

Weakly Supervised Instance Segmentation: For instance segmentation, bounding boxes have been widely used as weak labels. Since a bounding box informs the exact location and scale of an object, weakly supervised models using box labels focus mainly on estimating object shapes. For example, in [22], GraphCut is incorporated with generic boundary detection [48] to better estimate object shapes by considering boundaries. Also, in [41], an object shape estimator is trained by adversarial learning [14] so that a pseudo image generated by cutting and pasting the estimated object area to a random background looks realistic. Meanwhile, weakly supervised instance segmentation with image-level class labels has been rarely studied since this is a significantly ill-posed problem where supervision does not provide any instance-specific information. To tackle this challenging problem, a recent approach [50] detects peaks of class attentions to identify individual instances and combines them with high-quality segmentation proposals [2] to reveal entire instance areas. However, the performance of the method heavily depends on that of the segmentation proposals, which have to be trained with extra data with high-level supervision. In contrast, our approach requires neither off-the-shelf proposals nor additional supervision and it surpasses the previous work [50] by a substantial margin.

Pixel-wise Prediction of Instance Location: Pixel-wise prediction of instance location has been proven to be effective for instance segmentation in literature. In [26] the coordinates of the instance bounding box each pixel belongs to are predicted in a pixel-wise manner so that pixels with similar box coordinates are clustered as a single instance mask. This idea is further explored in [21, 35], which predict instance centroids instead of box coordinates. Our approach based on the displacement field share the same idea with [21, 35], but it requires only image-level supervision while the previous approaches are trained with instance-wise segmentation labels.

Semantic Affinities Between Pixels: Pairwise semantic affinities between pixels have been used to enhance the quality of semantic segmentation. In [4, 6], CNNs for semantic segmentation are incorporated with a differentiable module computing a semantic affinity matrix of pixels, and trained in an end-to-end manner with full supervision. In [4], a predicted affinity matrix is used as a transition probability matrix for random walk, while in [6], it is embedded into a convolutional decoder [34] to encourage local pixels to have the same labels during inference. Recently, a weakly supervised model has been proposed to learn pairwise semantic affinities with image-level class labels [1]. This model predicts a high-dimensional embedding vector for each pixel, and the affinity between a pair of pixels is defined as the similarity between their embedding vectors.

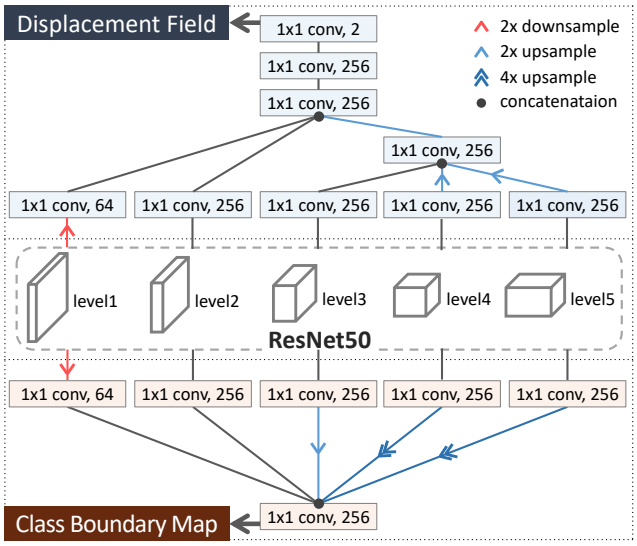

Figure 2. Overall architecture of IRNet.

Our approach shares the same motivation with [1], but our IRNet can learn and predict affinities more effectively and efficiently by detecting class boundaries.

3. Class Attention Maps

CAMs play two essential roles in our framework. First, they are used to define seed areas of instances, which are propagated later to recover the entire instance areas as in [1, 24]. Second, they are a source of supervision for learning IRNet; by exploiting CAMs carefully, we extract reliable inter-pixel relations, from which IRNet is trained. To generate CAMs for training images, we adopt the method of [49] using an image classification CNN with global average pooling followed by a classification layer. Given an image, the CAM of a groundtruth class c is computed by

$$M_c(\mathbf{x}) = \frac{\phi_c^\top f(\mathbf{x})}{\max_{\mathbf{x}} \phi_c^\top f(\mathbf{x})}, \tag{1}$$

where f is a feature map from the last convolution layer of the CNN, $\mathbf{x}$ is a 2D coordinate on f, and ϕ_c is the classification weights of the class c. Also, CAMs for irrelevant classes are fixed to a zero matrix. We adopt ResNet50 [18] as the classification network, and reduce the stride of its last downsampling layer from 2 to 1 to prevent CAMs from further resolution drop. As a result, the width and height of CAMs are $1/16$ of those of the input image.

4. Inter-pixel Relation Network

IRNet aims to provide two types of information: a displacement vector field and a class boundary map, both of which are in turn used to estimate pseudo instance masks from CAMs. This section describes the IRNet architecture

and the strategy for learning the model using CAMs as supervision. How to use IRNet for pseudo label generation will be illustrated in Sec. 5.

4.1. IRNet Architecture

IRNet has two output branches that predict a displacement vector field and a class boundary map, respectively. Its architecture is illustrated in Fig. 2. The two branches share the same ResNet50 backbone, which is identical to that of the classification network in Sec. 3. As inputs, both branches take feature maps from all the five levels[1] of the backbone. All the convolution layers of both branches are followed by group normalization [47] and ReLU except the last layer. Details of both branches are described below.

Displacement Field Prediction Branch: A 1×1 convolution layer is first applied to each input feature map, and the number of channels is reduced to 256 if it is larger than that. On top of them, we append a top-down path way [28] to merge all the feature maps iteratively in such a way that low resolution feature maps are upsampled twice, concatenated with those of the same resolution, and processed by a 1×1 convolution layer. Finally, from the last concatenated feature map, a displacement field is decoded through three 1×1 convolution layers, whose output has two channels.

Boundary Detection Branch: We first apply 1×1 convolution to each input feature map for dimensionality reduction. Then the results are resized, concatenated, and fed into the last 1×1 convolution layer, which produces a class boundary map from the concatenated features.

4.2. Inter-pixel Relation Mining from CAMs

Inter-pixel relations are the only supervision for training IRNet, thus it is important to collect them reliably. We define two kinds of relations between a pair of pixels: the displacement between their coordinates and their class equivalence. The displacement can be easily computed by a simple subtraction, but the class equivalence is not since pixel-wise class labels are not given in our weakly supervised setting.

Thus, we carefully exploit CAMs to predict pixel-wise pseudo class labels and obtain reliable class equivalence relations from them. The overall procedure of our method is illustrated in Fig. 3. Since CAMs are blurry and often inaccurate, we first identify areas with confident foreground/background attention scores. Specifically, we collect pixels with attention scores larger than 0.3 as foreground pixels, and smaller than 0.05 as background pixels. Note that we do not care pixels outside of confident areas during the process. Each confident area is then refined by dense CRF [25] to better estimate object shapes. After that, we construct a pseudo class map $\hat{M}$ by choosing the class with the best score for each pixel. Finally, we sample pairs of neighboring pixels from the refined confident areas, and categorize them into two sets $\mathcal{P}^+$ and $\mathcal{P}^-$ according to their class equivalence by

$$\mathcal{P} = \big\{(i,j) \mid \|\mathbf{x}_i - \mathbf{x}_j\|_2 < \gamma, \forall i \neq j\big\}, \quad (2)$$

$$\mathcal{P}^+ = \big\{(i,j) \mid \hat{M}(\mathbf{x}_i) = \hat{M}(\mathbf{x}_j), (i,j) \in \mathcal{P}\big\}, \quad (3)$$

$$\mathcal{P}^- = \big\{(i,j) \mid \hat{M}(\mathbf{x}_i) \neq \hat{M}(\mathbf{x}_j), (i,j) \in \mathcal{P}\big\}, \quad (4)$$

where γ is a radius limiting the maximum distance of a pair. We further divide $\mathcal{P}^+$ into $\mathcal{P}^+_{\text{fg}}$ and $\mathcal{P}^+_{\text{bg}}$, a set of foreground pairs and that of background pairs, respectively.

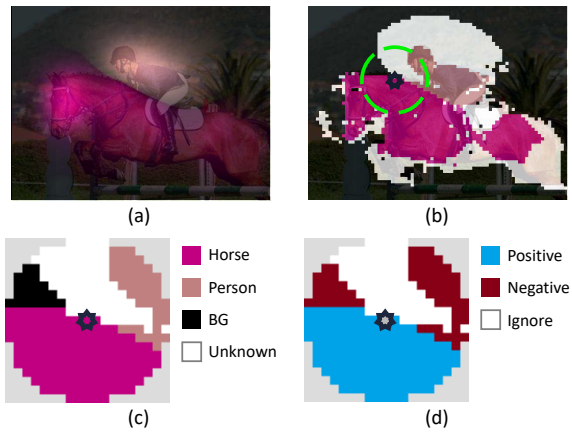

Figure 3. Visualization of our inter-pixel relation mining process. (a) CAMs. (b) Confident areas of object classes. (c) Pseudo class label map within a local neighborhood. (d) Class equivalence relations between the center and the others.

4.3. Loss for Displacement Field Prediction

The first branch of IRNet predicts a displacement vector field $\mathcal{D} \in \mathbb{R}^{w\times h\times 2}$, where each 2D vector points at the centroid of the associated instance. Although ground truth centroids are not given in our setting, we argue that $\mathcal{D}$ can be learned implicitly with displacements between pixels of the same class. There are two conditions for $\mathcal{D}$ to be a displacement field. First, for a pair of pixel locations $\mathbf{x}_i$ and $\mathbf{x}_j$ belonging to the same instance, their estimated centroids must be identical, *i.e.*, $\mathbf{x}_i+\mathcal{D}(\mathbf{x}_i) = \mathbf{x}_j+\mathcal{D}(\mathbf{x}_j)$. Second, by the definition of centroid, $\sum_{\mathbf{x}} \mathcal{D}(\mathbf{x}) = 0$ for each instance.

To satisfy the first condition, we first assume that a pair of nearby pixels $(i,j) \in \mathcal{P}^+$ is likely to be of the same instance since they are sampled within a small radius γ. Then, given such a pair (i,j), our goal is to approximate their image coordinate displacement $\hat{\delta}(i,j) = \mathbf{x}_j - \mathbf{x}_i$ with their difference in $\mathcal{D}$ denoted by $\delta(i,j) = \mathcal{D}(\mathbf{x}_i) - \mathcal{D}(\mathbf{x}_j)$. In the ideal case where $\delta = \hat{\delta}$, it will hold that $\mathbf{x}_i + \mathcal{D}(\mathbf{x}_i) = \mathbf{x}_j + \mathcal{D}(\mathbf{x}_j)$ for all (i,j) of the same instance. This implies that $\mathcal{D}(\mathbf{x})$ is the displacement vector indicating the corresponding centroid. For learning $\mathcal{D}$ with the inter-pixel relations obtained in Sec. 4.2, we minimize L_1 loss between

[1] A level means a group of residual units sharing the same output size in [18]. However, in our backbone, the output sizes of level4 and level5 are identical since the stride of the last downsampling layer is reduced to 1.

$\delta(i,j)$ and $\hat{\delta}(i,j)$:

$$\mathcal{L}_{\text{fg}}^{\mathcal{D}} = \frac{1}{|\mathcal{P}_{\text{fg}}^{+}|} \sum_{(i,j)\in\mathcal{P}_{\text{fg}}^{+}} \left|\delta(i,j) - \hat{\delta}(i,j)\right|. \tag{5}$$

The second condition, on the other hand, is not explicitly encouraged by Eq. (5). However, we argue that IRNet can still learn to predict displacement vectors pointing to rough centroids of instances due to the randomness of initial network parameters. Intuitively speaking, initial random displacement vectors are already likely to satisfy the second condition, and the training of IRNet converges to a local minimum that still satisfies the condition. A similar phenomenon is observed in [35]. Displacement vectors are then further refined by subtracting the mean of $\mathcal{D}$ from $\mathcal{D}$.

Also, we eliminate trivial centroid estimation from background pixels since the centroid of background is indefinite and may interfere with the above process. For the purpose, we minimize the following loss for background pixels:

$$\mathcal{L}_{\text{bg}}^{\mathcal{D}} = \frac{1}{|\mathcal{P}_{\text{bg}}^{+}|} \sum_{(i,j)\in\mathcal{P}_{\text{bg}}^{+}} |\delta(i,j)|. \tag{6}$$

4.4. Loss for Class Boundary Detection

Given an image, the second branch of IRNet detects boundaries between different classes, and the output is denoted by $\mathcal{B} \in [0,1]^{w\times h}$. Although no ground truth labels for class boundaries are given in our setting, we can train the second branch with class equivalence relations between pixels through a Multiple Instance Learning (MIL) objective. The key assumption is that a class boundary exists somewhere between a pair of pixels with different pseudo class labels.

To implement this idea, we express the semantic affinity between two pixels in terms of the existence of a class boundary. For a pair of pixels $\mathbf{x}_i$ and $\mathbf{x}_j$, we define their semantic affinity a_{ij} as:

$$a_{ij} = 1 - \max_{k\in\Pi_{ij}} \mathcal{B}(\mathbf{x}_k) \tag{7}$$

where Π_{ij} is a set of pixels on the line between $\mathbf{x}_i$ and $\mathbf{x}_j$. We utilize class equivalence relations between pixels as supervision for learning a_{ij}. Specifically, the class equivalence between two pixels is represented as a binary label whose value is 1 if their pseudo class labels are the same and 0 otherwise. The affinity is then learned by minimizing cross-entropy between the one-hot vector of the binary affinity label and the predicted affinity in Eq. (7):

$$\mathcal{L}^{\mathcal{B}} = -\sum_{(i,j)\in\mathcal{P}_{\text{fg}}^{+}} \frac{\log a_{ij}}{2|\mathcal{P}_{\text{fg}}^{+}|} - \sum_{(i,j)\in\mathcal{P}_{\text{bg}}^{+}} \frac{\log a_{ij}}{2|\mathcal{P}_{\text{bg}}^{+}|} - \sum_{(i,j)\in\mathcal{P}^{-}} \frac{\log(1-a_{ij})}{|\mathcal{P}^{-}|} \tag{8}$$

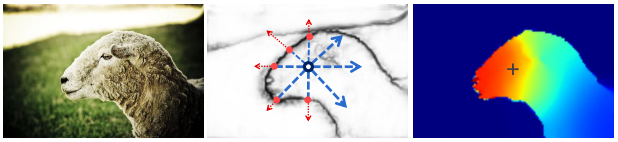

Figure 4. Deriving pairwise semantic affinities from a class boundary map. (left) Input Image. (center) A class boundary map. (right) Label propagation from the center after random walks.

Figure 5. Detecting instance centroids. (left) Input image. (center) An initial displacement field. (right) A refined displacement field and detected centroids.

where three separate losses are aggregated after normalization since populations of $\mathcal{P}_{\text{fg}}^{+}$, $\mathcal{P}_{\text{bg}}^{+}$, and $\mathcal{P}^{-}$ are significantly imbalanced in general. Through the loss in Eq. (8), we can learn $\mathcal{B}$ implicitly with inter-pixel class equivalence relations. In this aspect, Eq. (8) can be regarded as a MIL objective where Π_{ij} is a bag of potential boundary pixels.

4.5. Joint Learning of the Two Branches

The two branches of IRNet are jointly trained by minimizing all the losses we defined previously at the same time:

$$\mathcal{L} = \mathcal{L}_{\text{fg}}^{\mathcal{D}} + \mathcal{L}_{\text{bg}}^{\mathcal{D}} + \mathcal{L}^{\mathcal{B}}. \tag{9}$$

Note that the above loss is class-agnostic since $\mathcal{P}^{+}$ and $\mathcal{P}^{-}$ only consider class equivalence between pixels rather than their individual class labels. This allows our approach to utilize more inter-pixel relations per class and helps to improve the generalization ability of IRNet.

5. Label Synthesis Using IRNet

To synthesize pseudo instance labels, the two outputs $\mathcal{D}$ and $\mathcal{B}$ of IRNet are converted to a class-agnostic instance map and pairwise affinities, respectively. Among them, semantic affinities can be directly derived from $\mathcal{B}$ by Eq. (7) as illustrated in Fig. 4, while the conversion of $\mathcal{D}$ is not straightforward due to its inaccurate estimation. This section first describes how $\mathcal{D}$ is converted to an instance map, then how to generate pseudo instance segmentation labels with the instance map and semantic affinities.

5.1. Generating Class-agnostic Instance Map

A class-agnostic instance map I is a $w \times h$ 2D map, each element of which is the instance label associated with the element. If $\mathcal{D}$ is estimated with perfect accuracy, I can be obtained simply by grouping pixels whose displacement vectors point at the same centroid. However, $\mathcal{D}$ often fails to

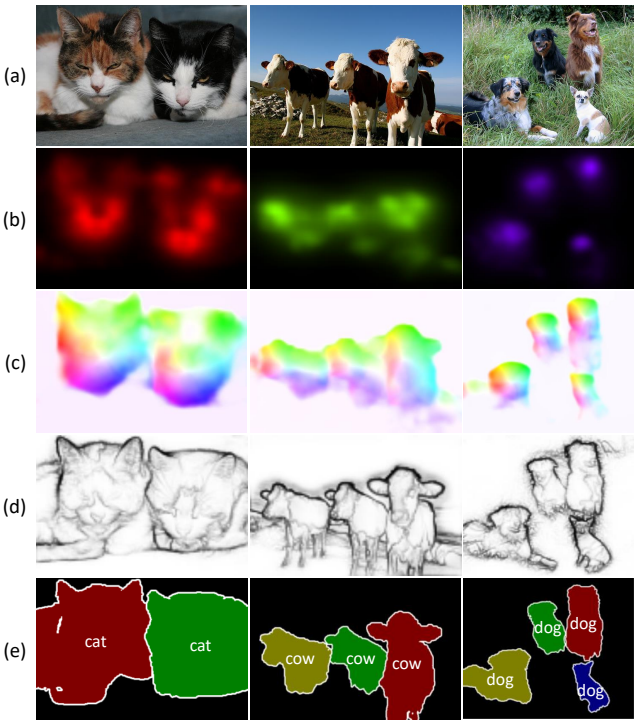

Figure 6. Examples of pseudo instance segmentation labels on the PASCAL VOC 2012 *train* set. (a) Input image. (b) CAMs. (c) Displacement field. (d) Class boundary map. (e) Pseudo labels.

predict the exact offsets to centroids since IRNet is trained with incomplete supervision derived from CAMs. To address this issue, $\mathcal{D}$ is refined iteratively by

$$\mathcal{D}_{u+1}(\mathbf{x}) = \mathcal{D}_u(\mathbf{x}) + \mathcal{D}\left(\mathbf{x} + \mathcal{D}_u(\mathbf{x})\right) \quad \forall \mathbf{x}, \tag{10}$$

where u is an iteration index and $\mathcal{D}_0$ is the initial displacement field given by IRNet. Each displacement vector is refined iteratively by adding the displacement vector at the currently estimated centroid location. As displacement vectors near centroids tend to be almost zero in magnitude, the refinement converges within a finite number of iterations. The effect of the refinement is demonstrated in Fig. 5.

Since centroids estimated via the refined $\mathcal{D}$ are still scattered in general, we consider a small group of neighboring pixels, instead of a single coordinate, as a centroid. To this end, we first identify pixels whose displacement vectors in $\mathcal{D}$ have small magnitudes, and regard them as candidate centroids since pixels around a true centroid will have near zero displacement vectors. Then each connected component of the candidates is considered as a centroid. Note that the candidates tend to be well grouped into a few connected components since displacement vectors change smoothly within a local neighborhood as can be seen in Fig. 5.

5.2. Synthesizing Instance Segmentation Labels

For generating pseudo instance masks, we first combine CAMs with a class-agnostic instance map as follows:

$$\bar{M}_{ck}(\mathbf{x}) = \begin{cases} M_c(\mathbf{x}) & \text{if } I(\mathbf{x}) = k, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $\bar{M}_{ck}$ is the instance-wise CAMs of class c and instance k. Each instance-wise CAM is refined individually by propagating its attention scores to relevant areas. Specifically, the propagation is done by random walk, whose transition probability matrix is derived from the semantic affinity matrix $A = [a_{ij}] \in \mathbb{R}^{wh \times wh}$ as follows:

$$T = S^{-1}A^{\circ\beta}, \quad \text{where} \quad S_{ii} = \sum_j a_{ij}^{\beta} \tag{12}$$

and $A^{\circ\beta}$ is A to the Hadamard power of β and S is a diagonal matrix for row-normalization of $A^{\circ\beta}$. Also, $\beta > 1$ is a hyper-parameter for smoothing out affinity values in A. The random walk propagation with T is then conducted by

$$\text{vec}(\bar{M}^*_{ck}) = T^t \cdot \text{vec}(\bar{M}_{ck} \odot (1 - \mathcal{B})), \tag{13}$$

where t denotes the number of iterations, $\odot$ is the Hadamard product, and $\text{vec}(\cdot)$ means vectorization. We penalize scores of boundary pixels by multiplying $(1 - \mathcal{B})$ since those isolated pixels do not propagate their scores to neighbors and have overly high scores compared to the others in consequence. Then an instance segmentation label is generated by choosing the combination of c and k that maximizes $\bar{M}^*_{ck}(\mathbf{x})$ for each pixel $\mathbf{x}$. If the maximum score is less than bottom 25%, the pixel is regarded as background.

6. Experiments

The effectiveness of our framework is demonstrated on the PASCAL VOC 2012 dataset [13], where our framework generates pseudo labels for training images and trains a fully supervised model with the images and their pseudo labels. We evaluate the quality of our pseudo labels as well as the performance of the model trained with them. The evaluation is done for both instance segmentation and semantic segmentation since our pseudo labels can be used to train semantic segmentation models as well.

6.1. Experimental Setting

Dataset: We train and evaluate our framework on the PASCAL VOC 2012 [12] dataset. Although the dataset contains labels for semantic segmentation and instance segmentation, we only exploit image-level class labels. Following the common practice, the training set is expanded by adding image set proposed in [15]. In total, 10,582 images are used for training, and 1,449 images are kept for validation.

Method	mIoU
CAM	8.6
CAM + Class Boundary	34.1
CAM + Displacement Field + Class Boundary (Ours)	37.7

Table 1. Quality of our pseudo instance segmentation labels in AP^r_{50}, evaluated on the PASCAL VOC 2012 *train* set.

CAM	Prop. w/ AffinityNet [1]	Prop. w/ IRNet (Ours)
48.3	59.3	66.5

Table 2. Quality of pseudo semantic segmentation labels in mIoU, evaluated on the PASCAL VOC 2012 *train* set. "Prop" means the semantic propagation using predicted affinities.

Hyperparameter Settings: The radius that limits the search space of pairs γ in Eq. (2) is set to 10 when training, and reduced to 5 at inference for conservative propagation. The number of random walk iterations t in Eq. (13) is fixed to 256. The hyperparameter β in Eq. (12) is set to 10. The iterative update of $\mathcal{D}$ in Eq. (10) is done 100 times.

Network Parameter Optimization: We adopt the stochastic gradient descent for network optimization. Learning rate is initially set to 0.1, and decreases at every iteration with polynomial decay [32]. The backbone of IRNet is frozen during training, and gradients that displacement field branch receives are amplified by a factor of 10.

Comparison to AffinityNet: For a fair comparison, we modified AffinityNet [1] by replacing its backbone with ResNet50 as in our IRNet. Then we compare IRNet with the modified AffinityNet in terms of the accuracy of pseudo segmentation labels (Table 2) and performance of DeepLab [5] trained with these pseudo labels (Table 4).

6.2. Analysis of Pseudo Labels

Instance Segmentation labels: A few qualitative examples of pseudo instance segmentation labels are presented in Fig. 6, and the contribution of each branch of IRNet to the quality of the labels is analyzed in Table 1. In the case of "CAM" in Table 1, we directly utilize raw CAMs to generate pseudo labels by thresholding their scores and applying connected component analysis while assuming that there are no instances of the same class attached to each other. In the case of "CAM + Class Boundary" in Table 1, pseudo labels are obtained in the same manner, but we enhance CAMs by the semantic propagation based on the class boundary map before generating pseudo labels. We evaluated the performance of each method in terms of average precision (AP). For evaluating APs, the score of each detected instance is given as the maximum class score within its mask. As shown in the table, exploiting a class boundary map effectively improves the quality of pseudo labels by more than 25% as it helps to recover the entire areas of objects missing in CAMs. Exploiting a displacement field further improves the performance by 3.6% as it helps to distinguish different instances of the same class.

Method	Sup.	Extra data / Information	AP^r_{50}	AP^r_{70}
PRM [50]	$\mathcal{I}$	MCG [2]	26.8	-
SDI [22]	$\mathcal{B}$	BSDS [33]	44.8	-
SDS [16]	$\mathcal{F}$	MCG [2]	43.8	21.3
MRCNN [17]	$\mathcal{F}$	MS-COCO [29]	69.0	-
Ours-ResNet50	$\mathcal{I}$	-	46.7	23.5

Table 3. Instance segmentation performance on the PASCAL VOC 2012 *val* set. The supervision types (Sup.) indicate: $\mathcal{I}$–image-level label, $\mathcal{B}$–bounding box, and $\mathcal{F}$–segmentation label.

Method	Sup.	Extra Data / Information	*val*	*test*
SEC [24]	$\mathcal{I}$	-	50.7	51.7
AffinityNet [1]	$\mathcal{I}$	-	58.7	-
PRM [50]	$\mathcal{I}$	MCG [2]	53.4	-
CrawlSeg [19]	$\mathcal{I}$	YouTube Videos	58.1	58.7
MDC [46]	$\mathcal{I}$	Ground-truth Backgrounds	60.4	60.8
DSRG [20]	$\mathcal{I}$	MSRA-B [31]	61.4	63.2
ScribbleSup [27]	$\mathcal{S}$	-	63.1	-
BoxSup [8]	$\mathcal{B}$	-	62.0	64.6
SDI [22]	$\mathcal{B}$	BSDS [33]	65.7	67.5
Upperbound	$\mathcal{F}$	-	72.3	72.5
Ours-ResNet50	$\mathcal{I}$	-	63.5	64.8

Table 4. Semantic segmentation performance on the PASCAL VOC 2012 *val* and *test* sets. The supervision type (Sup.) indicates: $\mathcal{I}$–image-level label, $\mathcal{B}$–bounding box, $\mathcal{S}$–scribble, and $\mathcal{F}$–segmentation label.

Semantic Segmentation Labels: A reduced version of our framework, which skips the instance-wise CAM generation step, produces pseudo labels for semantic segmentation. In this aspect, we compare our framework with the previous state-of-the-art in semantic segmentation label synthesis, AffinityNet [1], in terms of mean Intersection-over-Union (mIoU). Similar to ours, AffinityNet also conducts the semantic propagation to enhance CAMs using predicted pairwise semantic affinities. Table 2 compares the quality of our pseudo segmentation labels to that of AffinityNet [1]. The accuracy of our pseudo labels is substantially higher than that of AffinityNet thanks to the superior quality of pairwise semantic affinities predicted by IRNet.

6.3. Mask R-CNN for Instance Segmentation

We evaluate the performance of an instance segmentation network trained with pseudo labels generated by our framework. For evaluation, we adopt Mask R-CNN [17], which is one of the state-of-the-art instance segmentation networks, with ResNet-50-FPN [28] as its backbone. Fig. 7 shows qualitative results of the Mask-RCNN trained with our pseudo labels, and Table 3 compares its performance to those of previous approaches in AP^r[2] [16]. As shown in Table 3, ours largely outperforms PRM [50], which is the state-of-the-art that also uses image-level supervision. Our approach even outperforms SDI [22], which uses bounding box supervision, by 1.9%, and SDS [16], which uses full supervision, by 2.9% in AP^r_{50}.

[2] AP^r means average precision of masks at different IoU thresholds.

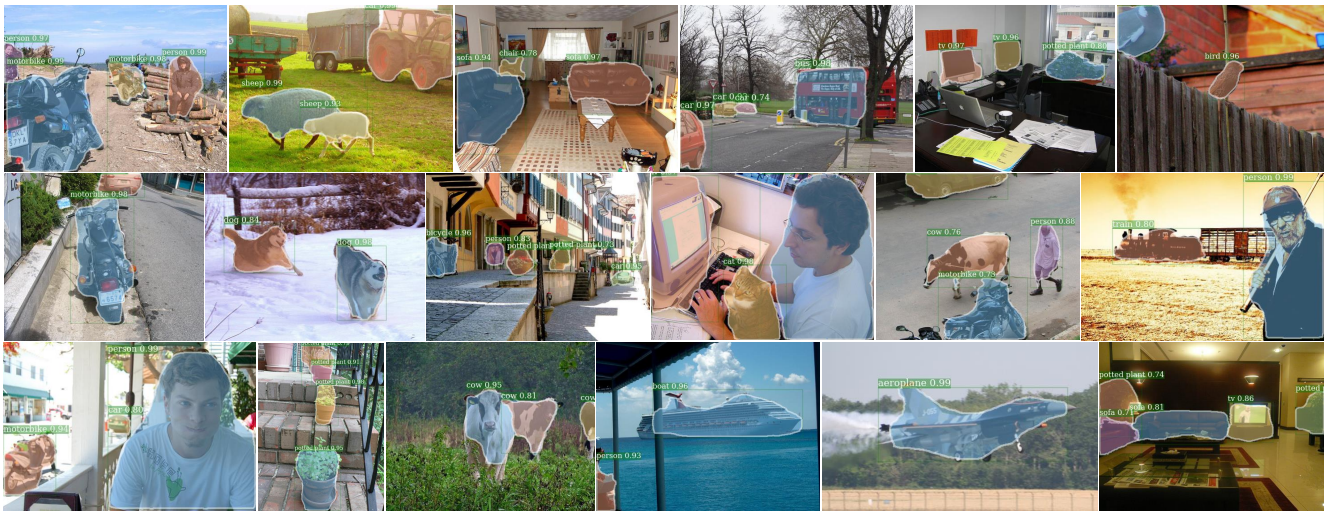

Figure 7. Qualitative results of our instance segmentation model on the PASCAL VOC 2012 *val* set.

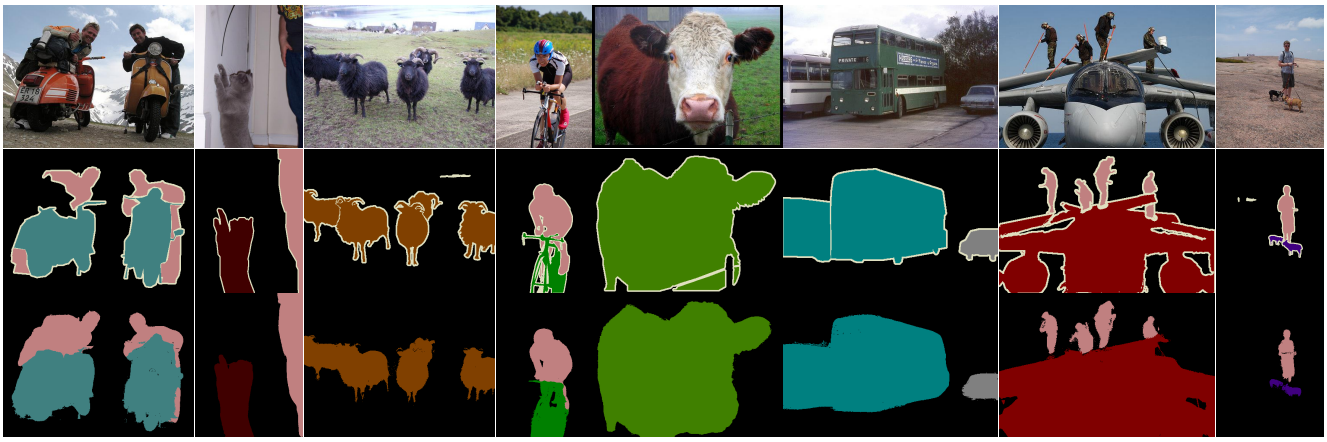

Figure 8. Qualitative results of smenatic segmentation on the PASCAL VOC 2012 *val* set. (*top*) Input images. (*middle*) Groundtruth semantic segmentaton. (*bottom*) Results of Ours-ResNet50.

6.4. DeepLab for Semantic Segmentation

We further explore the effectiveness of our framework by training DeepLab v2-ResNet50 [5] with our pseudo semantic segmentation labels. Fig. 8 visualizes semantic segmentation results obtained by our approach and Table 4 compares ours with other weakly supervised approaches. Our approach outperforms previous arts relying on the same level of supervision, and is even competitive with BoxSup [8], which utilizes stronger bounding box supervision. Also it recovers 88% of its fully supervised counterpart, the upper bound that it can achieve.

7. Conclusion

Weakly supervised instance segmentation with image-level supervision is a significantly ill-posed problem due to the lack of instance-specific information. To tackle this challenging problem, we propose IRNet, a novel CNN architecture that identifies individual instances and estimates their rough boundaries. Thanks to the evidences provided by IRNet, simple class attentions can be significantly improved and used to train fully supervised instance segmentation models. On the Pascal VOC 2012 dataset, models trained with our pseudo labels achieve the state-of-the-art performance in both instance and semantic segmentation.

Acknowledgement: This work was supported by Korea Creative Content Agency (KOCCA), Ministry of Culture, Sports, and Tourism (MCST) of Korea, Basic Science Research Program, and Next-Generation Information Computing Development Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT (NRF-2018R1C1B6001223, NRF-2018R1A5A1060031, NRF-2017M3C4A7066316). It was also supported by the DGIST Start-up Fund Program (2018010071).

References

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

[2] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[3] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the Point: Semantic Segmentation with Point Supervision. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[4] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi. Convolutional random walk networks for semantic image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

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

[7] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[8] J. Dai, K. He, and J. Sun. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

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

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 2010.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systems (NIPS)*, 2014.

[15] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2011.

[16] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 297–312, 2014.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly supervised semantic segmentation using web-crawled videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2224–2232, 2017.

[20] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[23] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother. Instancecut: From edges to instances with multicut. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[25] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proc. Neural Information Processing Systems (NIPS)*, 2011.

[26] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan. Proposal-free network for instance-level semantic object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

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

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

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, 2014.

[30] S. Liu, X. Qi, J. Shi, H. Zhang, and J. Jia. Multi-scale patch aggregation (MPA) for simultaneous detection and segmen-

tation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum. Learning to detect a salient object. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[32] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[33] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2001.

[34] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[35] D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi. Semi-convolutional operators for instance segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[36] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele. Exploiting saliency for object segmentation from image level labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[38] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[39] C. S. Pavel Tokmakov, Karteek Alahari. Learning semantic segmentation with weakly-annotated videos. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[40] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] T. Remez, J. Huang, and M. Brown. Learning to segment via cut-and-paste. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[45] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[47] Y. Wu and K. He. Group normalization. In *Proc. European Conference on Computer Vision (ECCV)*, September 2018.

[48] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[50] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao. Weakly supervised instance segmentation using class peak response. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.