Weakly Supervised Semantic Segmentation for Social Images

Anonymous CVPR submission

Paper ID 503

Abstract

We tackle the problem of weakly supervised semantic segmentation for social images, whose labels are provided by Internet users. This is an extremely difficult task because the labels are not pixel-level but image-level and usually imprecise/incomplete. In this paper, we present a joint conditional random field model leveraging various contextual relations to address this issue. More specifically, we utilize feature representation generated by convolutional neural network and latent semantic concept model as well as label correlations captured by visual contextual cues and label co-occurrence statistics to handle the noisy annotations. Experimental results on two challenging image datasets PASCAL VOC 2007 and SIFT-flow show that our method outperforms state-of-the-art weakly supervised approaches 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 approaches heavily rely on a sufficiently huge amount of annotated samples in training. However, there are no 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 annotated by image-level labels but no pixel-level annotation is given [23, 24, 25, 26, 29, 30, 31]. The existing weakly supervised semantic segmentation methods are based on an unrealistic assumption that image-level labels are provided by professional annotators, and thus are correct and complete.

With the prevalence of photo sharing websites and collaborative image tagging system, such as Flickr, a large number of social images with user provided labels are available from the Internet. These labels are usually image-level; moreover, the quality of labels is not satisfactory: they are often imprecise and incomplete! Figure 1 illustrates a set of representative social images and its associated labels. We can observe that only limited labels accurately describe the visual content of the image, while other labels are imprecise. Moreover, some important labels, which are highly associated with the image, are missing. It is challenging but attractive to learn an effective semantic segmentation model from such social images.

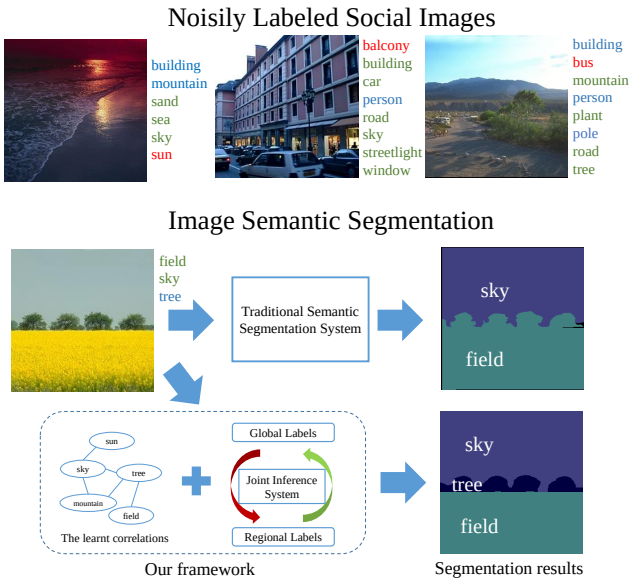

Figure 1. Given a set of social images and their associated labels where label may be precise (green), incorrect (red) or missing (blue), we learn a joint model that segments and recognizes visual concept in images. Best viewed in color.

In this paper, we present a weakly supervised method that, for the first time, overcomes the challenge posed by noisy annotations. The proposed method learns a joint conditional random field (CRF) model 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 neighborhood cues, and the label consistency between image-level and pixel-level. More specifically, we extract

global features for the whole image and local features for the superpixels in multiple scales 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.

To illustrate both robustness and effectiveness of our method, we demonstrate experimental results on two challenging datasets, PASCAL VOC 2007 and SIFT-flow datasets, which are representative of the hardness of annotation noise occurred in social images. Our method outperforms previous state-of-the-art approaches on standard datasets, demonstrating that the image-level annotation, especially potential relationships, is 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 imprecise or incomplete.
- 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 explore an effective strategy that cooperatively captures label co-occurrence statistics as well as visual contextual cues to model the label correlations, in addition to refine noisy labels.

2. Related Work

Being such a fundamental problem in computer vision community, numerous methods have been proposed for the semantic segmentation task in the fully supervised settings. Shotton *et al*. [18] formulate semantic segmentation as a CRF model over image pixels incorporating shape-texture color, location and edge clues in a single unified model. This model is further extended in series papers [10, 12, 13]. For instance, Kohli *et al*. utilize the higher order potentials [10] as a soft decision to ensure that pixels constituting a particular segment have the same semantic concept. Ladicky *et al*. extend the higher order potentials to hierarchical structure in [12] by using multiple segmentations and further integrate label co-occurrence statistics in [13]. However, these methods heavily rely on pixel-level annotations during the training stage.

Comparison with fully supervised semantic segmentation, there has been little work in the weakly supervised settings due to the fact that it is more challenging than the fully supervised task. Verbeek and Triggs [23] make the first attempt to learn a semantic segmentation model from image-level tagged data. They leverage several appearance descriptors to learn the latent aspect model via probabilistic Latent Semantic Analysis (pLSA) [8]. Furthermore, the spanning tree structure and Markov Fields are integrated with the aspect model in order to take the spatial information into consideration. In [24], Vezhnevets and Buhmann cast the weakly supervised task as a multi-instance multi-task learning problem with the framework of Semantic Texton Forest (STF) [17]. Based on [24], Vezhnevets *et al*. [25, 26] integrate the latent correlations among the superpixels belong to different images which share the same labels into CRF. Xu *et al*. [29] simplify 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.

However, all these approaches are based on the assumption that the initial image-level labels are clean and complete. It is not a practical requirement in many real-world applications. Although sharing similarities with both fully supervised and weakly supervised semantic segmentation, we address the issue of noisy annotations (*e.g*., labels can be incorrect and incomplete), which makes the task more challenging and intractable. To tackle this problem, we investigate label correlations, which are neglected by previous weakly supervised methods [23, 24, 25, 26, 29], based on not only label co-occurrence statistics but also the visual contextual cues.

Besides, we take latent semantic concept model generated by an unsupervised method as a mid-level representation of superpixels, while other methods (*e.g*., [25, 29]) only use the appearance model as a low-level representation, to narrow down the gap between semantic space and feature space, in the meantime, to make the whole framework more stable under the noisy condition. Unlike previous weakly supervised methods (*e.g*., [26, 29]), we also utilize multiple scale segmentations trying to avoid the weakness of choice of segmentation which cannot cover all the quantization level of objects.

3. The Proposed Model

We 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 may be incorrect or incomplete. In the test set, $\boldsymbol{y}$ is unknown. For each image, we employ the existing multi-scale segmentation algorithm and 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 provided in training/test stage.

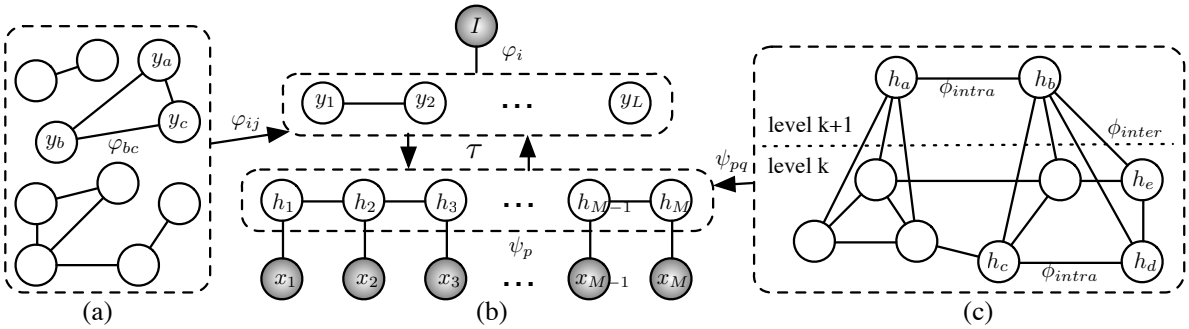

Figure 2. Graphical representation of our model, shadow nodes are observed variables and blank nodes are latent variables. (a) shows the learnt label correlations. (b) represents the joint inference framework. (c) shows the hierarchical model and spatial constraints.

Our goal is to infer the most suitable semantic label for each superpixel in an image and the adjacent superpixels sharing the same semantic label are fused as the whole one. In order to achieve this, we build a CRF on the image-level label variables $\boldsymbol{y}$ and the superpixel-level label variables $\boldsymbol{h}$. We connect each superpixel variables to its neighbors to encode a local smoothness constraint. Specifically, let $\mathcal{N}$ denote the neighborhood system among the superpixels, we define an energy function E with five types of potential as follows:

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

where φ_i and ψ_p encode the unary potential of image-level and superpixel-level constraints respectively, φ_{ij} impose label correlation and co-occurrence, ψ_{pq} are the spatial context constraints for each superpixel, and τ ensure the consistency between global and regional labels. The posterior distribution $P(\boldsymbol{y},\boldsymbol{h}|\boldsymbol{I})$ of the CRF can be written 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 constant. Thus, the most probable labeling configuration $\boldsymbol{y}^\star,\boldsymbol{h}^\star$ of the random field can be defined as $\boldsymbol{y}^\star,\boldsymbol{h}^\star = \arg\min_{\boldsymbol{y},\boldsymbol{h}} E(\boldsymbol{y},\boldsymbol{h},\boldsymbol{I})$. The details of each potential will be described in the following sections, and a graphical representation of the energy function is illustrated in Figure 2.

3.1. Unary Potential

Image-Level Potential We model input image by two aspects: On the one hand, we utilize CNN to represent the appearance model of each image, especially the first fully-connected layer containing 4096 neurons of a 19-layers model [19]. On the other hand, for latent semantic concept model, we model the CNN representation as a finite mixture of latent semantic concepts by pLSA [20], which can recover visual models of semantic labels in a completely unsupervised manner. Similar to [28], we regard fully-connected layer as input considering each neuron as a visual word w_i and each image as a document d_j, then the occurrence frequency of image on w_i is the i-th dimension of d_j. In addition, there is a hidden semantic topic variable t_k associated with all the visual words. We treat each topic t_k as a latent semantic concept. The pLSA optimizes the joint probability $P(w_i,d_j,t_k)$. Marginalizing over the latent concept t_k determines the conditional probability $P(w_i|t_k)$:

$$P(w_i|d_j) = \sum_{k=1}^{K} P(t_k|d_j)P(w_i|t_k) \tag{2}$$

where $P(t_k|d_j)$ is the probability of latent semantic concept t_k occurring in image j. Formally, we formulate image-level feature as $\boldsymbol{I}$ by concatenating the appearance feature d_j and latent semantic concept distribution $P(\boldsymbol{t}|d_j)$, and define the i-th image-level label presence/absence potential φ_i as follows:

$$\varphi_i(y_i = l,\boldsymbol{I}) = -\log f_i^l(\boldsymbol{I}) \tag{3}$$

where $f_i^l(\boldsymbol{I})$ is the linear support vector machine score function associated with label i for state $l \in \{0,1\}$.

Superpixel-Level Potential Similar with image-level potential, we both consider the appearance model and latent semantic concept model in order to narrow down the gap between low-level feature space and high-level semantic space, in the meantime, to reduce the negative impact of inaccurate image-level labels. Formally, let $\boldsymbol{x_p} = (\boldsymbol{a_p},\boldsymbol{c_p})$ indicates the appearance feature and latent semantic concept distribution vectors extracted from the superpixels, we encode the unary potential of superpixel-level as follows:

$$\psi_p(h_p = l,\boldsymbol{x_p}) = -\log\left\{w_1\boldsymbol{a_p}^\top\boldsymbol{\theta_a^l} + w_2\boldsymbol{c_p}^\top\boldsymbol{\theta_c^l}\right\} \tag{4}$$

where $\boldsymbol{\theta_a^l},\boldsymbol{\theta_c^l}$ denote the parameters for state $l \in \{1,2,...,L\}$ with respect to appearance model and latent

semantic concept model, w_1, w_2 are the weighting coefficients for the unary terms. The details of learning $\boldsymbol{\theta_a}$ and $\boldsymbol{\theta_c}$ will be illustrated in Section 3.3.

3.2. Pairwise Potential

Label Correlation We model inter-label co-occurrence by constructing inter-label correlation matrix which characterizing the interaction between semantic concepts and helps to model the relationship among feature space of superpixels. Due to the unknown semantic annotation of superpixels, learning these latent information is an unsupervised learning problem. Moreover, the context contain some useful latent information which can be learned for semantic label noise reduction.

Here we consider two aspects to construct inter-label matrix, on the one hand, we construct co-occurrence matrix A to capture the inter-label correlation. A is an $L \times L$ symmetric matrix and its entry $A(i,j)$, measuring the co-occurrence of concepts pair (i,j) based on statistics, can be defined as follows:

$$A(i,j) = 1 - (1 - P(i|j))(1 - P(j|i)) \tag{5}$$

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

On the other hand, we determine the inter-label correlation via visual contextual cues. Here, we capture such visual cues by calculating the intersection-over-union (IoU) overlap of discriminative regions between concepts pair. The discriminative region of concept indicates the most informative sub-window of each image within a multi-class classification framework. The basic idea is to analyze the variation of classification score when artificially removing different regions of the image, which means to black-out specific area of raw image. We observe that discarding a discriminative sub-window causes a massive confusion for one-vs-one classifier, especially in cluster condition.

Then we produce a set of sub-windows, which are deemed likely to contain the discriminative region for specific semantic label, from initial segmentation of each image. We finally utilize clustering technique, which merges regions according to their relative drop in classification score in addition to criteria of size of the selected area, to generate discriminative region hypotheses for semantic categories. In this way, we can construct the inter-label correlation matrix by using both the available image-level label co-occurrence and visual contextual clues. More concretely, we define the label correlation potential φ_{ij} as follows:

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

where $R(i,j)$, scaled to $[0,1]$, is calculated from the average overlapping area of discriminative regions between category i and j, $A(i,j)$ measures statistics co-occurrence and $\mathbf{1}(\cdot)$ is the indicator function.

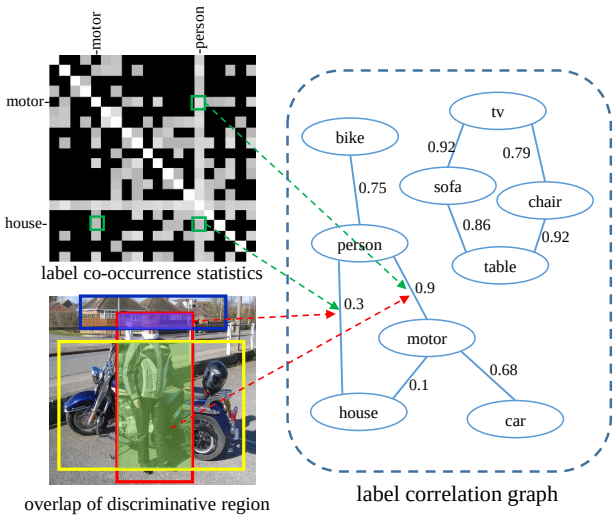

Figure 3. There are two aspects involved in our label correlations: label co-occurrence based on statistics and overlapping area of discriminate region based on visual cues.

As an illustration, Figure 3 demonstrates the graphical description of label correlation model. The top left shows co-occurrence matrix A displaying the interaction between semantic concepts. The brighter the block is, the stronger the co-occurrence probability is. The bottom left illustrates an example of visual contextual clues, as known as overlapping area of discriminative regions. The larger the overlap is, the closer the labels pair relationship 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 huge overlap between motor and person, strongly suggests the closer relationship between these two concepts, compared to no overlap between motor and house. The figure on the right side clarifies the label correlation graph that integrates two former models. The interdependency of concepts pair is quantized on the edge between two label nodes. The bigger the value is, the higher the correlation is.

Hierarchical Model and Spatial Constraints Considering the weakness of the single choice of segmentation, we utilize multiple segmentations to disambiguate low-level segmentation cues. We divide the superpixels into different quantization level according to the particular segmentation scale we chose. Then we include the inter-level energy cost ϕ_{inter} to investigate the most suitable segmentation scale each object belongs to. Besides, we integrate the intra-level energy cost ϕ_{intra}, which could discourage superpixel-level noise, to smooth the object boundaries. Let the two neighboring superpixels (inter-level or intra-level) be x_p and x_q (*i.e.*, $(p,q) \in \mathcal{N}$), we define the pairwise potential ψ_{pq} as

follows,

$$\psi_{pq}(h_p, h_q) = \begin{cases} \phi_{inter}(h_p, h_q) & \text{if } |l_p - l_q| = 1, \\ \phi_{intra}(h_p, h_q) & \text{if } l_p = l_q, \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where l_p indicates the quantization level that the superpixel x_p belongs to. The inter-level energy cost ϕ_{inter} is defined as:

$$\phi_{inter}(h_p, h_q) = \gamma \cdot O(x_p, x_q) \cdot \mathbf{1}(h_p \neq h_q) \quad (8)$$

where $O(x_p, x_q)$ refers to the intersection (overlapping area) of two superpixels, $\mathbf{1}(\cdot)$ is the indicator function and γ is the weighting coefficient. This formulation is based on the higher order constraints [10, 12] that superpixels lying within the same clique are more likely to take the same label. And the intra-level energy cost ϕ_{intra} is defined as:

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

where $S(x_p, x_q) \in [0, 1]$ measures the visual similarity between superpixel x_p and x_q, $R(h_p, h_q) \in [0, 1]$ is an inter-concept correlation between label h_p and h_q. Hence, we pay a high cost for the similar superpixels if they were assigned different labels and for the superpixels which were assigned an irrelevant label to the context.

Label Consistency We require that the superpixel labels be consistent with the image labels: if any superpixel x_p takes the label i, then image label indicator $y_i = 1$; otherwise $y_i = 0$. Such constraints can be encode by the following potential:

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

where $\mathbf{1}(\cdot)$ is the indicator function and C is a positive constant that penalizes any inconsistency between the image-level and superpixel-level labels. It is worth noting that such label consistency potential is a soft constraint. Thus, we can further simultaneously refine superpxiel label and image label via an iterative process.

3.3. Learning Parameters

Due to the fact that pixel-level labels are not available during the training stage, we cannot use cross-validation [10] to learn the weights for each potential. Inspired by [25], we scale the pairwise potentials so as to make them comparable with unary potentials. After selecting the weights of each potential, we can learn the parameters of appearance model $\boldsymbol{\theta_a}$ and latent semantic concept model $\boldsymbol{\theta_c}$ via an alternating optimization [25]: 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 appearance parameters $\boldsymbol{\theta_a}$ and latent semantic concept parameters

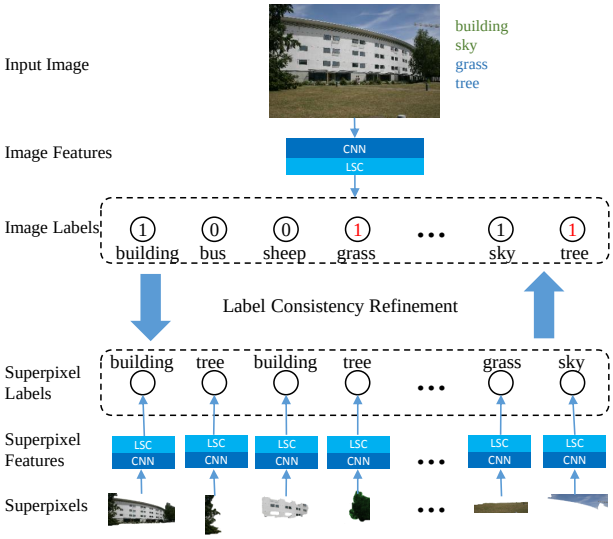

Figure 4. Illustration of our joint inference system. We simultaneously optimize the image-level label as well as superpixel-level label in a unified model so as to obtain the most suitable label configurations.

$\boldsymbol{\theta_c}$ can be found efficiently via the existing supervised methods (*e.g.*, [18]). The second step is a discrete optimization problem and we provide the details in Section 3.4.

3.4. Joint Inference with Alternating Procedure

Given an image I, our task is to assign each pixel a predefined semantic label. We achieve this as an energy minimization problem, in which our inference algorithm searches for optimal configuration of image-level label $\boldsymbol{y}^\star$ and superpixel-level label $\boldsymbol{h}^\star$. To efficiently minimize the energy function (1), we solve it in the following two alternating optimization steps, which can iteratively refine superpixel labels and image labels:

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

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

As a standard binary CRF problem, the first subproblem in Equation (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 Equation (12) reduces to an energy minimization for a multi-class CRF. Although finding the global

optimum for this energy function has been proved to be a NP-hard problem, there are various approximate methods for fast inference, such as approximate maximum a posteriori (MAP) methods (*e.g.*, graph-cuts [2]). 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 obtain by α-expansion has been proved to be within a known factor of the global optimum [2]. Considering the two alternate optimization steps together, we summarize our joint inference system in Algorithm 1.

Algorithm 1 Energy Minimization Inference

Input: a image I and its representation of superpixels $\{x_p\}$
Output: the image-level label variables $\boldsymbol{y}$ and the superpixel-level label variables $\boldsymbol{h}$

1: Construct the graphical model according to the energy function (1).
2: Initialize $\boldsymbol{y}$ and $\boldsymbol{h}$ with the highest unary potential according to Equation (3) and (4), respectively.
3: **for** iteration $t = 1$ to T **do**
4: fix $\boldsymbol{y}$, optimize $\boldsymbol{h}$ via Equation (12)
5: fix $\boldsymbol{h}$, refine $\boldsymbol{y}$ via Equation (11)
6: **end for**
7: Return the final configuration $\boldsymbol{y}$ and $\boldsymbol{h}$.

4. Experiments

In this section, we evaluate the effectiveness of our proposed approach for weakly supervised semantic segmentation based on two sets of experiments. The first set of experiments compares our approach with state-of-the-art algorithms on two standard datasets. The second set of experiments verifies the robustness of our approach under the noisy condition, in the meantime, evaluates the individual components that contribute to noisy reduction.

4.1. Experimental Setup

We utilize a linear maximum margin classifier leveraging both convolutional neural network feature and latent semantic concepts distribution in order to construct $\varphi_i(y_i, \boldsymbol{I})$ for label completion. In particular, we extract a 4296 dimensional global feature vector for each image by concatenating appearance feature and latent semantic concept distribution. The 4296 dimensional feature vector includes: the second to last layer of convolutional neural network [19] pre-trained on ImageNet [16], as the 4096-dimensional appearance feature, and topic distribution learned from pLSA [8], as the 200-dimensional latent semantic concept distribution. We use the publicly available implementation *Caffe* [9] to compute the CNN features, and a one-vs-one linear support vector machine [5] per class for initial forecast for image-level annotations.

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

4.2. Comparison with State-of-the-art

In this section, we compare our approach with the existing state-of-the-art weakly supervised semantic segmentation methods as well as fully supervised semantic segmentation algorithms on two challenging datasets: PASCAL VOC 2007 [4] and SIFT-flow [15].

PASCAL VOC 2007 This dataset was used for the PASCAL Visual Object Category segmentation contest 2007. It is especially challenging for the presence of background clutter, illumination effect and occlusions. It contains 5011 training images, and 4952 test images. Within the training set, for a subset of 422 images which are suitable for evaluation of the segmentation task, the object in these images are marked at pixel level. The objects in the other images only have the bounding boxes indicating the location of the object and rough boundaries. And there are 20 foreground and 1 background classes in this dataset used for the task of classification, detection, and segmentation.

SIFT-flow The SIFT-flow dataset[15] is derived from the LabelMe subset and contains 2688 images of resolution 256x256 pixels, accompanied with a hand labeled segmentation of 33 unique semantic categories. This dataset has been widely adopted for semantic segmentation and it is also very challenging for the reason that there are large number of classes in the dataset and 4.43 labels per image. Besides, the frequency of classes is distributed with a power-law. For fair comparison, we use standard dataset split (2488 images for training and 200 images for testing) provided by [15].

Quantitative and Qualitative Results Comparisons of our performances against other methods (both fully supervised and weakly supervised) are given in Tables 1 and 2. The results on the PASCAL VOC 2007 dataset show that our approach outperforms the other state-of-the-art weakly supervised methods, demonstrating that the image-level annotations are more efficiently utilized by our method. In the meantime, the results conducted by our framework are comparable with the fully supervised method even though we use much less supervised information than these methods. Similar results are obtained on the SIFT-flow dataset, which is more challenging than the previous one. It is worth noting that our method demonstrates a comparable performance in noisy condition as well, more detail will be discussed in the

	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 [27]	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.* [30]	24	–	48	20	26	25	3	7	23	13	38	19	15	39	17	18	25	47	9	41	17	33
Supervised	Ours	27	66	26	15	61	12	15	51	30	38	6	29	19	25	29	26	19	12	18	4	28	28

Table 1. Quantitative analysis of VOC2007 results [4], intersection vs. union measure, define as $\frac{TP}{TP+FN+FP}$, in comparison with state-of-the-art methods. The results of fully supervised methods are taken from [4].

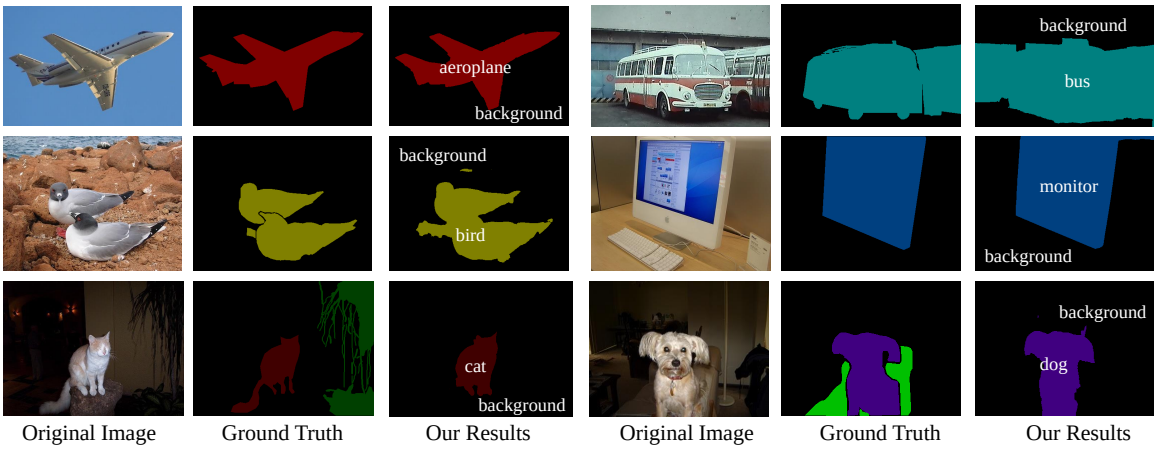

Figure 5. Qualitative results on the VOC-2007 data set. Successful segmentations (top 2 rows) and failure cases (bottom).

following section.

Figures 5 and 6 show the successful and failure cases of two datasets respectively. 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 by intra-class variability which remains very challenging in computer vision community.

4.3. Performance under the Noisy Condition

To verify both robustness and effectiveness of our method to noisy annotation condition, we try to reproduce the real-world noise distribution to the initial image-level labels for SIFT-flow dataset. More concretely, for a certain image in the dataset, each image-level label might be omitted or replaced by other incorrect labels. Here we suppose the probability of missing label is $p_{missing}$, which controls the missing labels set proportion of the whole label set.

We construct label pair confusion matrix where each entry p_{ij} indicate the probability of labeling category i instead of category j. This matrix is determined by manual observation on the empirical evidence from the collaborative image tagging system. In particular, We utilize Flickr API using queries for predefined semantic concepts and collect the number of incorrect labeling error occurred. After normalization we obtain the representative inter-label confu-

Supervision	Methods	Per-Class (%)
Fully Supervised (pixel-level)	Liu *et al.* [15]	24
	Tighe *et al.* [21]	29.4
	Tighe *et al.* [22]	39.2
Weakly Supervised (image-level no noise)	Vezhnevets *et al.* [25]	14
	Vezhnevets *et al.* [26]	21
	Zhang *et al.* [30]	26
	Zhang *et al.* [31]	27.7
	Xu *et al.* [29]	27.9
	Ours (0% noise)	32.3
Weakly Supervised (image-level with noise)	Ours (10% noise)	32.8
	Ours (25% noise)	32.4
	Ours (50% noise)	29.8
	Ours (75% noise)	22.3

Table 2. Quantitative results on the SIFT-flow dataset [15], average per-class recall measure, defined as $\frac{TP}{TP+FN}$, in comparison with state-of-the-art methods.

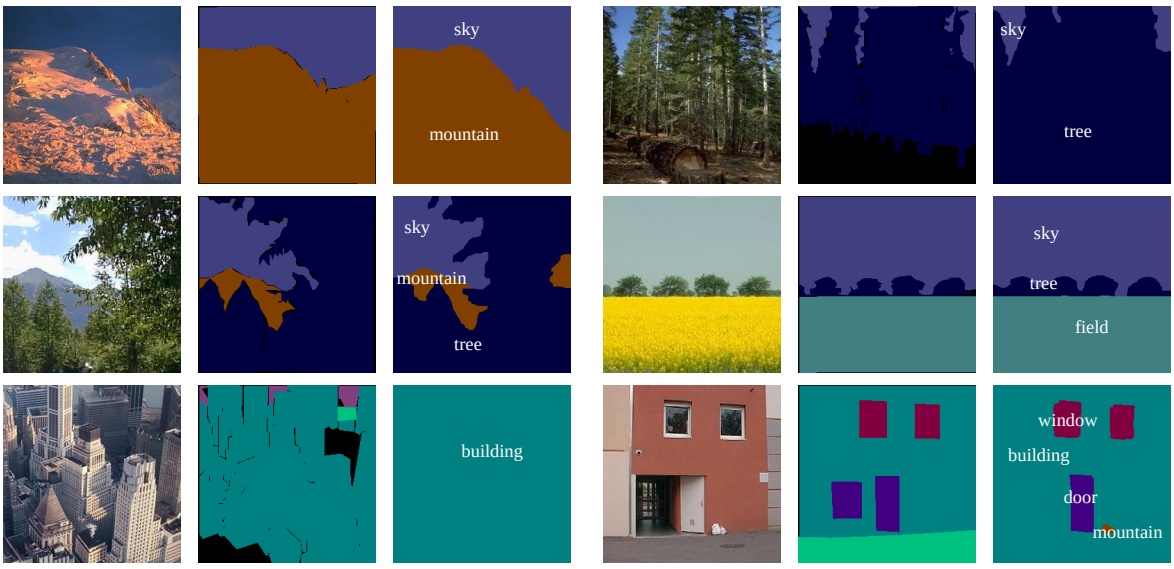

Figure 6. Qualitative results on the SIFT-flow data set. Successful segmentations (top 2 rows) and failure cases (bottom).

Noisily Labeled Images (%)	10	25	50	75
Noisy Labels per Image	1.3	1.4	1.7	2.4
Per-Class Accuracy (%)	32.8	32.4	29.8	22.3
Label Refinement (%)	93.7	93.4	92.3	90.4

Table 3. Extra statistics on noisily labeled SIFT-flow dataset as well as quantitative results on semantic segmentation and label refinement of our approach.

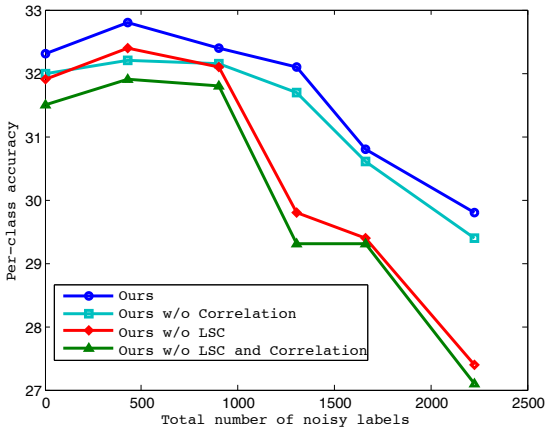

Figure 7. Evaluation of the individual components on noisily labeled SIFT-flow dataset.

sion probability distribution from statistics of social images.

We fix the probability p_{ij}, and set different values of $p_{missing}$ to obtain a set of noisy labeled datasets as shown in Table 3. The level of noise strength is determined by the percent of image with noisy annotations. It can be observed that our method can perform better or comparable results with respect to the state-of-the-art approaches.

To justify the individual components that contribute to noisy reduction, we conduct a control test as shown in Figure 7. It displays the performance degradation caused by removing these components and demonstrates the indispensability of these conponents in our approach especially under the noisy condition.

5. Conclusions

In this paper, we have investigated the challenging but realistic problem of weakly supervised segmentation for social images, where the only source of annotation are image-level labels, or even worse, the annotation can be imprecise/incomplete. To tackle this issue, we present a unified conditional random field incorporating various contextual relations, for instance, the associations between semantic concepts and visual appearance, label correlations, spatial neighborhood clues, and label consistency between image-level and pixel-level. We show both robustness and effectiveness of our method which can put up with social images with incorrect or incomplete annotations. Experiments on two real-world datasets illustrate improvement over existing weakly supervised semantic segmentation techniques on standard datasets.

References

[1] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 6

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001. 6

[3] E. Dinits. Algorithm of solution to problem of maximum flow in network with power estimates. *Doklady Akademii Nauk SSSR*, 1970. 5

[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. 6, 7

[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

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 7

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

[8] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999. 2, 6

[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. 6

[10] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009. 2, 5

[11] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *PAMI*, 2004. 6

[12] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *CVPR*, 2009. 2, 5

[13] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*. 2010. 2

[14] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008. 7

[15] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 2011. 6, 7

[16] 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. 6

[17] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 2

[18] 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. 2, 5

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 6

[20] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005. 3

[21] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*. 2010. 7

[22] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013. 7

[23] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, 2007. 1, 2

[24] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, 2010. 1, 2

[25] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 2011. 1, 2, 5, 7

[26] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012. 1, 2, 7

[27] V. Viitaniemi and J. Laaksonen. Techniques for image classification, object detection and object segmentation. In *VISUAL*, 2008. 7

[28] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*. 2014. 3

[29] J. Xu, A. G. Schwing, and R. Urtasun. Tell me what you see and i will show you where it is. In *CVPR*, 2014. 1, 2, 7

[30] K. Zhang, W. Zhang, Y. Zheng, and X. Xue. Sparse reconstruction for weakly supervised semantic segmentation. In *IJCAI*, 2013. 1, 7

[31] 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. 1, 7