

FeaBoost: Joint Feature and Label Refinement for Semantic Segmentation

Yulei Niu,¹ Zhiwu Lu,^{1*} Songfang Huang,² Xin Gao,³ Ji-Rong Wen¹

¹Beijing Key Laboratory of Big Data Management and Analysis Methods,
School of Information, Renmin University of China, Beijing 100872, China

²IBM China Research Lab, Beijing, China

³Computer, Electrical and Mathematical Sciences and Engineering Division,
King Abdullah University of Science and Technology (KAUST), Thuwal, Jeddah 23955, Saudi Arabia
nyl92@163.com, zhiwu.lu@gmail.com

Abstract

We propose a novel approach, called FeaBoost, to image semantic segmentation with only image-level labels taken as weakly-supervised constraints. Our approach is motivated from two evidences: 1) each superpixel can be represented as a linear combination of basic components (e.g., predefined classes); 2) visually similar superpixels have high probability to share the same set of labels, i.e., they tend to have common combination of predefined classes. By taking these two evidences into consideration, semantic segmentation is formulated as joint feature and label refinement over superpixels. Furthermore, we develop an efficient FeaBoost algorithm to solve such optimization problem. Extensive experiments on the MSRC and LabelMe datasets demonstrate the superior performance of our FeaBoost approach in comparison with the state-of-the-art methods, especially when noisy labels are provided for semantic segmentation.

Introduction

Image semantic segmentation (Arbelaez et al. 2012; Carreira et al. 2012) is one of the most fundamental challenges in computer vision. This task consists of two parts, image segmentation and superpixel annotation (see Figure 1), that is, grouping pixels into superpixels and then assigning each superpixel to one of the predefined classes. In fact, the solutions of these two subproblem can boost each other. On one hand, we can extract the features of superpixels and train a classifier to predict superpixel-level labels. On the other hand, the segmentation results can be upgraded using the results of superpixel annotation, since superpixels with the same label can be assembled into a complete region.

Recently, some methods have shown promising results in image semantic segmentation (Shotton et al. 2006; Yang, Meer, and Foran 2007; Kohli, Torr, and others 2009; Russell et al. 2009; Ladicky et al. 2010; Lucchi et al. 2012; Tighe and Lazebnik 2010; Long, Shelhamer, and Darrell 2015; George 2015). Most of these works concentrate on the fully supervised setting, where each pixel in the training images need to be annotated in advance. However, it is hard to widely apply fully supervised methods in real-world applications because collecting pixel-level labels costs enormous

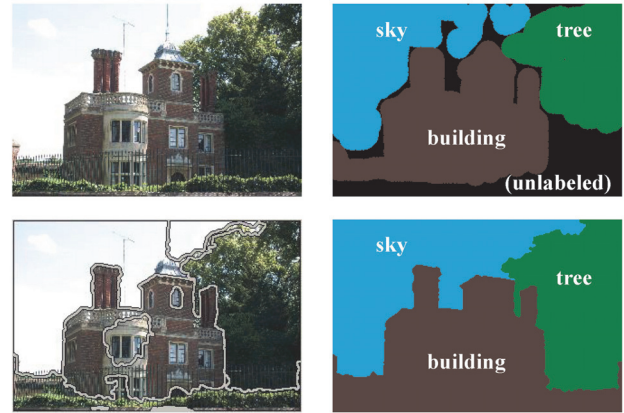


Figure 1: *Top Left*: original image, *top right*: ground truth, *bottom left*: oversegmented superpixels, *bottom right*: the result of semantic segmentation.

time and labour. For this reason, recent works start to focus on the weakly supervised setting, where only image-level labels are available for training (Verbeek and Triggs 2007; Vezhnevets and Buhmann 2010; Vezhnevets, Ferrari, and Buhmann 2011; 2012; Liu et al. 2013; Zhang et al. 2013). Since lots of users share their photos with tags on social websites (e.g., Flickr), collecting plenty of images along with image-level labels can be made automatic for weakly supervised methods.

Although extracting weak supervision from social images is less time-consuming, image-level labels provided by social users may be incorrect or incomplete (Tang et al. 2009). Hence, semantic segmentation with noisy labels becomes challenging, which has been rarely considered by most of weakly supervised methods. Some latest works (Li et al. 2015; Niu et al. 2015) start to focus on this challenging problem and achieve promising results. Although we also focus on semantic segmentation with noisy labels, our later experimental results show that our approach obviously outperforms (Li et al. 2015; Niu et al. 2015).

In this paper, we focus on proposing a novel framework for semantic segmentation, which is motivated from two evidences. Firstly, each superpixel can be represented as a lin-

*Corresponding author

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

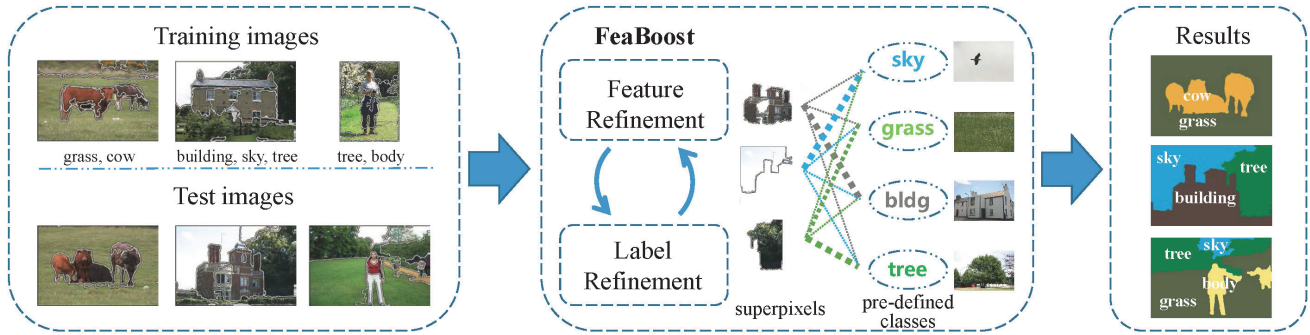


Figure 2: The flowchart of our FeaBoost approach to semantic segmentation.

ear combination of basic components (predefined classes are considered here). Secondly, visually similar superpixels tend to share the same set of labels (i.e., have common combination of predefined classes). By considering these two evidences together, we formulate the problem of semantic segmentation as joint feature and label refinement over superpixels. Hence, the proposed framework includes two main components with respect to superpixels: (1) oversegmentation and feature extraction; (2) joint feature and label refinement (see Figure 2). Specifically, each image is oversegmented into multiple superpixels, and their R-CNN (Girshick et al. 2014) features are extracted at both global and local levels. We further develop an efficient algorithm for joint feature and label refinement. In this algorithm, the Laplacian regularization is used to guarantee that visually similar superpixels share common combination of classes, and the L_1 -norm constraint is used for noise reduction. In addition, several one-vs-all SVM classifiers are trained using LIBLINEAR (Fan et al. 2008) to further improve the results of semantic segmentation.

To evaluate the performance of our approach, we conduct extensive experiments on the MSRC (Shotton et al. 2006) and LabelMe (Liu, Yuen, and Torralba 2011) datasets. The proposed approach is shown to achieve best results than the state-of-the-art methods under weakly supervised setting, and even outperform some fully supervised methods. In addition, plenty of experiments are conducted with noisy labels on the LabelMe dataset to evaluate the robustness of our approach. The experimental results demonstrate that our approach can make a good balance between nice semantic segmentation and effective noise reduction.

The main contributions of our work include:

- We have proposed a robust approach to semantic segmentation with noisy labels, which has been rarely considered in the literature.
- We have made the first attempt to formulate the problem of semantic segmentation as joint feature and label refinement over superpixels.
- We have developed an efficient algorithm to solve the joint feature and label refinement problem.

Related Work

Fully supervised semantic segmentation: In the past years, most of related works focus on the fully supervised setting for image semantic segmentation. (Shotton et al. 2006) used a Conditional Random Field (CRF) model to incorporate shape, texture, color, location and edge cues into a single unified model. Based on this typical method, lots of extensions are proposed to modify the CRF model. (Kohli, Torr, and others 2009) employed higher order potentials defined over image segments based on color, location, texture, and smoothness. (Russell et al. 2009) proposed a hierarchical random field model that allows integration of features computed at different levels of the quantization hierarchy. (Ladicky et al. 2010) further considered label co-occurrence statistics. However, these fully supervised methods heavily rely on pixel-level annotations, which are time-consuming and labor-sensitive to collect.

Weakly supervised semantic segmentation: Different from fully supervised methods, weakly supervised methods for semantic segmentation just require image-level labels for training. (Verbeek and Triggs 2007) presented a model by combining the global label coupling of Probabilistic Latent Semantic Analysis (PLSA) with local Markov Random Field (MRF) label interactions. (Vezhnevets and Buhmann 2010) used Semantic Texton Forest as the basic framework and extended it for the multiple instance learning setting. Furthermore, appearance similarity of superpixels were considered in their extended works (Vezhnevets, Ferrari, and Buhmann 2011; 2012). (Liu et al. 2013) proposed a coherent framework to cluster superpixels and assign a suitable label to each cluster. (Xu, Schwing, and Urtasun 2014) proposed a graphical model to encode the presence and absence of a class as well as the assignments of semantic labels to superpixels. Based on this work, (Xu, Schwing, and Urtasun 2015) proposed a unified approach that incorporates image-level tags, bounding boxes, and partial labels.

Semantic segmentation with noisy labels: The above weakly supervised methods for semantic segmentation assume that image-level labels are correct and complete in the training stage. However, in real-world applications, labels of social images provided by users may be incorrect or incomplete. Meanwhile, it is impracticable to clean image-level labels manually. Recently, some weakly supervised methods

start to focus on how to exploit noisy labels of images for semantic segmentation. (Li et al. 2015) extended the standard Restricted Boltzmann Machines (RBM) to a weakly supervised version by considering the weak supervision of image-level labels and the similarity of superpixels. (Niu et al. 2015) proposed a weakly supervised approach from a low-rank matrix factorization viewpoint for direct noise reduction over superpixel-level labels.

In this paper, we also focus on image semantic segmentation with noisy labels. The main difference between the present work and the latest works (Li et al. 2015; Niu et al. 2015) is that we have made the first attempt to formulate semantic segmentation as joint feature and label refinement over superpixels. Furthermore, the refined labels of superpixels can be used to train another typical classifier (e.g., SVM) over superpixels to further improve the semantic segmentation results.

The Proposed Framework

In this paper, we propose a novel framework to solve the problem of image semantic segmentation. As shown in Figure 2, the proposed framework consists of several components. Firstly, Blobworld method (Carson et al. 2002) is adopted to automatically oversegment each image into superpixels, and extract their R-CNN features at both local and global levels. We will introduce this component in detail in Section 5. Secondly, the features and labels of superpixels are jointly refined for semantic segmentation. Finally, several one-vs-all SVM classifiers are trained with the refined labels of superpixels to further improve the semantic segmentation results. In this section, we focus on joint feature and label refinement, also called FeaBoost.

Initial Label Estimation

Since pixel-level labels are unknown under the weakly supervised setting, we first infer the superpixel-level labels from the annotations of all the images (the annotations of test images can be predicted in advance). Supposed that all the images have been oversegmented into N superpixels, we extract a M -dimensional feature vector of each superpixel and obtain a set of feature vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in R^{M \times N}$. Let $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}^T \in R^{N \times C}$ represent the initial labels of superpixels, where C denotes the number of predefined classes. For any superpixel s_i belonging to an image I , if I is annotated with label j , then $y_{ij} = 1$, otherwise $y_{ij} = 0$. We further smooth Y by the area of superpixels. Let ρ_i denote the ratio of superpixel s_i occupying image I , then we define the smoothed labels of superpixels $\tilde{Y} = \{\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \dots, \tilde{\mathbf{y}}_N\}^T$ as

$$\tilde{y}_{ij} = \rho_i y_{ij}$$

which means that larger superpixels play an more important role in semantic segmentation. The smoothing procedure can sufficiently decrease the effect of tiny superpixels.

Feature Refinement

The main evidence of feature refinement is that each superpixel can be regarded as a linear combination of ba-

sic components, here predefined classes are used. The feature vector set of predefined classes is denoted as $U = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_C\} \in R^{M \times C}$. For any superpixel s_i , the linear combination can be represented as

$$\mathbf{x}_i \approx \sum_{k=1}^C \mathbf{u}_k v_{ik} \quad (1)$$

where $\sum_{k=1}^C \mathbf{u}_k v_{ik}$ denotes the refined feature vector of s_i , and $\mathbf{v}_i = \{v_{i1}, v_{i2}, \dots, v_{iC}\}$ denotes the coefficient vector. For example, a superpixel s_i from “tree” can be represented as the combination of “tree”, “grass” and “sky”. It is obvious that “tree” is quite different from “sky” visually, but looks alike to “grass” from color. This means “grass” denotes more than “sky” to represent the feature of s_i , but much less than “tree” from texture and position. Since the coefficient of “tree” is maximal in the vector v_i , this superpixel can be annotated with label “tree”. We further represent Eq. (1) in matrix form as $X \approx UV^T$, and define the feature refinement term as

$$\|X - UV^T\|_F^2 \quad (2)$$

where the Frobenius-norm constraint guarantees that the refined features UV^T should not change too much from X .

Label Refinement

The set of superpixels is modeled as a graph $\mathcal{G} = \{\mathcal{V}, W\}$, with the vertex set \mathcal{V} being defined as X and the similarity matrix $W = \{w_{ij}\}_{N \times N}$. The element of W can be calculated based on the Gaussian kernel as

$$w_{ij} = \begin{cases} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\sigma^2}) & , \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0 & , \text{otherwise} \end{cases}$$

where $\mathcal{N}_k(\mathbf{x})$ denotes the k -nearest set of superpixel \mathbf{x} . Based on the graph model, we define the Laplacian matrix L of the graph as $L = D - W$, where D is an $N \times N$ diagonal matrix with $D_{ii} = \sum_{j=1}^N w_{ij}$.

The evidence of label refinement is that visually similar superpixels have higher probability to share the same set of labels. We thus define a Laplacian regularization term as

$$\frac{1}{2} \sum_{i,j=1}^N w_{ij} \|\mathbf{v}_i - \mathbf{v}_j\|_2^2 = \text{tr}(V^T L V) \quad (3)$$

which means that similar superpixels should share common combination of predefined classes. In addition, we define a L_1 -norm fitting constraint term as

$$\|V - \tilde{Y}\|_1 \quad (4)$$

which can perform direct noise reduction over \tilde{Y} due to L_1 -norm regularization (Elad and Aharon 2006; Mairal, Elad, and Sapiro 2008; Wright et al. 2009).

Joint Feature and Label Refinement

Now the features and labels of superpixels can be jointly refined by considering Eqs. (2-4) together. Specifically, the

objective function of joint feature and label refinement is defined as follows:

$$\mathcal{F}(U, V, F) = \|X - UV^T\|_F^2 + \lambda_1 \text{tr}(V^T LV) + \lambda_2 \|V - F\|_F^2 + \lambda_3 \|F - \tilde{Y}\|_1 \quad (5)$$

where λ_1 , λ_2 and λ_3 are positive hyperparameters, and $U \geq 0, V \geq 0, F \geq 0$. Here, we define F as an intermediate representation of V for fast optimization. Since the above objective function is not convex over U, V and F simultaneously, we will develop an efficient algorithm for solving $\min_{U, V, F} \mathcal{F}(U, V, F)$ in Section 4.

Efficient FeaBoost Algorithm

To minimize the objective function $\mathcal{F}(U, V, F)$ in Eq. (5), we divide it into two subfunctions:

$$\mathcal{F}_1(U, V, F) = \|X - UV^T\|_F^2 + \lambda_1 \text{tr}(V^T LV) + \lambda_2 \|V - F\|_2^2 \quad (6)$$

$$\mathcal{F}_2(V, F) = \lambda_2 \|V - F\|_F^2 + \lambda_3 \|F - \tilde{Y}\|_1 \quad (7)$$

The optimization problem $\min_{U, V, F} \mathcal{F}(U, V, F)$ can be solved in two alternate steps as follows:

$$U^*, V^* = \arg \min_{U, V} \mathcal{F}_1(U, V, F^*) \quad (8)$$

$$F^* = \arg \min_F \mathcal{F}_2(V^*, F) \quad (9)$$

with F^* being initialized as \tilde{Y} . We first focus on how to minimize $\mathcal{F}_1(U, V, F^*)$, which can be rewritten as:

$$\mathcal{F}_1 = \text{tr}((X - UV^T)(X - UV^T)^T) + \lambda_1 \text{tr}(V^T LV) + \lambda_2 \text{tr}((V - F^*)(V - F^*)^T) \quad (10)$$

$$\begin{aligned} &= \text{tr}(XX^T) - 2\text{tr}(XVU^T) + \text{tr}(UV^T VU^T) \\ &\quad + \lambda_1 \text{tr}(V^T LV) + \lambda_2 \text{tr}(VV^T) - 2\lambda_2 \text{tr}(VF^{*T}) \\ &\quad + \lambda_2 \text{tr}(F^* F^{*T}) \end{aligned} \quad (11)$$

Updating U, V : Inspired by (Cai et al. 2011), let $A = \{\alpha_{ik}\}_{M \times C}$ and $B = \{\beta_{jk}\}_{N \times C}$ denote the Lagrange multipliers for $u_{ik} \geq 0$ and $v_{jk} \geq 0$, and the Lagrange \mathcal{L}_1 can be written as:

$$\begin{aligned} \mathcal{L}_1 &= \text{tr}(XX^T) - 2\text{tr}(XVU^T) + \text{tr}(UV^T VU^T) \\ &\quad + \lambda_1 \text{tr}(V^T LV) + \lambda_2 \text{tr}(VV^T) - 2\lambda_2 \text{tr}(VF^{*T}) \\ &\quad + \lambda_2 \text{tr}(F^* F^{*T}) + \text{tr}(AU^T) + \text{tr}(BV^T) \end{aligned} \quad (12)$$

The partial derivatives of \mathcal{L}_1 with respect to U and V are:

$$\frac{\partial(\mathcal{L}_1)}{\partial(U)} = -2XV + 2UV^T V + A \quad (13)$$

$$\begin{aligned} \frac{\partial(\mathcal{L}_1)}{\partial(V)} &= -2X^T U + 2VU^T U + 2\lambda_1 LV \\ &\quad + 2\lambda_2 V - 2\lambda_2 F^* + B \end{aligned} \quad (14)$$

Considering the KKT conditions $\alpha_{ik} u_{ik} = 0$ and $\beta_{jk} v_{jk} =$

Algorithm 1: FeaBoost

Input: The feature matrix X , the initial labels of superpixels \tilde{Y} , the Laplacian matrix L

Output: The refined labels V^*

```

1 Initialize  $F^* = \tilde{Y}$ ;
2 repeat
3   repeat
4      $u_{ik} \leftarrow u_{ik} \frac{(XV)_{ik}}{(UV^T V)_{ik}},$ 
5      $v_{jk} \leftarrow v_{jk} \frac{(X^T U + \lambda_1 W V + \lambda_2 F^*)_{jk}}{(VU^T U + \lambda_1 D V + \lambda_2 V)_{jk}};$ 
6   until The best  $U^*$  and  $V^*$  are found;
7    $F^* = \text{soft\_thr}(V^*, \tilde{Y}, \lambda_3/\lambda_2)$ 
8 until The convergence criterion is satisfied;
9 return  $V^*$ ;
```

0, we can get the following equations

$$-(XV)_{ik} u_{ik} + (UV^T V)_{ik} u_{ik} = 0 \quad (15)$$

$$-(X^T U)_{jk} v_{jk} + (VU^T U)_{jk} v_{jk} + \lambda_1 (LV)_{jk} v_{jk} + \lambda_2 V_{jk} v_{jk} - \lambda_2 F_{jk}^* v_{jk} = 0 \quad (16)$$

Now the best U^* and V^* can be obtained using the following updating rules:

$$u_{ik} \leftarrow u_{ik} \frac{(XV)_{ik}}{(UV^T V)_{ik}} \quad (17)$$

$$v_{jk} \leftarrow v_{jk} \frac{(X^T U + \lambda_1 W V + \lambda_2 F^*)_{jk}}{(VU^T U + \lambda_1 D V + \lambda_2 V)_{jk}} \quad (18)$$

Updating F : As a basic L_1 -norm optimization problem, Eq.(9) has an explicit solution as:

$$F^* = \text{soft_thr}(V^*, \tilde{Y}, \gamma) \quad (19)$$

where $\gamma = \lambda_3/\lambda_2$. Here, $z = \text{soft_thr}(x, y, \gamma)$ is a soft-thresholding function:

$$z = \begin{cases} z_1 = \max(x - \gamma, y) & , f_1 \leq f_2 \\ z_2 = \max(0, \min(x + \gamma, y)) & , f_1 > f_2 \end{cases}$$

where $f_1 = (z_1 - x)^2 + 2\gamma|z_1 - y|$ and $f_2 = (z_2 - x)^2 + 2\gamma|z_2 - y|$. The complete FeaBoost algorithm for semantic segmentation is outlined in Algorithm 1.

Oversegmentation and Feature Extraction

In Section 3, we assume that all the images have been oversegmented into multiple superpixels in advance. In this section, we will introduce how to oversegment each image into multiple superpixels and then extract the features of each superpixel for our FeaBoost algorithm.

The Blobworld method (Carson et al. 2002) is first adopted to group pixels within an image into superpixels. Specifically, we extract a 6-dimensional feature vector of color and texture for each pixel, and then model each image as a Gaussian mixture model. During grouping all the pixels, the number of superpixels can be automatically detected

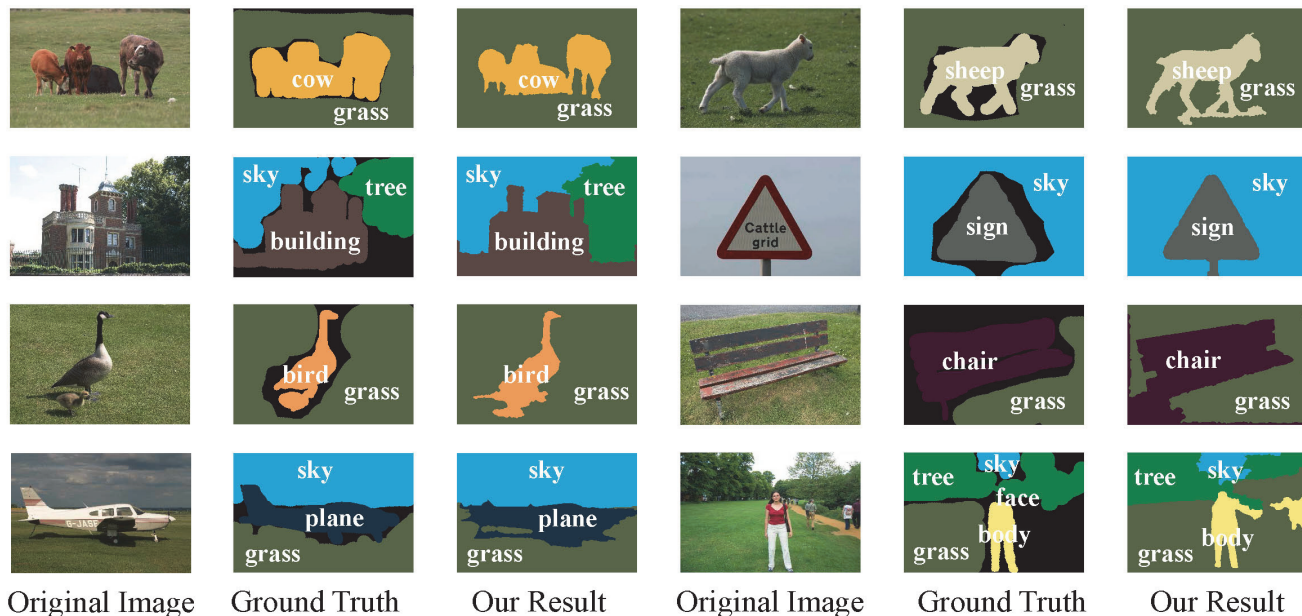


Figure 3: Example results obtained by our approach on the MSRC dataset. Black regions are unlabeled in ground truth.

by a model selection principle. To ensure the oversegmentation, the original Blobworld method is slightly modified: 1) the number of superpixels is initially set to a relatively large value; 2) model selection is made to be less important during oversegmentation. After oversegmentation, we will focus on feature extraction over superpixels.

For each superpixel, we extract the local and global R-CNN features (Girshick et al. 2014) just as (Xu, Schwing, and Urtasun 2015). Specifically, we first capture a 8,192-dimensional feature of each superpixel within the bounding box as well as the masked box, containing local texture and shape information. In addition, the bounding box is expanded to the whole image as well as the masked box to take the global context, superpixel size and location information into consideration. This gives another 8,192-dimensional feature vector. Totally, we obtain a 16,384-dimensional feature vector for each superpixel.

Experimental Evaluation

In this section, the performance of our approach is evaluated on two datasets: MSRC (Shotton et al. 2006) and LabelMe (Liu, Yuen, and Torralba 2011). Two group of experiments are conducted for performance evaluation: semantic segmentation with predicted labels, and semantic segmentation with noisy labels.

Experimental Setup

The MSRC dataset consists of 591 images and 21 classes. This dataset is standardly split into 276 training images and 256 test images. Experiments are also conducted on a more challenging dataset - LabelMe (also called as SIFT Flow). The LabelMe dataset contains 2,688 outdoor images with 33 outdoor classes, including sky, tree, grass, building,

Table 1: Comparison to the state-of-the-art methods for semantic segmentation on the MSRC dataset.

Methods	Supervision	Accuracy (%)
(Shotton et al. 2008)	FS	67
(Russell et al. 2009)	FS	75
(Lucchi et al. 2012)	FS	76
(Yao et al. 2012)	FS	79
(Zhang et al. 2013)	WS	69
(Liu et al. 2013)	WS	71
(Xie et al. 2014)	WS	73
(Niu et al. 2015)	WS	71
(Xu et al. 2015)	WS	73
Ours (FeaBoost)	WS	77
Ours (FeaBoost+SVM)	WS	78

and the standard split of 2,488 training images and 200 test images is used for this dataset. We totally obtain about 7,000 superpixels for the MSRC dataset and 33,000 superpixels for the LabelMe dataset by oversegmentation.

It should be noted that only image-level labels of the training set are known, and *the annotations of the test images are unseen* during the training and test stages. To infer the annotations of the test images, a 4,096-dimensional CNN feature is extracted from each image, and C one-vs-all SVM classifiers are trained using LIBLINEAR for prediction. Since *the pixel-level labels are unknown under the weakly supervised setting*, it is impossible to select the hyperparameters by cross-validation. In this paper, the hyperparameters are uniformly set as $k = 30$, $\lambda_1 = 900$, $\lambda_2 = 1$ and $\lambda_3 = 0.15$ for the two datasets.

Table 2: Comparison to the state-of-the-art methods for semantic segmentation on the LabelMe dataset.

Methods	Supervision	Accuracy (%)
(Liu et al. 2011)	FS	24
(Myeong et al. 2012)	FS	32
(Tighe et al. 2014)	FS	39
(Yang et al. 2014)	FS	49
(Long et al. 2015)	FS	51
(Vezhnevets et al. 2012)	WS	22
(Liu et al. 2013)	WS	26
(Niu et al. 2015)	WS	33
(Li et al. 2015)	WS	41
(Xu et al. 2015)	WS	41
Ours (FeaBoost)	WS	39
Ours (FeaBoost+SVM)	WS	42

Semantic Segmentation with Predicted Labels

An overview of the semantic segmentation results is given in comparison with the state-of-the-art in Tables 1 and 2, and some example results are also shown in Figure 3. Here, two main settings are considered: FS means the fully supervised setting with pixel-level labels, and WS means the weakly supervised setting with image-level labels. The typical quantitative measure is average per-class accuracy, i.e., computing the percentage of correctly classified pixels for each class and then averaging over all the classes. For fair comparison, the released results of the compared methods are directly cited. In addition, to further improve our FeaBoost, C one-vs-all SVM classifiers are trained using LIBLINEAR with the refined labels of superpixels.

We first make observations on MSRC according to Table 1. Our approach is shown to achieve at least 5% relative improvements over the other weakly supervised methods (Zhang et al. 2013; Liu et al. 2013; Xie, Peng, and Xiao 2014; Niu et al. 2015; Xu, Schwing, and Urtasun 2015), and even perform better than most of fully supervised methods (Shotton, Johnson, and Cipolla 2008; Russell et al. 2009; Lucchi et al. 2012). On the more challenging LabelMe dataset, our approach is also shown to be competitive to the state-of-the-art (Li et al. 2015; Xu, Schwing, and Urtasun 2015) (see Table 2). Note that only image-level labels of the training set are taken as weakly supervision, and infer the annotations of the test set, different from (Li et al. 2015) that exploits the ground-truth labels of all the images for semantic segmentation.

Semantic Segmentation with Noisy Labels

To verify the robustness of our approach, extensive experiments are further conducted on the LabelMe dataset with noisy image-level labels. Specifically, we randomly select $p \in \{20, 40, \dots, 100\}$ percents of images and add $r \in \{1, 2, 3\}$ random labels to each selected image, which is mostly related to (Li et al. 2015; Niu et al. 2015). Since each image in the LabelMe dataset is annotated with average 4.4 labels according to the ground-truth, the randomly added labels means strong noise in semantic segmentation.

Table 3: Comparison with different percentages p of images with r noisy tags per-image on the LabelMe dataset.

Methods	r	Noise ratio p (%)					
		0	20	40	60	80	100
Ours	1	47	43	42	41	38	35
	2	47	42	36	33	29	24
	3	47	37	35	32	26	20
(Li et al. 2015)	1	41	40	37	34	32	30
	2	41	36	33	30	25	21
	3	41	35	31	27	22	18
(Niu et al. 2015)	1	33	31	30	29	28	27
	2	33	30	28	26	24	21
	3	33	28	26	23	22	18

For fair comparison, we exploit the ground-truth labels of all the images for the clean setting ($p = 0$), just as (Li et al. 2015; Niu et al. 2015). The comparison results are presented in Table 3, and we can make several observations as follows. Firstly, our approach achieves 47% accuracy with none noisy tags, 6% higher than (Li et al. 2015) and 14% higher than (Niu et al. 2015). This observation gain verifies the effectiveness of our approach in the clean setting. Secondly, with the noise ratio increasing, all the three methods encounter the performance degradation as expected, but our approach always performs the best in all cases. This observation confirms that our approach is the most effective in noise reduction. Thirdly, our approach achieves 35% accuracy with one noisy label added into all the images, even better than (Niu et al. 2015) with none noisy labels. Finally, our approach with 3 noisy tags even performs slightly better than (Li et al. 2015) with 2 noisy tags under different noise ratios. These observations show that our approach can make a good balance between nice semantic segmentation and effective noise reduction.

Conclusion

In this paper, we propose a novel framework FeaBoost for image semantic segmentation under weakly supervised setting. A new approach is provided to semantic segmentation by jointly refining the features and labels of superpixels. The extensive experiments on two benchmark datasets demonstrate the promising performance of our approach. In addition, the experimental results with noisy labels show the robustness of our approach. In the future work, we will extend our approach to other challenging problems (e.g., image annotation) for joint feature and label refinement.

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), the Outstanding Innovative Talents Cultivation Funded Programs 2016 of Renmin University of China, and IBM Global SUR Award Program.

References

- Arbelaez, P.; Hariharan, B.; Gu, C.; Gupta, S.; Bourdev, L.; and Malik, J. 2012. Semantic segmentation using regions and parts. In *CVPR*, 3378–3385.
- Cai, D.; He, X.; Han, J.; and Huang, T. S. 2011. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(8):1548–1560.
- Carreira, J.; Caseiro, R.; Batista, J.; and Sminchisescu, C. 2012. Semantic segmentation with second-order pooling. In *ECCV*, 430–443.
- Carson, C.; Belongie, S.; Greenspan, H.; and Malik, J. 2002. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(8):1026–1038.
- Elad, M., and Aharon, M. 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing* 15(12):3736–3745.
- Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9(Aug):1871–1874.
- George, M. 2015. Image parsing with a wide range of classes and scene-level context. In *CVPR*, 3622–3630.
- Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 580–587.
- Kohli, P.; Torr, P. H.; et al. 2009. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision* 82(3):302–324.
- Ladicky, L.; Russell, C.; Kohli, P.; and Torr, P. H. 2010. Graph cut based inference with co-occurrence statistics. In *ECCV*, 239–253.
- Li, Y.; Liu, J.; Wang, Y.; Lu, H.; and Ma, S. 2015. Weakly supervised RBM for semantic segmentation. In *IJCAI*, 1888–1894.
- Liu, Y.; Liu, J.; Li, Z.; Tang, J.; and Lu, H. 2013. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR*, 2075–2082.
- Liu, C.; Yuen, J.; and Torralba, A. 2011. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(12):2368–2382.
- Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.
- Lucchi, A.; Li, Y.; Smith, K.; and Fua, P. 2012. Structured image segmentation using kernelized features. In *ECCV*, 400–413.
- Mairal, J.; Elad, M.; and Sapiro, G. 2008. Sparse representation for color image restoration. *IEEE Transactions on Image Processing* 17(1):53–69.
- Niu, Y.; Lu, Z.; Huang, S.; Han, P.; and Wen, J.-R. 2015. Weakly supervised matrix factorization for noisily tagged image parsing. In *IJCAI*, 3749–3755.
- Russell, C.; Kohli, P.; Torr, P. H.; et al. 2009. Associative hierarchical CRFs for object class image segmentation. In *ICCV*, 739–746.
- Shotton, J.; Winn, J.; Rother, C.; and Criminisi, A. 2006. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 1–15.
- Shotton, J.; Johnson, M.; and Cipolla, R. 2008. Semantic texton forests for image categorization and segmentation. In *CVPR*, 1–8.
- Tang, J.; Yan, S.; Hong, R.; Qi, G.-J.; and Chua, T.-S. 2009. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM Multimedia*, 223–232.
- Tighe, J., and Lazebnik, S. 2010. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*, 352–365.
- Verbeek, J., and Triggs, B. 2007. Region classification with Markov field aspect models. In *CVPR*, 1–8.
- Vezhnevets, A., and Buhmann, J. M. 2010. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, 3249–3256.
- Vezhnevets, A.; Ferrari, V.; and Buhmann, J. M. 2011. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 643–650.
- Vezhnevets, A.; Ferrari, V.; and Buhmann, J. M. 2012. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 845–852.
- Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2):210–227.
- Xie, W.; Peng, Y.; and Xiao, J. 2014. Semantic graph construction for weakly-supervised image parsing. In *AAAI*, 2853–2859.
- Xu, J.; Schwing, A. G.; and Urtasun, R. 2014. Tell me what you see and I will show you where it is. In *CVPR*, 3190–3197.
- Xu, J.; Schwing, A. G.; and Urtasun, R. 2015. Learning to segment under various forms of weak supervision. In *CVPR*, 3781–3790.
- Yang, J.; Price, B.; Cohen, S.; and Yang, M.-H. 2014. Context driven scene parsing with attention to rare classes. In *CVPR*, 3294–3301.
- Yang, L.; Meer, P.; and Foran, D. J. 2007. Multiple class segmentation using a unified framework over mean-shift patches. In *CVPR*, 1–8.
- Zhang, K.; Zhang, W.; Zheng, Y.; and Xue, X. 2013. Sparse reconstruction for weakly supervised semantic segmentation. In *IJCAI*, 1889–1895.