International Journal of Computer Vision
https://doi.org/10.1007/s11263-020-01293-3

Weakly-Supervised Semantic Segmentation by Iterative Affinity Learning

Xiang Wang[1,2] · **Sifei Liu**[3] · **Huimin Ma**[4] · **Ming-Hsuan Yang**[5]

Received: 21 April 2019 / Accepted: 5 January 2020

Abstract
Weakly-supervised semantic segmentation is a challenging task as no pixel-wise label information is provided for training. Recent methods have exploited classification networks to localize objects by selecting regions with strong response. While such response map provides sparse information, however, there exist strong pairwise relations between pixels in natural images, which can be utilized to propagate the sparse map to a much denser one. In this paper, we propose an iterative algorithm to learn such pairwise relations, which consists of two branches, a unary segmentation network which learns the label probabilities for each pixel, and a pairwise affinity network which learns affinity matrix and refines the probability map generated from the unary network. The refined results by the pairwise network are then used as supervision to train the unary network, and the procedures are conducted iteratively to obtain better segmentation progressively. To learn reliable pixel affinity without accurate annotation, we also propose to mine confident regions. We show that iteratively training this framework is equivalent to optimizing an energy function with convergence to a local minimum. Experimental results on the PASCAL VOC 2012 and COCO datasets demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods.

Keywords Weakly-supervised learning · Semantic segmentation · Affinity

1 Introduction

Semantic segmentation aims to predict a label for each pixel from a set of pre-defined object classes. With the advances of Deep Neural Networks (DNNs), significant progress has been made in semantic segmentation (Long et al. 2015; Zhao et al. 2017; Chen et al. 2018, 2017; Zhou et al. 2019). However, fully-supervised methods require a large amount of pixel-wise annotations, which is time-consuming and expensive. To make semantic segmentation more practical, a number of weakly-supervised methods have been proposed in recent years based on partial information of each image, such as bounding boxes (Dai et al. 2015; Khoreva et al. 2017), scribbles (Lin et al. 2016), points (Bearman et al. 2016), and even class labels (Pathak et al. 2015; Wang et al. 2018b; Ahn and Kwak 2018; Huang et al. 2018; Wei et al. 2018). In this paper, we present a weakly-supervised semantic segmentation algorithm based only on class labels of an image.

Weakly-supervised semantic segmentation based on class labels is challenging as no pixel in an image is annotated (i.e., an image is only annotated with class labels as shown in Fig. 1a). Recently, the Class Activation Map (CAM) method (Zhou et al. 2016) has been developed to generate discriminative object seed regions with classification networks. Since coarse response maps are generated (Fig. 1d), these regions cannot be directly used to train an accurate segmentation network. As data redundancy often exists in natural images (Kersten 1987), significant statistical dependencies among pixels in images can be exploited. We can learn similarities or affinities from images, and propagate sparse and

Communicated by Kristen Grauman.

✉ Huimin Ma
mhmpub@ustb.edu.cn

Xiang Wang
andyxwang@tencent.com

Sifei Liu
sifeil@nvidia.com

Ming-Hsuan Yang
mhyang@ucmerced.edu

[1] Tencent Research, Beijing, China

[2] Tsinghua University, Beijing, China

[3] Nvidia, Santa Clara, CA, USA

[4] University of Science and Technology Beijing, Beijing, China

[5] University of California at Merced, Merced, CA, USA

Published online: 30 January 2020

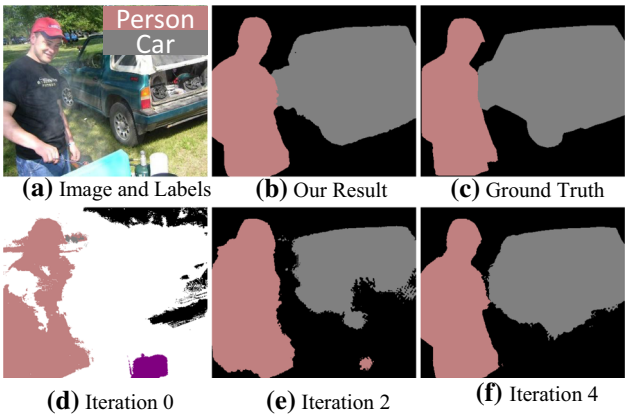


Fig. 1 Top row: Given training images and their class labels, our framework generates accurate segmentation results. Bottom row: By iteratively learning affinity, our framework progressively generates better segments for supervising the segmentation network. The seed regions generated by the CAM method (Zhou et al. 2016) are shown in **d** where white color pixels denote image locations with unknown labels

noisy labels of object regions to generate dense and accurate annotations. With weak supervision, this is challenging as there are no accurate pixel-wise annotations and the region labels from the CAM method are noisy and sometimes inaccurate. To address these issues, we mine confident regions from the coarse pixel labels and then learn pixel affinities from them to refine the coarse labels. Iteratively, we mine confident regions from the refined results and learn more robust affinities until convergence.

In this paper, we propose an iterative affinity learning framework, which consists of two major branches (see Fig. 2): a unary segmentation network which learns the pixel-wise probability of semantic categories from produced labels, and a pairwise network which refines the current labels by learning the affinity matrix and propagating the labels. The refined results by the pairwise network provide better "ground truth" to retrain the unary segmentation network in the next iteration. The above procedures are conducted iteratively until convergence to obtain better segmentation progressively. Figure 1 shows one example. Given training images and the class labels, the proposed framework can generate accurate semantic segmentation results. This is achieved by the iterative optimization strategy which learns reliable affinity and generates better masks for supervising the segmentation network.

The key ingredient of our framework is learning affinities between pixels, which determines the amount of improvements achieved at each iteration. However, under weak supervision, we do not have accurate annotations to learn pixel affinities. To address this issue, we propose to mine confident regions from the output results of the unary network, and then use them to supervise the pairwise affinity network. Our motivation is that, to learn the affinity, we only need to know some pixel samples, which indicate the pixels belonging to the same (their pixel affinity should be high) or different classes (their pixel affinity should be low). Even with a small amount of pixel samples, we are able to learn segmentation by propagating and mining more labels via learning the affinity. We also show that iteratively training the proposed framework is equivalent to optimizing an energy function with an EM-like approach. Furthermore, we show that this process always converges to a local minimum due to that the energy function is differentiable with respect to both the output labels and the network parameters.

The main contributions of this work are summarized as follows:

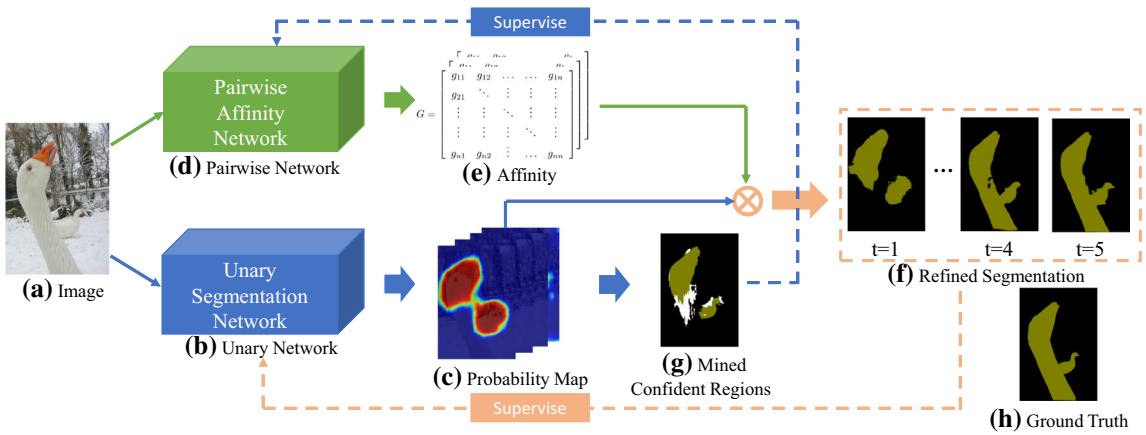


Fig. 2 Illustration of the proposed framework. The framework consists of two branches: a **b** unary network which predicts a **c** probability map of the input image, and a **d** pairwise network, which learns the **e** affinity matrix from the **g** mined confident regions. The learned affinities are then applied to the probability map from the unary network to generate the **f** refined segmentation. The refined results are then used as supervision signals to retrain the unary network. These procedures are conducted iteratively to learn more robust affinity progressively and produce more accurate segmentation.

- We present an iterative affinity learning framework to progressively generate better segmentation, and show that it is equivalent to optimizing an energy loss function. We show that it always converges to a local minimum.
- We propose a method to learn reliable affinity from inaccurate annotations by mining confident regions.
- We demonstrate that the proposed weakly-supervised semantic segmentation algorithm performs favorably against the state-of-the-art methods on the PASCAL VOC 2012 and COCO datasets.

2 Related Work

In this section, we discuss related methods for weakly-supervised semantic segmentation and learning affinity for segmentation.

2.1 Weakly-Supervised Semantic Segmentation

Weakly-supervised semantic segmentation based on class labels has drawn much attention in recent years due to low annotation costs. Early methods (Pathak et al. 2014, 2015; Pinheiro and Collobert 2015) mainly formulate this problem as a multi-instance learning (MIL) problem. Pathak et al. (2014) propose to add a max-pooling layer on top of FCN (Long et al. 2015) and design a multi-class MIL loss for training the network. Based on this framework, several methods have been developed (Pathak et al. 2015; Pinheiro and Collobert 2015). Pathak et al. (2015) add several constraints on foreground and suppression schemes to the MIL framework for weakly-supervised semantic segmentation. Pinheiro and Collobert (2015) replace the max-pooling layer of the MIL framework with a new *Log-Sum-Exp* layer which can consider more information of the feature layers.

Recent methods (Kolesnikov and Lampert 2016; Wei et al. 2017a; Wang et al. 2018b; Ahn and Kwak 2018; Huang et al. 2018; Wei et al. 2018) tackle weakly-supervised semantic segmentation by a two-stage procedure, which first generates initial object labels with class activation maps (Zhou et al. 2016), and then trains segmentation networks based on the response maps. Kolesnikov and Lampert (2016) present an end-to-end framework with three modules (seed, expand and constrain) as loss functions, and the class activation maps are used as supervisory signals. A number of methods are developed to expand object regions based on the class activation maps. Wei et al. (2017a) propose to progressively erase most significant regions in the activation maps and then generate more regions. These regions are then used as ground truth to train a segmentation network. Wang et al. (2018b) develop a bottom-up and top-down framework which iteratively mines common object features to expand initial object regions from the class activation maps. Ahn and Kwak (2018) learn the pixel affinity from the activation maps and then apply the random walk method to refine them. Huang et al. (2018) design a deep seeded region growing algorithm which improves the seed regions to supervise the network.

Among the above-mentioned approaches, Wei et al. (2017a) and Wang et al. (2018b) also use iterative strategies to refine segmentation results. However, the method of Wei et al. (2017a) heavily relies on the CAM network to progressively produce the most significant regions in the remaining images. Consequently, less discriminative object regions are usually missing. In addition, this method is not able to suppress noisy regions well. Wang et al. (2018b) expand object regions by mining common object features. However, common features are only learned from each superpixel region and the pixel-wise context information is not exploited. In contrast, our method can learn and propagate pixel-wise affinities to achieve better segmentation results. We note Ahn and Kwak (2018) also use the pixel affinities to refine segmentation results. However, the affinities are only learned from the coarse response map of the CAM method. In the proposed framework, the pixel affinities are iteratively optimized, which are more reliable and lead to better segmentation results.

2.2 Learning Pixel Affinity for Segmentation

An affinity matrix measures the similarities between pixels and has been widely used in object segmentation. Some early methods directly define similarity functions to compute affinity matrices. Hagen and Kahng (1992) propose a spectral methods for ratio cut (Wei et al. 1989) which captures both min-cut and equipartition to locate natural clusters. Shi and Malik (2000) formulate image segmentation as a graph partitioning problem and present the normalized cut algorithm. This algorithm considers both the dissimilarity between different groups and the similarity within the same group.

In recent years, with the advances of DNNs, numerous algorithms have been proposed to learn the affinity end-to-end with deep networks (Liu et al. 2017; Maire et al. 2016; Bertasius et al. 2017). Maire et al. (2016) present the affinity CNN which directly learns an affinity matrix to model pairwise relations for figure and ground embedding. Liu et al. (2017) design the spatial propagation network (SPN) which directly learns pixel affinities and a spatial linear propagation module. The SPN takes images and coarse masks as input and learns pixel affinities end-to-end to refine the coarse masks. Bertasius et al. (2017) develop a random walk layer on top of the semantic segmentation network to learn the pixel affinities.

These methods all learn the pixel affinities under full supervision to refine segmentation results. In contrast, our method aims to learn pixel affinities to refine object regions from coarse and inaccurate labels without pixel-wise anno-

tations. To address this challenging problem, we propose an iterative optimization framework which progressively mines confident regions for learning reliable affinity and generates better segmentation results.

3 Proposed Algorithm

We solve the weakly-supervised semantic segmentation problem with an iterative optimization algorithm which progressively learns robust pixel affinities and propagates label information for accurate results. We present an EM approach that alternatively learns the network parameters for both unary segmentation and pairwise affinity networks, and maximizes the likelihood of the "ground-truth" labels. This is different from the fully-supervised approaches where the supervision requires the ground-truth labels.

3.1 Formulation

Let x denote an image. The proposed framework consists of two major branches (Fig. 2): (a) a unary network $F = f(x, W_f)$ parameterized by W_f that learns the label probability with respect to each pixel in x, and (b) a pairwise network $G = g(x, W_g)$ that learns the pixel affinities, where $G \in \mathcal{R}^{N\times N}$, N is the number of pixels, and W_g is the parameter of the pairwise network. In addition, we denote α as the hidden state of the output labels. We use the subscript t to denote the t^{th} step in the iterative process.

We represent each image as an undirected weighted graph $\mathcal{G} = (V, E)$, with the vertex set $V = \{v_1, \ldots, v_N\}$, where each edge between v_i and v_j has a weight w_{ij}. The adjacency matrix is $W = (w_{ij})_{i,j=1,\ldots,N}$. The degree matrix D is a diagonal matrix with the degrees $d_1, \ldots, d_N$ as elements, where $d_i = \sum_{j=1}^{N} w_{ij}$. The semantic segmentation problem is then to minimize the following energy loss function:

$$\alpha^* = \arg\min_{\alpha} J(\alpha, W_f, W_g) = \arg\min_{\alpha} \alpha^\top L\alpha, \tag{1}$$

where $L = D - W$ is the Laplacian matrix, and

$$\begin{aligned}\alpha^\top L\alpha &= \alpha^\top (D - W)\alpha \\ &= \sum_{i=1}^{N} d_i\alpha_i^2 - \sum_{i,j=1}^{N} \alpha_i\alpha_j w_{i,j} \\ &= \frac{1}{2}\left(\sum_{i=1}^{N} d_i\alpha_i^2 - 2\sum_{i,j=1}^{N} \alpha_i\alpha_j w_{i,j} + \sum_{j=1}^{N} d_j\alpha_j^2\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{N} w_{i,j}(\alpha_i - \alpha_j)^2. \end{aligned} \tag{2}$$

That is, to minimize the loss function (2) is to enforce pixels with high similarities (their affinity $w_{i,j}$ is high) to have similar labels. This allows us to propagate label information for accurate results.

Instead of designing similarity metric and solve α as an optimization problem (Levin et al. 2008), we propose an iterative learning method to refine the probability map and the networks via an EM formulation. We denote $G = I - D + W = I - L$ as the affinity transformation matrix (Liu et al. 2017), which is learnable by the pairwise network $g(x, W_g)$, and α^u, α^p as the output of the unary network and the pairwise network, respectively.

The EM procedure are as follows:

- **Initialization**: We train the unary network and the pairwise network with object seeds Y_0 from class activation maps (Zhou et al. 2016) to obtain the initial parameter $\{W_f, W_g\}_0$, the unary response map α_0^u (Fig. 2c).
- **E-step**: We refine the unary probability by minimizing J_t w.r.t α_t^u given W_f, W_g, where:

$$\frac{\partial J}{\partial \alpha_t^u} = L_t\alpha_t^u = (I - G_t)\alpha_t^u, \tag{3}$$

  and compute the refined map as $\alpha_t^p = \alpha_t^u - \Delta\alpha_t^u$ (i.e., α_t^p is the output of the pairwise network in step t). From (3) we have $\alpha_t^p = G_t\alpha_t^u$, where the corresponding network implementation is described in Sect. 3.2.2.
- **M-step**: In this step, we minimize J_t by learning both W_f and W_g, through training the network $f_{t+1}(x, W_f)$ and $g_{t+1}(x, W_g)$ with the supervision signal extracted from α_t^p (Sects. 3.2.1, 3.2.2, 3.3).

It is straightforward to show the above procedures always converge to a local minimum, due to that J is differentiable with respect to both α and the network parameters. However, to validate the M-step, we need to validate the link between the E-step and M-step, i.e., how we use the network response from step t to train $f_{t+1}(x, W_f)$ and $g_{t+1}(x, W_g)$ to minimize the energy function (1).

For the unary network, in the $t+1$ step, it uses segmentation results of α_t^p as supervision to generate label probability α_{t+1}^u. For training the pairwise network in the $t+1$ step, we consider the softmax cross-entropy loss function with α_{t+1}^u as supervision:

$$H(\alpha_{t+1}^p) = -{\alpha_{t+1}^u}^\top \log \alpha_{t+1}^p. \tag{4}$$

Since $\log(\cdot)$ is a monotonic increasing and convex function, optimizing α_{t+1}^p is to learn G_{t+1} and $\alpha_{t+1}^p = G_{t+1}\alpha_{t+1}^u$. Therefore, minimizing H is equivalent to minimize $-{\alpha_{t+1}^u}^\top\alpha_{t+1}^p = -{\alpha_{t+1}^u}^\top G_{t+1}\alpha_{t+1}^u = -{\alpha_{t+1}^u}^\top(I - L_{t+1})\alpha_{t+1}^u$. As the first term $-{\alpha_{t+1}^u}^\top\alpha_{t+1}^u$ is a constant, to optimize α_{t+1}^p is to minimize the second term:

Algorithm 1 Procedures of the proposed approach

Input:

Generate object seeds from CAM, set it as Y_0.

Training images x.

Initialize:

Train networks $f_0(x, W_f)$ and $g_0(x, W_g)$ with Y_0 to obtain the parameters $\{W_f, W_g\}_0$, the affinity matrix G_0 and the output of the unary network α_0^u.

E-step:

1: Propagate α_t^u with G_t: $\alpha_t^p = G_t \alpha_t^u$ (Sect. 3.2.2).

M-step:

2: Train $f_{t+1}(x, W_f)$ with α_t^p as supervision to obtain α_{t+1}^u (Sect. 3.2.1).

3: Mine confident regions Y_{t+1} from the output of $f_{t+1}(x, W_f)$ (Sect. 3.3).

4: Train $g_{t+1}(x, W_g)$ with Y_{t+1} as supervision to obtain G_{t+1} (Sect. 3.2.2).

$$L_{t+1} = \arg\min_{L_{t+1}^*} {\alpha_{t+1}^u}^\top L_{t+1} \alpha_{t+1}^u, \tag{5}$$

which is consistent with (1). By using α_t^p as supervision to train the unary network and using α_{t+1}^u to supervise the pairwise network with the softmax cross-entropy loss, it is equivalent to minimize the original energy loss function. Namely, the objective of the M-step is to minimize the energy loss function.

However, in the stage of the pairwise network, if we use α_{t+1}^u to supervise the pairwise network to learn affinity and then refine itself, there is no information gain over iterations and the optimization will come to convergence to a relative low performance with very few steps (Sect. 4.5.2). To obtain more accurate supervision in each step, we propose to mine confident regions from the output of the unary network α_{t+1}^u. These confident regions contain pixels belonging to object regions with high precision from which we can learn reliable affinity matrices (Sect. 3.3). We denote it as Y_{t+1}, and expect it to have lower energy (Sect. 4.5.2):

$$Y_{t+1}^\top L_{t+1} \alpha_{t+1}^u \leq {\alpha_{t+1}^u}^\top L_{t+1} \alpha_{t+1}^u. \tag{6}$$

With mining confident regions, our algorithm converges to a lower energy and obtains better segmentation results.

The proposed EM procedures are summarized in Algorithm 1.

3.2 Network Architecture and Training

Figure 2 shows the architecture of the proposed framework. The framework consists of two major branches, a unary network $F = f(x, W_f)$ that learns the label probability of each pixel, and a pairwise network $G = g(x, W_g)$ that learns the affinity. The learned affinities are applied to the output probability map of the unary network to refine it and obtain better segmentation results.

3.2.1 Unary Network

The unary network aims to generate a probability map given a coarse segmentation mask. In this work, we use the DeepLab (Chen et al. 2018) model as the unary segmentation network. To initialize the framework, we first generate object seed regions using the CAM method (Zhou et al. 2016) in a way similar to (Ahn and Kwak 2018). The CAM method generates object regions for all classes, including background, and pixels with weak response are labelled as unknown, as shown in Fig. 1d. We then use them as pseudo ground truth to train the unary segmentation network. The training process is the same as fully-supervised methods with a softmax loss as the objective function. With this segmentation network, probability maps are generated for all classes.

3.2.2 Pairwise Network

The pairwise network aims to learn the pixel affinities from object regions and then applies to the probability maps to refine the segmentation results. In this work, we use the Spatial Propagation Network (SPN) (Liu et al. 2017) to learn pairwise affinities. The SPN learns the affinity transformation matrix from an image x to refine the coarse probability maps α_t^u and generates better segmentation α_t^p. It is an end-to-end framework which simultaneously learns the affinity transformation matrix G and outputs the refined segmentation $\alpha_t^p = G_t \alpha_t^u$. When learning the affinity, we raster scan the pixels from four directions: left-to-right, top-to-bottom, and vice versa. Since we use all three RGB image channels, we learn 12 affinity matrices. More details regarding the spatial propagation network can be found in Liu et al. (2017).

The spatial propagation network has been shown to perform well in pixel labelling under full supervision (Liu et al. 2017). However, under weak supervision, it is challenging to train the pairwise affinity network as no pixel-wise annotations are provided. A straight-forward approach is to use the segmentation result at time t as ground truth to supervise the pairwise affinity network. However, as the segmentation results are not accurate, the affinity matrix cannot be learned well. To address this issue, we first mine confident regions from the segmentation results and then learn the affinity matrix from the mined confident regions. If we can obtain some confident regions which have high precision to identify the class each region belongs to, we know that pixels within same class should have high affinities and pixels of different classes should have low affinities. Thus, we can also learn reliable affinity matrices. For regions with low confidence scores, we mark them with unknown labels when training the pairwise affinity network. Namely, when computing the softmax loss function, these regions are ignored. The details of mining confident regions are introduced in Sect. 3.3. We denote the mined confident regions as Y_t. To

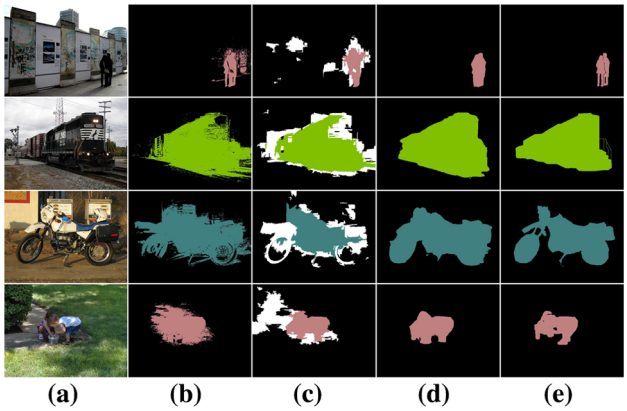


Fig. 3 Some examples of mining confident regions from segmentation results of the unary network: **a** images, **b** segmentation results of the unary network, **c** mined confident regions, **d** refined results of the pairwise affinity network, **e** ground truth. White color pixels denote image locations with unknown labels.

train the pairwise affinity network, we utilize the softmax loss:

$$\mathcal{L}_a = -Y_t^\top \log \alpha_t^p. \tag{7}$$

To learn accurate affinity matrices, we also introduce a *region smoothness loss*. Our motivation is that a good affinity matrix should have similar values for pixels in the same object regions such that the refined results can be smooth and have clear boundaries. To achieve this, we average the learned affinity matrix G within each superpixel region and denote it as G_s. The objective function is then to minimize the difference between G and G_s:

$$\mathcal{L}_s = \|G - G_s\|_2^2. \tag{8}$$

3.3 Mining Confident Regions

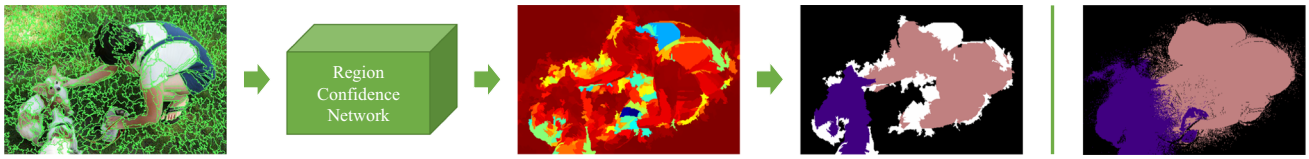
The key issue in our framework is how to learn reliable affinity matrices without accurate annotations. Our solution is to mine confident regions. We expect these confident regions contain pixels belonging to object regions with high precision from which we can learn reliable affinity matrices.

Our method is based on statistical learning. The object regions generated by the unary network contain noisy results. For example, some background pixels may be recognized as parts of an object, as shown in Fig. 3b. We can learn a confidence score for each region with the segmentation results of the unary network as training samples. For a region with a certain class label, its initial confidence score is set as 1 for this object, and 0 for other objects. By training a multi-class classification network with these regions, each region is assigned with a new confidence score. For region pixels that have high similarity to one object, they will receive high confidence scores. For region pixels different from one object (i.e., noisy regions), they will receive low confidence scores. With this procedure, we can remove noisy regions and select confident regions with high confidence scores. In this paper, we set the threshold as 0.7 based on our empirical observations. Some examples are shown in Fig. 3c. With these confident regions, we can learn reliable affinity matrices from them and thus generate more accurate segmentation results (Fig. 3d).

We first segment images into superpixel regions (Felzenszwalb and Huttenlocher 2004) $\mathcal{S} = \{S_{i,j}\}$, where $S_{i,j}$ denotes the j-th superpixel in the i-th image. For each region, its class label is obtained from the segmentation results. If more than 80% pixels of a superpixel is marked with a certain class c in the segmentation results, then this superpixel is considered as a sample of class c. This scheme is formulated with the one-hot encoding, namely, $L_{i,j} = [0, \ldots, 1, \ldots, 0]$, where $L_{i,j}(c) = 1$, $L_{i,j}(k) = 0$ $(k = 0, \ldots, C, k \neq c)$, and C is the number of classes. With the superpixel regions and corresponding labels $\mathcal{D} = \{S, L\}_{i,j}$, we can train a region classification network f_c^m parameterized by θ_m to obtain a confidence score for each region with the cross-entropy loss function:

$$\mathcal{L}_m = -\sum_{i,j,c} L_{i,j}(c) \log f_c^m(S_{i,j}|\theta_m). \tag{9}$$

We train the region confidence network with the architecture proposed by (Wang et al. 2018a) which is a variant of the fast R-CNN model with a mask pooling scheme. Similar to recent weakly-supervised learning methods (Pathak et al. 2015; Kolesnikov and Lampert 2016; Ahn and Kwak 2018; Huang et al. 2018), we initialize this network with the weights of a pre-trained model based on the ImageNet. The model is trained with $\mathcal{D} = \{S, L\}_{i,j}$ using (9) as the loss function, where the superpixel region S is the input and the corresponding class label L is the supervisory signal. With this region confidence network, we extract features of all superpixel regions of an image in one forward pass, and then recognize their classes. Figure 4 shows the process of mining confident regions. With the trained region confidence network, we can re-predict each superpixel region of images, and obtain confidence scores for all regions (Fig. 4c). To extract regions with high precision for learning reliable affinities, we select regions with high confidence scores (e.g., > 0.7 in this work), and leave others as unknown (Fig. 4d). Namely, we do not use unknown regions for training the pairwise affinity network.



(a) Image regions **(b)** Region confidence network **(c)** Region confidence scores **(d)** Mined confident regions **(e)** Unary segmentation result

Fig. 4 Illustration of mining confident regions. Given an **a** input image, we segment it into superpixel regions, and apply the learned **b** region confidence network to predict object classes, and generate **c** confidence scores for all regions. By selecting regions with high confidence score, we can obtain the **d** mined confident regions. White color pixels denote image regions with unknown labels. Compared with the **e** unary segmentation result, the noisy regions are mostly removed and some regions are corrected. Thus, regions with high precision are extracted and used to better supervise the pairwise network

4 Experimental Results

4.1 Settings

We evaluate the proposed method on the PASCAL VOC 2012 (Everingham et al. 2010) and COCO (Lin et al. 2014) datasets. The PASCAL VOC 2012 dataset contains 20 object classes and 1 background class with 1464 training images, 1449 validation images, and 1456 testing images. Same as the recent work (Wei et al. 2017a; Ahn and Kwak 2018; Huang et al. 2018; Wei et al. 2018; Wang et al. 2018b), we use the augmented set with 10582 images from (Hariharan et al. 2011) for training. For the COCO dataset, it contains more complex scenes and more classes (80 classes plus 1 background class) with 80k images for training and 40k images for validation. We iteratively train our framework on the training set using only class labels. For inference, we forward the input images to the trained networks in the last iteration to obtain segmentation results, and the process is still efficient. We evaluate the proposed algorithm against the state-of-the-art methods using the mean intersection-over-union (mIoU) metric.

4.2 Training Process

The CAM Network is trained with the PyTorch framework and other models are trained with the Caffe package (Jia et al. 2014). Similar to recent weakly-supervised learning methods (Pathak et al. 2015; Kolesnikov and Lampert 2016; Ahn and Kwak 2018; Huang et al. 2018), all networks are initialized with the weights of a pre-trained model based on the ImageNet. All the source code and trained models will be made available online.

CAM Model The CAM model is used to generate object seed regions from images based on the implementation by Ahn and Kwak (Ahn and Kwak 2018). To train this CAM model, the input data is the training images and the supervisory signals are the corresponding class labels. Similar to the CAM model by Zhou et al. (Zhou et al. 2016), we use random cropping to augment data. For each crop, we take the class labels corresponding to the original images before cropping as supervision, and no additional supervisory signals are required.

Unary Network We use the polynomial decay policy for the learning rate to train the model (Chen et al. 2018). The learning rate of the k-th iteration, α_k, is:

$$\alpha_k = \alpha_b \times \left(1 - \frac{k}{K}\right)^{\tau}, \tag{10}$$

where the base learning rate $\alpha_b = 0.001$, $\tau = 0.9$, and maximal iterations $K = 20,000$. The momentum parameter is set to be 0.9.

Pairwise Network We use the polynomial decay policy for the learning rate as described in (10) to train the model, where $\alpha_b = 0.00001$, $\tau = 0.5$, $K = 20{,}000$, the momentum parameter is set as 0.9.

Mining Confident Regions Network We use the step learning rate decay policy. For the k-th iteration, the learning rate is:

$$\alpha_k = \alpha_b \times \gamma^{\lfloor \frac{k}{S} \rfloor}, \tag{11}$$

where the base learning rate $\alpha_b = 0.001$, $S = 20,000$, $\gamma = 0.1$, and $\lfloor \cdot \rfloor$ is the floor function. The momentum parameter is also set to be 0.9.

4.3 Performance Evaluation

We evaluate the proposed algorithm on the PASCAL VOC 2012 dataset against the state-of-the-art weakly-supervised segmentation methods including MIL-FCN (Pathak et al. 2014), CCNN (Pathak et al. 2015), MIL-sppxl (Pinheiro and Collobert 2015), EM-Adapt (Papandreou et al. 2015), BFBP (Saleh et al. 2016), DCSM (Shimoda and Yanai 2016), AF-SS (Qi et al. 2016), AF-MCG (Qi et al. 2016), SEC (Kolesnikov and Lampert 2016), STC (Wei et al. 2017b), CBTS (Roy and Todorovic 2017), AE-PSL (Wei et al. 2017a), MCOF (Wang et al. 2018b), PSA (Ahn

Fig. 5 Visual comparisons with the state-of-the-art methods on the PASCAL VOC 2012 *val* set

and Kwak 2018), DSRG (Huang et al. 2018), MDC (Wei et al. 2018) and AISI (Fan et al. 2018). Table 1 shows the experimental results by all the evaluated methods using the VGG16 (Simonyan and Zisserman 2014) model as the backbone network. The proposed algorithm achieves 62.0% and 62.4% on the *val* and *test* sets, respectively, with performance gain over the MDC (Wei et al. 2018) method by 1.6%. We note the PSA (Ahn and Kwak 2018) model also uses affinities to refine object regions. However, as this method only learns affinities from coarse masks generated from CAM, the improvement by the affinity propagation is limited. The proposed algorithm performs favorably against the PSA method by 3.6% and 1.9% on the *val* and *test* sets, respectively. We also note that the AISI (Fan et al. 2018) model recently achieves similar performance as the proposed algorithm (61.3% on *val*, 62.1% on *test*). However, this method uses the S^4Net (Fan et al. 2019) to generate salient instances, which is trained with full supervision using pixel-wise annotations. Table 2 shows the results when the ResNet (He et al. 2016) is used as the backbone model. The proposed algorithm achieves performance gain over PSA (Ahn and Kwak 2018) by 2.6% and 1.7% and AISI (Fan et al. 2018) by 0.7% and 0.9% on *val* and *test* sets, respectively. Figure 5 shows some segmentation results. Overall, the segmentation results by the proposed algorithm contain fewer noisy segments.

Table 1 Comparisons with the state-of-the-art weakly-supervised semantic segmentation methods on the PASCAL VOC 2012 *val* set and *test* set

Methods	Training images	*val*	*test*
MIL-FCN (ICLR'15)	10K	25.7	24.9
CCNN (ICCV'15)	10K	35.3	35.6
MIL-sppxl (CVPR'15)	700K	36.6	35.8
EM-Adapt (ICCV'15)	10K	38.2	39.6
BFBP (ECCV'16)	10K	46.6	48.0
DCSM (ECCV'16)	10K	44.1	45.1
AF-SS (ECCV'16)	10K	52.6	52.7
AF-MCG[†] (ECCV'16)	10K	54.3	55.5
SEC (ECCV'16)	10K	50.7	51.7
STC (PAMI'17)	50K	49.8	51.2
CBTS (CVPR'17)	10K	52.8	53.7
AE-PSL (CVPR'17)	10K	55.0	55.7
MCOF (CVPR'18)	10K	56.2	57.6
PSA (CVPR'18)	10K	58.4	60.5
DSRG (CVPR'18)	10K	59.0	60.4
MDC (CVPR'18)	10K	60.4	60.8
AISI[a] (ECCV'18)	10K	61.3	62.1
Ours	10K	**62.0**	**62.4**

All methods use the VGG16 model as the backbone network
[a] indicates methods implicitly use full supervision
Better results are given in bold

Table 2 Evaluation results when using the ResNet as the backbone model on the PASCAL VOC 2012 dataset

Methods	Training images	*val*	*test*
MCOF (CVPR'18)	10K	60.3	61.2
PSA (CVPR'18)	10K	61.7	63.7
DSRG (CVPR'18)	10K	61.4	63.2
AISI[a] (ECCV'18)	10K	63.6	64.5
Ours	10K	**64.3**	**65.4**

[a] indicates methods implicitly use full supervision)
Better results are given in bold

4.4 Comparison with Iterative PSA

We note that the PSA method (Ahn and Kwak 2018) also refines the confident regions from the CAM model based on affinities for semantic segmentation. However, this approach differs significantly from our method in finding confident regions and learning affinities. Different from the PSA model, the proposed method is optimized iteratively. To analyze the performance of the proposed method, we design an alternative PSA approach for evaluation. In this alternative method, confident regions are minded from the PSA model and the affinities are iteratively learned. We show the evaluation results[1] on the PASCAL VOC 2012 training set in Table 3. The segmentation results of the alternative PSA model are not further refined as the number of iterations is increased. We analyze the mined confident regions by both approaches in terms of the precision metric. Table 4 shows that the precision of the confident regions by the alternative PSA does not increase with the number of iterations. This can be attributed that the PSA method determines confident object and background regions by strengthening foreground and weakening background activation maps. We note that this approach is effective for the coarse CAM results as it can remove noisy regions. However, this operation also removes numerous object regions when the results are dense (e.g., large objects and complex scenes). Consequently, such object regions cannot be identified with more iterations. With the learned affinities from the spatial propagation network, our method mines confident regions with a confidence network, which can remove ambiguous regions and correct noisy regions to obtain regions with higher precision. As shown in Sect. 3.1 and (6), confident regions with higher precision help the framework converge to lower energy and obtain better segmentation results, such that our approach can gradually improve with more iterations until convergence.

[1] We use the code provided by the authors. The authors report results on the original training set (1464 images) of the PASCAL VOC 2012 dataset. Here we present results on the augmented training set (10582 images) as all models are trained on the augmented training set.

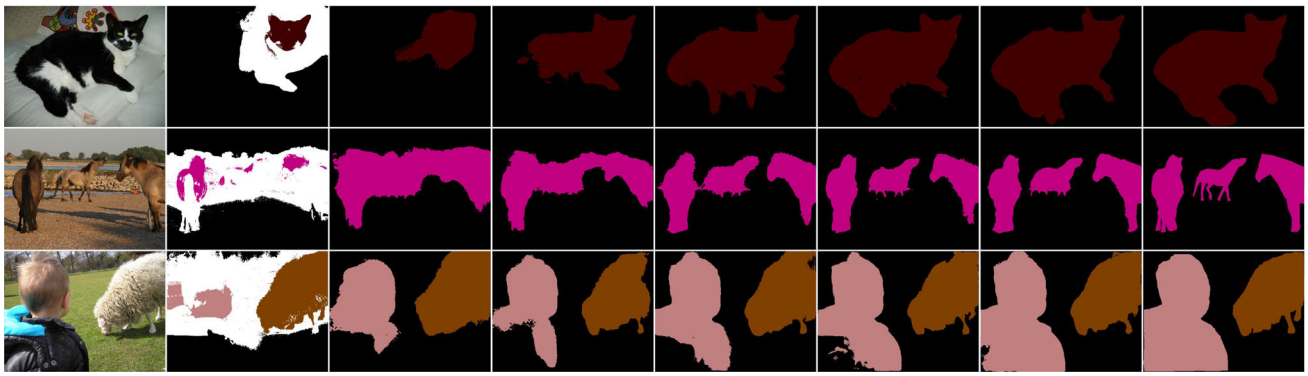


Fig. 6 Visual segmentation results of each iteration of our framework on the PASCAL VOC 2012 training set. The initial object seeds are very coarse, by iteratively learning affinity, the segmentation results become better from coarse to fine. **a** Images, **b** initial object seeds, **c**–**g** produced segmentation results of iterations $t = 1, \ldots, 5$, **h** ground truth

Table 3 Comparisons with the PSA when it is also refined iteratively

	Step 1	Step 2	Step 3	Step 4	Step 5
PSA	**55.6**	54.9	52.1	49.8	48.1
Ours	55.2	**59.5**	**61.4**	**62.7**	**63.1**

The results show the mIoU on the PASCAL VOC 2012 training set
Better results are given in bold

Table 4 Analyze the accuracy of the confident regions of the iterative PSA and ours

	Step 1	Step 2	Step 3	Step 4	Step 5
PSA	**76.2**	73.7	71.7	70.2	68.7
Ours	73.4	**78.1**	**81.0**	**81.2**	**81.2**

The results show the precision on the PASCAL VOC 2012 training set
Better results are given in bold

4.5 Ablation Studies

We conduct ablation studies to analyze the contribution of each module in the proposed framework. All experiments are carried out on the PASCAL VOC 2012 dataset with the VGG16 model as the backbone network.

4.5.1 Iterative Affinity Learning

To demonstrate the effectiveness of the proposed iterative affinity learning method, we show the intermediate results on the PASCAL VOC 2012 training and validation sets in Table 5. We analyze the segmentation results of the training process using the IoU and precision metric. As the performance of the proposed method reaches a plateau after 5 iterations, we use the networks trained at the 5-th iteration for inference. With the affinity matrix being optimized in the first 5 iterations, the performance of the unary network increases gradually from 51.5 to 61.7%, and that of the pairwise network increases from 55.2 to 63.1%. At each step, the mIoU of the pairwise network results is higher than that of the unary network, which demonstrates that the learned affinity matrix is effective in refining the unary segmentation network. The

Table 5 Intermediate results of the proposed framework on the PASCAL VOC 2012 training set

		train mIoU	*train* Precision	*val* mIoU
Step 0	Seeds	46.2	62.2	–
	Unary network	51.5	72.7	
Step 1	Mined conf. regions	49.8	73.4	
	Pairwise network	55.2	73.1	51.3
	Unary network	56.6	76.3	
Step 2	Mined conf. regions	54.1	78.1	
	Pairwise network	59.5	80.5	57.2
	Unary network	59.3	79.2	
Step 3	Mined conf. regions	56.4	81.0	
	Pairwise network	61.4	79.7	59.9
	Unary network	60.8	80.7	
Step 4	Mined conf. regions	56.9	81.2	
	Pairwise network	62.7	80.9	61.6
	Unary network	61.7	79.7	
Step 5	Mined conf. regions	57.5	81.2	
	Pairwise network	63.1	80.8	62.0
	Unary network	61.6	77.7	
Step 6	Mined conf. regions	57.6	81.0	
	Pairwise network	62.8	80.6	61.8
	Unary network	61.8	78.9	
Step 7	Mined conf. regions	57.7	81.3	
	Pairwise network	63.2	80.9	62.0

Table 6 Comparisons with method that eliminates the procedure of mining confident regions

	Step 1	Step 2	Step 3	Step 4	Step 5
Without mining conf.	52.8	56.6	56.5	56.6	56.4
With mining conf.	**55.2**	**59.5**	**61.4**	**62.7**	**63.1**

The results show the mIoU on the PASCAL VOC 2012 training set
Better results are given in bold

Table 7 Energy of each iteration without and with the procedure of mining confident regions

	Step 1	Step 2	Step 3	Step 4	Step 5
Without mining conf.	0.092	0.065	0.061	0.053	0.048
With mining conf.	**0.061**	**0.044**	**0.042**	**0.034**	**0.029**

Better results are given in bold

main reason that the performance is increased with more iterations is that the proposed method learns robust affinities from the mined confident regions. Under weak supervision, we do not have integral and accurate object masks. As we mentioned in Sect. 3.2.2, to learn robust affinities we only need some confident regions which have high precision to identify the class each region belongs to. As shown in Table 5, at each step, as ambiguous regions are removed, the mined confident regions are less integral than that of the unary network (i.e., lower mIoU), but the precision is higher, which provides more accurate supervision for learning affinities robustly. We also show segmentation results at each iteration in Fig. 6. With more iterations, our framework gradually generates more accurate segmentation results.

4.5.2 Mining Regions with High Confidence Scores

To validate the proposed mining method for confident regions, we compare with the alternative without this procedure. We show the output of the pairwise network at each iteration in Table 6. Without mining confident regions, the framework converges after 2 iterations and achieves lower segmentation performance. With the mined confident regions for learning reliable affinity matrices, the proposed method performs better over the iterations. The results demonstrate the importance of the proposed mining method.

As stated in Sect. 3.1, by mining regions with high confidence scores, we expect they have a lower energy than original (formula (6)). To validate this claim, we compare the energy before and after mining the confident regions. We randomly select 500 images as samples and compute their average energy with (1). Table 7 shows the intermediate results at each step. With the proposed mining confident regions, the energy is decreased, which indicates that formula (6) can be satisfied.

Table 8 Comparisons with the alternative method without the pairwise affinity network

	Step 1	Step 2	Step 3	Step 4	Step 5
Without learning affinity	49.6	51.7	53.5	54.2	54.3
With learning affinity	**55.2**	**59.5**	**61.4**	**62.7**	**63.1**

The segmentation results on the PASCAL VOC 2012 training set are presented using the mIoU
Better results are given in bold

4.5.3 Pairwise Affinity Learning

The pairwise affinity network aims to learn the pixel affinities to refine the segmentation results with spatial propagation. To validate the effectiveness and necessity of learning the pairwise network, we remove it from our framework. Table 8 shows the segmentation results with and without using pixel affinities. Without learning the affinity network, the performance at each iteration is much lower than the proposed method. Finally, the mIoU is lower than our method by 8.8% on the training set and 9.2% on the val set. These results demonstrate the importance of learning the affinity network. Although the proposed algorithm is able to obtain regions with high precision by mining confident regions, it misses some regions of objects, as shown in Fig. 3c. If we directly use the mined confident regions to supervision the unary segmentation network, some segments are likely missing, and thus affect the performance. By learning the pairwise affinity network, we can propagate the pixel labels from confident regions to regions with unknown labels. As such, we can achieve better object segmentation results.

4.5.4 Region Smoothness Constraint on Affinity

To validate the proposed region smoothness loss for training the pairwise affinity network, we show the learned pixel affinities in Fig. 7. As mentioned in Sect. 3.2.2, we learn 12 affinity matrices (4 directions for 3 image channels), each affinity matrix has the same channels with the input probability maps. For presentation clarity, here we show some channels of the first learned affinity matrix. The results are similar for other matrices. With the region smoothness constraint, the learned affinity values inside object regions are smoother with more clear object boundaries. For segmentation, the region smoothness constraint can improve the final results from 61.2 to 62.0% on the PASCAL VOC 2012 *val* set.

4.6 Results on the COCO Dataset

We conduct experiments on the more challenging COCO dataset, and compare with some recent methods including

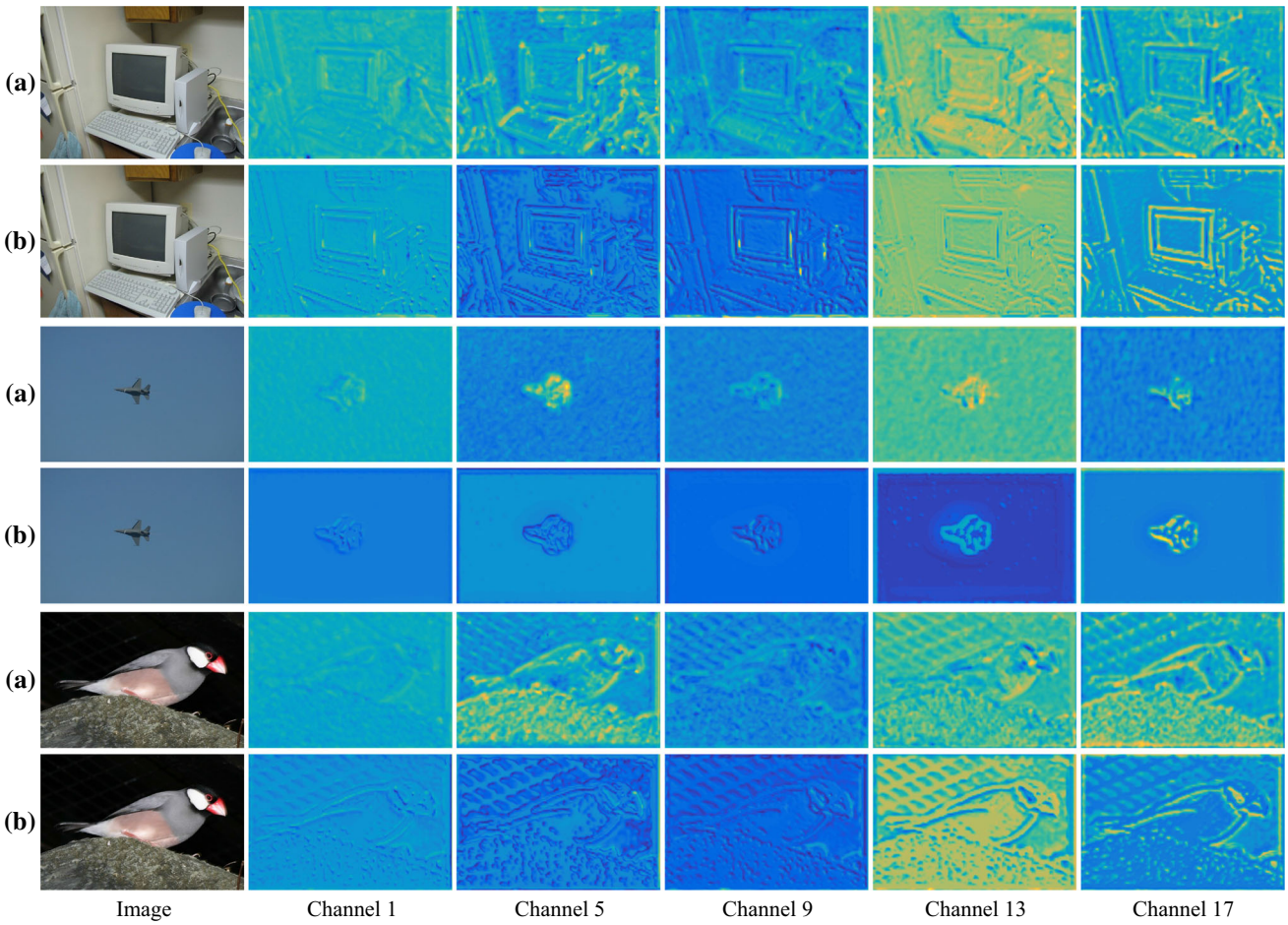


Fig. 7 Visualization of the learned affinity without and with region smoothness constraint when training the pairwise affinity network. For each image, the first row **a** show results without the region smoothness constraint, and the second row **b** is the results with the region smoothness constraint. With the region smoothness constraint, the learned affinity values inside object regions are smoother with more clear object boundaries. Best viewed in color (Color figure online)

Table 9 Evaluation results on the COCO dataset

Methods	mIoU
SEC (ECCV'16) (Kolesnikov and Lampert 2016)	22.4
BFBP (ECCV'16) (Saleh et al. 2016)	20.4
DSRG (CVPR'18) (Huang et al. 2018)	26.0
Ours	**27.7**

Better result is given in bold

SEC (Kolesnikov and Lampert 2016), BFBP (Saleh et al. 2016), and DSRG (Huang et al. 2018). Table 9 shows the results on the *val* set, where all methods use the VGG16 network as the backbone model. The proposed algorithm achieves 27.7% on mIoU and performs favorably against the state-of-the-art methods.

5 Conclusions

In this paper, we propose a weakly-supervised semantic segmentation algorithm using an iterative affinity learning framework. Starting from the coarse annotations from the class activation maps, we exploit data redundancies in natural images to learn pixel affinities and propagate labels iteratively. Our framework consists of a unary segmentation network to predict the class probability map, and a pairwise affinity network to learn affinity and refine the results of the unary network. We propose to mine confident regions for learning the reliable affinity. The refined results are then considered as supervisory signals to retrain the unary network. The procedures are conducted iteratively to learn more robust affinity and generate better segmentation progressively. Experimental results on both the PASCAL VOC 2012 and COCO datasets demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods.

Acknowledgements This work is supported by National Key Basic Research Program of China (No. 2016YFB0100900), Beijing Science and Technology Planning Project (No. Z191100007419001), National Natural Science Foundation of China (No. 61773231), and National Science Foundation (CAREER No. 1149783).

References

Ahn, J., & Kwak, S. (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4981–4990).

Bearman, A., Russakovsky, O., Ferrari, V., & Fei-Fei, L. (2016). What's the point: Semantic segmentation with point supervision. In *Proceedings of European conference on computer vision (ECCV)* (pp. 549–565).

Bertasius, G., Torresani, L., Stella, X. Y., & Shi, J. (2017). Convolutional random walk networks for semantic image segmentation. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 858–866).

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *40*(4), 834–848.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587

Dai, J., He, K., & Sun, J. (2015) Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of IEEE international conference on computer vision (ICCV)* (pp. 1635–1643).

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, *88*(2), 303–338.

Fan, R., Cheng, M. M., Hou, Q., Mu, T. J., Wang, J., & Hu, S. M. (2019). S4net: Single stage salient-instance segmentation. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 6103–6112).

Fan, R., Hou, Q., Cheng, M. M., Yu, G., Martin, R. R., & Hu, S. M. (2018). Associating inter-image salient instances for weakly supervised semantic segmentation. In *Proceedings of European conference on computer vision (ECCV)* (pp. 367–383).

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, *59*(2), 167–181.

Hagen, L., & Kahng, A. B. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *11*, 1074–1085.

Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., & Malik, J. (2011). Semantic contours from inverse detectors. In *Proceedings of IEEE international conference on computer vision (ICCV)* (pp. 991–998).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Huang, Z., Wang, X., Wang, J., Liu, W., & Wang, J. (2018). Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7014–7023).

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of ACM international conference on Multimedia (ACM MM)* (pp. 675–678).

Kersten, D. (1987). Predictability and redundancy of natural images. *JOSA A*, *4*(12), 2395–2400.

Khoreva, A., Benenson, R., Hosang, J., Hein, M., & Schiele, B. (2017). Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 876–885).

Kolesnikov, A., & Lampert, C. H. (2016). Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proceedings of European conference on computer vision (ECCV)* (pp. 695–711).

Levin, A., Lischinski, D., & Weiss, Y. (2008). A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *30*, 228–242.

Lin, D., Dai, J., Jia, J., He, K., & Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3159–3167).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Proceedings of European conference on computer vision (ECCV)* (pp. 740–755).

Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M. H., & Kautz, J. (2017). Learning affinity via spatial propagation networks. In *Proceedings of annual conference on neural information processing systems (NeurIPS)* (pp. 1520–1530).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3431–3440).

Maire, M., Narihira, T., & Yu, S. X. (2016). Affinity CNN: Learning pixel-centric pairwise relations for figure/ground embedding. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 174–182).

Papandreou, G., Chen, L. C., Murphy, K. P., & Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of IEEE international conference on computer vision (ICCV)* (pp. 1742–1750).

Pathak, D., Krahenbuhl, P., & Darrell, T. (2015). Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of IEEE international conference on computer vision (ICCV)* (pp. 1796–1804).

Pathak, D., Shelhamer, E., Long, J., & Darrell, T. (2014). Fully convolutional multi-class multiple instance learning. arXiv preprint arXiv:1412.7144.

Pinheiro, P. O., & Collobert, R. (2015). From image-level to pixel-level labeling with convolutional networks. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1713–1721).

Qi, X., Liu, Z., Shi, J., Zhao, H., & Jia, J. (2016). Augmented feedback in semantic segmentation under image level supervision. In *Proceedings of European conference on computer vision (ECCV)* (pp. 90–105).

Roy, A., & Todorovic, S. (2017). Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3529–3538).

Saleh, F., Aliakbarian, M. S., Salzmann, M., Petersson, L., Gould, S., & Alvarez, J. M. (2016). Built-in foreground/background prior for weakly-supervised semantic segmentation. In *Proceedings of European Conference on Computer Vision (ECCV)* (pp. 413–432).

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *22*(8), 888–905.

Shimoda, W., & Yanai, K. (2016). Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *Proceedings of European conference on computer vision (ECCV)* (pp. 218–234).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Wang, X., Ma, H., Chen, X., & You, S. (2018a). Edge preserving and multi-scale contextual neural network for salient object detection. *IEEE Transactions on Image Processing (TIP)*, *27*(1), 121–134.

Wang, X., You, S., Li, X., & Ma, H. (2018b). Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1354–1362).

Wei, Y. C., Cheng, C. K., et al. (1989) Towards efficient hierarchical designs by ratio cut partitioning. In *IEEE international conference on computer-aided design* (pp. 298–301).

Wei, Y., Feng, J., Liang, X., Cheng, M. M., Zhao, Y., & Yan, S. (2017a). Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1568–1576).

Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M. M., Feng, J., et al. (2017b). STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *39*(11), 2314–2320.

Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., & Huang, T. S. (2018). Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7268–7277).

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2881–2890).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2921–2929).

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., et al. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision (IJCV)*, *127*(3), 302–321.