Semantic Segmentation in Real-World Images

Anonymous CVPR submission

Paper ID ****

Abstract

In this paper, we tackle the problem of semantic segmentation in real-world images. Semantic segmentation is the task of assigning each superpixel to one of semantic categories. Although similar problems such as fully supervised and weakly supervised semantic segmentation have been previously studied, we focus on performing semantic segmentation in real-world setting, where only source of annotation are image-level labels encoding which categories are present in the image, and worse, the annotation can be noisy. To address these issues, we present a joint TBD:approach. In experiments on three real-world datasets, our method outperforms previous state-of-the-art weakly supervised approaches and achieves accuracy comparable with fully supervised methods. In addition, we also verify both robustness and effectiveness of our method to noisy annotation condition.TBD:conclusion

1. Introduction

Semantic segmentation, an attractive but challenging task in computer vision community, is an efficient way to handle an explosive growth in the volume of image in real world. Aiming to assign each pixel to one of predefined semantic categories, machine learning methods are used to learn classifier from labeled training images. Most state-of-the-art approaches heavily relies on extensive guidance in training, using a sufficiently huge amount of annotated samples, while the truth is only a subset of large-scale image dataset can be manual labeled, due to its time-consuming and labor-intensive. Recent works have begun to address the semantic segmentation problem under the weakly supervised settings, where each training image is annotated by image-level labels specifying which classes are present but no pixel-level annotation is given [25, 27, 28, 29, 30, 33, 34]. With the prevalence of photo sharing websites and collaborative image tagging system, such as Flickr, which host vast of digital images with user provided tags, such weakly supervised methods are more flexible in real-world applications since the image-level annotated images are much easier to obtain.

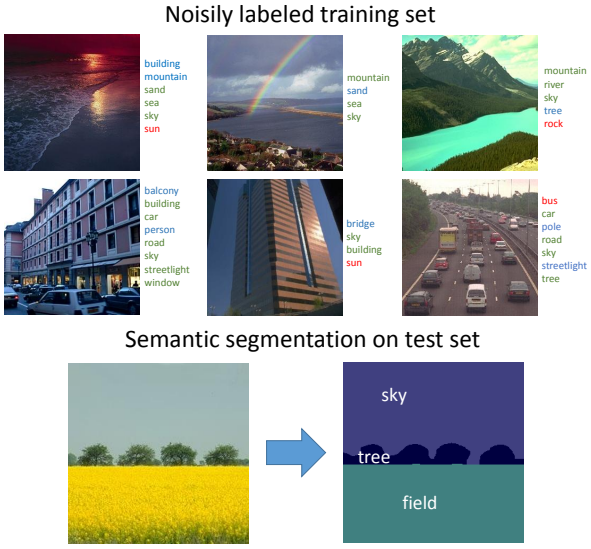

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

Inspired by these works, we focus on the problem of , which further relaxes the prerequisites for annotations(*e.g.* labels must be precise and complete). We tackle semantic segmentation in real-world settings where only source of annotation are image-level labels encoding which categories are present in the image, and worse, the annotation can be noisy. It is worth noting that the annotations of collaboratively-tagged images may not be accurate (incorrect or incomplete) in practice, but such noisily tagged annotation has been ignored in recent work. It is not a realistic assumption in many real-world settings because collecting large-scale images with clean labels is still a labor-intensive task. Figure 1 illustrates a set of representative real-world images and its associated tags. We can observe that only limited tags accurately describe the visual content of the image, while other tags are imprecise. Meanwhile, some important tags, which are highly associated with the

image, are missing.

Aiming to overcome the challenge posed by noisy annotations, we present a weakly supervised method which achieves results competitive with fully supervised methods. TBD:approach

To illustrate both robustness and effectiveness of our method, we present experiment results on three challenging datasets which are representative of the difficulties of annotation noise present in real-world images. Our method outperforms previous state-of-the-art approaches on standard datasets, demonstrating that the image-level annotation are more efficiently utilized by our method. Moreover, our approach TBD:Noise Refine and Prediction

The main contributions of this paper are summarized as follows:

1. We propose an weakly supervised semantic segmentation framework in noisily annotation condition.

2. We design a novel CRF model that jointly models various contextual relations in a single framework. It can be investigated from different perspectives: both appearance model and latent semantic categories distribution, inter-class label co-relation, image-level and pixel-level label consistency.

3. We propose a efficient method to model that jointly captures label statistics and discriminative regions of each category.

2. Related Work

Being such a fundamental problem in computer vision community, numerous methods have been proposed for the semantic segmentation task. Here we review the works that most related to ours.

In the fully supervised settings, Shotton *et al*. [20] formulate semantic segmentation as an Conditional Random Fields (CRF) over image pixels incorporating shape-texture color, location and edge cues in a single unified model. This model is further extended in series papers [9, 11, 12]. For instance, Kohli *et al*. utilize the higher order potentials [9] 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 [11] by using multiple segmentations and further integrate label co-occurrence statistics in [12]. However, these methods require pixel-level annotations during the training stage.

Unlike 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 [25] 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) [7]. Furthermore, the spanning tree structure and Markov Fields are integrated with the aspect model in order to take the spatial information into consideration.

In [27], Vezhnevets and Buhmann cast the weakly supervised task as a multi-instance multi-task learning problem with the framework of Semantic Texton Forest (STF) [19]. Based on [27], Vezhnevets *et al*. [28, 29] integrate the latent correlations among the superpixels belong to different images which share the same labels into Conditional Random Field (CRF). Xu *et al*. [30] 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.

Other works use cluster-based or classifier-based methods. For example, Liu *et al*. [14] leverage the spectral clustering and discriminative clustering techniques to investigate the relationship between feature space and semantic space, and solve it as an optimization problem within weakly supervised constraints. Zhang *et al*. [33] address weakly supervised problem by proposing a classifier evaluation criterion and replacing training stage with evaluating stage to obtain the superpixel-level classifiers.

However, all these approaches are based on the assumption that the initial image-level labels are clean and complete. And this assumption does not always hold in real-world applications. Different from these weakly supervised methods, our apporach imposes no more extra prerequisites on the initial image-level labels (*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 [25, 27, 28, 29, 30], based on discriminative areas of each category as well as label co-occurrence statistics.

Besides, we take topic model generated by an unsupervised method as a mid-level representation of superpixels, while other methods (*e.g*. [28, 30]) 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. In our approach, 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 formulate the semantic segmentation problem as an Conditional Random Field (CRF) over image superpixels incorporating label correlation, appearance model and topic model ... in a unified model. TBD:same as main contribution 2

More formally, suppose we have a set of weakly labeled

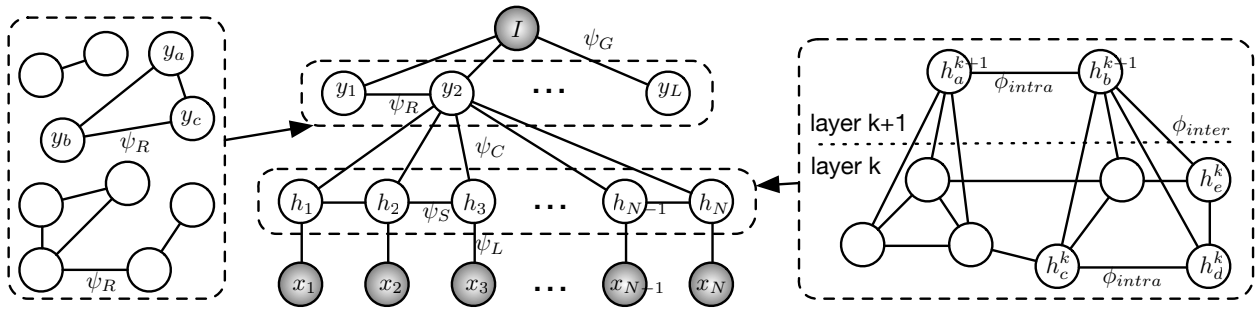

Figure 2. Example of a short caption, which should be centered.

images $\mathcal{I} = \{I_k\}_{k=1}^N$, where N donates the total number of training images. Each image I is associated with a vector of the L binary variables $\boldsymbol{y} = (y_1, ..., y_L)^{\mathrm{T}}$, *i.e.* $y_i \in \{0, 1\}$, where $y_i = 1$ indicates that the i-th category is present in this image, and 0 otherwise.

We adopt a superpixel representation of each image I, where the superpixels are generated from multi-scale segmentation algorithm. Then we associate every superpixel x_p ($p \in \mathcal{V}$) with a random variable $h_p \in \mathcal{C}$ to represent its semantic category. Here, $\mathcal{V} = \{1, ..., M\}$ is a set of all the superpixels in image I and M indicates the total number of superpixels. Besides, $\mathcal{C} = \{1, ..., L\}$ denotes the set of predefined labels.

Our goal is to find the accurate semantic label for each pixel in an image and the adjacent pixels sharing the same semantic label are fused as the whole one. To tackle this problem, we build a Conditional Random Field (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 $\mathcal{E}$ donate the neighborhood system among the superpixels, we define an energy function E with five types of potential as follows:

$$\begin{aligned} E(\boldsymbol{y}, \boldsymbol{h}, I) = &\sum_{i=1}^{L} \psi_G(y_i, I) + \sum_{1 \le i,j \le L} \psi_R(y_i, y_j) \\ &+ \sum_{p \in \mathcal{V}} \psi_L(h_p, x_p) + \sum_{(p,q) \in \mathcal{E}} \psi_S(h_p, h_q) \\ &+ \psi_C(\boldsymbol{y}, \boldsymbol{h}) \end{aligned} \tag{1}$$

where ψ_G and ψ_L encode the unary potential of global and regional constraints respectively, ψ_R impose labels' correlation and co-occurrence, ψ_S are the spatial context constraints for each superpixel, and ψ_C ensure the consistency between global and regional labels. The details of each potential will be described in the following sections. The posterior distribution $P(\boldsymbol{y}, \boldsymbol{h}|I)$ of the CRF can be written as $P(\boldsymbol{y}, \boldsymbol{h}|I) = \frac{1}{Z(I)} \exp\{-E(\boldsymbol{y}, \boldsymbol{h}, I)\}$, where $Z(I)$ is the normalizing constant. Thus, the most probable labeling configuration $\boldsymbol{y}^\star, \boldsymbol{h}^\star$ of the random field can be defined as $\boldsymbol{y}^\star, \boldsymbol{h}^\star = \arg\min_{\boldsymbol{y},\boldsymbol{h}} E(\boldsymbol{y}, \boldsymbol{h}, I)$. In the following subsections, we explain the details of each term, and a graphical representation of the framework is shown in Figure 2

3.1. Image-level

Multiple labels present correlatively and influence each other at semantic space (as shown in Figure 2 (a)). Different from the multi-label learning framework for fully supervised semantic segmentation, many different types of contextual cues cannot be utilized since the pixel-level annotation is not given. It is challenging to capture the inter-label correlation due to large appearance variations in cluttered backgrounds, in addition, noisy image annotation. TBD:what we done

We model each images as a finite mixture of latent semantic concepts by probabilistic latent semantic analysis (pLSA), which can recover visual models of semantic labels in a completely unsupervised manner. Probabilistic latent semantic analysis (pLSA) is a probabilistic model that is well suited to weakly supervision. Each image has its own mixing proportions whereas the topics are shared by all images. In document analysis, the pLSA usually takes the histogram of occurrence frequency on words as input. Here we regard fully-connected layer as input when we consider each neuron as a visual word and each image as a document. We denote each visual word (neuron) as w_i, then the occurrence frequency of image j 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 category in a semantic label. The pLSA optimizes the joint probability $P(w_i, d_j, t_k)$. Marginalizing over the latent category 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 topic t_k occurring in image j. Just by concatenating the appearance feature

d_j and topic distribution $P(w_i|d_j)$, we formulate image's global feature as I. Then, we define the label presence potential ψ_G as follow:

$$\psi_G(y_i, I) = -\log f_i(I) \tag{3}$$

where $f_i(I)$ is an SVM score function associated label i.

Due to the unknown label of superpixels, learning these latent information is an unsupervised learning problem. The context contain some useful latent information which can be learned for semantic label noise reduction. The inter-label correlation matrix is constructed to characterize the interdependency between semantic concepts and helps to model the inter-label co-occurrence among feature space of superpixels. Here we consider two aspect to construct inter-label matrix.

Inspired by [17]

On the one hand, we

The key idea is to analyze the change in the classification scores when artificially blackout different regions of the image. We observe that blackout a region that contains an discriminative region causes a massive confusion in cluster condition. This produces for each image a set of sub-windows from segmentation that are deemed likely to contain the discriminative region for specific semantic label. After localizing discriminative region in cluster condition, we can construct the inter-label correlation matrix by using both the available image-level label and region-level overlap. More concretely, we define the label correlation potential ψ_R as follow:

$$\psi_R(y_i, y_j) = R(i,j) \cdot Cooc(i,j) \cdot I(y_i = y_j) \tag{4}$$

where $R(i,j)$, scaled to $[0,1]$, is calculated from the overlapping area of discriminative region between category i and j, $Cooc(i,j) = 1 - (1 - P(i|j))(1 - P(j|i))$ measures the label co-occurrence based on statistics and $I(\cdot)$ is the indicator function.

3.2. Superpixel-Level

Similar with image level prediction, we both consider the appearance model and topic model in order to narrow down the gap between low-level feature space and high-level semantic space, in the meantime, to reduce the influence of inaccurate image-level labels. general description of superpixel unary Formally, we encode the unary potential of regions as follows:

$$\begin{aligned}\psi_L(h_p, x_p) = -\log \{ & w_1 \phi_a(h_p, a_p, \theta_a) \\ & + w_2 \phi_t(h_p, t_p, \theta_t) \}\end{aligned} \tag{5}$$

where a_p, t_p are the appearance and topic feature vectors extracted from the superpixels, θ_a, θ_t donate the parameters with respect to appearance model and topic model, $\{w_i\}|_{i=1}^2$ are the weighting coefficients for the unary terms. We define the appearance model $\phi_a(h_p, a_p, \theta_a) = f_{h_p}(a_p, \theta_a)$ and topic model $\phi_t(h_p, t_p, \theta_t) = g_{h_p}(t_p, \theta_t)$ measuring how well the local appearance a_p and topic t_p matches the semantic label h_p.

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 (either inter-level or intra-level) be x_p and x_q (*i.e.*, $(p,q) \in \mathcal{E}$), we define the pairwise potential ψ_S as follows,

$$\psi_S(h_p, h_q) = \begin{cases} \phi_{inter}(h_p, h_q) & \text{if } |l_p - l_q| = 1, \\ \phi_{intra}(h_p, h_q) & \text{if } l_p = l_q, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

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

$$\phi_{inter}(h_p, h_q) = \gamma \cdot O(x_p, x_q) \cdot I(h_p \neq h_q) \tag{7}$$

where $O(x_p, x_q)$ refers to the intersection (overlapping area) of two superpixels, $I(\cdot)$ is the indicator function and γ is the weighting coefficient. This formulation is based on the higher order constraints [9, 11] 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:

$$\phi_{intra}(h_p, h_q) = Sim(x_p, x_q) \cdot (1 - R(h_p, h_q)) \tag{8}$$

where $Sim(x_p, x_q) \in [0,1]$ measures the visual similarity between superpixel x_p and x_q, $R(h_p, h_q) \in [0,1]$ is a learnt correlation between label h_p and h_q TBD: explanation and details. 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.

3.3. 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:

$$\psi_C(\boldsymbol{y}, \boldsymbol{h}) = C \cdot \sum_{i,p} I(y_i = 0 \text{ and } h_p = i) \tag{9}$$

where $I(\cdot)$ is the indicator function and C is a positive constant that penalizes any inconsistency between the global and local labels.

3.4. Learning Parameters

Due to the fact that pixel-level labels are not available during the training stage, we cannot use cross-validation [9] to learn the weights for each potential. Inspired by [28], we scale the pairwise potential by median of maximum per unary term contribution of all the pairwise potentials in order to make them comparable to unary potentials. After selecting the weights of each potential, we can learn the parameters of appearance model θ_a and topic model θ_t via an alternating optimization [28]: 1) fix $\boldsymbol{h}$ and learn θ_a, θ_t; 2) fix θ_a, θ_t and infer $\boldsymbol{h}$. The first step corresponds to a continues optimization problem, hence the optimal appearance parameters θ_a and topic parameters θ_t can be found efficiently via the existing supervised methods (*e.g.* [20]). The second step is a discrete optimization problem and we provide the details in Section 3.5.

3.5. 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 (1), 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, we solve it in the following two alternating optimization steps:

$$\begin{aligned}\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} \sum_i \psi_G(y_i, I) + \frac{1}{2}\psi_C(\boldsymbol{y}, \boldsymbol{h}^*) \\ + \sum_{1\le i,j\le L} \psi_R(y_i, y_j),\end{aligned} \tag{10}$$

$$\begin{aligned}\boldsymbol{h}^* = \arg\min_{\boldsymbol{h}} \sum_p \psi_L(h_p, x_p) + \frac{1}{2}\psi_C(\boldsymbol{y}^*, \boldsymbol{h}) \\ + \sum_{(p,q)\in\mathcal{E}} \psi_S(h_p, h_q).\end{aligned} \tag{11}$$

As a standard binary CRF problem, the first subproblem in Equation (10) has an explicit solution which utilizes min-cut/max-flow algorithms (*e.g.* the Dinic algorithm [4]) to obtain the global optimal label configuration. And the second subproblem in Equation (11) 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, 10] 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 XXXX in Algorithm 1.

Algorithm 1 Energy minimization

1: 123

4. Experiments

In particular, we extract a 4296 dimensional feature vector for each image by concatenating appearance feature and topic distribution. The 4296 dimensional feature vector includes: the second to last layer of convolutional neural network[21] pre-trained on ImageNet [17], as the appearance feature, and topic distribution which learned from pLSA [7], as topic distribution.

As an illustration, Figure shows the inter-label correlation matrix illustrating the inter-dependency between 33 categories on the SiftFlow dataset. The brighter the block is, the stronger co-occurrence between labels exists. The dark blocks indicate the concepts pairs without correlation on the dataset. Among these pairs of the different concepts, we find that the concept pair of television and sofa have strong correlation.

4.1. 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 three real-world datasets: MSRC-21 [20], PASCAL VOC 2007 [5] and SIFT-flow [13].

MSRC-21 The MSRC segmentation dataset contains 591 images of resolution 320x213 pixels, accompanied with a hand labeled object segmentation of 21 categories [20]. This dataset has been widely adopted for semantic segmentation. Pixels on the boundaries of objects are usually labeled as background and not taken into consideration in these segmentations. For fair comparison, we use standard dataset split (276 images for training and 256 images for testing) provided by [20].

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, while 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[13] is derived from the LabelMe subset and contains 33 unique semantic labels. It has 2688 images, 2488 used for training and 200 for testing. This dataset is very challenging for the reason that there are large number of classes in the dataset and 4.43 classes

	Methods	average	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat
	Shotton *et al.* [20]	58	62	98	86	58	50	83	60	53	74	63	75	63	35	19	92	15	86	54	19	62	7
Fully	Yang *et al.* [31]	62	63	98	90	66	54	86	63	71	83	71	80	71	38	23	88	23	88	33	34	43	32
Supervised	Shotton *et al.* [19]	67	49	88	79	97	97	78	82	54	87	74	72	74	36	24	93	51	78	75	35	66	18
	Ladicky *et al.* [11]	75	80	96	86	74	87	99	74	87	86	87	82	97	95	30	86	31	95	51	69	66	9
	Lucchi *et al.* [15]	76	59	90	92	82	83	94	91	80	85	88	96	89	73	48	96	62	81	87	33	44	30
	Verbeek and Triggs [25]	50	45	64	71	75	74	86	81	47	1	73	55	88	6	6	63	18	80	27	26	55	8
Weakly	Vezhnevets *et al.* [28]	67	5	80	58	81	97	87	99	63	91	86	98	82	67	46	59	45	66	64	45	33	54
Supervised	Zhang *et al.* [33]	69	63	93	92	62	75	78	79	64	95	79	93	62	76	32	95	48	83	63	38	68	15
	Ours																						

Table 1. Quantitative results on the MSRC-21 dataset [20], average per-class recall measure, defined as $\frac{TP}{TP+FN}$, in comparison with state-of-the-art methods.

	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	24	3	1	45	34	16	20	0	68	58	11	0	44	8	1	2	59	37	0	6	19	63
Supervised	MPI	28	3	30	31	10	41	7	8	73	56	37	11	19	2	15	24	67	26	9	3	5	55
	TKK	30	23	19	21	5	16	3	1	78	1	3	1	23	69	44	42	0	65	30	35	89	71
	UoCTTI	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	27	66	26	15	61	12	15	51	30	38	6	29	19	25	29	26	19	12	18	4	28	28

Table 2. Quantitative analysis of VOC2007 results [5], 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 [5].

per image. Besides, the frequency of classes is distributed with a pow-low.

Quantitative and Qualitative Results Comparisons of our performances against other methods (both fully supervised and weakly supervised) are given in Tables 1, 2 and 3. The results on the MSRC dataset show that our approach outperforms the other state-of-the-art weakly supervised methods, demonstrating that the image-level annotation are more efficiently utilized by our method. In the meantime, the results conducted by our framework is comparable with the fully supervised method even though we use much less supervised information than these methods. Similar results are obtain on VOC2007 and SIFT-flow dataset, which are more challenging than the previous one. Figures 4, 5 and 6 show the success and failure cases of three dataset respectively.

4.2. Performance under the Noisy Condition

To verify both robustness and effectiveness of our method to noisy annotation condition, we try to add some noise to the initial image-level labels for SiftFlow dataset. More concretely, for a certain image in the dataset, each image-level label might be missing or replaced by other incorrect labels. Suppose the probability of missing label is $p_{missing}$, and the probability of labeling category i instead of category j is p_{ij} which is based on the empirical evidence from the collaborative image tagging system, *e.g.* Flickr. We fix the probability p_{ij}, and set different values of $p_{missing}$ to obtain a set of noisy labeled datasets. Then we perform a controlled test

Methods	Supervision	Per-Class (%)
Liu *et al.* [13]	full	24
Tighe *et al.* [23]	full	29.4
Tighe *et al.* [24]	full	39.2
Vezhnevets *et al.* [28]	weak	14
Vezhnevets *et al.* [29]	weak	21
Zhang *et al.* [33]	weak	26
Zhang *et al.* [34]	weak	27.7
Xu *et al.* [30]	weak	27.9
Ours	weak	30.2

Table 3. Quantitative results on the SIFT-flow dataset [13], average per-class recall measure, defined as $\frac{TP}{TP+FN}$, in comparison with state-of-the-art methods.

5. Conclusion

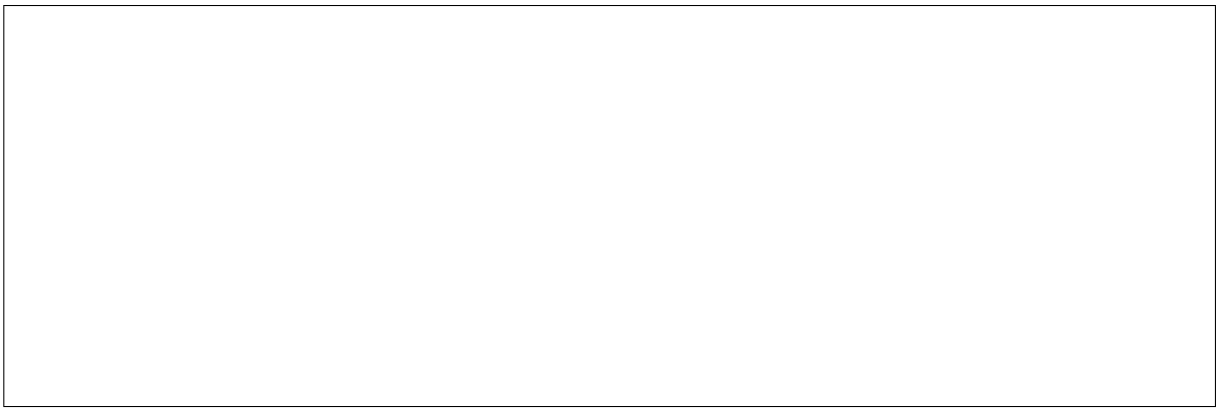

Figure 4. Qualitative results on the MSRC data set. Successful segmentations (top 2 rows) and failure cases (bottom).

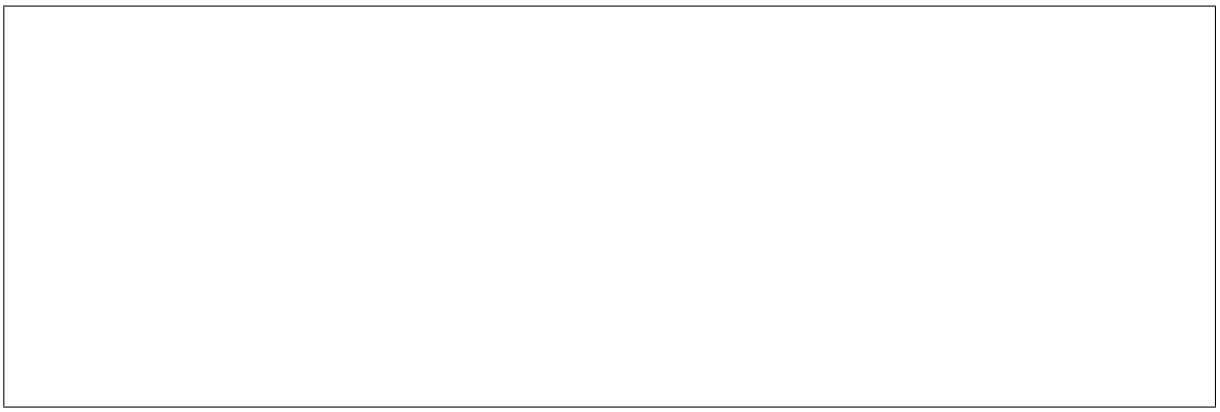

Figure 5. Qualitative results on the VOC-2007 data set. Successful segmentations (top 2 rows) and failure cases (bottom).

Experiment Setting	exp1	exp2	exp3	exp4
Noisy Image (%)	23.2	33.8	53.4	77.8
Noisy Labels per Image	1.4	1.5	1.8	2.4
Per-Class accuracy (%)	29.8	29.1	28.1	22.3

Table 4. Quantitative results

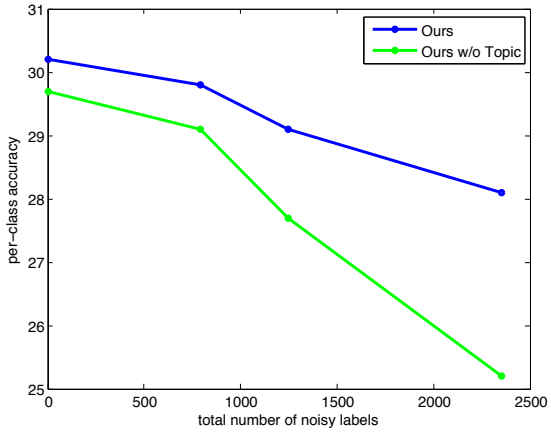

Figure 3.

References

[1] P. Agrawal, R. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *ECCV*. 2014.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001. 5

[3] G. Csurka and F. Perronnin. An efficient approach to semantic segmentation. *IJCV*, 2011.

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

[5] 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, 6

[6] J. M. Gonfaus, X. Boix, J. Van De Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez. Harmony potentials for joint clas-

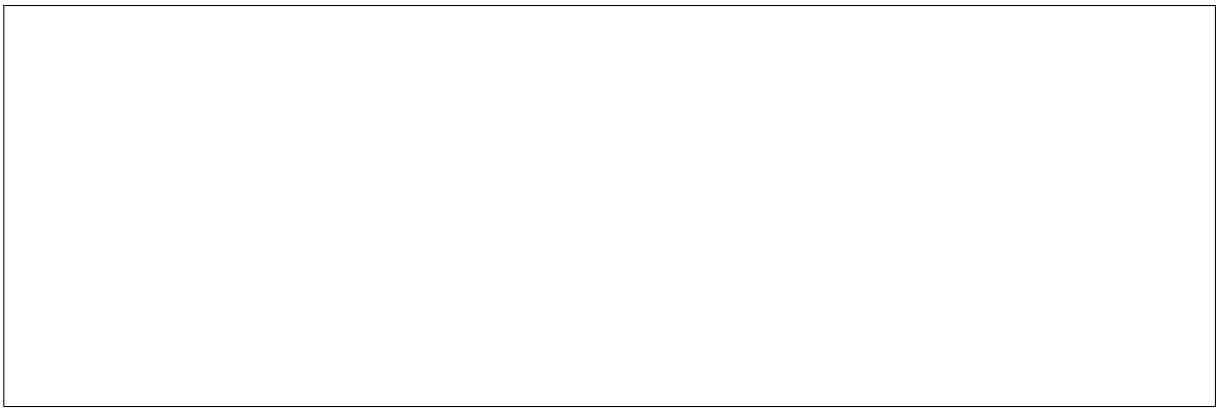

Figure 6. Qualitative results on the SIFT-flow data set. Successful segmentations (top 2 rows) and failure cases (bottom).

sification and segmentation. In *CVPR*, 2010.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999. 2, 5

[8] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[9] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009. 2, 4, 5

[10] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *PAMI*, 2004. 5

[11] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *CVPR*, 2009. 2, 4, 6

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

[13] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 2011. 5, 6

[14] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR*, 2013. 2

[15] A. Lucchi, Y. Li, K. Smith, and P. Fua. Structured image segmentation using kernelized features. In *ECCV*. 2012. 6

[16] S. Nowozin, P. V. Gehler, and C. H. Lampert. On parameter learning in crf-based approaches to object class image segmentation. In *ECCV*. 2010.

[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. 4, 5

[18] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[19] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 2, 6

[20] 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, 6

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[22] G. Singh and J. Košecká. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *CVPR*, 2013.

[23] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*. 2010. 6

[24] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013. 6

[25] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, 2007. 1, 2, 6

[26] J. Verbeek and W. Triggs. Scene segmentation with crfs learned from partially labeled images. *NIPS*, 2007.

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

[28] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 2011. 1, 2, 5, 6

[29] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012. 1, 2, 6

[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. 1, 2, 6

[31] L. Yang, P. Meer, and D. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *CVPR*, 2007. 6

[32] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012.

[33] K. Zhang, W. Zhang, Y. Zheng, and X. Xue. Sparse reconstruction for weakly supervised semantic segmentation. In *IJCAI*, 2013. 1, 2, 6

[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. 1, 6