Pattern Recognition 59 (2016) 234–244

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning to segment with image-level annotations

Yunchao Wei [a,b,d], Xiaodan Liang [c,d], Yunpeng Chen [d], Zequn Jie [d], Yanhui Xiao [e], Yao Zhao [a,b,*], Shuicheng Yan [d]

[a] Institute of information Science, Beijing Jiaotong University, Beijing 100044, China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China
[c] School of Advanced Computing, Sun Yat-Sen University, Guangzhou 510006, China
[d] Vision and Machine Learning Lab, National University of Singapore, Singapore 117583, Singapore
[e] People's Public Security University of China, Beijing 100038, China

ARTICLE INFO

Article history:
Received 19 July 2015
Received in revised form
17 January 2016
Accepted 18 January 2016
Available online 27 January 2016

Keywords:
Semantic segmentation
Weakly supervised
Deep learning

ABSTRACT

Recently, deep convolutional neural networks (DCNNs) have significantly promoted the development of semantic image segmentation. However, previous works on learning the segmentation network often rely on a large number of ground-truths with pixel-level annotations, which usually require considerable human effort. In this paper, we explore a more challenging problem by learning to segment under image-level annotations. Specifically, our framework consists of two components. First, reliable hypotheses based localization maps are generated by incorporating the hypotheses-aware classification and cross-image contextual refinement. Second, the segmentation network can be trained in a supervised manner by these generated localization maps. We explore two network training strategies for achieving good segmentation performance. For the first strategy, a novel multi-label cross-entropy loss is proposed to train the network by directly using multiple localization maps for all classes, where each pixel contributes to each class with different weights. For the second strategy, the rough segmentation mask can be inferred from the localization maps, and then the network is optimized based on the single-label cross-entropy loss with the produced masks. We evaluate our methods on the PASCAL VOC 2012 segmentation benchmark. Extensive experimental results demonstrate the effectiveness of the proposed methods compared with the state-of-the-arts.

1. Introduction

During the past few years, many compositional and hierarchical models [1–6] have been proposed to address computer vision issues. For instance, Felzenszwalb et al. [1,3] propose deformable part models to address the object detection task. In [5], a hierarchical Markov Random Field model is proposed for the human action segmentation task. In addition, Lin et al. [4] present an novel And-Or graph model for the object shape detection task. All these previous works have made tremendous contributions for the computer vision community. In this paper, we focus on a more challenging computer vision task, called semantic image segmentation, which aims to assign a semantic label to each pixel from a pre-defined class set. Recently, tremendous advances in semantic segmentation [7–16] have been made by taking advantage of the powerful recognition ability of deep convolutional neural networks (DCNNs) [17–20]. These methods usually pre-train DCNNs with a large-scale image classification dataset [21], and then transfer the pre-trained parameters to the segmentation task. However, these methods need a large number of pixel-level annotated data for training. The burden of annotation collection for pixel-wise segmentation masks is very heavy, which requires considerable financial expenses as well as human efforts.

To alleviate the demand for the expensive pixel-level annotated images, some weakly supervised approaches [22–28] have been proposed to solve semantic image segmentation. Among them, some methods [22,23] make use of annotated bounding boxes to train the network for semantic segmentation. Although bounding box annotations are much easier to obtain compared with pixel-level annotations, it still requires considerable human effort. To further reduce the reliance on these costly annotations as supervision, e.g., pixel-level annotated masks or labeled bounding boxes, some multiple instance learning methods [24,25] and Expectation-Maximization (EM) methods [23] adopt a more challenging setting where only image-level labels are used as the supervision, for pixel-level prediction. These previous works on image-level annotation based segmentation only utilize the single image information to train the DCNN model. However, due to high

* Corresponding author.
E-mail address: yzhao@bjtu.edu.cn (Y. Zhao).

http://dx.doi.org/10.1016/j.patcog.2016.01.015
0031-3203/

intra-class variation (e.g. diverse appearance, viewpoints, scales) within the objects, it may be difficult to learn a good DCNN by only relying on the single image cues. We argue that the cross-image contextual information can better help infer more reasonable object proposals or masks and effectively reduce the possible noisy labels by incorporating more contextual relations.

In this paper, we propose a novel weakly supervised framework for semantic segmentation under image-level annotations. Two components are included in our framework as illustrated in Fig. 1. First, given the image-level annotation(s) of each image, a hypothesis-based localization map for each class can be generated by incorporating the hypotheses information and cross-image contextual cues. Specifically, for each image, we first extract the class-interdependent object proposal and then predict the classification scores for each proposal belonging to a class based on the state-of-the-art Hypothesis-CNN-Pooling (HCP) [30] method. The cross-image contextual refinement is then performed to select more reliable proposals with high predictive scores. The localization maps for each class can thus be generated by combining all selected proposals for each class. Second, two network training strategies are explored to train the segmentation network based on the generated localization maps for each class. For the first strategy, a novel multi-label cross-entropy loss is introduced for network training by directly using multiple localization maps of all classes. In this way, each pixel in the image can adaptively contribute to each class with different weights, which are naturally embedded in each localization map. For the second strategy, the rough mask for each image can be inferred by combining all the localization maps for all classes, and then the single-label cross-entropy loss for each pixel is used to optimize the network based on the generated mask.

The main contributions of this work are summarized as follows:

- Our novel framework investigates how to use the image-level annotations and cross-image contextual cues to learn a good segmentation network. The hypothesis-based localization map generation is proposed by incorporating the hypothesis-based classification and cross-image refinement.
- We propose a novel multi-label cross-entropy loss function to train the network based on multiple localization maps. Each pixel adaptively contributes to each class according to the predicted weights embedded in the localization map.
- Based on the generated localization maps, we propose a simple but effective method to predict the rough mask of the given training image, and thus the single-label cross-entropy loss for each pixel can be used to optimize the segmentation network.
- We evaluate the methods on the PASCAL VOC 2012 segmentation benchmark [31]. Our weakly supervised methods achieve new state-of-the-art results compared with previous methods under the same supervised setting.

The rest of the paper is organized as follows. We briefly review the related work of semantic image segmentation in Section 2. Section 3 presents the details of the proposed methods for weakly supervised segmentation. Finally the experimental results and conclusions are provided in Sections 4 and 5, respectively.

2. Related work

2.1. Segmentation with pixel-level annotations

Most recently, great progress has been made in image semantic segmentation with the development of deep convolutional neural networks (DCNNs). Most existing CNN-based semantic segmentation methods, such as CFM [8], FCN [10], DeepLab-CRF [7], rely on pixel-level annotations as the supervision for training. Specifically, Dai et al. [8] proposed to exploit shape information through

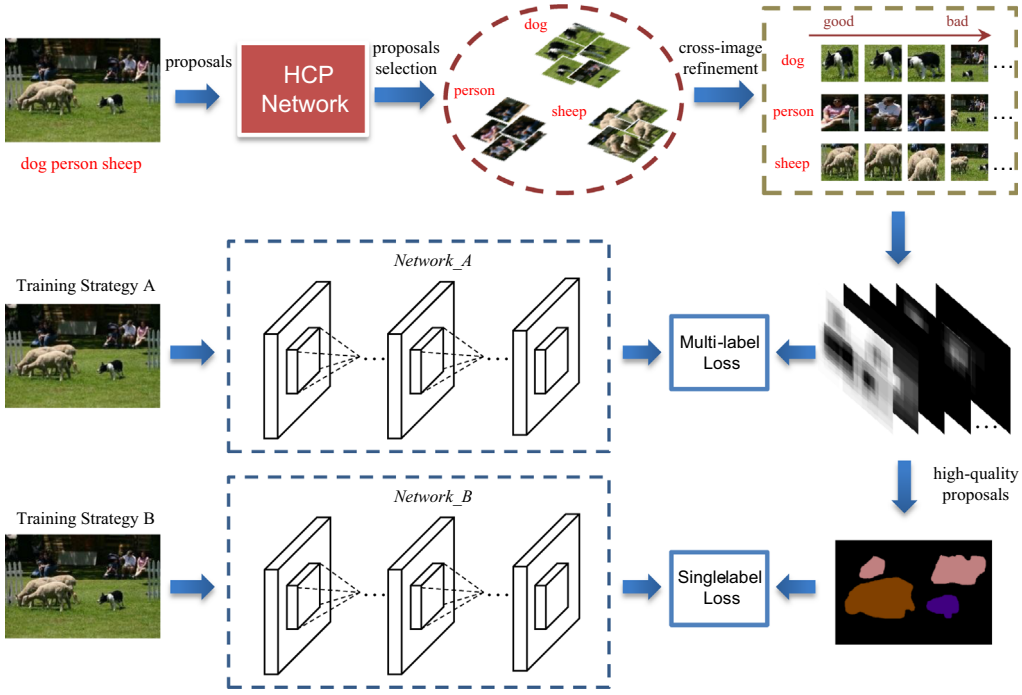

Fig. 1. Overview of the proposed framework. Give an image with image-level labels, we utilize MCG [29] to generate proposals as the inputs of the pre-trained HCP classification network [30]. The proposals with high predictive scores of the ground-truth labels are selected for cross-image refinement. We use the refined proposals to generate a localization map for each class. Furthermore, the rough segmentation mask is generated according to the localization maps and proposals with high predictive scores of the image-level labels. The segmentation *Network_A* is trained based on localization maps with multi-label loss and the segmentation *Network_B* is trained based on the generated mask with single-label loss on each pixel.

convolutional feature masking to train classifiers for segmentation. Long et al. [10] proposed to build a fully convolutional network which takes the image of an arbitrary size as the input and produces the segmentation result of a corresponding size with efficient inference and learning. Based on a fully convolutional network, Chen et al. [7] proposed to refine the pixel-wise prediction from the last DCNN layer with a fully connected Conditional Random Field (CRF) and achieved better segmentation results. However, the annotation collection for pixel-level segmentation masks usually requires much money as well as human effort.

2.2. Segmentation with bounding box annotations

Some existing segmentation methods [32–35,22,23,27] use bounding box annotations instead of pixel-level annotations. Xia et al. [32] introduced a voting scheme to estimate shape guidance for each bounding box, and then the derived shape guidance was used in the subsequent graph-cut-based segmentation. Dai et al. [22] and Papandreou et al. [23] estimated segmentation masks by extracting region proposals on the annotated bounding boxes. Xu et al. [27] proposed a unified approach to incorporate various forms of weak supervision information for semantic segmentation. Although bounding box annotations are much easier to obtain compared with pixel-level annotations, it still requires considerable human effort.

2.3. Segmentation with image-level labels

A more challenging setting of semantic image segmentation is to train the segmentation network with only image-level labels, which has attracted much interest in the literature. Some recent works adopt Multiple Instance Learning (MIL) methods based on DCNN architectures for the weakly supervised learning with image-level annotations. Specifically, Pinheiro et al. [24] proposed a MIL framework for the DCNN training and utilized smoothing prior to refine the predicted results. Besides, Papandreou et al. [23] presented an alternative training procedure based on the Expectation-Maximization (EM) algorithm for DCNN training supervised by image-level labels. Most recently, Pathak et al. [36] introduced the constrained convolutional neural network for weakly supervised segmentation. Although [37–39,26,27,40] have obtained promising results on some simple datasets, they have not demonstrated the performance on the challenging PASCAL VOC benchmark.

Compared with these previous image-level supervised works, our approach has some unique characteristics in the following aspects. First, the hypothesis-based localization map generation is proposed to incorporate the hypothesis-based classification and cross-image contextual cues to generate reliable maps for all classes. Second, different from previous segmentation training strategies, where each pixel is assigned to one class, we propose a multi-label cross-entropy loss function that each pixel may be assigned to multiple classes with different weights embedded in the generated localization maps. Finally, relying on the learned localization information for each class of a given image, we train the segmentation network with the produced rough masks or multiple localization maps for all classes as the supervision.

3. Proposed methods

3.1. Training the hypothesis-CNN-pooling classification network

In [30], a flexible deep network called Hypothesis-CNN-Pooling (HCP) is proposed to address the multi-label classification problem. As can be seen from Fig. 2, HCP is a proposal based method that takes an arbitrary number of object hypotheses (proposals) as the input. Then, a shared CNN is connected with each hypothesis to make a prediction. Finally, to aggregate the single-label CNN predictions from different hypotheses into multi-label results, a cross-hypothesis max-pooling layer is integrated into the shared CNN model for the ultimate multi-label predictions.

We choose HCP as the basis network to predict the category of each proposal for the following two reasons. Firstly, the training process of HCP network is based on proposals rather than images, which gives HCP network a better discriminative ability for proposals compared with those networks directly fine-tuned with whole images. Secondly, although HCP is trained based on proposals, no ground-truth bounding box information is required for training. Therefore, using HCP to predict the category of the given proposal is not in conflict with our segmentation setting, where only image-level labels are utilized in training.

For HCP training, we follow the steps as detailed in [30]. To efficiently train the HCP network, following [30], only 10 specific proposals are selected. During the prediction stage, we take all proposals of a training image as the inputs. We adopt a state-of-the-art region proposal method, i.e., Multiscale Combinatorial Grouping (MCG) [29], to generate about 2000 proposals per image for prediction. As illustrated in Fig. 1, for a training image, MCG is firstly utilized to generate proposals as the inputs. Then, we use the pre-trained HCP network to predict the category of each proposal. Finally, proposals with high predictive confidences on the ground-truth labels are then taken as the object candidates.

We utilize the scores after the softmax layer as the predictive confidences for each proposal. For each ground-truth label of a given image, we rank the proposals based on the confidence values of this label in a descending order, and those proposals with

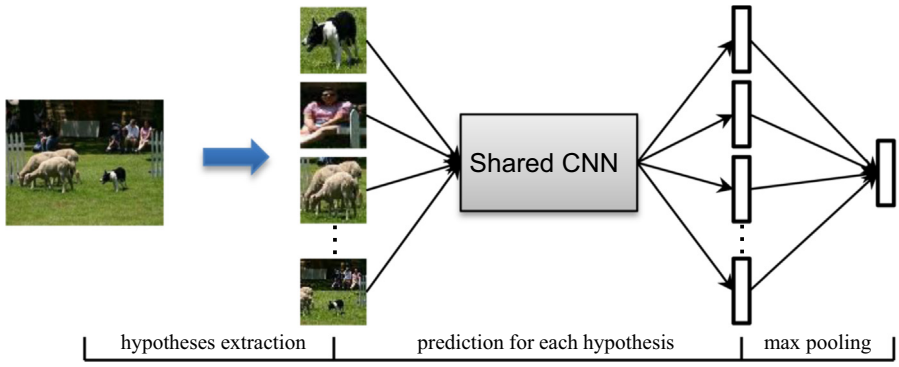

Fig. 2. Brief illustration of the HCP framework. For a given multi-label image, we feed some hypotheses into the shared CNN and fuse the outputs into the final prediction with cross-hypothesis max-pooling operation. The shared CNN is firstly pre-trained on the ImageNet [21], and then fine-tuned with images and hypotheses from multi-label dataset.

the confidences larger than 0.5 are considered as the object candidates.

3.2. Cross-image contextual refinement

Based on the selected proposals of a given image, we choose those in which object instances are tightly included. In most cases, the proposals selected by the HCP network contain the object of the specific class. However, directly using these proposals to generate the localization map for the specific class may lead to unsatisfactory results. As shown in Fig. 3, the proposals selected by the HCP network may have the following two problems: some proposals only contain part of the target object(s) or some proposals contain a lot of background pixels. Both problems have negative impacts on localization map generation. We consider the assumption that the feature representations between the proposals in different images (belonging to the same class), in which the objects are tightly included, share some similar characteristics. In contrast, the feature representations of the proposals from different images, which are part of target objects or contain many background pixels, differ from each other. Therefore, to address the two problems, we propose to utilize cross-image information inspired by Multiple Instance Learning (MIL) [41] to refine the selected proposals for localization map generation.

Denote the number of classes as c in the training image dataset. For the k-th ($k=1,2,\ldots,c$) class, there are N^k training images. We denote n_i^k as the number of the selected proposals by the HCP network for the i-th ($i=1,2,\ldots,N^k$) image on the k-th class, and denote $\boldsymbol{x}_{ij}^k$ as the feature vector of the j-th ($j=1,2,\ldots,n_i^k$) proposal. Then, the average distance from the j-th proposal in the i-th image to the proposals in other images can be defined as follows,

$$dis(\boldsymbol{x}_{ij}^k)=\frac{1}{N^k-1}\sum_{i'\neq i}\left(\frac{1}{n_{i'}^k}\sum_{j'=1}^{n_{i'}^k}\|\boldsymbol{x}_{ij}^k-\boldsymbol{x}_{i'j'}^k\|^2\right). \quad (1)$$

We re-rank the proposals of the i-th image following the calculated average distance in a descending order. After the re-ranking, the top half of proposals are selected for the localization map generation. It can be observed from Fig. 3 that those proposals which tightly include an object are ranked on the top.

For the feature representation of each proposal, we firstly extract its 4,096-dimensional CNN feature from the second last fully connected layer via the HCP network, and then perform principal component analysis (PCA) to reserve 98% energy of the original feature, which can reduce the dimensionality from 4096 to 512 to speed up the operation.

3.3. Segmentation network training

We explore two training strategies for segmentation network training. (1) Based on the generated localization map for each class, we treat the multiple localization maps as the supervision to train the segmentation network. (2) Based on the generated localization map, we infer a rough pixel-wise segmentation mask by taking both selected proposals and their confidences for each ground-truth label from HCP into account for the segmentation network training.

3.3.1. Training the network based on class-wise saliency maps

For semantic segmentation, each pixel is classified into $c+1$ classes (c *object* classes and one *background* class). We denote Ω as the label set of the given training image. Denote l_k as the localization map of the k-th ($k\in\{1,2,\ldots,c\}$) class. The score of each pixel $p^k(i,j)(1\le i\le h, 1\le j\le w)$ in the localization map is

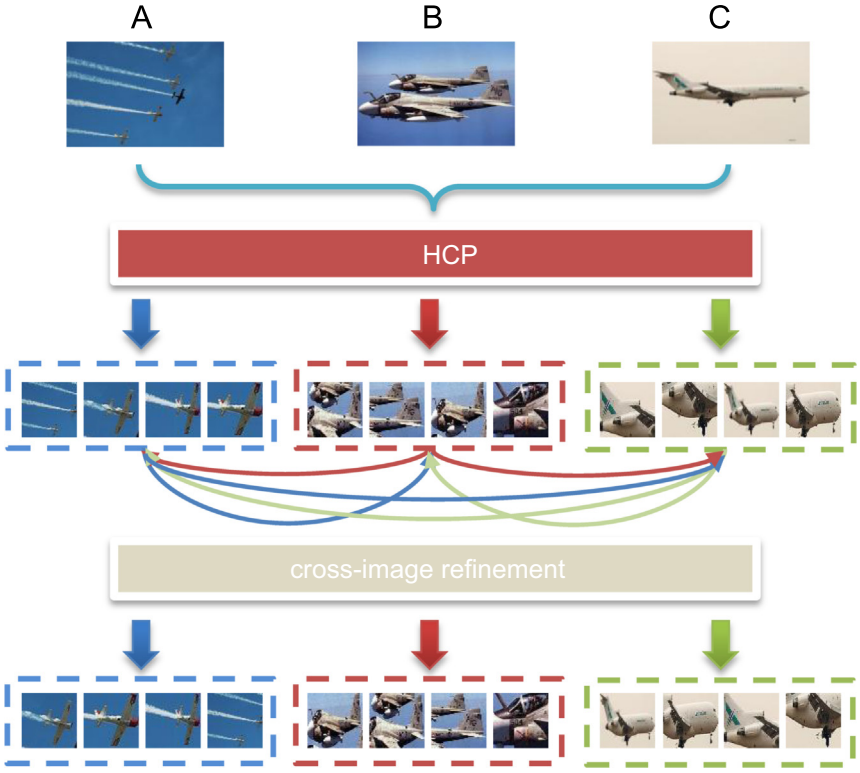

Fig. 3. Illustration of cross-image refinement. The refinement is performed among images that share the same image-level label(s), e.g., *airplane*. We calculate the average distance between each selected proposal from the set A and other selected proposals from images which have the same image-level label with A, e.g., B and C. Then, we rank the selected proposals in a descending order according to the calculated distances.

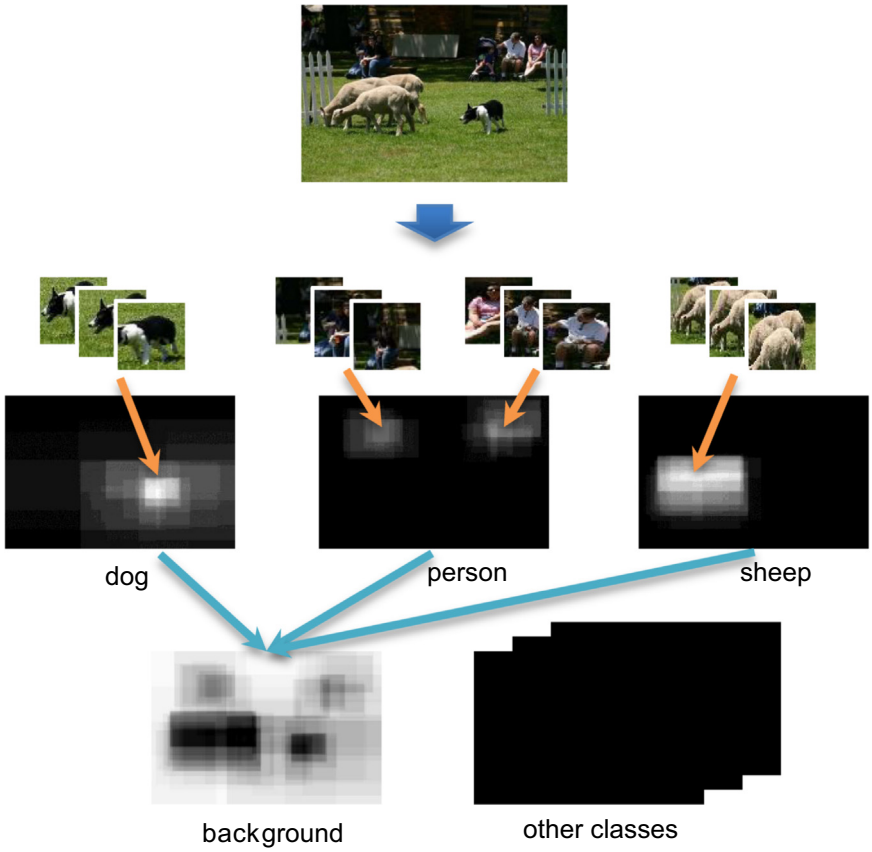

Fig. 4. Illustration of localization map generation for each class. The localization map of each class (excluding *background*) is obtained by adding the refined proposals on the *black* map at their locations. Then, the localization map of *background* can be obtained by excluding the localization regions of the ground truth classes. The localization maps of other classes are considered as *black* maps.

Fig. 5. Illustration of the generated localization maps and the rough masks for training samples. We fuse multiple localization maps of multi-label images into one localization map for brief. Some difficult training samples (with heavy occlusion or small objects) are shown in the last row.

initialized as 0 (i.e., *black* maps), where h and w are the height and the width of the given training image, respectively. For each $k \in \Omega$, each $p^k(i,j)$ in l_k is assigned by summing up all refined proposals of the k-th class at this location and then dividing the number of the refined proposals. The higher the probability is, the more likely this pixel is to belong to the k-th class. Based on the obtained localization maps of *object* classes, we compute the localization map of *background* with the following criterion:

$$p^{bg}(i,j) = \max\left(0, 1 - \sum_{k \in \Omega} p^k(i,j)\right), \tag{2}$$

where $p^{bg}(i,j)$ is the pixel value of the *background* localization map and $p^k(i,j)$ refers to the pixel value of l_k. Fig. 4 illustrates the generated localization maps of the given image with image-level labels.

Under the weakly supervised scheme, it is very challenging to obtain accurate pixel-level labels only with the image-level supervision. Therefore, based on the localization maps, we propose to train the segmentation network with a soft label. Specifically, each pixel is assigned to several candidate classes with different probabilities. We adopt the DeepLab-CRF [7] as the basic structure due to its competitive accuracy and efficiency. Denote $A = a_t \mid t = 1, \ldots, N$ as the image training set. We denote the segmentation network filtering by $f(\cdot)$, where all the layers filter the given image a_t. The $f(\cdot)$ produces a $\hat{h} \times \hat{w} \times (c+1)$ dimensional output of activations, where $\hat{h}$ and $\hat{w}$ are the height and the width of the feature map for each channel. The softmax function is used to compute the posterior probability of each pixel belonging to the k-th ($k \in \{1, 2, \ldots, c+1\}$) class of the given image a_t, i.e.,

$$p_t^k(i,j) = \frac{\exp\left(f_{(i,j)}^k(a_t)\right)}{\sum_{m=1}^{c+1} \exp\left(f_{(i,j)}^m(a_t)\right)}, \tag{3}$$

where $f_{(i,j)}^k(a_t)$ is the activation value of the image a_t at the location (i,j) $(1 \le i \le \hat{h}, 1 \le j \le \hat{w})$ for the k-th class. Denote the ground-truth probability for the k-th class of the image a_t at the location (i,j) as $\hat{p}_t^k(i,j)$, which is obtained from the generated localization map and normalized with cross-channel information. Given the network prediction in Eq. (3), the loss function is then defined as

$$J = -\eta \sum_{t=1}^{N} \sum_{i=1}^{\hat{h}} \sum_{j=1}^{\hat{w}} \sum_{m=1}^{c+1} \hat{p}_t^m(i,j) \log\left(p_t^m(i,j)\right), \tag{4}$$

where η is the weight parameter, which is set as $1/(N \times \hat{h} \times \hat{w})$ in this paper. Based on this loss function, the network parameter is expected to be learned through those high confident pixels, so that the segmentation mask of a new image could be inferred.

3.3.2. Training the network based on rough mask

In this section, we present a simple but effective method to predict rough masks for segmentation network training. We utilize the generated localization maps in Section 3.3.1 to roughly locate the object of interest. To obtain the rough segmentation mask for each *object* class, we adopt the region proposals generated by MCG. In addition, the predicted score of each region proposal from the HCP network is also employed to refine proposals so that they would not be small regions with high confidence. Denote $\boldsymbol{r}$ and $s_{\boldsymbol{r}}^k \in [0,1]$ $(k \in \{1, 2, \ldots, c\})$ as the candidate region and the predicted score of the k-th class from the HCP network. Then, the confidence $v_{\boldsymbol{r}}$ that the candidate proposal is selected into the segmentation mask can be calculated as follows:

$$v_{\boldsymbol{r}} = s_{\boldsymbol{r}}^k + \frac{1}{|\boldsymbol{r}|} \sum_{(i,j) \in \boldsymbol{r}} p^k(i,j), \tag{5}$$

where $|\boldsymbol{r}|$ denotes the number of pixels within the region $\boldsymbol{r}$ and $p^k(i,j)$ is the value of the pixel (i,j) from the k-th localization map. We rank the candidate regions in a descending order of the corresponding confidences and combine the top 10 candidate regions as the segmentation mask of the k-th class. Specifically, for a training image with multiple labels, if there is an overlap between any two segmentation regions of different classes, the category of each pixel in the overlap region is decided by the pixel values on localization maps of the corresponding classes. The regions, which are not selected by any ground-truth class, are considered as *background*.

By exploring the information from both global (the first term) and local (the second term) points of view, high quality regions belonging to a specific category can be more reasonably selected for training. Based on the generated rough masks, we adopt the DeepLab-CRF [7] method to train the network for semantic segmentation.

4. Experimental results

4.1. Dataset

The proposed weakly supervised methods are evaluated on the PASCAL VOC 2012 segmentation benchmark [31]. The performance is measured in terms of pixel intersection-over-union (IoU) averaged on 21 classes (20 *object* classes and one *background* class). The segmentation part of the original PASCAL VOC 2012 dataset contains 1464 *train*, 1449 *val* and 1456 *test* images, respectively. Hariharan et al. [42] provided extra annotated images with the number of 10,582 (train_aug) for training. In our experiment, the training process is implemented based on the 10,269 images, which is an intersection set between *trainval* of the image classification task and train_aug. Extensive evaluations of the proposed methods are primarily conducted on the PASCAL VOC 2012 *val* set and we also report the performance on the *test* set (in which the ground truth masks are not released) by submitting the results to the official PASCAL VOC 2012 server.

Table 1
Comparison of different training settings in terms of IoU (%) on PASCAL VOC 2012.

Categories	Results on *val* Set				Results on *test* Set			
	AN_A	SN_A	AN_B	SN_B	AN_A	SN_A	AN_B	SN_B
bkg	69.4	74.6	**81.0**	80.7	70.4	74.9	**82.4**	82.1
plane	37.6	40.8	**56.6**	54.6	38.0	42.4	**54.1**	53.6
bike	15.6	**17.3**	7.1	10.7	16.6	**17.4**	7.8	12.4
bird	26.6	32.0	**56.9**	55.6	27.8	30.0	**58.2**	53.5
boat	26.4	28.5	**39.2**	37.5	23.6	24.8	**31.1**	29.5
bottle	35.5	41.0	39.5	**51.8**	36.1	36.9	39.0	**41.6**
bus	**63.9**	58.0	41.7	46.3	**63.8**	59.6	36.9	46.9
car	**52.9**	51.0	29.8	42.6	**51.3**	49.6	30.0	46.3
cat	54.1	**54.8**	44.7	48.0	48.3	49.4	48.5	**50.3**
chair	14.1	16.0	**16.5**	16.0	15.0	15.7	**17.2**	16.8
cow	43.9	**47.9**	47.3	46.3	44.0	45.8	48.6	**48.7**
table	**31.3**	19.5	18.2	10.0	**42.7**	27.4	14.2	17.2
dog	47.9	51.7	**54.9**	54.6	52.4	53.0	59.0	**60.6**
horse	39.7	43.5	**48.3**	45.9	44.9	45.8	50.0	**51.8**
mbike	45.0	**50.0**	46.6	47.5	54.3	58.5	57.4	**61.7**
person	**48.3**	41.6	34.2	34.4	**47.9**	41.6	39.1	36.4
plant	**27.9**	25.2	23.1	24.5	**31.4**	29.8	29.4	25.2
sheep	**54.4**	53.5	53.0	53.7	55.9	**58.4**	56.0	58.3
sofa	25.2	**25.5**	24.4	23.0	**31.0**	24.8	30.0	19.3
train	43.3	**48.3**	42.0	47.8	36.4	42.7	43.4	**48.5**
tv	28.3	30.8	47.7	**48.6**	28.4	29.6	43.9	**45.5**
mIoU	39.7	40.6	40.6	**41.9**	41.0	40.9	41.7	**43.2**

4.2. Training strategies

Both HCP network and the segmentation network are initialized by the publicly released VGG-16 model [18], which is pre-trained on the ImageNet classification dataset [21]. We replace the 1000-way ImageNet classifier in the last layer of VGG-16 with the 20-way one for classification and the 21-way one for segmentation. We optimize the objective function with respect to the parameters at all weighted layers by the standard SGD procedure of [17].

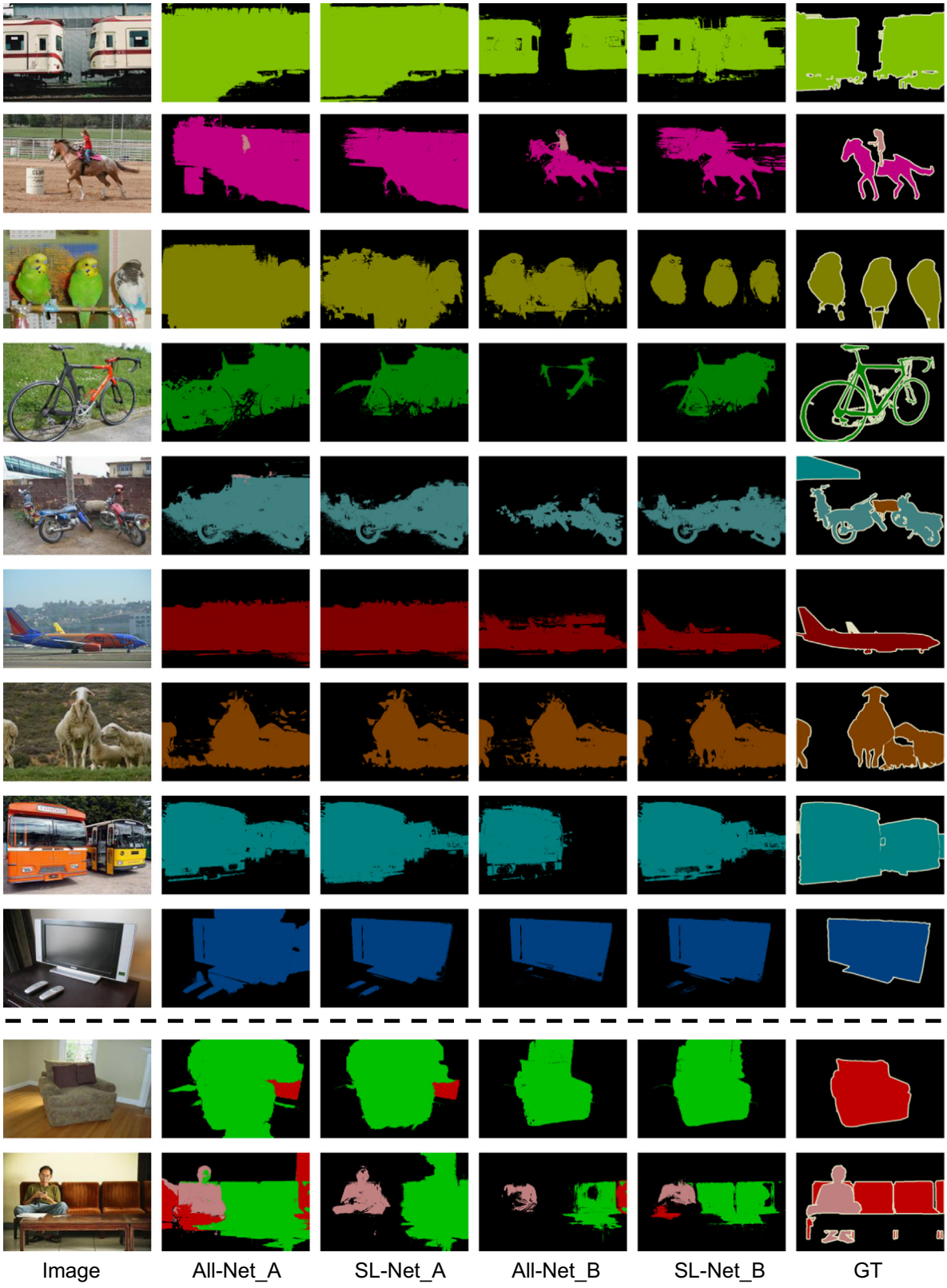

Fig. 6. Illustration of segmentation results on the PASCAL VOC 2012 *val* set with the proposed methods. GT indicates ground truth segmentation mask. Some difficult examples are shown in last two rows.

For the training of HCP network, we follow a similar setting as detailed in [30]. In the image-fine-tuning step, the initial learning rates of the last layer and other layers are set as 0.01 and 0.0001, respectively. In the hypothesis-fine-tuning step, the initial learning rates of the last layer and other layers are set as 0.001 and 0.0001, respectively. For both steps, we use a mini-batch size of 30 images and multiply the learning rate by 0.1 after every 10 epochs. We take the momentum of 0.9 and the weight decay of 0.0005. The HCP network training is performed for about 30 epochs.

For the training of the segmentation network, we use a mini-batch size of 8 images. The initial learning rate is set as 0.001 (0.01 for the final layer) and divided by 10 after every 5 epochs. The momentum and the weight decay are set as 0.9 and 0.0005. The network training is performed for about 15 epochs. We denote the segmentation networks trained with the generated localization maps and the rough masks as Network_A (N_A for short) and Network_B (N_B for short), respectively. Considering that the predicted localization region (or object segmentation) of one class may be affected by another class in the case that the given image is annotated with multiple labels, we try two kinds of training strategies, i.e., using all (10,269) training images and using single-label (6628) images. In summary, four kinds of segmentation networks implemented in this paper are listed as follows:

- AN_A: The network is trained by taking the predicted localization maps as the supervision based on all the images from the training set.
- SN_A: The network is trained by taking the predicted localization maps as the supervision based on the single-label images from the training set.
- AN_B: The network is trained by taking the rough masks as the supervision based on all the images from the training set.
- SN_B: The network is trained by taking the rough masks as the supervision based on the single-label images from the training set.

Each segmentation network takes about half of a day to train based on a NVIDIA GeForce Titan GPU with 6GB memory. All the experiments are conducted using DeepLab code [23], which is implemented based on the publicly available Caffe framework [43].

4.3. Results

4.3.1. Comparisons of different training schemes

Table 1 reports the comparison results of different training schemes, i.e., AN_A, SN_A, AN_B and SN_B.

First, it can be observed that the networks trained with rough masks perform better than that trained with localization maps on the segmentation task. The reason can be explained as follows. By using localization maps as the supervision, many background pixels tend to be assigned with some weights for foreground object(s). Therefore, some background pixels are more likely to be predicted as foreground object(s), which may decrease the IoU score. Fig. 6 shows some segmentation results of the proposed methods on the PASCAL VOC 2012 *val* set. We can see that the foreground objects predicted by N_A usually contain more background pixels compared with the results from N_B.

Second, using the single-label images for training can achieve better results compared with those that are based on all images. In Table 2, we show the number of objects on different training sets. Although the number of objects in the single-label images is less than half of the number in all images, better performance can be obtained by only utilizing the single-label images, e.g., 41.9% *vs.* 40.6% on the *val* set and 43.2% *vs.* 41.7% on the *test* set. The reason may be that the complexity of the multi-label images may have a negative effect upon network training. In some multi-label images, the different compositions and interactions between objects, like partial visibility and occlusion, may decrease the accuracy of the predicted localization maps or masks. As shown in Fig. 5, the generated localization maps and masks of the single-label images are always much better than those of the multi-label images. Specifically, many pixels of foreground objects are incorrectly predicted in the last row of Fig. 5.

In addition, the performance of several classes (e.g., *table* and *sofa*) under SN_* schemes is worse than that of AN_* schemes in some cases. The reason may be the insufficiency of training samples. If we increase the number of training samples in these classes, the performance may also be boosted.

Table 2
The number of objects on different training sets.

Training Set	All Images	Single-label Images
aero	564	479
bike	470	151
bird	677	623
boat	444	274
bottle	634	153
bus	366	129
car	1062	443
cat	979	790
chair	1028	171
cow	246	205
table	494	26
dog	1160	781
horse	424	208
mbike	462	156
person	3763	1108
plant	467	121
sheep	285	223
sofa	455	80
train	481	340
tv	522	167
Total	14,983	6628

Table 3
Justifications of cross-image contextual refinement.

Categories	AN_A		SN_A	
	w/o	w	w/o	w
bkg	66.6	69.4	72.3	**74.6**
plane	42.3	37.6	**43.4**	40.8
bike	**17.3**	15.6	14.9	17.3
bird	29.5	26.6	**32.5**	32.0
boat	**32.6**	26.4	30.6	28.5
bottle	35.5	35.5	35.3	**41.0**
bus	59.6	**63.9**	56.4	58.0
car	43.3	**52.9**	49.8	51.0
cat	46.4	54.1	50.1	**54.8**
chair	**16.2**	14.1	15.3	16.0
cow	44.5	43.9	43.6	**47.9**
table	29.8	**31.3**	15.4	19.5
dog	44.1	47.9	48.1	**51.7**
horse	34.0	39.7	36.3	**43.5**
mbike	46.9	45.0	46.2	**50.0**
person	39.5	**48.3**	47.5	41.6
plant	25.4	27.9	**30.0**	25.2
sheep	42.6	**54.4**	40.3	53.5
sofa	24.2	25.2	24.1	**25.5**
train	46.7	43.3	47.5	**48.3**
tv	24.8	28.3	**31.8**	30.8
mIoU	37.7	39.7	38.6	**40.6**

4.3.2. Justifications of cross-image contextual refinement

To validate the effectiveness of the proposed cross-image contextual refinement, we conduct the segmentation experiments without using this step. We first utilize the method as detailed in Section 3.1 to select object candidates for each training image, and class-wise localization maps as illustrated in Section 3.3.1 are then generated for the network training.

We mainly compare the results between with and without the cross-image contextual refinement based on the N_A scheme. Table 3 shows the comparison results on the *val* set. It can be observed that without the cross-image contextual refinement step, the mean IoU scores for AN_A and SN_A drop by almost 2%. The reason is that the proposals selected by HCP often contain either part of target objects or a lot of background pixels. Both cases will have negative impacts on the localization maps generation. With the refinement step, noisy object candidates can be reduced, which will be beneficial for producing more precise localization maps for training.

4.3.3. Comparison with the state-of-the-art methods

We mainly compare our method with four state-of-the-art methods, i.e., MIL-FCN [25], EM-Adapt [23], CCNN [36] and MIL-ILP-* [24]. Tables 4 and 5 report the comparison results of different weakly supervised methods on the *val* set and the *test* set of PASCAL VOC 2012, respectively. It can be observed that the proposed method is much better than most of the state-of-the-arts.

Specifically, both EM-Adapt and the proposed method are trained based on the DeepLab-CRF model. The subtle difference is that our method utilizes fewer training samples compared with EM-Adapt. Both EM-Adapt [23] and our methods try to learn the segmentation network based on the generated masks for the training samples. For EM-Adapt, the generated mask of each training image dynamically changes during the training process and no other information is utilized to refine the evaluated mask. In contrast, based on the classification confidences of proposals, we generate the fixed rough masks as supervision by exploring the cross-image relationship to train the segmentation network. From the experimental results, we can see that our method is more effective, which can achieve 3.7% and 4.2% improvements on *val* set and *test* set, respectively. For MIL-ILP-* [24], the weakly supervised segmentation network is trained with 760,000 images of 21 classes, whose number is much larger than ours (10,269), taken from ILSVRC 2013. In addition, image-level classification prior (ILP), and many complex smooth priors, i.e., superpixels (-suppxl), BING boxes and MCG segments (-seg), are utilized for post-processing to further boost the segmentation results. From Table 4, we can note that our method is much better than ILP, ILP-suppxl and ILP-bb, and achieves similar performance to ILP-seg, i.e., 41.9% *vs.* 42.0%, on the *val* set. However, from Table 5, it can be noticed that our method can outperform ILP-seg by 2.6% on the *test* set.

Qualitative segmentation results from the proposed methods are shown in Fig. 6. Two failure cases are shown in the last two rows of Fig. 6. We can see that many pixels belonging to *chair* are predicted as *sofa*, which has a similar appearance to *chair*. Some

Table 5
Comparison of the state-of-the-art methods in terms of IoU (%) on PASCAL VOC 2012 *test* set.

Methods	EM-Adapt	CCNN	ILP-sppxl	ILP-bb	ILP-seg	SN_B
bkg	76.3	–	74.7	76.2	78.7	**82.1**
plane	37.1	21.3	38.8	42.8	48.0	**53.6**
bike	**21.9**	17.7	19.8	20.9	21.2	12.4
bird	41.6	22.8	27.5	29.6	31.1	**53.5**
boat	26.1	17.9	21.7	25.9	28.4	**29.5**
bottle	38.5	38.3	32.8	38.5	35.1	**41.6**
bus	50.8	51.3	40.0	40.6	**51.4**	46.9
car	44.9	43.9	50.1	51.7	**55.5**	46.3
cat	48.9	51.4	47.1	49.0	**52.8**	50.3
chair	16.7	15.6	7.2	9.1	7.8	**16.8**
cow	40.8	38.4	44.8	43.5	**56.2**	48.7
table	**29.4**	17.4	15.8	16.2	19.9	17.2
dog	47.1	46.5	49.4	50.1	53.8	**60.6**
horse	45.8	38.6	47.3	46.0	50.3	**51.8**
mbike	54.8	53.3	36.6	35.8	40.0	**61.7**
person	28.2	**40.6**	36.4	38.0	38.6	36.4
plant	30.0	**34.3**	24.3	22.1	27.8	25.2
sheep	44.0	36.8	44.5	44.5	51.8	**58.3**
sofa	**29.2**	20.1	21.0	22.4	24.7	19.3
train	34.3	32.9	31.5	30.8	33.3	**48.5**
tv	46.0	38.0	41.3	43.0	**46.3**	45.5
mIoU	39.6	35.5	35.8	37.0	40.6	**43.2**

Table 4
Comparison of the state-of-the-art methods in terms of IoU (%) on PASCAL VOC 2012 *val* set.

Methods	MIL-FCN	EM-Adapt	ILP	ILP-sppxl	ILP-bb	ILP-seg	CCNN	SN_B
bkg	–	–	73.2	77.2	78.6	79.6	65.9	**80.7**
plane	–	–	25.4	37.3	46.9	50.2	23.8	**54.6**
bike	–	–	18.2	18.4	18.6	**21.6**	17.6	10.7
bird	–	–	22.7	25.4	27.9	40.6	22.8	**55.6**
boat	–	–	21.5	28.2	30.7	34.9	19.4	**37.5**
bottle	–	–	28.6	31.9	38.4	40.5	36.2	**51.8**
bus	–	–	39.5	41.6	44.0	45.9	**47.3**	46.3
car	–	–	44.7	48.1	49.6	**51.5**	46.9	42.6
cat	–	–	46.6	50.7	49.8	**60.6**	47.0	48.0
chair	–	–	11.9	12.7	11.6	12.6	**16.3**	16.0
cow	–	–	40.4	45.7	44.7	**51.2**	36.1	46.3
table	–	–	11.8	14.6	14.6	11.6	**22.2**	10.0
dog	–	–	45.6	50.9	50.4	**56.8**	43.2	54.6
horse	–	–	40.1	44.1	44.7	**52.9**	33.7	45.9
mbike	–	–	35.5	39.2	40.8	44.8	44.9	**47.5**
person	–	–	35.2	37.9	38.5	**42.7**	39.8	34.4
plant		–	20.8	28.3	26.0	**31.2**	29.9	24.5
sheep	–	–	41.7	44.0	45.0	**55.4**	33.4	53.7
sofa	–	–	17.0	19.6	20.5	21.5	22.2	**23.0**
train	–	–	34.7	37.6	36.9	38.8	38.8	**47.8**
tv	–	–	30.4	35.0	34.8	36.9	36.3	**48.6**
mIoU	25.7	38.2	32.6	36.6	37.8	**42.0**	34.5	41.9

post-processing strategies, such as using image-level classification prior for refinement, may mitigate this kind of issue.

5. Conclusion and future work

In this paper, we proposed a weakly supervised framework by only using image-label annotations for semantic segmentation. Specifically, we proposed to train the segmentation DCNN supervised by multiple localization maps, where each pixel can be assigned to multiple classes with different weights. The localization maps can be obtained via a proposals voting technique with only image-level labels. Furthermore, based on the generated localization maps, we proposed a simple but effective method to predict rough masks to train the segmentation DCNN. Experimental results on the PASCAL VOC 2012 segmentation benchmark well demonstrated the effectiveness of our proposed methods. In the future, we plan to further improve the segmentation performance by exploring more images with image-level annotations.

Conflict of interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

This work is supported in part by National Basic Research Program of China (No. 2012CB316400), National NSF of China (61210006, 61532005, 61502506), the Program for Changjiang Scholars, Innovative Research Team in University under Grant IRT201206 and National High Technology Research and Development Program of China (No. 2013AA013801).

References

[1] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, Cascade object detection with deformable part models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2241–2248.
[3] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Recognit. Mach. Intell. 32 (9) (2010) 1627–1645.
[4] L. Lin, X. Wang, W. Yang, J.-H. Lai, Discriminatively trained and-or graph models for object shape detection, IEEE Trans. Pattern Recognit. Mach. Intell. 37 (5) (2015) 959–972.
[5] J. Lu, J.J. Corso, et al., Human action segmentation with hierarchical supervoxel consistency, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3762–3771.
[6] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, Exemplar based deep discriminative and shareable feature learning for scene image classification, Pattern Recognit., http://dx.doi.org/10.1016/j.patcog.2015.02.003, in press.
[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062.
[8] J. Dai, K. He, J. Sun, Convolutional feature masking for joint object and stuff segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3992–4000.
[9] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 891–898.
[10] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[11] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, arXiv preprint arXiv:1411.5752.
[12] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. Torr, Conditional random fields as recurrent neural networks, arXiv preprint arXiv:1502.03240.
[13] F. Liu, G. Lin, C. Shen, Crf learning with cnn features for image segmentation, Pattern Recognit., http://dx.doi.org/10.1016/j.patcog.2015.04.019, in press.
[14] X. Liang, Y. Wei, X. Shen, J. Yang, L. Lin, S. Yan, Proposal-free network for instance-level object segmentation, arXiv preprint arXiv:1509.02636.
[15] X. Liang, Y. Wei, X. Shen, Z. Jie, J. Feng, L. Lin, S. Yan, Reversible recursive instance-level object segmentation, arXiv preprint arXiv:1511.04517.
[16] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, S. Yan, Human parsing with contextualized convolutional neural network, in: IEEE International Conference on Computer Vision, 2015, pp. 1386–1394.
[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Neural Information Processing Systems, 2012, pp. 1097–1105.
[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.
[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, arXiv preprint arXiv:1409.4842.
[20] X. Jin, C. Xu, J. Feng, Y. Wei, J. Xiong, S. Yan, Deep learning with s-shaped rectified linear activation units, arXiv preprint arXiv:1512.07030.
[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[22] J. Dai, K. He, J. Sun, Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation, arXiv preprint arXiv:1503.01640.
[23] G. Papandreou, L.-C. Chen, K. Murphy, A.L. Yuille, Weakly and semi-supervised learning of a dcnn for semantic image segmentation, arXiv preprint arXiv:1502.02734.
[24] P.O. Pinheiro, R. Collobert, From image-level to pixel-level labeling with convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1713–1721.
[25] D. Pathak, E. Shelhamer, J. Long, T. Darrell, Fully convolutional multi-class multiple instance learning, arXiv preprint arXiv:1412.7144.
[26] J. Xu, A.G. Schwing, R. Urtasun, Tell me what you see and i will show you where it is, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3190–3197.
[27] J. Xu, A.G. Schwing, R. Urtasun, Learning to segment under various forms of weak supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3781–3790.
[28] W. Yang, P. Luo, L. Lin, Clothing co-parsing by joint image segmentation and labeling, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3182–3189.
[29] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, J. Malik, Multiscale combinatorial grouping for image segmentation and object proposal generation, arXiv preprint arXiv:1503.00848.
[30] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, S. Yan, Cnn: Single-label to multi-label, arXiv preprint arXiv:1406.5726.
[31] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. 111 (1) (2014) 98–136.
[32] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, S. Yan, Semantic segmentation without annotating segments, in: IEEE International Conference on Computer Vision, 2013, pp. 2176–2183.
[33] L.-C. Chen, S. Fidler, R. Urtasun, Beat the mturkers: Automatic image labeling from weak 3d supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3198–3205.
[34] J. Zhu, J. Mao, A.L. Yuille, Learning from weakly supervised data by the expectation loss svm (e-svm) algorithm, in: Neural Information Processing Systems, 2014, pp. 1125–1133.
[35] M. Guillaumin, D. Küttel, V. Ferrari, Imagenet auto-annotation with segmentation propagation, Int. J. Comput. Vis. 110 (3) (2014) 328–348.
[36] D. Pathak, P. Krähenbühl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, arXiv preprint arXiv:1506.03648.
[37] J. Verbeek, B. Triggs, Region classification with Markov field aspect models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[38] A. Vezhnevets, J.M. Buhmann, Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3249–3256.
[39] A. Vezhnevets, V. Ferrari, J.M. Buhmann, Weakly supervised semantic segmentation with a multi-image model, in: IEEE International Conference on Computer Vision, 2011, pp. 643–650.
[40] X. Liu, W. Yang, L.-C. Lin, Q. Wang, Z. Cai, J.-S. Lai, Data-driven scene understanding with adaptively retrieved exemplars, IEEE Multimed., http://dx.doi.org/10.1109/MMUL.2015.22, in press.
[41] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, Neural Inf. Process. Syst. (1998) 570–576.
[42] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: IEEE International Conference on Computer Vision, 2011, pp. 991–998.
[43] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: ACM Multimedia, 2014, pp. 675–678.

Yunchao Wei is a Ph.D. student from the Institute of Information Science, Beijing Jiaotong University, China. He is currently working at National University of Singapore as a Research Intern. His research interests mainly include semantic segmentation, object detection and classification in computer vision and multi-modal analysis in multimedia.

Xiaodan Liang is a Ph.D. student from Sun Yat-sen University, China. She is currently working at National University of Singapore as a Research Intern. Her research interests mainly include semantic segmentation, object/action recognition and medical image analysis.

Yunpeng Chen received his bachelor degree in Huazhong University of Science and Technology in 2015. He currently is a Ph.D. student in the Department of Electrical and Computer Engineering at the National University if Singapore. His research interests include computer vision and machine learning.

Zequn Jie is currently a Ph.D. student from Vision and Machine Learning Group, directed by Professor Shuicheng Yan of National University of Singapore. His current research interests mainly include object localization related topics in computer vision, such as object proposal, object detection.

Yanhui Xiao received the Ph.D. degree in signal and information processing from Beijing Jiaotong University (BJTU), in 2014, and is currently a Lecturer at People's Public Security University of China. His current research interests include intelligence analysis, face recognition, computer vision, and machine learning.

Yao Zhao received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor at BJTU in 1998 and became a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He is currently leading several national research projects from the 973 Program, 863 Program, and the National Science Foundation of China. He serves on the editorial boards of several international journals, including as an Associate Editor of the IEEE Transactions on Cybernetics, Associate Editor of the IEEE Signal Processing Letters, Area Editor of Signal Processing: Image Communication (Elsevier), and Associate Editor of Circuits, System, and Signal Processing (Springer). He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013.

Shuicheng Yan is currently an Associate Professor at the Department of Electrical and Computer Engineering at National University of Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). Dr. Yan's research areas include machine learning, computer vision and multimedia, and he has authored/co-authored nearly 400 technical papers over a wide range of research topics, with Google Scholar citation $>17,000$ times. He is an ISI highly cited Researcher 2014, and IAPR Fellow 2014. He has been serving as an Associate Editor of IEEE TKDE, CVIU and TCSVT. He received the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM'12 (Best Demo), PCM'11, ACM MM'10, ICME'10 and ICIMCS'09, the runner-up prize of ILSVRC'13, the winner prizes of the classification task in PASCAL VOC 2010–2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.