Graph-Boosted Convolutional Neural Networks for Semantic Segmentation

Guangzhen Liu, Peng Han, Yulei Niu, Wenwu Yuan, Zhiwu Lu, and Ji-Rong Wen
Beijing Key Laboratory of Big Data Management and Analysis Methods
School of Information, Renmin University of China
Beijing 100872, China
Email: luzhiwu@ruc.edu.cn

***Abstract*—This paper investigates the problem of weakly-supervised semantic segmentation, where image-level labels are used as weak supervision. Inspired by the successful use of Convolutional Neural Networks (CNNs) for fully-supervised semantic segmentation, we choose to directly train the CNNs over the oversegmented regions of images for weakly-supervised semantic segmentation. Although there are a few studies on CNNs-based weakly-supervised semantic segmentation, they have rarely considered the noise issue, i.e., the initial weak labels (e.g., social tags) may be noisy. To cope with this issue, we thus propose graph-boosted CNNs (GB-CNNs) for weakly-supervised semantic segmentation. In our GB-CNNs, the graph-based model provides the initial supervision for training the CNNs, and then the outcomes of the CNNs are used to retrain the graph-based model. This training procedure is iteratively implemented to boost the results of semantic segmentation. Experimental results demonstrate that the proposed model outperforms the state-of-the-art weakly-supervised methods. More notably, the proposed model is shown to be more robust in the noisy setting for weakly-supervised semantic segmentation.**

I. Introduction

Semantic segmentation is a fundamental and challenging task in computer vision, which aims at assigning semantic labels to the pixels of images. The state-of-the-art approaches to semantic segmentation require a large amount of pixel-level labeled training images. However, it is time-consuming and laborious to annotate pixel-level (or superpixel-level) labels. Recent works focus on semantic segmentation in the weakly-supervised settings, where the training set has access only to image-level labels [1], [2], [3], [4], [5], [6]. However, most of the weakly-supervised semantic segmentation methods are based on the assumption that image-level labels are accurate, without considering noisy labels as weak supervision.

Due to the popularity of online photo sharing websites (e.g., Flickr), we can easily collect images with labels, which can be used as weak supervision for semantic segmentation. It should be noted that the image-level labels from online photo sharing websites might be noisy: some labels do not belong to an image, or some labels are missing from the ground truth. Figure 1 shows several example images with noisy labels. The noisy labels add further challenges to the problem of weakly-supervised semantic segmentation.

In this paper, we thus propose graph-boosted convolutional neural networks (GB-CNNs) to overcome the challenges posed by noisy image-level labels in weakly-supervised semantic segmentation. We choose to train our GB-CNNs directly over the oversegmented regions of images for semantic segmentation, rather than over the whole images. It should be noted that the CNNs have been shown to achieve the state-of-the-art results in the large scale object classification tasks [7]. This is mainly due to that the CNNs can learn sufficiently general features of images and even the well-trained CNNs on large datasets (e.g. ImageNet) can be directly used to extract features. Moreover, the CNNs have also been shown to well fit the task of fully-supervised semantic segmentation [8], [9]. Although there have been a few studies [6], [10], [11] on CNNs-based weakly-supervised semantic segmentation, they have rarely considered the noise issue, i.e., the initial weak labels (e.g., social tags) may be noisy.

In contrast, we focus on weakly-supervised semantic segmentation with noisy labels in the present work. Our main motivation of model design is to take the advantages of the CNNs-based model and graph-based model simultaneously: 1) The graph-based model has the ability of noise reduction by formulating its training process as an L_1-optimization problem, but it lacks the function of label prediction; 2) The CNNs-based model can provide a powerful classifier for label prediction, but it tends to be misled by the initial noisy labels during its training process. Hence, we choose a boosting training strategy for our GB-CNNs: the graph-based model provides the initial supervision for training the CNNs-based model and then the outcomes of the CNNs-based model are used to retrain the graph-based model, as shown in Figure 1. In fact, this training procedure can be iteratively implemented to boost the semantic segmentation results to the largest extent possible. Our later experimental results show that our GB-CNNs can effectively suppress the noise in the weak supervision for semantic segmentation.

Our main contributions are summarized as follows:

- We have presented a novel model, called GB-CNNs, for weakly-supervised semantic segmentation with noisy labels. In fact, the noise issue has been rarely considered in previous work on CNNs-based weakly-supervised semantic segmentation.
- We have exploited no external information for training our GB-CNNs, unlike other CNNs-based weakly-supervised semantic segmentation methods that directly train parts of CNNs on ImageNet.

978-1-5090-6182-2/17/$31.00 ©2017 IEEE

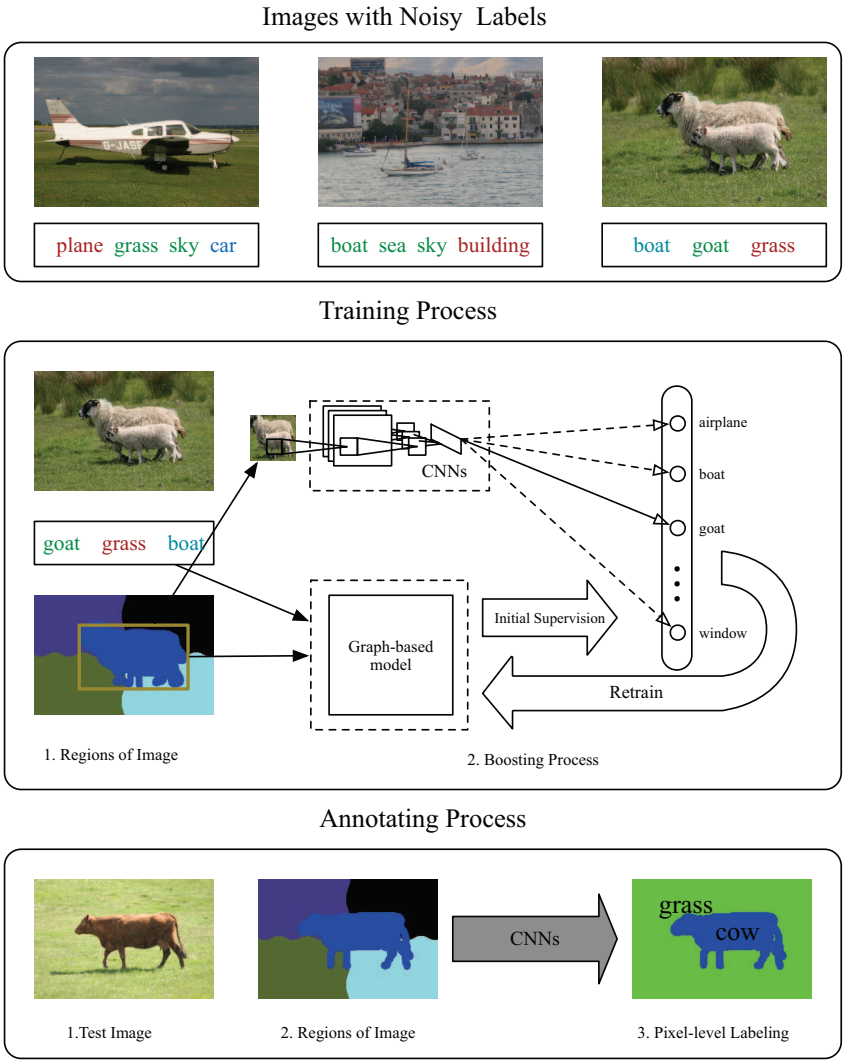

Fig. 1. **Illustration of our GB-CNNs.** *Top*: Images with noisy labels: incorrect (blue), correct (green) or missing (red). *Middle*: A boosting procedure for training our GB-CNNs. *Bottom*: Annotation process of test images.

- We have proposed a boosting training strategy for our GB-CNNs, which can improve the semantic segmentation results to the largest extent possible.

The reminder of this paper is organized as follows. Section II gives a brief overview of related work. In Section III, we present our GB-CNNs for weakly-supervised semantic segmentation with noisy labels. In Section IV, we conduct experiments on benchmark datasets to compare our GB-CNNs with the state-of-the-art approaches to semantic segmentation. We draw our conclusions in Section V.

II. Related Work

A. Semantic Segmentation

In the past years, the problem of semantic segmentation has been widely studied in computer vision. Most of the existing methods have been proposed for the fully-supervised setting, requiring access to a pixel-level labeled training set. [12] utilized the higher-order potentials as a soft decision to ensure that pixels constituting a particular segment have the semantic concept. [13] proposed a nonparametric system based on the probabilistic Markov random field model. These semantic segmentation algorithms require pixel-level labels for training, which are very difficult to obtain in practice.

Recently, more attention has been paid to the weakly-supervised setting for semantic segmentation. [1] proposed an unsupervised learning method by combining low-level texture, color and position cues with spatial random field models that capture the local coherence of region labels. [2] cast the weakly-supervised task as a multiple-instance learning problem and used Semantic Texton Forest (STF) as the basic framework. In [14], a multi-edge graph was first established to simultaneously consider both images and superpixels, and was then used to obtain superpixel labels through a majority voting strategy. However, these methods are all based on

the assumption that the image-level labels of the training set should be correct and complete, which is not practical for real-world applications.

B. Graph-Based Methods

In the recent years, the graph-based models have shown excellent performance in the task of weakly-supervised semantic segmentation. [15] claimed that the affinity graph over regions, which was generally constructed in relatively simple settings in many existing methods, was of crucial importance to the performance due to the fact that the weakly-supervised semantic segmentation problem cannot be solved within a single image and the affinity graph enables label propagation among multiple images. In [16], a weakly-supervised matrix factorization approach was developed for semantic segmentation by following the idea of graph-regularized matrix factorization.

Since the graph-based model can exploit the contextual information by discovering the manifolds hidden among the data, it has the ability of noise reduction. However, the graph-based model lacks the function of label prediction for semantic segmentation. Given the test set, the graph needs to be rebuilt and thus the test process is very time-consuming.

C. CNNs-Based Methods

In the latest years, convolutional neural networks have been widely used in the tasks of image classification and object detection. For example, [7] proposed deformable deep convolutional neural networks for generic object detection, which perform well on ImageNet.

There have also been latest studies on fully-supervised semantic segmentation using CNNs. [8] used multi-scale convolutional networks to capture texture, shape and contextual information and then combined CNNs with a segmentation tree. [9] proposed fully convolutional networks for fully-supervised semantic segmentation, which only adopt convolutional networks.

For weakly-supervised semantic segmentation, [5] presented a joint conditional random field model to leverage various contexts. More specifically, they extracted the global and local features at multiple scales by the CNNs and topic model. In [11], a well-trained CNN model on ImageNet was used to extract features for the bounding boxes and images. In [10], a CNNs-based model is used to handle the task of weakly-supervised semantic segmentation, which consists of ten convolutional layers and a log-sum-exp layer. The first six convolutional layers of their CNNs are parts of Overfeat which has been well-trained on ImageNet. This means that they actually exploited external information for weakly-supervised semantic segmentation.

Different from [10], we take the regions of image as the inputs of our CNNs, instead of the whole images. In fact, we train our CNNs only with image-level labels, i.e., no external information has been used in the training process. Due to this training strategy, our CNNs can be easily combined with the graph-based model and thus maintain stable in the noisy setting (i.e. noisy image-level labels are provided).

III. Graph-Boosted CNNs

In this section, we first give the model formulation for our GB-CNNs which perform a boosting training procure between the graph-based model and CNNs-based model. We further provide the details of the graph-based model and CNNs-based model. Finally, we give the annotation process for test images using the trained CNNs.

A. Model Formulation

Assuming that each training and test image has been over-segmented into a set of regions and each region has been assigned a label that can be noisy, our objective is to identify and correct the noisy region labels. Formally, we are given a large set of regions $\mathcal{X} = \{x_1, ..., x_N\}$ and their initial labels $Y = \{y_{ij} : y_{ij} \in \{0, 1\}\}_{N \times C}$, where N is the total number of regions and C is the number of object categories. These regions are represented as feature vectors x_i $(i = 1, ..., N)$ by extracting color and texture features (see the experiment part), while the initial labels $\{y_{i1}, y_{i2}, ..., y_{iC}\}$ of them are inferred from the image-level labels. The details of initial label assignment are presented in the following.

Given that each training and test image has a set of image-level labels (provided for training images in advance and estimated for test images just as [6]), we can assign the initial labels of regions by propagating the labels of images to the regions. Specifically, the initial labels of regions $Y = \{y_{ij}\}_{N \times C}$ are estimated as: $y_{ij} = 1$ if region x_i belongs to an image which is labelled with category j and $y_{ij} = 0$ otherwise. Note that the initial labels of regions cannot be accurately estimated by such simple inference. The noise issue becomes even severer when the image-level labels are noisy to begin with. This is true for the estimated labels of test images and can also be true if the labels of training images are obtained from user-provided tags. In this paper, we will develop a graph-based model to suppress the noise in Y and thus generate the initial supervision for our CNNs.

As shown in Figure 2, the training process of our GB-CNNs is a boosting procedure, where the graph-based model provides the initial supervision for training the CNNs and then the outcomes of the CNNs are used to retrain the graph-based model. Our main motivation is to take the advantages of the CNNs-based model and graph-based model simultaneously. In fact, this training procedure can be iteratively implemented to boost the semantic segmentation results to the largest extent possible. In the next two subsections, we will describe the two models respectively.

B. Graph-Based Model

First of all, we compute the weight matrix $W = \{w_{ij}\}_{N \times N}$ of the affinity graph over the regions as follows:

$$w_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|_2^2}{\sigma^2}), & \text{if } x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\ 0, & \text{otherwise} \end{cases}$$

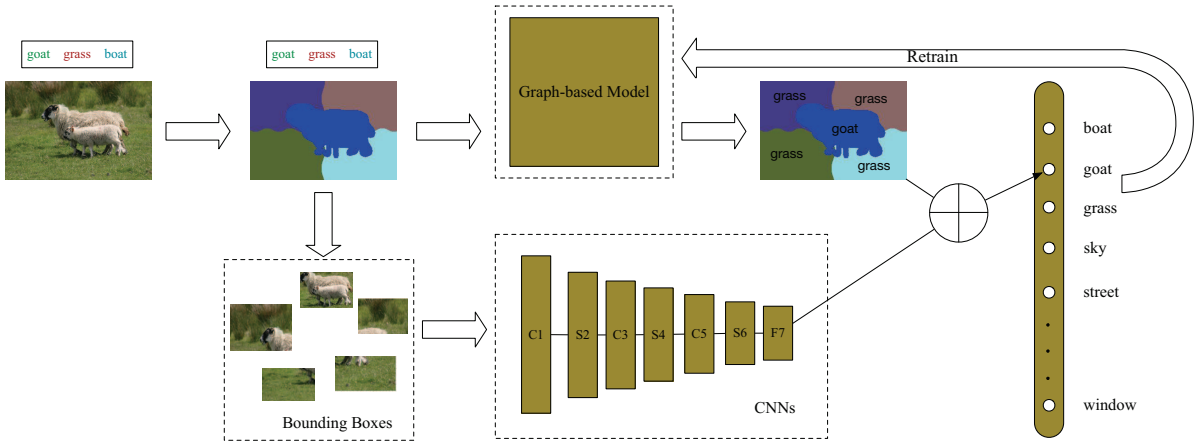

Fig. 2. **Training process of our GB-CNNs.** We train our CNNs using the outputs of graph-based model and then retrain the graph-based model using the outcomes of CNNs. The regions of images are set as the inputs of our CNNs.

where $\mathcal{N}_k(x)$ is the set of k-nearest regions of x. Here, we actually define the weight matrix based on the Gaussian kernel. The normalized Laplacian matrix of W is given by

$$\mathcal{L} = I - D^{-1/2} W D^{-1/2}$$

where I is an $n \times n$ identity matrix and D is an $n \times n$ diagonal matrix with its i-th diagonal element being the sum of the i-th row of W. To reduce the computational complexity, we further derive a new matrix B from $\mathcal{L}$:

$$B = \Sigma^{\frac{1}{2}} V^{\mathrm{T}}$$

where V is an $n \times n$ orthonormal matrix with each column being an eigenvector of $\mathcal{L}$, and Σ is an $n \times n$ diagonal matrix with its diagonal element Σ_{ii} being an eigenvalue of $\mathcal{L}$ (sorted as $0 \leq \Sigma_{11} \leq \cdots \leq \Sigma_{nn}$).

To suppress the noise in Y, we formulate semantic segmentation as the following L_1-optimization problem:

$$\min_{F,\hat{Y}} \frac{1}{2}\|F - \hat{Y}\|_F^2 + \lambda \|BF\|_1 + \gamma \|\hat{Y} - Y\|_1$$

where λ and γ denote the positive regularization parameters, $\hat{Y}$ denotes the ideal region label assignment, and F denotes an intermediate label assignment.

In fact, we can solve this L_1-optimization problem directly using the standard algorithm [17]. Moreover, we find that our graph-based model is not much sensitive to the parameters and thus we uniformly set the parameters as $k = 200, \lambda = 0.01, \gamma = 0.1$ in this paper.

C. CNNs-Based Model

The structure of our CNNs is illustrated in Figure 2. In total, our CNNs have three convolutional layers, three max pooling layers, and two full connection layers. Some details of the network structure are given as follows:

- For every convolutional layer, we add a non-saturating nonlinearity as Rectified Linear Unit (ReLU). Hence, every unit x in the convolutional layer is set to $\max(0, x)$. In [7], this way is shown to be faster than saturating nonlinearities such as $f(x) = \tanh(x)$ and $f(x) = (1 + e^{-x})^{-1}$.
- We also adopt "dropout" [18], which chooses to set to zero the output of each hidden neuron with probability 0.5. The neurons which are "dropped out" in this way do not contribute to the forward pass and do not participate in back-propagation. Hence, when every time an input is presented, the neural network samples a different architecture, but all of the architectures share weights. This technique reduces complex co-adaptations of neurons, since a neuron cannot rely on the presence of particular other neurons. It is, therefore, forced to learn more robust features that are useful in conjunction with many different random subsets of the other neurons. At the test time, we take all the neurons into account. We use dropout in the first fully-connected layer.
- For the output layer, we add a softmax layer. The output layer is then transformed into a probability distribution $\mathbf{p} \in \mathbb{R}^C$ for all the objects of C categories, and the cross entropy is used to measure the prediction loss of the network. For the unbalanced problem, we choose to set different weights for different categories. Formally, we can define the loss function L using the cross entropy:

$$L = -\sum_i \lambda_i t_i \log(p_i),$$

$$p_i = \frac{\exp(h_i)}{\sum_i \exp(h_i)},$$

where $i = 1, \ldots, C$. The gradients of the deep CNNs are calculated via back-propagation:

$$\frac{\partial L}{\partial h_i} = p_i - \lambda_i t_i,$$

where $t = \{t_i | t_i \in \{0, 1\}, i = 1, \ldots, C, \sum_{i=1}^{C} t_i = 1\}$ denotes the true label of the sample x, and λ_i denotes the i-th category's weight.

The process of training CNNs is also shown in Figure 2. Here, each image is first divided into several regions, and we assume that all the pixels in one region have the same label. For every region, we use a bounding box as the input and the bounding box is resized into 124×124. The initial labels of all the regions are generated by our graph-based model, which are further used as the initial supervision for training our CNNs. To reduce the influence of unbalance problem, different categories are set with different weights in the process of back-propagation. That is, an object category that has more regions is set with a smaller weight.

D. Annotation Process for Test Images

At the test time, each test image is first divided into several regions, and we then bound the regions with boxes. The obtained bounding boxes are resized into 124×124. In the annotation process, we no longer consider the graph-based model. That is, we only utilize the trained CNNs to predict the labels of regions. Finally, we assign all the pixels in a region with the same predicted label.

IV. Experimental Results

In this section, we evaluate the performance of the proposed model for weakly supervised semantic segmentation. We first compare the proposed model with the state-of-the-art approaches on two benchmark datasets. Furthermore, we conduct another group of experiments to verify the robustness of our model under the noisy condition.

A. Experimental Setup

Benchmark Datasets: The SIFT-Flow dataset [19] consists of 2688 images of resolution 256×256 pixels, which have been thoroughly labeled by the LabelMe users, and split into 2488 training images and 200 test images. Synonym correction is considered to obtain 33 object categories. To further testify the capacity of the proposed model, we also evaluate it on the MSRC dataset, which has 591 images from 21 object categories.

Performance Measures: We evaluate the performance of semantic segmentation in terms of both total per-pixel accuracy (a_p) and average per-class accuracy (a_c), which have been widely used for performance evaluation in the literature. Here, a_p measures the percentage of totally correctly classified pixels, while a_c measures the percentage of correctly classified pixels with respect to a class and then averaging over all classes.

Feature Extraction: For each image, we use the Blobworld method [20] for over-segmentation. On average, this generates 12.2 and 14.3 regions per image for the two benchmark datasets, respectively. For the graph-based model, we further extract a 137-dimension feature vector for every region, which includes three mean color features with their standard deviations (6-dimension), three mean texture features with their standard deviations (6-dimension), and 125-dimension color histogram. Finally, we apply Gaussian kernel over all regions to define the weight matrix W in our graph-based model.

TABLE I
Comparison to the state-of-the-art on SIFT-Flow. DL denotes deep learning used for semantic segmentation.

Supervision	Methods	DL	a_p (%)	a_c (%)
Fully supervised	[21]	No	77.0	30.1
	[13]	No	76.7	24.0
	[22]	No	77.1	32.5
	[8]	Yes	78.5	29.6
	[23]	No	79.2	33.8
	[24]	No	78.6	39.3
	[25]	No	79.8	48.7
	[9]	Yes	**85.2**	**51.7**
Weakly supervised	[3]	No	N/A	14
	[4]	No	51	21
	[26]	No	N/A	26.3
	[27]	No	N/A	27.7
	[6]	Yes	N/A	27.9
	[5]	Yes	N/A	32.3
	[16]	No	N/A	33
	[28]	No	N/A	41
	[11]	Yes	**62.7**	41.4
	Ours	Yes	60.2	**42.1**

TABLE II
Comparison to the state-of-the-art on MSRC.

Supervision	Methods	a_p (%)	a_c (%)
Fully supervised	[29]	72	67
	[30]	**86**	75
	[31]	77	75
	[32]	79	**78**
	[33]	82	76
Weakly supervised	[34]	N/A	57
	[2]	N/A	37
	[3]	67	67
	[27]	N/A	71
	[35]	N/A	62
	[16]	N/A	69
	[11]	**70**	73
	Ours	69	**74**

Training Settings: In the experiments, we train our CNNs using Torch7. Specifically, we adopt stochastic gradient descent with a learning rate of 0.001, a batch size of 1 example, momentum of 0, and weight decay of 0.

B. Comparison to the State-of-the-Art

We present the comparison to the state-of-the-art semantic segmentation methods on the two benchmark datasets in Tables 1 and 2, respectively. Both total per-pixel accuracy (a_p) and average per-class accuracy (a_c) are used as performance measures for semantic segmentation. We can make the following observations: (1) Our model generally outperforms the other weakly-supervised methods except [11]. In particular, although our model is only comparable to [11], our later experiments show that our model is much more robust against noisy labels. (2) When average per-class accuracy is considered as the performance measure, our model even performs better than (at least comparably to) some fully-supervised methods.

We also show some example results of semantic segmentation obtained by our model on the SIFT-Flow dataset in Figure 3. From these qualitative results, we find that our model can assign the labels of pixels well based on the trained CNNs and maintain the shape of most objects.

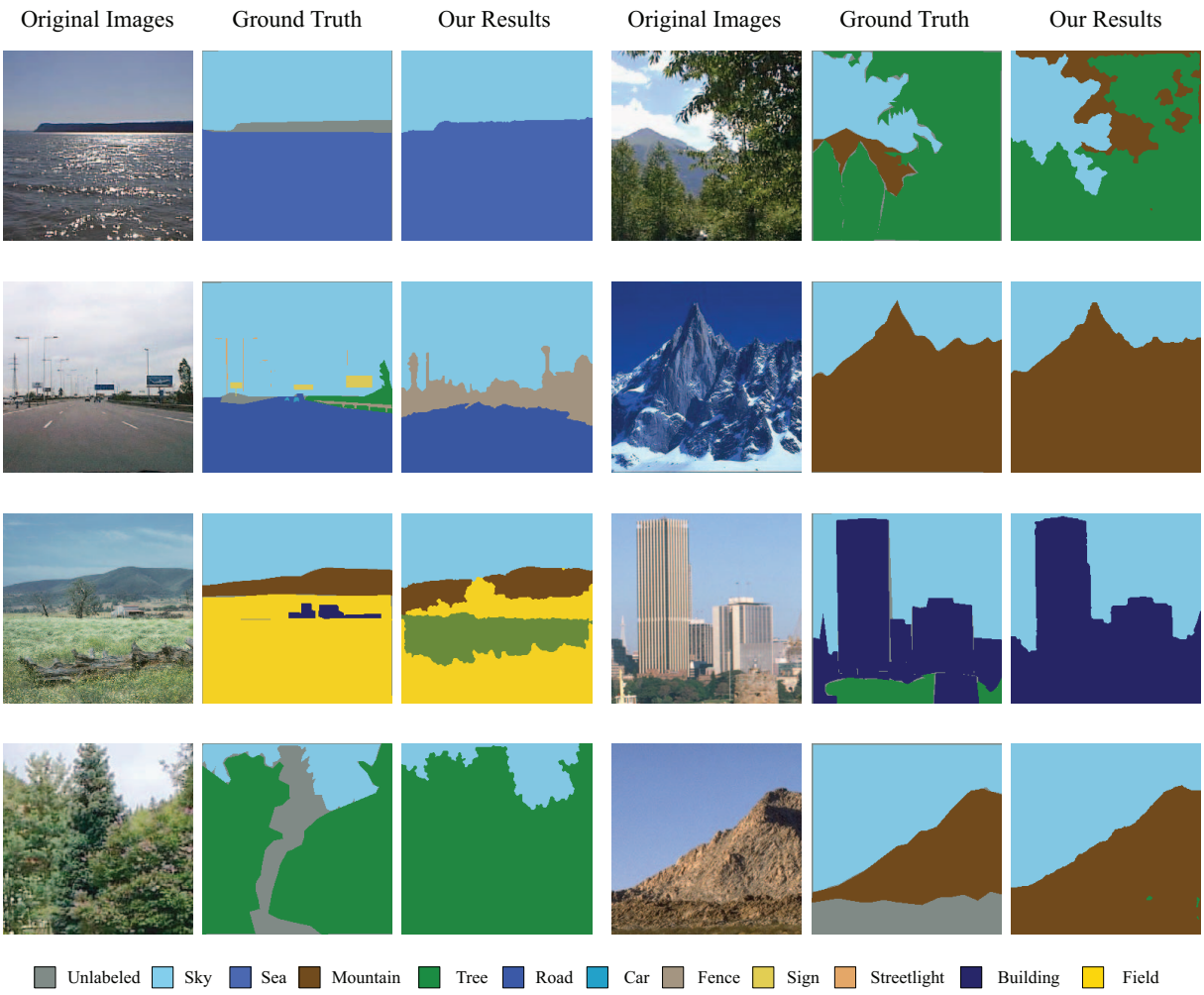

Fig. 3. Some example results obtained by our GB-CNNs for semantic segmentation on the SIFT-Flow dataset.

C. Performance under Noisy Condition

To evaluate the robustness of our model under the noisy condition, we assign noisy labels to the images in SIFT-Flow. For each noisily labeled image, we randomly add a wrong extra label to it or remove one existing label from it. The percentage of noisily labeled images is set as: 0%, 25%, 50%, 75%, and 100%. In the experiments, we compare different models with the same noise percentage.

We make comparison among the following four models for semantic segmentation with noisy labels:

- **Graph+CNNs (Ours):** The graph-based model provides the initial supervision for the CNNs that are trained to predict the region labels of test images.
- **Graph+SVM (Ours):** The graph-based model provides the initial supervision for SVM that is trained to predict the region labels of test images.
- **CNNs-only (Ours):** Only CNNs are trained over regions for weakly-supervised semantic segmentation. For each region, we initially assign it with all the labels of the image that it belongs to.

TABLE III
RESULTS UNDER THE NOISY CONDITION ON SIFT-FLOW. ONLY a_p (%) IS USED AS THE MEASURE.

Noisily Labeled Images	0 %	25%	50%	75%	100%
Graph+CNNs (Ours)	60.2	**58.6**	**58.3**	**57.8**	**57.6**
Graph+SVM (Ours)	51.9	52.3	50.2	49.8	43.5
CNNs-only (Ours)	52.6	51.5	50.6	49.3	47.6
CNNs-only (CaffeNet)	44.3	39.5	37.6	31.4	33.6
WSDC [26]	39.0	38.3	28.6	25.3	29.3
ILT [11]	**62.7**	48.4	47.1	42.5	39.4

- **CNNs-only (CaffeNet):** The same setting as CNNs-only (Ours), except that the CNNs are the CaffeNet pre-trained on the ImageNet dataset.
- **WSDC:** The Weakly-Supervised Dual Clustering (WSDC) model is proposed in [26] for semantic segmentation.
- **ILT:** The ILT model is proposed in [11] for semantic segmentation.

The comparison results are shown in Table 3. It can be seen that when increasing percentage of noisy labels are added to

the training set, the performance of our model degrades much more gracefully than the compared ones. This observation shows that our model indeed has better ability to cope with noisy labels due to the graph-based model used to provide the initial supervision for our CNNs. Moreover, the distinct advantages of our Graph+CNNs over Graph+SVM and CNNs-only verify that either component (graph-based model or CNNs) of our model plays an important role in semantic segmentation. In addition, the comparison between CNNs-only (ours) and CNNs-only (CaffeNet) shows that the architecture of our CNNs is better than that of CaffeNet in the noisy setting for semantic segmentation, even when the CaffeNet is pre-trained on the ImageNet dataset.

D. Discussion

It should be noted that the results of our model can be further refined by the probabilistic graphical models such as conditional random field. However, our extra experiments show that the use of the contextual information only leads to very limited improvements (less than 1%). This observation indirectly verifies that we have already trained a powerful model (i.e. GB-CNNs) for semantic segmentation.

V. Conclusion

In this paper, we have presented novel GB-CNNs for weakly-supervised semantic segmentation with noisy labels. To improve the semantic segmentation results to the largest extent possible, we have proposed a boosting training strategy for our GB-CNNs. Unlike other CNNs-based weakly-supervised semantic segmentation methods that directly train parts of CNNs on ImageNet, we have exploited no external information during training our GB-CNNs. Experimental results show that the proposed model outperforms most of the existing methods and even maintains more stable under noisy condition. In the future work, we will try other architectures of CNNs for semantic segmentation. Moreover, we will also apply our CNNs to other problems in computer vision.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (61573363 and 61573026), 973 Program of China (2014CB340403 and 2015CB352502), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (15XNLQ01), and the Outstanding Innovative Talents Cultivation Funded Programs 2016 of Renmin University of China.

References

[1] J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in *CVPR*, 2007, pp. 1–8.

[2] A. Vezhnevets and J. M. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *CVPR*, 2010, pp. 3249–3256.

[3] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *ICCV*, 2011, pp. 643–650.

[4] ——, "Weakly supervised structured output learning for semantic segmentation," in *CVPR*, 2012, pp. 845–852.

[5] W. Zhang, S. Zeng, D. Wang, and X. Xue, "Weakly supervised semantic segmentation for social images," in *CVPR*, 2015, pp. 2718–2726.

[6] J. Xu, A. G. Schwing, and R. Urtasun, "Tell me what you see and i will show you where it is," in *CVPR*, 2014, pp. 3190–3197.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *PAMI*, vol. 35, no. 8, pp. 1915–1929, 2013.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[10] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *CVPR*, 2015, pp. 1713–1721.

[11] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to segment under various forms of weak supervision," in *CVPR*, 2015, pp. 3781–3790.

[12] P. Kohli and P. H. Torr, "Robust higher order potentials for enforcing label consistency," *IJCV*, vol. 82, no. 3, pp. 302–324, 2009.

[13] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *PAMI*, vol. 33, no. 12, pp. 2368–2382, 2011.

[14] D. Liu, S. Yan, Y. Rui, and H. Zhang, "Unified tag analysis with multi-edge graph," in *ACMMM*, 2010, pp. 25–34.

[15] W. Xie, Y. Peng, and J. Xiao, "Semantic graph construction for weakly-supervised image parsing," in *AAAI*, 2014.

[16] Y. Niu, Z. Lu, S. Huang, P. Han, and J. Wen, "Weakly supervised matrix factorization for noisily tagged image parsing," in *IJCAI*, 2015, pp. 3749–3755.

[17] M. Schmidt, G. Fung, and R. Rosales, "Fast optimization methods for l1 regularization: A comparative study and two new approaches," in *ECML*. Springer, 2007, pp. 286–297.

[18] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[19] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *CVPR*, 2009, pp. 1972–1979.

[20] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *PAMI*, vol. 24, no. 8, pp. 1026–1038, 2002.

[21] J. Tighe and S. Lazebnik, "Superparsing: scalable nonparametric image parsing with superpixels," in *ECCV*, 2010, pp. 352–365.

[22] D. Eigen and R. Fergus, "Nonparametric image parsing using adaptive neighbor sets," in *CVPR*, 2012, pp. 2799–2806.

[23] G. Singh and J. Kosecka, "Nonparametric scene parsing with adaptive feature relevance and semantic context," in *CVPR*, 2013, pp. 3151–3157.

[24] J. Tighe and S. Lazebnik, "Finding things: Image parsing with regions and per-exemplar detectors," in *CVPR*, 2013, pp. 3001–3008.

[25] J. Yang, B. Price, S. Cohen, and M. Yang, "Context driven scene parsing with attention to rare classes," in *CVPR*, 2014, pp. 3294–3301.

[26] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, "Weakly-supervised dual clustering for image semantic segmentation," in *CVPR*, 2013, pp. 2075–2082.

[27] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation," in *CVPR*, 2013, pp. 1908–1915.

[28] Y. Li, J. Liu, Y. Wang, H. Lu, and S. Ma, "Weakly supervised RBM for semantic segmentation," in *IJCAI*, 2015.

[29] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *CVPR*, 2008, pp. 1–8.

[30] C. Russell, P. Kohli, P. H. Torr *et al.*, "Associative hierarchical crfs for object class image segmentation," in *ICCV*, 2009, pp. 739–746.

[31] J. M. Gonfaus, X. Boix, J. Van d. Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez, "Harmony potentials for joint classification and segmentation," in *CVPR*, 2010, pp. 3280–3287.

[32] A. Lucchi, Y. Li, X. Boix, K. Smith, and P. Fua, "Are spatial and global constraints really necessary for segmentation?" in *ICCV*, 2011, pp. 9–16.

[33] A. Lucchi, Y. Li, K. Smith, and P. Fua, "Structured image segmentation using kernelized features," in *ECCV*, 2012, pp. 400–413.

[34] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *IJCV*, vol. 81, no. 1, pp. 2–23, 2009.

[35] E. Akbas and N. Ahuja, "Low-level hierarchical multiscale segmentation statistics of natural images," *PAMI*, vol. 36, no. 9, pp. 1900–1906, 2014.