Weakly Supervised Semantic Segmentation for Social Images

Wei Zhang, Sheng Zeng, Dequan Wang, and Xiangyang Xue
Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University, Shanghai, China
{weizh, zengsheng, dqwang12, xyxue}@fudan.edu.cn

Abstract

Image semantic segmentation is the task of partitioning image into several regions based on semantic concepts. In this paper, we learn a weakly supervised semantic segmentation model from social images whose labels are not pixel-level but image-level; furthermore, these labels might be noisy. We present a joint conditional random field model leveraging various contexts to address this issue. More specifically, we extract global and local features in multiple scales by convolutional neural network and topic model. Inter-label correlations are captured by visual contextual cues and label co-occurrence statistics. The label consistency between image-level and pixel-level is finally achieved by iterative refinement. Experimental results on two real-world image datasets PASCAL VOC2007 and SIFT-Flow demonstrate that the proposed approach outperforms state-of-the-art weakly supervised methods and even achieves accuracy comparable with fully supervised methods.

1. Introduction

Semantic segmentation, *i.e.*, parsing image into several semantic regions, assigns each pixel (or superpixel) to one of the predefined semantic categories. Most state-of-the-art methods rely on a sufficiently huge amount of annotated samples in training. However, there are not enough labeled samples for this task because pixel-level (or superpixel-level) annotation is time-consuming and labor-intensive. Recent works have begun to address the semantic segmentation problem in the weakly supervised settings, where each training image is only annotated by image-level labels [24, 25, 26, 27, 30, 33, 34]. The existing weakly supervised semantic segmentation methods are based on one strict assumption that image-level labels are guaranteed to be precise by professional annotators.

With the prevalence of photo sharing websites and collaborative image tagging system, *e.g.*, Flickr, a large number of social images with user provided labels are available from the Internet. These labels are usually image-level; what's more, they might be noisy: There are either incorrect additional labels assigned to a training image or labels missing from the ground truth. Figure 1 shows several social images and the associated noisy labels. It is challenging but attractive to learn an effective semantic segmentation model from such social images.

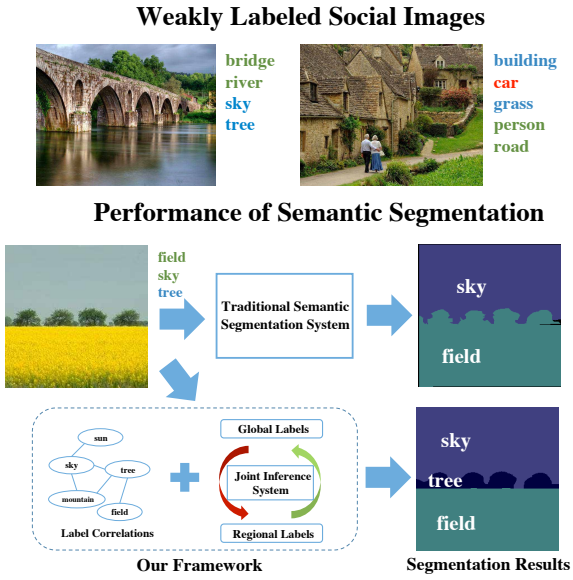

Figure 1. Several social images and the associated noisy labels which may be correct (green), incorrect (red) or missing (blue). We learn a joint model to simultaneously segment and recognize visual concept in images. Best viewed in color.

In this paper, we propose a weakly supervised semantic segmentation model to overcome the challenge posed by noisy image-level labels for training. We learn a joint conditional random field (CRF) from weakly labeled social images by sufficiently leveraging various contexts, *e.g.*, the associations between high-level semantic concepts and low-level visual appearance, inter-label correlations, spatial neighborhoods, and label consistency between image-level and pixel-level. More specifically, each image is segmented into superpixels with multiple quantization levels. Global

features for the whole image and local features for the superpixels in multiple scales are extracted by convolutional neural network (CNN) and latent semantic concept model (LSC). Then we capture the inter-label correlations by visual contextual cues as well as label co-occurrence statistics. The label consistency between image-level and pixel-level is finally achieved by iterative refinement in a flip-flop manner. We conduct experiments on two challenging datasets, PASCAL VOC 2007 and SIFT-Flow datasets. The proposed approach achieves comparable results or outperforms previous state-of-the-art methods, even though it is in the weakest supervision, which demonstrates that the image-level labels, especially potential relationships, are more efficiently utilized by our method.

The main contributions of this paper are summarized as follows:

- We propose a weakly supervised semantic segmentation model for social images, where only image-level labels are available for training, or even worse, the annotations can be noisy.
- We design a joint learning framework to sufficiently leverage various contexts including feature-label association, inter-label correlation, spatial neighborhood cues, and label consistency.
- We learn inter-label correlation not only by investigating label co-occurrence statistics from training samples but also by looking at the overlap of the most informative regions for different classes.

2. Related Works

In the past years, image semantic segmentation has attracted a lot of attentions. Most of the existing works model the task as a fully supervised problem [32]. Shotton *et al*. [19] implemented semantic segmentation by incorporating shape-texture color, location and edge clues in a CRF model over image pixels. This model is then extended in the follow-up works [10, 12, 13]. Kohli *et al*. utilized the higher order potentials as a soft decision to ensure that pixels constituting a particular segment have the same semantic concept [10]. Ladicky *et al*. extended the higher order potentials to hierarchical structure by using multiple segmentations in [12] and further integrated label co-occurrence statistics in [13]. However, these methods heavily rely on pixel-level annotations during the training stage.

In addition to fully supervised semantic segmentation, there have been several works in the weakly supervised settings as well recently. The method in [31] attempted to automatically annotate image regions by learning a correlative multi-label multi-instance model from image-level tagged data. Verbeek and Triggs [24] used several appearance descriptors to learn the latent aspect model via probabilistic Latent Semantic Analysis (pLSA) [8], and integrated the spanning tree structure and Markov Random Fields to capture spatial information. Vezhnevets and Buhmann [25] cast the weakly supervised task as a multi-instance multi-task learning problem with the framework of Semantic Texton Forest (STF) [18]. Based on [25], Vezhnevets *et al*. [26, 27] integrated the latent correlations among the superpixels belonging to different images which share the same labels into CRF. Xu *et al*. [30] simplified the previous complicated framework by a graphical model that encodes the presence/absence of a class as well as the assignments of semantic labels to superpixels. [33] performed semantic segmentation in weak supervision via classifier evaluation where the classifier parameters are firstly sampled at random and then the superpixel classifiers are evaluated by measuring the distance between the ground-truth negative samples and the predicted positive samples. It should be pointed out that all above approaches are based on the assumption that the given image-level labels for training are correct and complete, which is not practical in many real-world applications. It is a realistic problem where the end goal is pixel-level labels but the input is noisy image-level annotations.

To address the problem of having noise in the ground truth, we investigate label correlations based on both label co-occurrence statistics and visual contextual cues simultaneously, which differs from the existing weakly supervised methods [24, 25, 26, 27, 30]. In addition, to make the proposed framework more robust under the noisy condition, we take latent semantic concept model as a mid-level representation, which also helps to narrow down the gap between semantic space and feature space; in contrast, the previous methods (*e.g*., [26, 30]) only used the appearance model as a low-level representation. In comparison with the state-of-the-art weakly supervised methods (*e.g*., [27, 30]), we utilize multiple scale segmentations to overcome the weakness of single choice of segmentation which fails to cover different quantization levels of objects.

3. The Proposed Model

Suppose that each image I is associated with a label vector $\boldsymbol{y} = [y_1, ..., y_L]$, where L is the number of categories, and $y_i = 1$ indicates that the i-th category is present in this image, otherwise $y_i = 0$. In the training set, $\boldsymbol{y}$ is given; however, it might be noisy. In the test set, $\boldsymbol{y}$ is unknown. For each image, we firstly employ the existing multi-scale segmentation algorithm to get a set of superpixels $\{x_p\}_{p=1}^{M}$ over multiple quantization levels. Here, M is the total number of superpixels in image I. The label of superpixel x_p is denoted as $h_p \in \{1, 2, ..., L\}$, and the labels of all superpixels for image I are $\boldsymbol{h} = [h_1, ..., h_M]$, which are not available for training.

Our goal is to infer semantic label for each superpixel in an image and the adjacent superpixels sharing the same

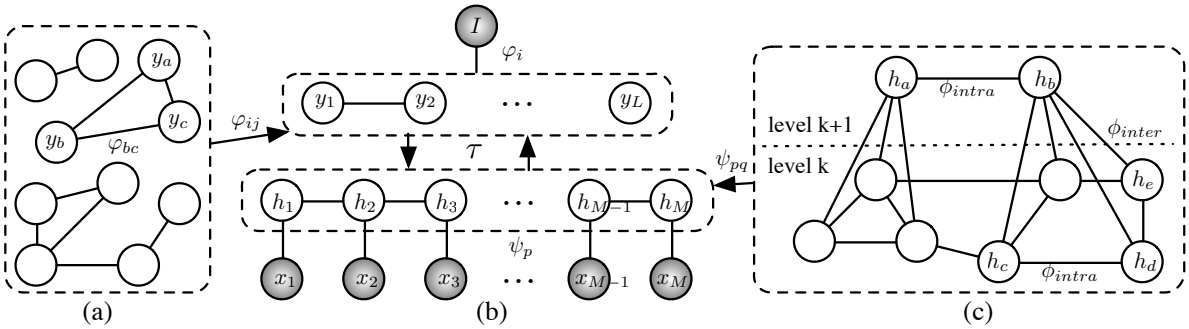

Figure 2. Illustration of the proposed model. (a) Inter-label correlations. (b) Feature-label associations on image-level (top) and superpixel-level (bottom), respectively. (c) Hierarchy of multi-scale segmentations and spatial context constraints for adjacent superpixels.

semantic label are fused as a whole one. We jointly build a conditional random field (CRF) over the image-level label variables $\boldsymbol{y}$ and the superpixel-level label variables $\boldsymbol{h}$. We leverage label-pair correlation and connect each superpixel to its neighbors to encode local smoothness constraints. Thus we formulate an energy function E with five types of potentials as follows:

$$\begin{aligned} E(\boldsymbol{y},\boldsymbol{h},\boldsymbol{I}) = &\sum_{i=1}^{L}\varphi_i(y_i,\boldsymbol{I}) + \sum_{1\le i,j\le L}\varphi_{ij}(y_i,y_j) \\ &+ \sum_{p=1}^{M}\psi_p(h_p,\boldsymbol{x_p}) + \sum_{(p,q)\in\mathcal{N}}\psi_{pq}(h_p,h_q) \\ &+ \tau(\boldsymbol{y},\boldsymbol{h}) \end{aligned} \tag{1}$$

where φ_i and ψ_p are the unary potentials for feature-label associations on image-level and superpixel-level respectively, φ_{ij} is the pairwise potential for label correlation, ψ_{pq} is the pairwise potential encoding the spatial context constraints for adjacent superpixels, $\mathcal{N}$ denotes the set of pairs of neighboring superpixels, and τ ensures the coherence between image-level labels and superpixel-level labels.

A graphical illustration of the energy function $E(\boldsymbol{y},\boldsymbol{h},\boldsymbol{I})$ is given in Figure 2, and the details of each potential will be described in the following subsections. The posterior distribution $P(\boldsymbol{y},\boldsymbol{h}|\boldsymbol{I})$ of the CRF can be defined as $P(\boldsymbol{y},\boldsymbol{h}|\boldsymbol{I}) = \frac{1}{Z(\boldsymbol{I})}\exp\{-E(\boldsymbol{y},\boldsymbol{h},\boldsymbol{I})\}$, where $Z(\boldsymbol{I})$ is the normalizing factor. Thus, the most probable labeling configuration $\boldsymbol{y}^\star,\boldsymbol{h}^\star$ of the random field can be obtained as $(\boldsymbol{y}^\star,\boldsymbol{h}^\star) = \arg\min_{\boldsymbol{y},\boldsymbol{h}} E(\boldsymbol{y},\boldsymbol{h},\boldsymbol{I})$.

3.1. Unary Potentials for Feature-Label Associations

Image-Level Potential We extract two kinds of global features for each image. On one hand, we learn 4096 dimensional features $\boldsymbol{d}$ for image $\boldsymbol{I}$ by using the first 16 layers of a 19-layers-deep convolutional neural network (CNN) introduced in [20]. On the other hand, we employ pLSA [21] to model each image as a mixture of latent semantic concepts (LSC) t_k, $(k = 1,...,K)$. We look on each image as a document, and consider each component of the learned CNN features $\boldsymbol{d}$ as a word w_j (j=1,...,4096). Like [29], we solve the conditional probability $P(t_k|\boldsymbol{d})$ of latent semantic concept t_k occurring in image from the equation $P(w_j|\boldsymbol{d}) = \sum_{k=1}^{K} P(t_k|\boldsymbol{d})P(w_j|t_k)$, where $P(w_j|\boldsymbol{d})$ and $P(w_j|t_k)$ are the probabilities of visual word w_j occurring in image represented by $\boldsymbol{d}$ and occurring in the concept t_k, respectively.

Thus we obtain the global feature of each image $\boldsymbol{I}$ by concatenating the appearance feature $\boldsymbol{d}$ and latent semantic concept distribution $P(\boldsymbol{t}|\boldsymbol{d}) = [P(t_1|\boldsymbol{d}),...,P(t_K|\boldsymbol{d})]$, and formulate the image-level potential for feature-label association $\varphi_i, (i = 1,...,L)$, as follows:

$$\varphi_i(y_i,\boldsymbol{I}) = -\ln\frac{\exp\{f_i(y_i,\boldsymbol{I})\}}{\exp\{f_i(0,\boldsymbol{I})\}+\exp\{f_i(1,\boldsymbol{I})\}} \tag{2}$$

where $f_i(y_i,\boldsymbol{I})$ is the linear support vector machine score for the semantic concept i with the $4096+K$ dimensional feature vector $I = [\boldsymbol{d};P(\boldsymbol{t}|\boldsymbol{d})]$. Although the labels of social images for training might be missing or incorrect, the potential for feature-label association is robust due to the features learned by the latent semantic concept model which is unsupervised.

Superpixel-Level Potential Similar with image-level potential, we also extract $4096+K$ dimensional features for each superpixel by simultaneously employing the CNN appearance model and latent semantic concept model, which helps to narrow the semantic gap and to alleviate the impact of noisy training image-level labels. Let $\boldsymbol{x_p} = [\boldsymbol{a_p};\boldsymbol{c_p}]$ be the feature vector concatenating the CNN feature and latent semantic concept distribution extracted from the superpixels. The superpixel-level potential for feature-label association is formulated as follows:

$$\psi_p(l,\boldsymbol{x_p}) = -\ln\frac{\exp\left\{\boldsymbol{a_p}^\top\boldsymbol{\theta_a^l}+\boldsymbol{c_p}^\top\boldsymbol{\theta_c^l}\right\}}{\sum_{i=1}^{L}\exp\left\{\boldsymbol{a_p}^\top\boldsymbol{\theta_a^i}+\boldsymbol{c_p}^\top\boldsymbol{\theta_c^i}\right\}} \tag{3}$$

where $\boldsymbol{\theta}_a^l$ and $\boldsymbol{\theta}_c^l$ $(l = 1, 2, ..., L)$ are the parameters for CNN and LSC features, respectively. The details of learning $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_c$ are given in Section 3.3.

3.2. Pairwise Potentials

Inter-Label Correlation To model the pairwise potential for inter-label correlation, we not only utilize label co-occurrence statistics but also capture visual contextual cues. For instance, since cars usually appear on roads, our model learns this regularity, and then if we see a road in an image then we will expect there may be a car in that image too. Like [13], we firstly leverage co-occurrence statistics from available labels. Let A be the $L \times L$ symmetric matrix whose entry $A(i, j)$ measures the co-occurrence of label pair (i, j) based on training dataset. It is reasonable to formulate $A(i, j)$ as follows:

$$A(i, j) = 1 - (1 - P(i|j))(1 - P(j|i)) \tag{4}$$

where $P(i|j)$ is the empirical probability of concept i occurring under the condition that concept j has occurred.

At the same time, we take advantage of visual contextual cues to learn inter-label correlations as well. Two objects that overlap one another in the same image tend to be correlated. We measure the overlap of two objects i, j by calculating the ratio of Intersection-over-Union (IoU) as follows:

$$R(i, j) = \frac{area(i \cap j)}{area(i \cup j)} \tag{5}$$

where $i \cap j$ and $i \cup j$ are intersection and union of the informative regions of objects i and j, respectively. $area(.)$ is the area of the regions. However, in our weakly supervised settings, the location of each object is not available for training, i.e., the regions of objects i and j are unknown. Inspired by [16], we use sub-windows to mask out different regions in each image and analyze the changes of recognition scores. Masking out a region which contains the concerned object leads to a significant drop in recognition. In this way we obtain a set of sub-windows which probably contain the discriminative region for the object. For each sub-window, we get its normalized score by calculating the ratio of the drop in score to the area of the sub-window. Finally we choose the sub-window whose absence causes the largest drop in normalized score as the center, select other sub-windows surrounding it, and generate a bounding box which covers all these sub-windows. The bounding box is then viewed as the informative region of the object.

For each pair of labels (i, j), $R(i, j)$ is averaged on the training data and normalized to $[0, 1]$. The label correlation potential φ_{ij} can be defined as follows:

$$\varphi_{ij}(y_i, y_j) = A(i, j)R(i, j)\mathbf{1}(y_i \neq y_j) \tag{6}$$

where $A(i, j)$ and $R(i, j)$ capture label correlations by label co-occurrence statistics and visual contextual cues, respectively, and $\mathbf{1}(\cdot)$ is the indicator function.

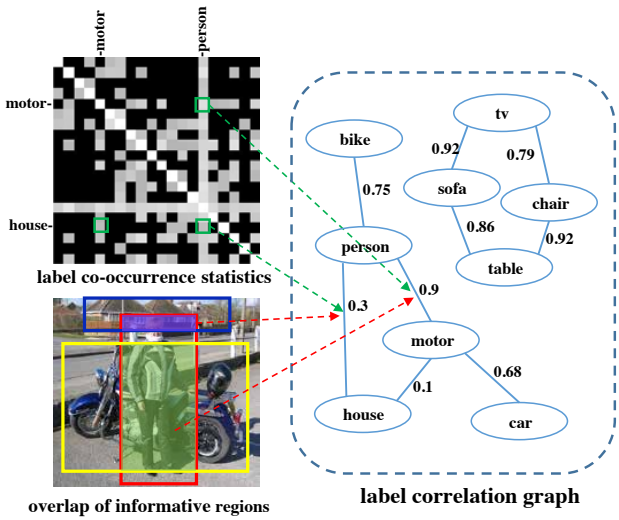

Figure 3. Illustration of the pairwise potential for label correlations which are leveraged from two aspects: label co-occurrence statistics and visual contextual cues.

Figure 3 illustrates the pairwise potential for label correlations. The top left visualizes the matrix A displaying the co-occurrence between concepts. The brighter the block is, the stronger the co-occurrence probability is. The bottom left gives an example of visual contextual clues, known as overlapping area of discriminative regions. The larger the overlap is, the closer the relationship between labels is. $Person$, $motor$ and $house$ are three annotated semantic concepts in this image, whose discriminative region is marked as bounding box in different colors. The large overlap between $motor$ and $person$ strongly suggests the close relationship between these two concepts. The graph on the right side shows the label correlation that integrates both cues. The interdependency between concepts is evaluated on the edge. The larger the value is, the higher the correlation is.

Although the given labels of social images might be noisy, label co-occurrence statistics on the dataset still makes sense. Moreover, visual contextual cues based on the overlap of different objects are learned without using any ground-truth superpixel-level labels for training. Due to the visual context containing some useful latent semantic information, the learned label correlations are immune against the impact of noisy labels.

Pairwise Potential for Superpixels Since there is not a common choice of quantization of an image space for all object categories, it is more suitable to segment one image at different levels of the quantization hierarchy [12]. As illustrated in Figure 2(c), we focus on adjacent superpixels in the same quantization level and overlapped superpixels in the neighboring levels, and define the pairwise potentials

for superpixels as follows:

$$\psi_{pq}(h_p, h_q) = \begin{cases} \phi_{inter}(h_p, h_q) & \text{if } |lev(p) - lev(q)| = 1, \\ \phi_{intra}(h_p, h_q) & \text{if } lev(p) = lev(q), \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $lev(p)$ indicates the quantization level for superpixel x_p. $|lev(p) - lev(q)| = 1$ indicates that superpixels x_p and x_q are from the the neighboring levels of the quantization hierarchy. Since superpixels lying within the same clique are more likely to take the same label [10, 12], the inter-level energy cost ϕ_{inter} can be formulated as:

$$\phi_{inter}(h_p, h_q) = \gamma \mathbf{1}(h_p \neq h_q) area(x_p \cap x_q) \tag{8}$$

where $area(x_p \cap x_q)$ refers to the area of intersection (overlapping region) of two superpixels, $\mathbf{1}(\cdot)$ is the indicator function and γ is the weighting coefficient. ϕ_{inter} can be used to find the proper segmentation scale for each object. As for the intra-level energy cost ϕ_{intra}, it is formulated as:

$$\phi_{intra}(h_p, h_q) = (1 - R(h_p, h_q))S(x_p, x_q) \tag{9}$$

where $S(x_p, x_q)$ measures the visual similarity between superpixels x_p and x_q, $R(h_p, h_q) \in [0, 1]$ is the inter-label correlation defined in Eq.(5). The penalty is large in case similar superpixels are assigned irrelevant labels. Hence, ϕ_{intra} encodes the spatial context constraints for adjacent superpixels, which helps to reduce superpixel noise and smooth the object boundaries.

Label Consistency It is naturally required that superpixel-level labels should be consistent with image-level labels: if any superpixel x_p takes the label i, then the image label indicator $y_i = 1$; otherwise $y_i = 0$. Such constraints can be encoded by the following potential:

$$\tau(\boldsymbol{y}, \boldsymbol{h}) = C \sum_{i,p} \mathbf{1}(y_i = 0 \wedge h_p = i) \tag{10}$$

where $\mathbf{1}(\cdot)$ is the indicator function and C is a cost that penalizes any inconsistency between the image-level and superpixel-level labels. Such label consistency potential is a soft constraint, and we can further refine superpixel label and image label via an iterative process.

3.3. Model Parameters Learning

Like [26], we scale the pairwise potentials so as to make them comparable with unary potentials. After selecting the weights of each potential, we learn the parameters of superpixel-level potentials ψ_p for feature-label associations in Eq.(3). The model parameters $\boldsymbol{\theta_a}$ for CNN features and $\boldsymbol{\theta_c}$ for LSC features can be learned via iteratively solving the optimization problem in an alternating manner: 1) Fix $\boldsymbol{h}$ and learn $\boldsymbol{\theta_a}$, $\boldsymbol{\theta_c}$; 2) Fix $\boldsymbol{\theta_a}$, $\boldsymbol{\theta_c}$ and infer $\boldsymbol{h}$. The first step corresponds to a continues optimization problem, hence the optimal $\boldsymbol{\theta_a}$ and $\boldsymbol{\theta_c}$ can be estimated efficiently via the existing supervised methods (*e.g.*, [19]). The second step is a discrete optimization problem, and there are some approximate maximum a posteriori (MAP) methods to infer $\boldsymbol{h}$. We provide the details of inference of $\boldsymbol{h}$ in Section 3.4.

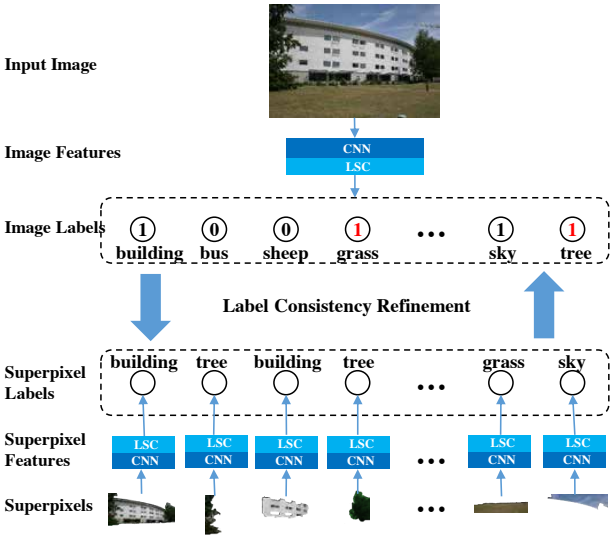

Figure 4. Joint inference of image-level labels and superpixel-level labels in a flip-flop manner.

3.4. Inference of Labels

Given an image I, our task is to assign each pixel a predefined semantic label. The inference algorithm is to search for optimal configuration of image-level label $\boldsymbol{y}^\star$ and superpixel-level label $\boldsymbol{h}^\star$ satisfying $(\boldsymbol{y}^\star, \boldsymbol{h}^\star) = \arg\min_{\boldsymbol{y},\boldsymbol{h}} E(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{I})$.

To efficiently minimize the energy function $E(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{I})$ in Eq.(1), we can iteratively solve the optimization problem in a flip-flop manner:

$$\begin{aligned} \boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} \sum_i \varphi_i(y_i, \boldsymbol{I}) + \frac{1}{2}\tau(\boldsymbol{y}, \boldsymbol{h}^*) \\ + \sum_{1 \leq i,j \leq L} \varphi_{ij}(y_i, y_j), \end{aligned} \tag{11}$$

$$\begin{aligned} \boldsymbol{h}^* = \arg\min_{\boldsymbol{h}} \sum_p \psi_p(h_p, \boldsymbol{x_p}) + \frac{1}{2}\tau(\boldsymbol{y}^*, \boldsymbol{h}) \\ + \sum_{(p,q) \in \mathcal{N}} \psi_{pq}(h_p, h_q). \end{aligned} \tag{12}$$

i.e., one of $\boldsymbol{y}$ and $\boldsymbol{h}$ is optimized while the other is fixed, using Eq.(11) and Eq.(12) alternatively. The joint inference of image-level label $\boldsymbol{y}$ and superpixel-level label $\boldsymbol{h}$ is summa-

rized in Algorithm 1. In this way, we can iteratively refine superpixel labels and image labels, as shown in Figure 4 .

Algorithm 1 Joint Inference of $\boldsymbol{y}$ and $\boldsymbol{h}$

Input: one image I and its superpixels $\{x_p\}$
Output: image-level labels $\boldsymbol{y}$ and superpixel-level labels $\boldsymbol{h}$

1: Initialize $\boldsymbol{y}$ and $\boldsymbol{h}$ with the largest unary potential according to Eq. (2) and (3), respectively.
2: **for** iteration $t = 1$ to T **do**
3: Fix $\boldsymbol{y}$, optimize $\boldsymbol{h}$ via Eq. (12)
4: Fix $\boldsymbol{h}$, refine $\boldsymbol{y}$ via Eq. (11)
5: **end for**
6: Return the final configuration $\boldsymbol{y}$ and $\boldsymbol{h}$.

As a standard binary CRF problem, the first subproblem in Eq.(11) has an explicit solution which utilizes min-cut/max-flow algorithms (*e.g.*, the Dinic algorithm [3]) to obtain the global optimal label configuration. And the second subproblem in Eq.(12) can be reduced to an energy minimization problem for a multi-class CRF. Although seeking the global optimum for this energy function has been shown to be a NP-hard problem, there are various approximate maximum a posteriori (MAP) methods for fast inference, such as *Loopy Belief Propagation, Mean Field Inference, α-Expansion, Linear Programming Relations*. In this paper, we adopt move making approach [2] that finds the optimal α-expansion [2, 11] by converting the problems into binary labeling problems, which can be solved efficiently using graph cuts techniques. The energy obtained by α-expansion has been proved to be within a known factor of the global optimum [2].

4. Experiments

In this section, we evaluate the performance of the proposed approach to weakly supervised image semantic segmentation. In the first experimental setting, we compare the proposed approach with state-of-the-art algorithms on two real-world image datasets. In the second setting, experimental results verify the robustness of our approach under the noisy condition.

We extract 4296 dimensional global features for each image by concatenating appearance feature and latent semantic concept distribution. The appearance feature vector is 4096 dimensional, and is extracted by the first 16 layers of a 19-layers-deep convolutional neural network [20] pretrained on ImageNet [17]. We employ the publicly available implementation *Caffe* [9] to compute the CNN features. The other type of feature represented by latent semantic concept distribution is 200 dimensional, and is learned by pLSA [8]. For unlabeled test images, linear support vector machines [5] are used to obtain the initial image-level labels.

We employ the Multiscale Combinatorial Grouping System [1] to obtain the multi-scale superpixel representation of each image. We use three segmentation scales to generate about 10, 30, 50 superpixels per image respectively. We represent each superpixel by its CNN feature and latent semantic concept distribution, which are extracted in the same way as the image global features .

4.1. Comparison with the State of the Art

We compare the proposed approach with the state-of-the-art weakly supervised semantic segmentation methods as well as fully supervised ones on two challenging datasets: PASCAL VOC 2007 [4] and SIFT-Flow [15].

PASCAL VOC 2007[4] is a publicly available dataset consisting of annotated consumer photographs collected from Flickr photo-sharing web-site. It is challenging for the presence of background clutter, illumination effect and occlusions. It contains 5011 training images, and 4952 test images. Within the dataset, a subset of 632 images are labeled at pixel level, and thus are suitable for evaluation of the segmentation task. We use 422 samples for training and 210 for test. There are 20 foreground and 1 background categories in this dataset.

SIFT-Flow[15] dataset is derived from the LabelMe subset and contains 2688 images of resolution 256x256 pixels, accompanied with a hand labeled segmentation of 33 semantic categories. This dataset is more challenging and has been widely used for semantic segmentation evaluation. There are 4.43 labels per image in average. For fair comparison, we use the standard dataset split (2488 images for training and 200 images for test) as in [15].

Quantitative and Qualitative Results Comparisons of our performances with other methods (both fully supervised and weakly supervised) are given in Tables 1 and 2. We compute the per-class average accuracy defined as $\frac{TruePositives}{TruePositives+FalseNegatives}$ and the mean average. The results on the PASCAL VOC 2007 and SIFT-Flow datasets show that our approach outperforms the state-of-the-art weakly supervised methods, demonstrating that the image-level annotations are more efficiently utilized by our method. In the meantime, the performance of our approach are comparable with the fully supervised method even though we use much less supervised information than these methods. It is worth noting that our approach achieves a promising performance in noisy condition as well, and more details will be discussed in the following section.

Some example results for semantic segmentation by our approach in comparison with the ground-truth on two datasets are shown in Figures 5 and 6, respectively. Not only successful results but also failure cases are given. In Figure 5, the typical failure is due to the cluttered background which shares high visual similarities with the undetected objects. And in Figure 6, the failure is mainly caused

	Methods	average	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tv/monitor
	Brookes	9	78	6	0	0	0	0	9	5	10	1	2	11	0	6	6	29	2	2	0	11	0
Fully	INRIA [7]	24	3	1	45	34	16	20	0	68	58	11	0	44	8	1	2	59	37	0	6	19	63
Supervised	MPI [14]	28	3	30	31	10	41	7	8	73	56	37	11	19	2	15	24	67	26	9	3	5	55
	TKK [28]	30	23	19	21	5	16	3	1	78	1	3	1	23	69	44	42	0	65	30	35	89	71
	UoCTTI [6]	21	3	24	53	0	2	16	49	33	1	6	10	0	0	3	21	60	11	0	26	72	58
Weakly	Zhang *et al*. [33]	24	–	48	20	26	25	3	7	23	13	38	19	15	39	17	18	25	47	9	41	17	33
Supervised	Ours	44.6	75	47	36	65	15	35	82	43	62	27	47	36	41	73	50	36	46	32	13	42	33

Table 1. Accuracies (%) of our approach on VOC2007, in comparison with state-of-the-art methods (fully supervised or weakly supervised). The results of fully supervised methods are reported in [4].

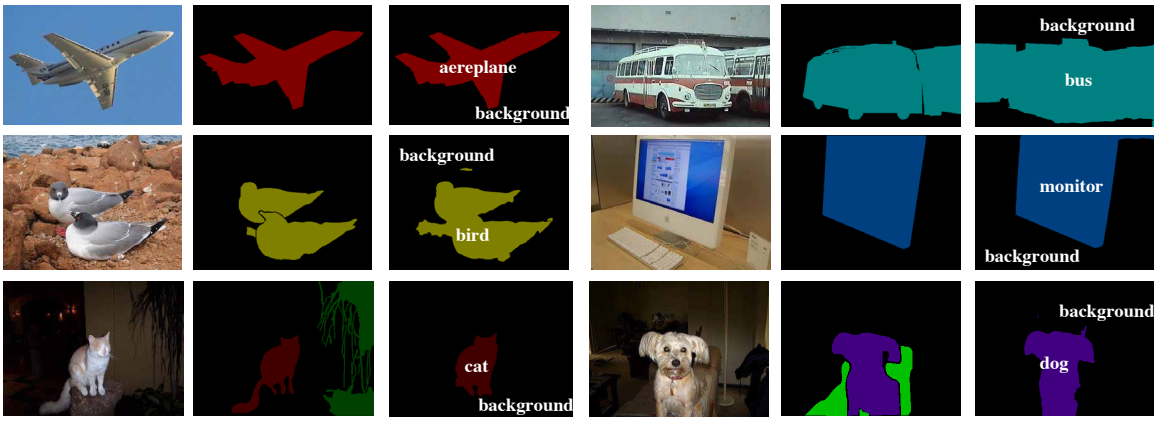

Figure 5. Some example results for semantic segmentation by our approach in comparison with the ground-truth on VOC-2007 dataset. Successful segmentations (top 2 rows) and failure cases (bottom).

by intra-class variability which remains very challenging in computer vision community.

Supervision	Methods	Accuracy (%)
Fully	Liu *et al*. [15]	24
Supervised	Tighe *et al*. [22]	29.4
(pixel-level)	Tighe *et al*. [23]	39.2
	Vezhnevets *et al*. [26]	14
Weakly	Vezhnevets *et al*. [27]	21
Supervised	Xu *et al*. [30]	27.9
(image-level	Zhang *et al*. [34]	27.7
w/o noise)	Ours (0% noise)	32.3
Weakly	Ours (10% noise)	32.8
Supervised	Ours (25% noise)	32.4
(image-level	Ours (50% noise)	29.8
with noise)	Ours (75% noise)	22.3

Table 2. Accuracies (%) of our approach on SIFT-Flow dataset [15], in comparison with state-of-the-art methods (fully supervised or weakly supervised).

4.2. Performance under Noisy Condition

To verify the robustness of our method in noisy annotation condition, we reproduce the real-world noise distribution to the initial image-level labels for SIFT-Flow dataset. For a certain image in the dataset, each image-level label might be missing or replaced by other incorrect labels. Let $P_{miss}(j)$ be the probability of missing the label j, and let $P(i|j)$ be the conditional probability of being annotated as the incorrect label i given that the label j is missing. $P(i|j)$ is empirically estimated from collaborative image tagging system. More specifically, with Flickr API, we query predefined semantic concepts, calculate the number of incorrect labels, and finally compute the normalized $P(i|j)$. By setting different values of $P_{miss}(j)$, we obtain a set of noisy labels as shown in Table 3. From Table 2 and 3, it can be observed that in spite of noisy condition the performance of our approach is still better than or comparable to the state-of-the-art.

In the proposed model, we make use of features extracted by convolutional neural network (CNN) and latent semantic concept distribution (LSC), and leverage label correlations

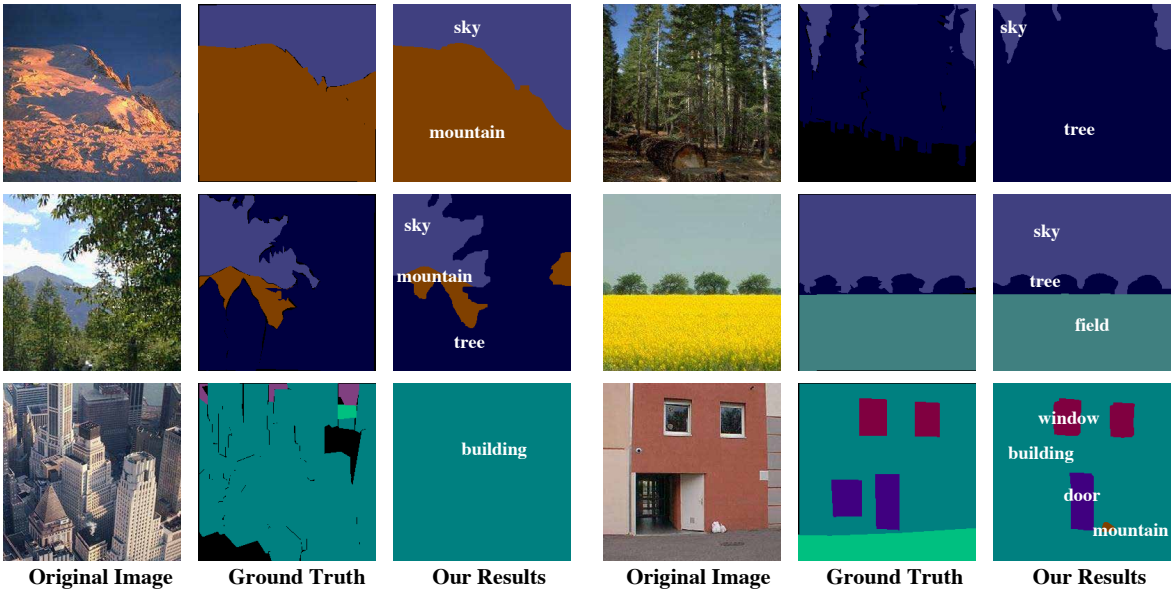

Figure 6. Some example results for semantic segmentation by our approach in comparison with the ground-truth on SIFT-Flow dataset. Successful segmentations (top 2 rows) and failure cases (bottom).

Noise (%)	10	25	50	75
Noisy Labels per Image	1.3	1.4	1.7	2.4
Average Accuracy (%)	32.8	32.4	29.8	22.3

Table 3. Statistics of noisy labels on SIFT-Flow dataset and the accuracy of our approach to semantic segmentation in different noisy conditions.

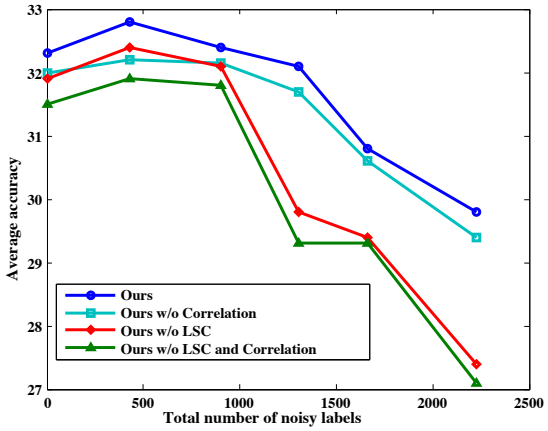

Figure 7. Evaluation of different potentials contributing to the overall performance on SIFT-Flow dataset in noisy conditions.

to encode the pairwise potentials. To investigate the contributions of different potentials to the overall performance, we evaluate several degenerated versions of our method: 1) without label correlations, 2) without LSC, 3) without LSC and label correlations, as shown in Figure 7. It illustrates the performance degradation caused by removing LSC or ignoring label correlations, and demonstrates the indispensability of these parts in our system under different noisy conditions.

5. Conclusions

In this paper, we propose a semantic segmentation algorithm that is trained from image-level labels instead of pixel-level labels and can handle noisy labels. We take advantage of a unified conditional random field to incorporate various contextual relations such as the associations between semantic concepts and visual appearance, label correlations, spatial neighborhood clues, and label consistency between image-level and pixel-level. Visual features are extracted by deep convolutional neural network and latent semantic concept distribution. Label correlations are learned by simultaneously exploiting how often two labels co-occur in the same image and what pairs of labels usually overlap. Experimental results on two real-world image datasets PASCAL VOC2007 and SIFT-Flow demonstrate that the proposed approach outperforms most of the existing methods and achieves a promising performance in noisy condition as well.

Acknowledgments

We would like to thank anonymous reviewers who gave us useful comments. This work was supported by Natural Science Foundation of China (No.61473091).

References

[1] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001.

[3] E. Dinits. Algorithm of solution to problem of maximum flow in network with power estimates. *Doklady Akademii Nauk SSSR*, 1970.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 2008.

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[7] V. Ferrari, L. Fevrier, C. Schmid, F. Jurie, et al. Groups of adjacent contour segments for object detection. *PAMI*, 2008.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, 2014.

[10] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009.

[11] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *PAMI*, 2004.

[12] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *CVPR*, 2009.

[13] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*. 2010.

[14] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008.

[15] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 2011.

[16] B. Loris, B. Alessandro, A. Dragomir, and T. Lorenzo. Self-taught object localization with deep networks. In *arXiv:1409.3964v2 [cs.CV] 24 Nov 2014*.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.

[18] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[19] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*. 2006.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[22] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*. 2010.

[23] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[24] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, 2007.

[25] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, 2010.

[26] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 2011.

[27] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012.

[28] V. Viitaniemi and J. Laaksonen. Techniques for image classification, object detection and object segmentation. In *VISUAL*, 2008.

[29] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*. 2014.

[30] J. Xu, A. G. Schwing, and R. Urtasun. Tell me what you see and i will show you where it is. In *CVPR*, 2014.

[31] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu. Correlative multi-label multi-instance image annotation. In *ICCV*. 2011.

[32] K. Zhang, W. Zhang, S. Zeng, and X. Xue. Semantic segmentation using multiple graphs with block-diagonal constraints. In *AAAI*. 2014.

[33] K. Zhang, W. Zhang, Y. Zheng, and X. Xue. Sparse reconstruction for weakly supervised semantic segmentation. In *IJCAI*, 2013.

[34] 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.