Saliency Guided Dictionary Learning for Weakly-Supervised Image Parsing

Baisheng Lai Xiaojin Gong*
College of Information Science & Electronic Engineering,
Zhejiang University, Hangzhou, Zhejiang, P. R. China,
{laibs, gongxj}@zju.edu.cn

Abstract

In this paper, we propose a novel method to perform weakly-supervised image parsing based on the dictionary learning framework. To deal with the challenges caused by the label ambiguities, we design a saliency guided weight assignment scheme to boost the discriminative dictionary learning. More specifically, with a collection of tagged images, the proposed method first conducts saliency detection and automatically infers the confidence for each semantic class to be foreground or background. These clues are then incorporated to learn the dictionaries, the weights, as well as the sparse representation coefficients in the meanwhile. Once obtained the coefficients of a superpixel, we use a sparse representation classifier to determine its semantic label. The approach is validated on the MSRC21, PASCAL VOC07, and VOC12 datasets. Experimental results demonstrate the encouraging performance of our approach in comparison with some state-of-the-arts.

1. Introduction

Image parsing is a fundamental but challenging problem that aims to predict a semantic label for each pixel in the image. In contrast to conventional fully supervised techniques [26, 11, 13, 7, 9, 19], in recent years, it has been attracting more and more research interest to infer labels from weak supervision, for which expensive pixel-level annotated training samples are not required. So far, various forms of weak supervision, such as image-level tags [27, 28, 29], bounding boxes [36], and points [23], have been taken into account. Considering that image-level tags are the cheapest [23] and most convenient to obtain, in this paper, we use them to supervise image parsing.

This paper formulates our task within the dictionary learning and sparse representation (SR) framework. Previous SR-based works [15, 16] mainly focus on sparse constraints while using raw image patches for representation. In light of the outstanding performance that dictionary learning methods have exhibited in many other applications, we attempt to incorporate this technique to boost parsing performance. However, most discriminative dictionary learning methods [39, 37, 10] work in a fully supervised manner. It is a challenge to extend them to weakly supervised data because of the label ambiguities. That is, when image-level tags are given, we only know whether an object is present or not within the image, but without location information. Therefore, we are not able to obtain pixel- or superpixel-level training instances to learn dictionary for each semantic class. In order to accomplish the task, we make the following contributions:

- For weakly supervised discriminative dictionary learning, we design an adaptive scheme to dynamically adjust the weights that a superpixel contributes to each semantic class. For instance, given an image labeled with 'grass' and 'cow', a superpixel within it is not equally important for the learning of 'grass' and 'cow' dictionaries. Instead, the weights are adaptively learned according to some constraints.

- We introduce a saliency prior to guide the learning of the weights. Intuitively, a saliency map provides us with certain information about foreground and background, which helps to reduce the label ambiguities. To make use of saliency, we propose an automatic way to evaluate the confidence for a semantic class to be foreground or background according to the co-occurrence of tags. A linear programming constraint is further formulated for the saliency guided weight assignment. Moreover, the procedure of saliency detection [41] often considers both local and global contexts within an image. Thus, incorporating the saliency prior implicitly introduces global information into our model.

- We also incorporate a smoothness prior into our model. This constraint encourages label consistency by enforcing superpixels that are similar in appearance to

*The corresponding author.

Table 1. A comparison of related works.

Supervision	Framework	Method	Annotation Form	Using Extra Dataset
Full	Conditional Random Field	TextonBoost [26]	Pixel-level Annotation	
		TextonForest [25]		
		HCRF [11]		
		NLT [13]		
	Deep Learning	C+ref [9]		ImageNet
Weak	Topic Model	Li *et al*. [12]	Noisy Image tags	
		Spatial-LTM [3]	Image-level Tags	
	Conditional Random Field	MIM [28]		Geometric context dataset
		GMIM [29]		
		Zhang *et al*. [40]		
	Deep Learning	MIL-ILP-seg [22]		ImageNet
		FCN [19]		
		DCNN-EM-Adapt [20]		
		STC [31]		ImageNet + Flickr
		Russakovsky *et al*. [23]	Boxes + Tags + Points	ImageNet
	Max-margin Clustering	ILT + transductive [36]	Boxes + Tags + Parts	
Weak + GT Tags	Graphical Model	WSG [14]	Image-level Tags	
		k-NN SG [33]		
		k-NN SG + HG [34]		
		PGC [38]		
		Xu *et al*. [35]		
	Sparse Representation	BiLayer [15]		
		BiLayer+Continuity [16]		
		LAS [17]		Internet image search
		WSDC [18]		

have the same representation coefficients. The incorporation of the smoothness prior benefits both dictionary learning and sparse representation results.

2. Related Work

Image parsing is also named semantic segmentation in many literatures. In this paper, following the way defined in [33], we use weakly supervised image parsing to refer those researches that propagate labels from images to pixels. It means that image-level tags (ILT) are available for all images. While most semantic segmentation deal with test images that have no ILT. We hereby briefly introduce fully and weakly supervised semantic segmentation first, and then present parsing works.

2.1. Fully Supervised Semantic Segmentation

In these works, pixel-level annotation is available for training. Therefore, traditional fully supervised methods often train parametric [26, 25, 11] or non-parametric [13] classifiers to segment pixels into different semantic categories, and meanwhile Conditional Random Field (CRF) frameworks are commonly employed to ensure label consistency between neighboring pixels. A major problem these methods suffer is that they perform poorly in rare classes if the training set contains unbalanced number of samples.

Recently, deep learning techniques have been applied to semantic segmentation. Researchers [9] usually take a network trained in ImageNet [4] as an initialization and fine-tune it in semantic segmentation data sets. Due to the use of extra data set and the power of deep structure, these methods demonstrate extremely outstanding performance.

2.2. Weakly Supervised Semantic Segmentation

Existing works take image-level tags, bounding boxes, or other forms of annotations for weak supervision. They often separate training data from test sets. Different techniques such as latent topic models [3, 12, 40], multiple instance learning[27], conditional random fields [28, 29], as well as deep learning [22, 20, 31, 19], have been applied to training data to learn models for predicting pixel- or superpixel-level labels. When testing a new image, they either adopt an image-level annotator [28, 35] to predict the image's tags first, or directly use the trained model to infer labels for superpixels. Due to the lack of ground truth tags in the test set, their performance is limited.

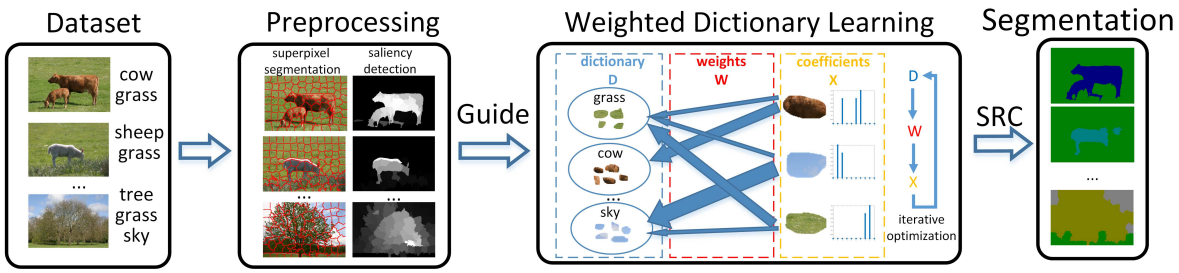


Figure 1. The overview of the proposed approach.

2.3. Weakly Supervised Image Parsing

As mentioned above, methods in this category assume that tags are available for all images. Their spirit is to propagate labels from image-level to pixel-level. Thus, various graph-based label propagation techniques [14, 33, 34] have been developed, which construct graphs or hypergraphs over superpixels regarding to k-NN or other criteria. Vertices' labels are propagated concerning superpixel consistency, incongruity and the weak supervision information. Another research line is using sparse representation techniques [15, 16, 17, 18]. Liu *et al*. [15, 16] proposed a bi-layer sparse model, in which each superpixel is sparsely reconstructed via the atomic superppixels selected from very few images. Labels are then propagated from images to the associated superpixels. Other SR-based methods are variant mainly on model construction and label propagation scheme. In contrast to existing SR methods, our work focus on dictionary learning and prior incorporation.

Table 1 lists most typical or state-of-the-art methods, whose formulation framework and supervision form are provided. It needs to mention that the distinction between two weakly supervised categories are ignored in some methods. Moreover, the name of each method is taken from their own papers if available, otherwise the authors are listed.

3. Problem Formulation

Assume we are given an image collection $\mathcal{I} = \{I_1, \cdots, I_i, \cdots, I_N\}$. Each image is tagged with a label set $\mathcal{Y}_i$, which is a subset of the full semantic labels $\mathcal{Y} = \{1, \cdots, N_l\}$. We first segment each image into n_i number of superpixels via SLIC [1] and extract their feature descriptors. All superpixels are then represented by $\mathbf{A} \in \mathbb{R}^{d \times N_s}$, where d is the feature dimension and $N_s = \sum_i^N n_i$ is the total number of superpixels. Meanwhile, we detect a saliency map [41] for each image to guide the learning of our dictionary $\mathbf{D} = [\mathbf{D}_1, ..., \mathbf{D}_l, ..., \mathbf{D}_{N_l}] \in \mathbb{R}^{d \times N_d}$, in which $\mathbf{D}_l = [\mathbf{D}_{l1}, \cdots, \mathbf{D}_{lM}]$ denotes the dictionary atoms associated with the l-th semantic class, M is the number of atoms for a class, and $N_d = M \cdot N_l$. All superpixels are sparsely represented by the dictionary, and the corresponding coefficient matrix is denoted by $\mathbf{X} \in \mathbb{R}^{N_d \times N_s}$. Once we obtain the coefficient matrix, the semantic label of each superpixel is determined by the Sparse Representation Classifier (SRC) [32]. Figure 1 illustrates the overview of our algorithm.

3.1. Weighted Dictionary Learning

Given all superpixels, the basic dictionary learning problem is formulated as

$$\arg\min_{\mathbf{X},\mathbf{D}} \frac{1}{2}\|\mathbf{A} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda_1\|\mathbf{X}\|_1, \tag{1}$$

where λ_1 is a parameter for balancing the two terms. In order to take the image-level tags into account, we convert the formulation into the following form

$$\arg\min_{\mathbf{X},\mathbf{D}} \frac{1}{2}\sum_{p=1}^{N_s}\|\mathbf{A}_p - \mathbf{D}diag(\mathbf{V}_p)\mathbf{X}_p\|_2^2 + \lambda_1\|\mathbf{X}\|_1, \tag{2}$$

in which $\mathbf{A}_p$ denotes the p-th superpixel; $\mathbf{X}_p$ is its representation coefficient; $diag()$ transforms a vector into a diagonal matrix; and $\mathbf{V}_p \in \mathbb{R}^{N_d}$ is a vector indicating whether a dictionary atom is used for representing the superpixel. Thus, it is defined as

$$\mathbf{V}_p[i] = \begin{cases} 1, & L(\mathbf{D}_i) \in \mathcal{Y}_{I(\mathbf{A}_p)} \\ 0, & otherwise \end{cases}. \tag{3}$$

Here, $\mathbf{V}_p[i]$ indicates the i-th element; $L(\mathbf{D}_i)$ gets the class of the dictionary atom $\mathbf{D}_i$; $I(\mathbf{A}_p)$ denotes the image to which the superpixel $\mathbf{A}_p$ belong; and $\mathcal{Y}_{I(\mathbf{A}_p)}$ is the associated label set.

The above formulation confines that a superpixel only impacts on the learning of the dictionary atoms which associate with its image-level tags. No matter which labels are tagged, the superpixel contributes to each labeled class equally. However, it is not desirable. For instance, if an image is labeled with 'grass' and 'cow', we wish a superpixel on 'cow' should play a more important role in learning the 'cow' dictionary than 'grass'. To this end, we introduce a dynamic weight vector $\mathbf{W}_p \in \mathbb{R}^{N_d}$ to replace $\mathbf{V}_p$. That is

$$\arg\min_{\mathbf{X},\mathbf{D},\mathbf{W}} \frac{1}{2}\sum_{p=1}^{N_s}\|\mathbf{A}_p - \mathbf{D}diag(\mathbf{W}_p)\mathbf{X}_p\|_2^2 + \lambda_1\|\mathbf{X}\|_1 \quad s.t.$$
$$\mathbf{W}_p \geq 0, \Pi_p(\mathbf{W}_p) = 0, \sum \overline{\Pi}_p(\mathbf{W}_p) = 1, p = 1, 2, ..., N_s, \tag{4}$$

in which $\mathbf{W}_p$ is assigned the same as $\mathbf{V}_p$ except that its nonzero entries are unknown variables summed up to be 1. Here, $\Pi_p(\ )$ is an operator to extract the zero part of a vector; $\overline{\Pi}_p(\ )$ gets the nonzero entries; and $\mathbf{W} = [\mathbf{W}_1, \cdots, \mathbf{W}_p, \cdots, \mathbf{W}_{N_s}]$.

The model defined in (4) provides a way to adjust the weights dynamically. It makes possible to assign a higher weight for a superpixel to its 'real' class if more information is included, and thus a more discriminative dictionary can be expected.

3.2. Dictionary Clustering

Before introducing how to adjust the weights, we first explore the structure of dictionary atoms. We expect that the dictionary atoms associated with the same label should be similar to each other. Thus, the spectral clustering technique [30, 18] is employed.

We define an affinity matrix $\mathbf{U}_D \in \mathbb{R}^{N_d \times N_d}$ as follows to indicate whether two atoms belong to a same class or not.

$$\mathbf{U}_D(i,j) = \begin{cases} 1, & L(\mathbf{D}_i) = L(\mathbf{D}_j) \\ 0, & otherwise \end{cases} \tag{5}$$

Then, dictionary clustering aims to minimize the term

$$tr(\mathbf{D}\mathbf{L}_D\mathbf{D}^T), \tag{6}$$

in which $tr()$ represents the trace of a matrix, $\mathbf{L}_D$ is the Laplace matrix computed by $\mathbf{L}_D = \mathbf{I} - \mathbf{B}^{-1/2}\mathbf{U}_D\mathbf{B}^{-1/2}$, and $\mathbf{B}$ is a diagonal matrix defined as $\mathbf{B}_{ii} = \sum_{j=1}^{N_d} \mathbf{U}_D(i,j)$.

3.3. Saliency Prior

In this subsection, we introduce how a saliency prior is integrated to guide weight assignment. The use of saliency is motivated by an observation that salient regions in an image are often roughly aligned with foreground objects, as shown in Figure 2. This property helps to reduce the ambiguities of superpixel labeling. For instance, in the simplest cases where an image is tagged only with two classes, such as 'cow' and 'grass' in Figure 2, we can assign salient regions to be the foreground class 'cow' and the remaining to be the background 'grass' with high confidence. When an image contains multiple classes like 'tree', 'grass', 'sky', 'building', the saliency map at least helps to distinguish between foreground and background classes.

However, a problem still remains: how to tell if a semantic class is foreground or background. Instead of manually determining it, we propose an automatic way to solve this problem. Let us denote $P(L_j|L_i)$ as the probability that label L_j occurs in an image conditioned on the occurrence of label L_i. This probability can be estimated from a given data set. If $P(L_j|L_i) > P(L_i|L_j)$, then label L_i is more likely to be foreground than L_j. The conclusion is based on a phenomenon that a foreground label often presents along with a certain background, while a background label may occur with different foreground objects. For example, 'cow' commonly occurs with 'grass', but 'grass' may occur with 'sheep', 'building', and other classes as well. In this case, we get $P(grass|cow) > P(cow|grass)$ and conclude that 'cow' is more likely to be foreground than 'grass'.

Thus, we define a *foreground-background score* to measure the confidence for a label L_i to be foreground or background in an image I_k. It is

$$fbs(L_i, I_k) = \begin{cases} -\frac{1}{2} + \frac{1}{1+\exp(-g(L_i,I_k))}, & L_i \in \mathcal{Y}_k \\ 0, & otherwise \end{cases} \tag{7}$$

where

$$g(L_i, I_k) = \frac{1}{|\mathcal{Y}_k|} \sum_{\substack{L_j \in \mathcal{Y}_k \\ L_j \neq L_i}} P(L_j|L_i) - P(L_i|L_j), \tag{8}$$

and $|\mathcal{Y}_k|$ is the cardinality. The range of the score is $[-0.5, 0.5]$, in which a positive value indicates a high confidence to be foreground and a negative value implies background.

Now, we design a guidance vector $\mathbf{G}_p \in \mathbb{R}^{N_d}$ that is

$$\mathbf{G}_p[i] = \begin{cases} -fbs(L(\mathbf{D}_i), I(\mathbf{A}_p)) \cdot S(\mathbf{A}_p), & S(\mathbf{A}_p) > T_s \\ fbs(L(\mathbf{D}_i), I(\mathbf{A}_p)), & B(\mathbf{A}_p) = 1 \\ -c, & otherwise \end{cases} \tag{9}$$

in which $S(\mathbf{A}_p)$ is the average saliency value of superpixel $\mathbf{A}_p$; $B(\mathbf{A}_p)$ indicates if the superpixel is on the image boundary or not; T_s is a threshold; $L(\mathbf{D}_i)$ and $I(\mathbf{A}_p)$ are defined in Eq.(3); and c is a constant experimentally determined.

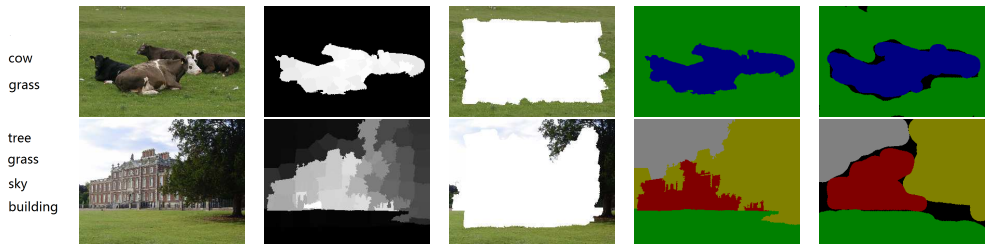
The designed guidance vector is used for guiding weight assignment. We propose to minimize

$$\begin{aligned} &\arg\min_{\mathbf{W}_p} \mathbf{G}_p^T \mathbf{W}_p \quad s.t. \\ &\mathbf{W}_p \geq 0, \Pi_p(\mathbf{W}_p) = 0, \sum \overline{\Pi}_p(\mathbf{W}_p). \end{aligned} \tag{10}$$

It is a linear programming problem that aims to assign large weights to the dictionary atoms associated with the labels having high foreground scores if a superpixel is salient. If a superpixel is on the image boundary, it is treated as background so that large weights are expected for background dictionary atoms. Meanwhile, the superpixels that are neither salient nor on boundary are assigned with equal weights for all tagged labels.

3.4. Smoothness Prior

The smoothness prior refers that two neighboring superpixels tend to have the same labels if they are similar in appearance and saliency. In the sparse representation framework, we are not able to enforce this constraint directly on



(a) Color Image (b) Saliency Map (c) Boundary Map (d) Our Result (e) Ground Truth

Figure 2. Typical examples on MSRC21.

the labels as in MRF [28]. Instead, we confine such superpixels to have the same representation coefficients.

Let $\mathbf{U}_X \in \mathbb{R}^{N_s \times N_s}$ denote a weighted affinity matrix. It is defined as follows:

$$\mathbf{U}_X(i,j) = \begin{cases} exp(-\|\mathbf{A}_i - \mathbf{A}_j\|_2^2 - \|S(\mathbf{A}_i) - S(\mathbf{A}_j)\|_2^2), \\ \qquad\qquad\qquad\qquad \mathbf{A}_i \in \mathcal{N}(\mathbf{A}_j) \\ 0, \qquad\qquad\qquad\qquad otherwise \end{cases} \tag{11}$$

in which $\mathcal{N}(\mathbf{A}_j)$ is the neighboring superpixel set of $\mathbf{A}_j$. Then, the smoothness constraint is formulated by minimizing

$$tr(\mathbf{X}\mathbf{L}_X\mathbf{X}^T), \tag{12}$$

where $\mathbf{L}_X$ is the Laplace matrix defined analogously to $\mathbf{L}_D$.

3.5. The Proposed Formulation

In summary, we get the entire model as follows:

$$\begin{aligned} &\arg\min_{\mathbf{X},\mathbf{D},\mathbf{W}} \frac{1}{2}\sum_{p=1}^{N_s}\|\mathbf{A}_p - \mathbf{D}diag(\mathbf{W}_p)\mathbf{X}_p\|_2^2 + \lambda_1\|\mathbf{X}\|_1 \\ &+ \frac{\lambda_2}{2}tr(\mathbf{D}\mathbf{L}_D\mathbf{D}^T) + \lambda_3\sum_{p=1}^{N_s}(\mathbf{G}_p^T\mathbf{W}_p) \\ &+ \frac{\lambda_4}{2}tr(\mathbf{X}\mathbf{L}_X\mathbf{X}^T) \quad s.t. \\ &\mathbf{W}_p \geq 0, \Pi_p(\mathbf{W}_p) = 0, \sum\overline{\Pi}_p(\mathbf{W}_p) = 1, p = 1, 2, ..., N_s, \end{aligned} \tag{13}$$

where $\lambda_1, \cdots, \lambda_4$ are parameters for balancing the terms.

4. Optimization

Before optimization, we first initialize the dictionary $\mathbf{D}$ and the weight matrix $\mathbf{W}$ as follows. Based on the definition in Eq.(9), we know that a small $\mathbf{G}_p[i]$ value indicates a high confidence for superpixel $\mathbf{A}_p$ belonging to the $L(\mathbf{D}_i)$ class. Therefore, for each class we collect the superpixels of the smallest guidance values and use k-means to cluster them. The obtained centroids are taken as the initial dictionary atoms of the corresponding semantic class. For $\mathbf{W}$, we simply assign normalized equal values for all non-zero entries.

With above initializations, we then employ an alternating scheme to solve $\mathbf{X}$, $\mathbf{D}$, and $\mathbf{W}$ iteratively.

4.1. Update X

The superscript t denotes the t-th iteration. At the $(t+1)$-th iteration, we fix $\mathbf{D}^t$ and $\mathbf{W}^t$ to solve $\mathbf{X}^{t+1}$. Thus, the problem in Eq. (13) is turned into the following form:

$$\begin{aligned} \mathbf{X}^{t+1} = \arg\min_{\mathbf{X}} &\frac{1}{2}\sum_{p=1}^{N_s}\|\mathbf{A}_p - \mathbf{D}^t diag(\mathbf{W}_p^t)\mathbf{X}_p\|_2^2 \\ &+ \lambda_1\|\mathbf{X}\|_1 + \frac{\lambda_4}{2}tr(\mathbf{X}\mathbf{L}_X\mathbf{X}^T). \end{aligned} \tag{14}$$

It consists of one L_1 term and two quadratic terms. Although the first two terms are separable in column-wise, the trace term is only separable in rows. Therefore this problem can not be solved in column-wise. We thus apply a general method, FISTA [2], to solve the whole matrix $\mathbf{X}$. It first computes a gradient step on the terms except the sparse one and then applies a soft-thresholding step on $\mathbf{X}$ to get the sparse results, so it is applicable to our model. In addition, FISTA is also efficient.

4.2. Update D

When fixing $\mathbf{X}^{t+1}$ and $\mathbf{W}^t$, we update $\mathbf{D}$ via solving

$$\begin{aligned} \mathbf{D}^{t+1} = \arg\min_{\mathbf{D}} &\sum_{p=1}^{N_s}\|\mathbf{A}_p - \mathbf{D}diag(\mathbf{W}_p^t)\mathbf{X}_p^{t+1}\|_2^2 \\ &+ \lambda_2 tr(\mathbf{D}\mathbf{L}_D\mathbf{D}^T). \end{aligned} \tag{15}$$

This is a quadratic problem so that we apply L-BFGS [24] to solve it efficiently.

4.3. Update W

We finally fix $\mathbf{X}^{t+1}$ and $\mathbf{D}^{t+1}$ to optimize $\mathbf{W}$. Note that $\|\mathbf{A}_p - \mathbf{D}^{t+1}diag(\mathbf{W}_p)\mathbf{X}_p^{t+1}\|_2^2$ can be rewritten as $\|\mathbf{A}_p - (\mathbf{D}^{t+1} \circ \mathbf{X}_p^{t+1})\mathbf{W}_p\|_2^2$, where $\mathbf{D} \circ \mathbf{X}_p = [\mathbf{D}_1 X_{p1}, \mathbf{D}_2 X_{p2}, ...]$. Therefore $\mathbf{W}$ can be solved column-

wisely as below.

$$\mathbf{W}_p^{t+1} = \arg\min_{\mathbf{W}_p} \frac{1}{2}\|\mathbf{A}_p - (\mathbf{D}^{t+1} \circ \mathbf{X}_p^{t+1})\mathbf{W}_p\|_2^2 + \lambda_3 \mathbf{G}_p^T \mathbf{W}_p$$
$$s.t. \quad \mathbf{W}_p \geq 0, \Pi_p(\mathbf{W}_p) = 0, \sum \overline{\Pi}_p(\mathbf{W}_p) = 1, \tag{16}$$

This is a standard quadratic programming problem that can be solved via an interior-point-convex algorithm [8].

4.4. Segmentation

Once we obtain the final coefficient matrix $\mathbf{X}$, the semantic label of each superpixel is determined by the Sparse Representation Classifier (SRC) [32]. It chooses the label via minimizing the representation residual, i.e.

$$\arg\min_l \|\mathbf{A}_p - \mathbf{D}diag(\mathbf{W}_p)\delta_l(\mathbf{X}_p)\|_2^2. \tag{17}$$

Here, $\delta_l(\mathbf{X}_p)$ sets the coefficients that are not associated with the class l to be zero while preserves others.

5. Experiments

In this section, we conduct a series of experiments to validate and analyze the performance of our approach.

5.1. Experimental Setup

The experiments are performed on three extensively used datasets: MSRC21 [26], PASCAL VOC07, and VOC12 [6]. Throughout all experiments, we empirically set the involved parameters as follows: $\lambda_1 = 10^{-3}$, $\lambda_2 = 5 \times 10^{-1}$, $\lambda_3 = 10^{-1}$, and $\lambda_4 = 10^{-2}$ in Problem (13); $T_s = 50$ and $c = 10^{-1}$ in Eq. (9). Moreover, the number of dictionary atoms for each class is set to be 30. The number of iterations for updating $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{D}$ is 5, which is enough for getting converged in almost all images.

On each dataset, we compare our results to the available results reported in some typical or state-of-the-art methods. Two criteria are used for comparison, which are, respectively, the average per-class accuracy (mAcc) and the average intersection-over-union score (mIOU) [5]. The former criterion measures the percentage of correctly labeled pixels for each class then averaged over all classes. It is commonly used for previous works to evaluate the performance on MSRC21 and VOC07. The mIOU is a standard measure for segmentation evaluation in PASCAL VOC12 challenges.

5.2. Experiments on MSRC

The MSRC21 dataset contains 591 images, accompanied by ground-truth segmentations of 21 classes. The scenarios range from simple objects to complicated road scenes. The average number of tagged labels is about 3.

We first investigate the performance of our saliency guided weight assignment scheme. Figure 3 illustrates a typical example, in which the superpixels coming from each class and their weights contributed to all dictionary atoms are presented. In this image, the *foreground-background score* (*fbs*) of 'tree', 'building', 'sky', and 'grass' are 0.0277, 0.0035, 0.0018, and -0.0329 respectively. According to Formulation (10) we know that if a superpixel is salient then its weights are assigned to the class with the largest *fbs*, like 'tree' in this case, and a boundary superpixel assigns its weights to the class with the smallest *fbs*, such as 'grass'. With the balance of the representation term, as defined in (16), the weights of 'sky' and 'building' superpixels are correctly assigned. Taking the 'building' superpixel as an example, although there are non-zero weights assigned to dictionary atoms of the other classes, the weights of 'building' dictionary atoms are dominant. Moreover, due to the sparse constraint on $\mathbf{X}$, the weights are also learned sparsely.

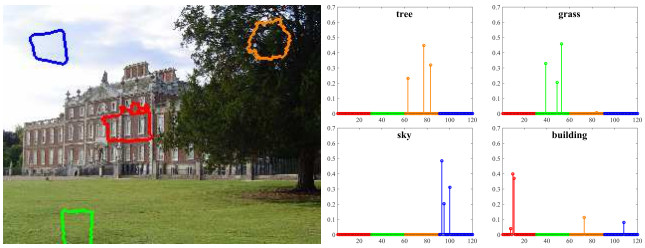
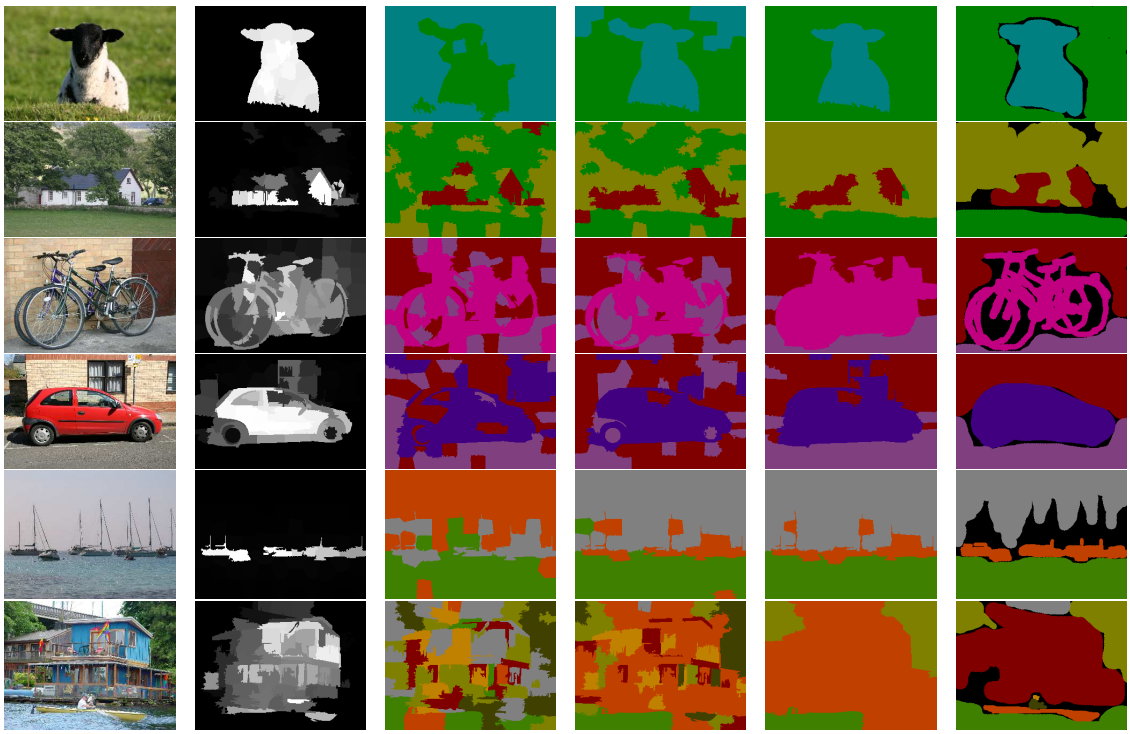


Figure 3. An illustration of superpixels and their weights learned for each label. In the right figures, the x-axis represents the indexes of dictionary atoms. The atoms belonging to 'building', 'grass', 'tree', and 'sky' are marked in red, green, orange, and blue respectively.

Then, we justify the effectiveness of each prior in our proposed model. We take the full formulation in (13), which is named the Saliency Guided Dictionary Learning (SGDL) model, as a reference, and leave the smoothness prior and the saliency prior out step by step. The model without the smoothness prior is referred to as SGDL-Sm and the one without both priors is denoted by SGDL-Sm-Sal. Experiments are conducted on MSRC21 for the three models and compared. Figure 4 presents some typical examples, from which we gain the following observations: 1) Without the guidance of saliency, the learned dictionary might be wrong even for simple objects, which results in totally wrong labeling results such as shown in the 'sheep' image. 2) The use of the smoothness prior greatly improves label consistency. 3) The full model obtains encouraging results in most cases. However, as shown in the last row, if the saliency order of regions do not match the estimated foreground-background scores, we may get wrong labeling results. Quantitative results listed in Table 2 show that SGDL greatly improves mAcc and mIOU in comparison with SGDL-Sm and SGDL-Sm-Sal.

We also compare our models to the classical or state-of-the-art techniques summarized in Table 1 if their results are



(a) Color Image (b) Saliency Map (c) SGDL-Sm-Sal (d) SGDL-Sm (e) SGDL (f) Ground Truth

Figure 4. Performance comparison of our models on MSRC21. SGDL refers to the entire model, SGDL-Sm is the model leaving the smoothness prior out, and SGDL-Sm-Sal is the one without both prior terms.

Table 2. Performance comparison on MSRC21.

Methods	bldg	grass	tree	cow	sheep	sky	plane	water	face	car	bike	flower	sign	bird	book	chair	road	cat	dog	body	boat	mAcc	mIOU
Fully-supervised																							
TextonBoost[26]	62	98	86	58	50	83	60	53	74	63	75	63	35	19	92	15	86	54	19	62	7	58	
HCRF[11]	80	96	86	74	87	99	74	87	86	87	82	97	95	30	86	31	95	51	69	66	9	75	
Weakly-supervised																							
MIM [28]	12	83	70	81	93	84	91	55	97	87	92	82	69	51	61	59	66	53	44	9	58	67	-
Weakly-supervised + Ground Truth Tags																							
WSG[14]	70	92	49	10	10	83	36	**82**	62	20	52	98	88	48	98	70	75	95	76	**43**	23	61	-
LAS [17]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67	-
BiLayer+Cont[16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	70	-
WSDC [18]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71	-
k-NN SG+HG[34]	**71**	89	60	64	57	93	**90**	76	**90**	**85**	**95**	**99**	**95**	83	**99**	**99**	66	**99**	**99**	34	25	80	-
SGDL-Sm-Sal	42	42	52	24	20	51	21	73	63	59	86	97	81	59	98	67	52	92	69	31	5	56	49
SGDL-Sm	52	86	76	**80**	**86**	**94**	74	75	83	56	64	96	87	**85**	97	89	43	98	91	31	**80**	77	60
SGDL	50	**96**	**89**	78	83	93	82	80	81	60	87	96	91	83	97	88	**83**	95	90	34	77	**82**	**70**

reported. Table 2 lists quantitative comparisons. Note that both fully supervised methods and weakly-supervised but without tags methods split the dataset into training and test parts. Their results are evaluated on the test set. While for methods in the category of 'weakly-supervised + Ground Truth Tags', results are on the entire dataset. From the results we see that the proposed approach outperforms both traditional fully supervised techniques and state-of-the-art weakly supervised methods.

Table 3. Experimental results on VOC07.

Methods	bkgd	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAcc	mIOU
Fully-supervised																							
TextonForest[25]	22	77	45	45	19	14	45	48	29	26	20	59	45	54	63	37	40	42	10	68	72	42	-
Weakly-supervised																							
Zhang et al. [40]	75	47	36	65	15	35	82	43	62	27	47	36	41	73	50	36	46	32	13	42	33	45	-
Weakly-supervised + Ground Truth Tags																							
WSG[14]	65	28	20	62	28	46	41	39	60	**25**	68	25	35	17	35	**56**	36	46	17	31	20	38	-
BiLayer[15]	82	24	25	40	25	32	35	27	45	16	49	24	32	13	25	**56**	28	17	16	33	18	32	-
k-NN SG+HG[34]	41	**77**	48	**87**	50	**56**	48	44	60	27	**76**	18	38	25	31	52	38	59	**31**	51	34	47	-
SGDL	**79**	65	**56**	75	**67**	27	**58**	**56**	**69**	**25**	54	**33**	**65**	**67**	**69**	28	**45**	**60**	23	**67**	**48**	**54**	31

Table 4. Experimental results on VOC12 val set.

Methods	bkgd	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAcc	mIOU
Weakly-supervised																							
MIL-FCN [19]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	26
CCNN[21]	66	24	18	23	19	36	47	47	47	16	36	22	43	34	45	40	30	33	22	39	36	-	35
MIL-ILP[22]	73	25	18	23	22	29	40	45	47	12	40	12	45	40	36	35	21	42	17	35	30	-	33
MIL-ILP-seg[22]	80	50	22	41	35	41	46	52	61	13	51	12	57	53	45	43	31	55	22	39	37	-	42
STC[31]	85	68	20	61	43	45	68	64	65	15	52	23	58	55	58	61	41	57	23	57	31	-	50
Weakly-supervised + Ground Truth Tags																							
SGDL	68	37	19	32	21	24	49	35	44	13	38	22	45	38	42	29	23	40	23	48	26	60	34

5.3. Experiments on VOC

We further validate the proposed method on the more challenging PASCAL VOC07 and VOC12 datasets [5]. The reason of choosing these two datasets is that most previous works have reported experimental results on VOC07 and the recent deep learning approaches are mainly performed on VOC12. VOC07 contains 632 segmented images of 21 labels. VOC12 has the same number of classes but more images. Its training, validation and test sets have 1464, 1449, and 1456 images respectively. We conduct our experiments on the training and validation sets, i.e. 2913 images in total, for evaluation because the ground truth segmentation of the test set is not available.

Table 1 presents the quantitative results of our full model, together with comparisons to the classical and state-of-the-art methods. It shows that our approach outperforms the others to a great extent, even for k-NN SG+HG[34] that is comparable to ours on MSRC21.

The results on VOC12 are also provided in Table 4. Among all the methods summarized in Table 1, only these powerful deep learning techniques published their results on this dataset. Therefore, although these techniques do not use the ground-truth tags and evaluate their results on the val set, we still include their results, only for the purpose of reference.

6. Conclusion

In this paper, we have presented a Saliency Guided Dictionary Learning (SGDL) method to conduct weakly-supervised image parsing. The spectral dictionary clustering, the saliency prior, and the smoothness prior are integrated into our model to learning dictionaries, weights, and sparse representations at the same time. Extensive experiments on three challenging datasets have validated the effectiveness of our approach.

Future work will focus on placing a group sparse constraint on the weights so that each superpixel only contributes to less semantic classes. We believe it will improve the performance of our approach further. Moreover, in the current model, some errors in saliency detection are unavoidably propagated to dictionary learning. How to reduce error propagation is also another direction we will take.

7. Acknowledge

This work was supported by State High-Tech Development Plan (863 Program) of China (No. 2014AA09A510), and the Fundamental Research Funds for the Central Universities.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[3] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 2007.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[5] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2014.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 35(8):1915–1929, 2013.

[8] N. Gould and P. L. Toint. Preprocessing for quadratic programming. *Math. Programming, Series B*, 100:95–132, 2004.

[9] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014.

[10] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent k-svd: learning a discriminative dictionary for recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2651–2664, 2013.

[11] L. Ladicky, C. Russell, and P. Kohli. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

[12] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2036–2043. IEEE, 2009.

[13] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1972–1979. IEEE, 2009.

[14] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu. Weakly supervised graph propagation towards collective image parsing. *Multimedia, IEEE Transactions on*, 14(2):361–373, 2012.

[15] X. Liu, B. Cheng, S. Yan, J. Tang, T. S. Chua, and H. Jin. Label to region by bi-layer sparsity priors. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 115–124. ACM, 2009.

[16] X. Liu, S. Yan, B. Cheng, J. Tang, T.-S. Chua, and H. Jin. Label-to-region with continuity-biased bi-layer sparsity priors. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 8(4):50, 2012.

[17] X. Liu, S. Yan, J. Luo, J. Tang, Z. Huang, and H. Jin. Nonparametric label-to-region by search. In *CVPR*, pages 3320–3327, 2010.

[18] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu. Weakly-supervised dual clustering for image semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2075–2082. IEEE, 2013.

[19] J. Long and E. S. andTrevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[20] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[21] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1796–1804, 2015.

[22] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015.

[23] O. Russakovsky, A. L. Bearman, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *arXiv*, 2015.

[24] M. Schmidt. minfunc: unconstrained differentiable multivariate optimization in matlab. *URL http://www. di. ens. fr/mschmidt/Software/minFunc. html*, 2012.

[25] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[26] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[27] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3249–3256. IEEE, 2010.

[28] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 643–650. IEEE, 2011.

[29] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 845–852. IEEE, 2012.

[30] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[31] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, and Y. Zhao. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI*, 2015.

[32] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[33] W. Xie, Y. Peng, and J. Xiao. Semantic graph construction for weakly-supervised image parsing. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[34] W. Xie, Y. Peng, and J. Xiao. Weakly-supervised image parsing via constructing semantic graphs and hypergraphs. In *Proceedings of the ACM International Conference on Multimedia*, pages 277–286. ACM, 2014.

[35] J. Xu, A. G. Schwing, and R. Urtasun. Tell me what you see and i will show you where it is. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3190–3197. IEEE, 2014.

[36] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *CVPR*, pages –, 2015.

[37] M. Yang, D. Zhang, and X. Feng. Fisher discrimination dictionary learning for sparse representation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 543–550. IEEE, 2011.

[38] 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 *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1908–1915. IEEE, 2013.

[39] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

[40] W. Zhang, S. Zeng, D. Wang, and X. Xue. Weakly supervised semantic segmentation for social images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2718–2726, 2015.

[41] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *CVPR*, 2014.