Image Piece Learning for Weakly Supervised Semantic Segmentation

Yi Li, Yanqing Guo, *Member, IEEE*, Yueying Kao, and Ran He, *Senior Member, IEEE*

***Abstract*—The task of semantic segmentation is to infer a predefined category label for each pixel in the image. For most cases, image segmentation is established as a fully supervised task. These methods all built on the basis of having access to sufficient pixel-wise annotated samples for training. However, obtaining the satisfied ground truth is not only labor intensive but also time-consuming, which severely hinders the generality of these fully supervised methods. Instead of pixel-level ground truth, weakly supervised approaches learn their models from much less prior information, e.g., image-level annotation. In this paper, we propose a novel conditional random field (CRF) based framework for weakly supervised semantic segmentation. Enlightened by jigsaw puzzles, we start the approach with merging superpixels from an image into larger pieces by a newly designed strategy. Then pieces from all the training images are gathered and associated with appropriate semantic labels by CRF. Thus, the piece library is constructed, achieving remarkable universality and flexibility. In the case of testing, we compare the superpixels with image pieces in the library and assign them the labels that minimize the potential energy. In addition, the proposed framework is fit for domain adaption and obtains promising results, which is of great practical value. Extensive experimental results on PASCAL VOC 2007, MSRC-21, and VOC 2012 databases demonstrate that our framework outperforms or is comparable to state-of-the-art segmentation methods.**

***Index Terms*—Conditional random field (CRF), image semantic segmentation, piece learning, weakly supervised.**

I. Introduction

THE HUMAN visual system can rapidly understand an image by recognizing objects and their background with a subtle boundary. In the similar case for machines, semantic segmentation manages to tackle the problem of assigning a label from a predefined category set to every pixel in the image. The task is fundamental in computer vision and benefits a variety of applications, ranging from biometric identification [1] to object recognition [2]. The problem is rather challenging, since a natural object may generate innumerable images with diverse appearances, poses, viewpoints, illumination as well as complicated background and the limited access of ideal training data makes it even worse. As a consequence, a promising solution is to explore segmentation methods with less supervision.

Unlike pixel-level ground truth, which is time and labor consuming, image-level annotation is more convenient to obtain. Recent developments in image classification and image annotation have brought inspiring results. On the other hand, with the boom of all kinds of social networks, even annotating extremely large scale images manually on the image level is more practical than on the pixel level. Thus it is appropriate for weakly supervised semantic segmentation. Fewer priors would make the training rather challenging and would make methods focus more on various structure information among images. The conditional random field (CRF) is one of the most widely used models in segmentation. Its intuitive meaning can be explained as: image regions (pixels, superpixels, or segments) that are visually alike or spatially close tend to share the same semantic label, while those regions disparate or remote are prone to diverse labels. To fully utilize structure priors, we propose a weakly supervised semantic segmentation framework based on CRF. In particular, the structure information includes spatial and visual characteristics of images as well as correlation among semantic labels.

To our knowledge, the existing weakly supervised semantic segmentation methods are mostly designed to train and test on the same database to achieve excellent performance. But if we train these methods on one database while test it on the other one, the results are often far from satisfying, even though the databases share some of the semantic categories. This fact results in the lack of universality. The reason lies in that these models are sensitive to parameters and depend much on training. For different databases, the customized parameters might differ a lot. To overcome the problem, we disassemble a database into categories, instead of regarding it as a whole like other methods and for a specific object category, capturing the

Manuscript received August 1, 2016; accepted October 19, 2016. Date of publication December 19, 2016; date of current version March 24, 2017. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61402079, in part by the Foundation for Innovative Research Groups of the NSFC under Grant 71421001, in part by the Open Project Program of the National Laboratory of Pattern Recognition, in part by the Youth Innovation Promotion Association Chinese Academy of Sciences under Grant 2015190, and in part by the State Key Development Program under Grant 2016YFB1001001. This paper was recommended by Associate Editor K. Huang. *(Corresponding author: Yanqing Guo.)*

Y. Li and Y. Guo are with the School of Information and Communication Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: liyi@mail.dlut.edu.cn; guoyq@dlut.edu.cn).

Y. Kao is with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yueying.kao@nlpr.ia.ac.cn).

R. He is with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Chinese Academy of Sciences Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China (e-mail: rhe@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2016.2623683

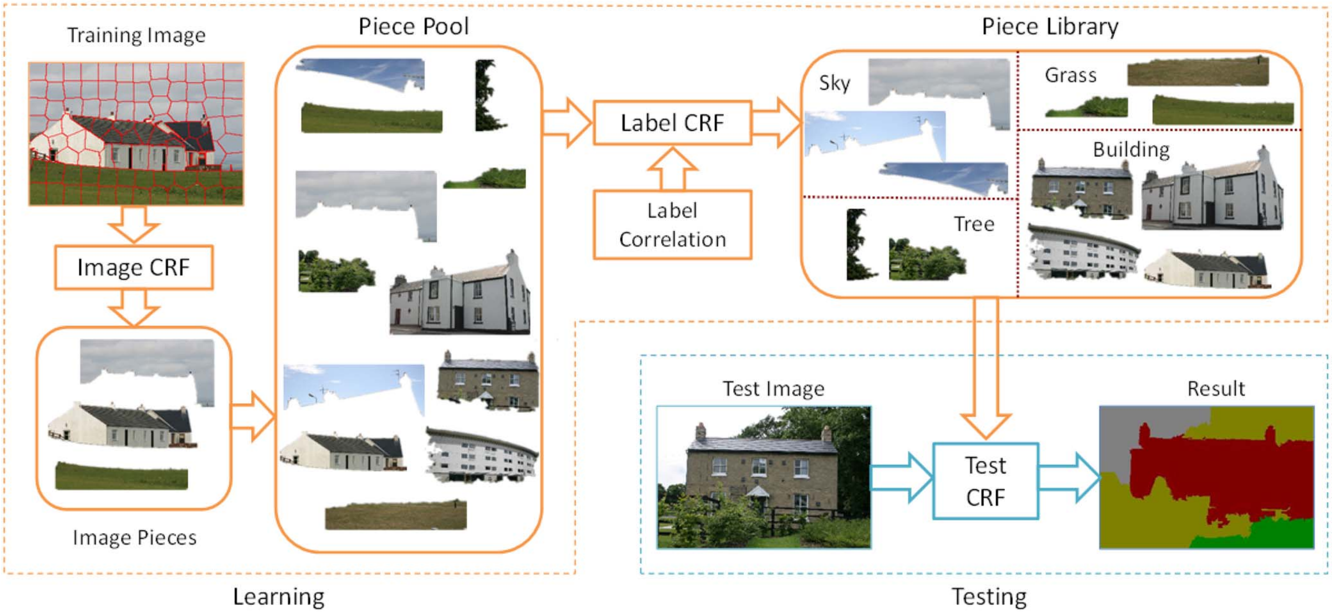

Fig. 1. System overview. We first merge superpixels into larger pieces. Then the pieces in the pool are associated with appropriate semantic labels by CRF, and thus the piece library is established. It is worth noticing that the library is easy to enrich and is able to be accessed by any data of the same form. Best viewed in color.

most representative information for segmentation is the aim of our framework. However, it does not mean that the categories are independent of each other. On the contrary, we consider the correlation between categories to be of significant importance for weakly supervised segmentation.

In this paper, we propose a novel CRF-based framework for weakly supervised semantic segmentation by learning image pieces. Enlightened by the knack of playing jigsaw puzzles, we first merge superpixels into larger pieces, since a superpixel is too trivial to contain much semantic information. The piece amount is expected to be as few as possible in one image, on the premise of that each of them is corresponding to merely one semantic category. Then pieces from all the images are gathered together to form the piece pool. Finally, a resource library for the categories is constructed by associating semantic labels with pieces by CRF. There are two main advantages of the piece library: 1) it could be accessed by any data of the same form, which achieves universality and 2) it is quite convenient to enrich, which achieves flexibility.

To the best of our knowledge, [3] and [4] are the works most related to ours. We try to solve similar problem but focus on different aspects, resulting in totally different frameworks. To address the issue of noisy labels from social images, Zhang *et al.* [3] formulated the problem as a joint CRF model with various contexts. It is an innovative CRF aiming at a specific issue, whereas we employ CRF in our framework as the resource to sufficiently utilize structural information. Hence, the typical second-order CRF is adequate in our approach, with no need for exclusive learning and inferring scheme. Pourian *et al.* [4] constructed a graph-based system to deal with the problem of partially provided labels. They directly build graph on regions (similar with superpixels in this paper) from all the images which need to find the nearest neighbors of each region. Even by using an optimized structure, their computational cost remains more expensive than ours. The reason is that we first merge superpixels into pieces in our framework. Note that in one image, the piece amount is set to be less than two times of semantic label amount empirically. This would significantly reduce the computational cost and make our framework train faster. Besides, since we focus on sifting out the representative pieces for each category and construct the piece library, our framework also achieves fair compatibility across databases, which brings more practical significance. An illustration of our model is exhibited in Fig. 1 and to verify the effectiveness of our framework, we conduct extensive experiments on three widely used databases: 1) PASCAL VOC 2007; 2) MSRC-21; and 3) VOC 2012.

A. Paper Contributions

In this paper, we propose a novel CRF based framework for weakly supervised semantic segmentation. Our main contributions are summarized as follows.

1) Enlightened by jigsaw puzzles, we propose a novel framework to incorporate various structure cues for weakly supervised semantic segmentation. The framework components are formulated as adapted CRFs which can be efficiently solved by existing optimization algorithms.
2) We design a strategy to merge superpixels into pieces and further associate these pieces with semantic labels by analyzing the relationship among training images. The merging process is able to significantly reduce the computational cost, which brings more practicability.
3) Our framework aims to construct an image piece library by assigning appropriate semantic label to each piece. The presented piece library is particularly suitable for domain adaption (i.e., pieces from all the databases are easy to share), achieving universality. Meanwhile, for new categories, it is quite convenient to enlarge, achieving flexibility.

II. Related Work

After years of development, the existing segmentation methods can be divided into three slices by their learning supervision, i.e., fully supervised, weakly supervised, and unsupervised.

1) Fully Supervised Methods: In the past decades, image semantic segmentation has been usually established as a fully supervised task [5]–[17]. The work in [5] examines the use of co-occurrence statistics in the likelihood model. Carreira and Sminchisescu [7] proposed to generate hypotheses by solving a sequence of constrained parametric min-cut problems and rank plausible ones for the spatial extent of objects. With the rise of the big data and deep learning, tremendous progress has been made in semantic segmentation. The problem becomes diverse and many additional tasks are solved along with segmentation. Chen *et al.* [12] tackled multi-instance object occlusions in segmentation. A novel algorithm for semantic part segmentation for animals is designed in [16]. In addition to these, there are methods that focus on segmenting a specific object rather than deal with the multiclass problem. The works in [6] and [9] concentrate on human segmentation in images and video sequences, respectively. And Yuan *et al.* [11] presented graph-based ranking and segmentation algorithms for traffic sign detection. However, the aforementioned methods are all built on the basis of having sufficient pixel-wise annotated samples for training. As the output of existing automatic systems is far from satisfactory, the vast majority of these annotations are obtained manually, which is tedious and time-consuming. Accordingly, the fully supervised approaches are typically unadaptable to general application in the reality.

2) Unsupervised Methods: On the other side, there are unsupervised semantic segmentation methods that utilize image data without any annotation for training [18], [19]. Note that different from simple image segmentation (see [20]) and unsupervised methods in other fields (see [21]), these works care for the category of each image pixel but in a unsupervised manner. Nevertheless, the importance of label correlation for performance improvement in semantic segmentation is well established [5], [22]. Without sufficient utilization of image-level annotations, the unsupervised methods tend to suffer from the under-constrained nature inherently and consequently impair their robustness toward variation. Moreover, with the development of social network service, acquisition of images with tags as labels has become more convenient and efficient than ever. For all these reasons, weakly supervised semantic segmentation is considered to be more suitable for most application scenario and has attracted more and more attention.

3) Weakly Supervised Methods: Instead of the ground truth of each pixel, weakly supervised semantic segmentation approaches [3], [4], [23]–[29] often require image-level annotations for training. Among these methods, [29] is an exceptional one since it utilizes object bounding boxes as supervision. Similar as our framework, Dai *et al.* [29] iteratively updated a pool of region proposals and assign them labels by training convolutional networks. However, their final target is the trained network while our aim is to construct a piece library with even less supervision. As for other methods, a graphical model called multi-image model is designed for recovering the pixel labels of the training images in [23]. Weakly-supervised dual clustering approach [24] collaboratively adopts spectral clustering and discriminative clustering to address the image segmentation and the tag alignment simultaneously. Zhang *et al.* [25] proposed probabilistic graphlet cut to efficiently exploit the distribution of spatially structured superpixel sets from image-level labels. And Zhang *et al.* [27] further augmented it by focusing on learning the semantic associations between the graphlets. The work in [26] develops a new way of evaluating classification models using sparse reconstruction and obtain the best parameters by iterative merging update algorithm. Pinheiro and Collobert [28] built a model based on convolutional neural network (CNN), with the constraint of putting more weight on the helpful pixels for image classification when training. Although the detailed location information is no longer necessary, most of these methods have the assumption that the exact labels for each image is available, which is another barrier we make through.

III. Proposed Framework

We propose a novel framework based on CRF that incorporates clues from clustering algorithm to accomplish the segmentation task. As is described in Fig. 1, superpixels from the same image are first merged into pieces, on the premise that each piece corresponds to only one semantic label. And the definition of image piece is enlightened by the concept in jigsaw puzzles. Second, pieces from all images are gathered into the piece pool. Finally, each piece is associated with an appropriate semantic label by integrating priors from its neighborhood and semantic label correlation. Thus, the piece library is constructed and is ready for testing. The detailed implementation is as follows.

A. Merging Superpixels Into Pieces

Suppose $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m\times n}$ is an image with n superpixels and $\boldsymbol{x}_i$ is the m-dimensional feature descriptor of the ith superpixel. The corresponding category labels of these superpixels are denoted by $\boldsymbol{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$ where $y_i \in \{1, \ldots, L\}$ with L representing the total number of object categories. However, for a training image in weakly supervised problem, the superpixel semantic labels $\boldsymbol{y} = [y_1, \ldots, y_n]$ are no longer available. Instead, we exploit the image-level labels, denoted by $\boldsymbol{l} = [l_1, \ldots, l_L]$ where $l_i \in \{0, 1\}$, and $l_i = 1$ indicates the presence of category i in the image while $l_i = 0$ indicates absence. Furthermore, $\boldsymbol{l}$ might be noisy or only partially provided. To correctly infer the superpixel semantic labels $\boldsymbol{y}$ for each image is the challenging task.

Although it is widely believed that superpixel has more expressive ability than pixel, it is too trivial to capture the high-level information of each category. Hence, we first merge superpixels into pieces. Each piece is expected to correspond to only one semantic label, while a semantic label can be distributed to multiple pieces. To that end, we build a graph $G = \{V, E\}$ on the image and use the CRF model to merge superpixels into pieces, where V refers to the set of nodes and E the edges. Specifically, every superpixel is defined as a node and two nodes (i, j) are connected if they are spatially close and/or visually alike. In this paper, we first connect two superpixels if their spacial distance is less than $(1/10)$ of the

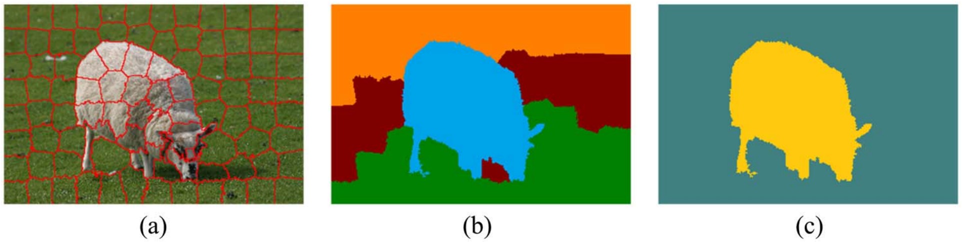

Fig. 2. (a) Over-segmented image, and every region (*red bordered*) is a superpixel. (b) Outcome of superpixel clustering, each color indicating a cluster. In (c), superpixels in the image are further polymerized into pieces by the CRF model, incorporating clues from (b). Then, pieces from the training set are to mapped to semantic labels, in Section III-B.

larger side, and then attach each edge with a weight which is positively related to the visual similarity of the two nodes.

Proposed by Lafferty *et al.* [30], the CRF model is first applied to natural language processing and now is broadly utilized to solve the problem of image segmentation. The essence of the model is that it directly builds the posterior distribution of the label field conditioned on the observation field. Given the observation, a CRF models the conditional posterior distribution of labels as a Gibbs distribution

$$P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) = \frac{1}{Z}\exp(-E(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta})) \tag{1}$$

where $\boldsymbol{\theta}$ is the parameters, Z denotes the normalization term, and $E(\cdot)$ is the energy function defined as the sum of potentials of all cliques in the graph G. In a typical second-order CRF, the energy function E can be further described as

$$E(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}) = \sum_{i\in V}\phi_u(y_i,\boldsymbol{X},\boldsymbol{\theta}) + \sum_{(i,j)\in E}\phi_p(y_i,y_j,\boldsymbol{X},\boldsymbol{\theta}) \tag{2}$$

where ϕ_u is the unary potential modeling the cost of assigning label y_i to node $\boldsymbol{x}_i$ and ϕ_p is the pairwise potential modeling the cost of assigning a pair of labels (y_i, y_j) to a pair of connected nodes $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Finally, the objective is to search for an optimal label assignment that maximizes the condition probability as is shown in (3), which is accordingly equivalent to minimizing the energy function E

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) = \arg\min_{\boldsymbol{y}} E(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}). \tag{3}$$

In this section, the unary potential for image pieces merging is indicated by ϕ_u^I. To formulate it, we cluster all the superpixels in an image into K groups by existing algorithm, e.g., K-means. The K here is determined by the image-level labels: $K \geq \|\boldsymbol{l}\|_0$, and we set $K = 2\|\boldsymbol{l}\|_0$ empirically. Although there are plenty of modern clustering algorithms, K-means is still one of the most widely used one for its advantages of simplicity and efficiency [31]. The unary potential (before normalization) is formulated as follows:

$$\phi_u^I(z_i,\boldsymbol{x}_i) = \left\|\boldsymbol{x}_i - \boldsymbol{c}_{z_i}\right\|_2 \tag{4}$$

where $z_i \in \{1,\ldots,K\}$ is the label indicating which image piece does $\boldsymbol{x}_i$ belong to, and $\boldsymbol{c}$ denotes the corresponding cluster center. And the pairwise potential is in the form of

$$\phi_p^I(z_i,z_j,\boldsymbol{x}_i,\boldsymbol{x}_j) = \lambda_1 I(z_i \neq z_j)\exp\left(-\frac{\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|_2^2}{\delta}\right) \tag{5}$$

where λ_1 weights the contribution of the pairwise potential, $I(\cdot)$ is an indicator function that equals 1 if the input is true and 0 otherwise, and δ is the parameter of Gaussian kernel. In this paper, we set $\delta = 1$ for all the employed Gaussian kernel, without loss of generality. Note that the amount of pieces P in an image might be less than that of clusters K. We accordingly update the center feature as $\boldsymbol{c}_k = (1/N_k)\sum \boldsymbol{x}_i I(z_i = k)$ and skip the absent ones. A sample illustration of the process is showed in Fig. 2. It is worth mentioning that if the pieces of an image is less than $\|\boldsymbol{l}\|_0$, even taking noisy labels into consideration, it might still lead to semantic category missing, which should be avoid by parameter setting.

B. Constructing Piece Library

After merging pieces on each image, we gather them together to form a piece pool, where each piece is to be associated with the most appropriate semantic label. Similar with Section III-A, we still utilize CRF to accomplish the task. In this section, each piece denoted by its center feature $\boldsymbol{c}$ in the pool is regarded as a node. Note that all the superpixels in one piece would share the same semantic label $s \in \{1,\ldots,L\}$. Since the relative location of a piece in its own image appears to be of little value when pieces are gather together, we connect a pair of nodes if they are visually alike.

The CRF model contributes to assigning closely related semantic labels to similar pieces while assigning diverse labels to disparate pieces. At the same time, we initialize the semantic label of each piece with its image-level label $\boldsymbol{l}$, to control the divergence between the assigned label and its priors. The unary potential for label mapping is formulated as

$$\phi_u^L(s_i,\boldsymbol{c}_i) = \exp\left(-\frac{\boldsymbol{l}_i(s_i)}{Z}\sum_{\boldsymbol{c}_j\in\mathcal{N}(\boldsymbol{c}_i)}\boldsymbol{l}_j(s_i)\right) \tag{6}$$

where $\boldsymbol{l}_i(s_i)$ indicates the s_ith element of $\boldsymbol{l}_i$ and Z is for normalization. $\mathcal{N}(\boldsymbol{c}_i)$ represents the neighborhood of $\boldsymbol{c}_i$ containing similar pieces with $\boldsymbol{c}_i$. It can also be obtained by the K-means algorithm. Note that $\boldsymbol{l}_i$ is binary and might be incorrect or missing. To solve the problem, we can replace zeros in $\boldsymbol{l}_i$ with ϵ $(0 < \epsilon < 1)$ and $\boldsymbol{l}_i$ becomes $\boldsymbol{l}_i'$. Here the ϵ is used to control the confidence of labels not corresponding to the piece $\boldsymbol{c}_i$. Thus it enhances the robustness of the model.

The exploitation of label correlation that assists in label mapping becomes vital, for the location of each label is unknown. To take full advantage of semantic label correlation, we integrate both co-occurrence statistics and label similarity into the pairwise potential ϕ_p^L. The value of label co-occurrence statistics for semantic segmentation has been explored by many publications [3], [5], [13]. The weakly

Algorithm 1 CRF-Based Image Piece Learning Framework

```
Input: $N$ images over-segmented into superpixels $\{\boldsymbol{X}^i\}_{i=1}^N$ and their image-level ground truth $\{\boldsymbol{l}^i\}_{i=1}^N$, piece number in each image $K$
Output: piece library containing piece centers $C = \{\boldsymbol{c}_i\}_{i=1}^P$ and their semantic labels $\{s_i\}_{i=1}^P$
$C^0 = \varnothing$
permute training data randomly
for $i = 1 \to N$ do
    Merging superpixels in image $\boldsymbol{X}^i$ into $K$ pieces $\{\boldsymbol{c}_j^i\}_{j=1}^K$ by CRF using Eq.(4) and Eq.(5)
    Update piece center set $C^i = C^{i-1} \cup \{\boldsymbol{c}_j^i\}_{j=1}^K$
end for
Converging all the pieces to form piece pool
Calculating label co-occurrence statistics matrix $\boldsymbol{A}$ and label similarity matrix $\boldsymbol{B}$
Constructing piece library by associating $\boldsymbol{c}_i$ with $s_i$ by CRF using Eq.(6) and Eq.(9)
```

supervised method in [3] propose to capture visual contextual cues with the help of [32]. Let $\boldsymbol{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_N]^{\mathrm{T}} \in \mathbb{R}^{N \times L}$ be the category labels of all the images in the training set with N indicating the total number of images. The label co-occurrence statistics matrix $\boldsymbol{A}$ is symmetric whose entry can be formulated by

$$\boldsymbol{A}(i,j) = \frac{\mathrm{count}(i \cap j)}{\mathrm{count}(i \cup j)} \tag{7}$$

where count$(\cdot)$ is the count of input, $i \cap j$ indicates the co-occurrence of l_i and l_j, and $i \cup j$ is the union set.

Inspired by Pourian *et al.* [4], we exploit the standard cosine similarity to measure the label similarity. Suppose $\boldsymbol{L}_{\langle i \rangle}$ is the ith column of $\boldsymbol{L}$, then $\boldsymbol{L}_{\langle i \rangle} \in \mathbb{R}^N$ can be regarded as a type of feature vector of label l_i. Hence, the entry of the label similarity matrix $\boldsymbol{B}$ is formulated as

$$\boldsymbol{B}(i,j) = \frac{\boldsymbol{L}_{\langle i \rangle} \cdot \boldsymbol{L}_{\langle j \rangle}}{|\boldsymbol{L}_{\langle i \rangle}|\,|\boldsymbol{L}_{\langle j \rangle}|}. \tag{8}$$

The same as the co-occurrence matrix $\boldsymbol{A}$, the similarity matrix $\boldsymbol{B}$ is also symmetric. And then, the pairwise potential is formulated as

$$\phi_p^L(s_i, s_j, \boldsymbol{c}_i, \boldsymbol{c}_j) = \frac{\lambda_2}{\boldsymbol{A}(s_i, s_j)}\left(1 - \boldsymbol{B}(s_i, s_j)\right) I(s_i \neq s_j) \exp\left(-\frac{\|\boldsymbol{c}_i - \boldsymbol{c}_j\|_2^2}{\delta}\right). \tag{9}$$

By minimizing the energy function, we associate each piece with a semantic label. These pieces and labels make up the piece library, which is quite convenient to enlarge. With persistent enrichment, the universality would become more and more prominent. We summarize the proposed weakly supervised learning framework in Algorithm 1.

C. Inference of Testing Image

As has been mentioned above, for a test image, the image-level label remains unavailable. The same as the training images, we first over-segment each test image into superpixels $\{\boldsymbol{x}_i\}$ and build a graph G^I upon them. The CRF we adopt for testing borrows elements from the previously used model: the graph structure of G^I and the pairwise potential for the label mapping. As for the unary potential, we formulate it based on the piece library out of the training phase, which can be written as

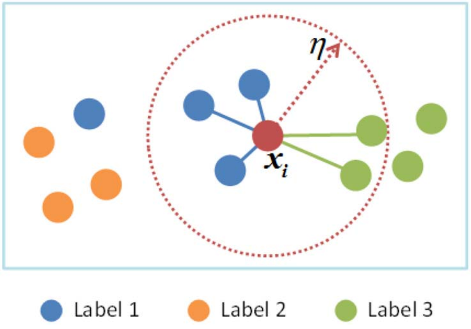

Fig. 3. Illustration of the unary potential for testing. Dots with different colors represent piece centers with different labels. The circle (*red dashed*) is the neighborhood of $\boldsymbol{x}_i$. Note the blue dot (*at the top left*) might be mislabeled, which should be excluded.

$$\phi_u^T(y_i, \boldsymbol{x}_i) = \frac{\frac{1}{N_{y_i}} \sum_{\boldsymbol{c}_j \in \mathcal{N}(\boldsymbol{x}_i)} \|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2 \, I(s_j = y_i)}{\sum_{y_i=1}^{L} \frac{1}{N_{y_i}} \sum_{\boldsymbol{c}_j \in \mathcal{N}(\boldsymbol{x}_i)} \|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2 \, I(s_j = y_i)}. \tag{10}$$

Note that similar local structures as shown in (10) can also be found in [33] and [34], but to address different problems. Considering the existence of few mislabeled pieces in the training and to avoid harm from them, the neighborhood $\boldsymbol{c}_j \in \mathcal{N}(\boldsymbol{x}_i)$ is obtained by setting an adapted threshold η to $\|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2$. In our experiments, we empirically set $\eta = \mathrm{mean}(\|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2)$. And N_{y_i} is the number of activated pieces with $s = y_i$ in the neighborhood. A graphical illustration of the unary potential for testing is given in Fig. 3

Eventually, the inference for a test image is to seek for the optimal solution that satisfying

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} \sum_{i \in V} \phi_u^T(y_i, \boldsymbol{x}_i) + \sum_{(i,j) \in E} \phi_p^L(y_i, y_j, \boldsymbol{x}_i, \boldsymbol{x}_j). \tag{11}$$

Therefore, $\boldsymbol{y}^*$ is the semantic label result for the input test image. Since our CRF model for testing is based on the image piece centers and their corresponding labels, it is suitable for cross-dataset testing.

IV. Experimental Studies

In this section, we evaluate the performance of our proposed approach by conducting extensive experiments on two commonly used databases for semantic segmentation: PASCAL VOC 2007 [38], MSRC-21 [35], and VOC 2012 [39]. Note that the ground truth segmentation is only used for evaluation. First, we compare the result of our method with state-of-the-art (both fully and weakly supervised) semantic segmentation approaches. And second, the cross-dataset performance of our model is exhibited. As for the measurement, we employ the widely used per-class accuracy defined as [#TP/(#TP + #FN)], and per-class intersection-over-union (IoU) score defined as [#TP/(#TP + #FN + #FP)], both on

TABLE I
ACCURACY (%) ON PASCAL VOC 2007. "FS" DENOTES THE FULLY SUPERVISED BASELINES, WHILE "WS" THE STATE-OF-THE-ART WEAKLY SUPERVISED RESULTS. THE MEAN WITHOUT BACKGROUND IS MARKED WITH †

	Method	bg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean/mean†
FS	STF [35]	20	**66**	6	**15**	6	**15**	**32**	19	7	7	**13**	**44**	31	**44**	27	39	35	12	7	39	23	24.1/24.3
	MPI [36]	3	30	**31**	10	**41**	7	8	73	**56**	**37**	11	19	2	15	24	**67**	26	9	3	5	55	25.3/26.5
	TKK [37]	**23**	19	21	5	16	3	1	**78**	1	3	1	23	**69**	**44**	**42**	0	**65**	**30**	**35**	**89**	**71**	**30.4/30.8**
WS	SR [26]	–	**48**	20	26	25	3	7	23	13	38	19	15	39	17	18	25	**47**	9	**41**	17	33	−/24.2
	SI [3]	**75**	47	**36**	**65**	15	**35**	**82**	**43**	62	27	**47**	**36**	41	**73**	50	36	46	32	13	42	33	44.6/43.1
	Ours	32	46	35	42	**45**	32	61	38	**65**	**60**	23	20	**67**	38	**67**	**61**	45	**77**	39	**60**	**48**	**47.7/48.5**

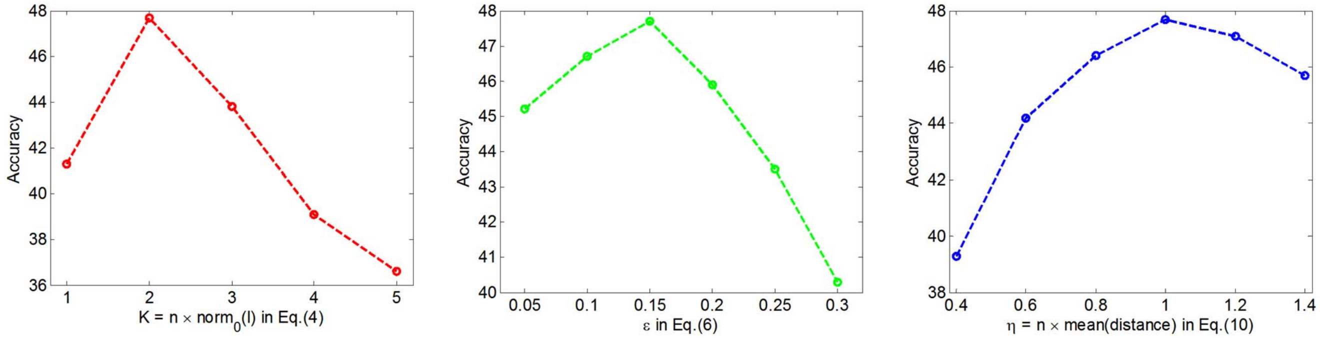

Fig. 4. Performance trend against different parameters on PASCAL VOC 2007.

pixel level. Here #TP, #FN, and #FP are the number of true positives, false negatives, and false positives, respectively. The metric mean over all classes is also calculated, for it effectively avoids bias toward categories with relatively large area and also penalizes the situation predicting few labels overall [4].

1) PASCAL VOC 2007: The database is a publicly available database consisting of photographs collected from photo-sharing website Flickr. There are all together 5011 images for training and 4952 for testing. However, a subset of 632 images are annotated at pixel level, including 422 images for training and 210 for testing. The ground truth is labeled with 20 object categories and a background class.

Although the database seems a little outdated and is relatively obsoleted for having been solved around 90% by fully supervised approaches, the conclusion is definitely inapplicable to the weakly supervised circumstance. For the limitation of much less priors, the database remains rather challenging for weakly supervised segmentation due to its complex variation of background, illumination, and occlusion.

2) MSRC-21: This is a popular multiclass benchmark for semantic segmentation and mainstream approaches for weakly supervised semantic segmentation are evaluated on MSRC-21. The database contains 591 images of size 320×213 with annotations from 21 object categories. Different from PASCAL VOC 2007, the ground truth label set includes not only object (e.g., chair, table, and book) but also stuff (e.g., ground, grass, and sky), both having their own characteristics. Contrary to the object categories, the stuff does not have a fixed shape. In our experiments, we follow the standard split dividing the database into training/validation/test subsets for fair comparison. In the ground truth images, pixels on the boundaries or in the background are labeled with void. Therefore, we add an extra "background" class to overcome the partially labeled problem.

3) PASCAL VOC 2012: The same as VOC 2007, the 2012 version also have all together 21 categories, but its scale is much larger than the former. PASCAL VOC 2012 is widely regarded as one of the main semantic segmentation benchmarks nowadays. For the segmentation task, there are 1464, 1449, and 1456 images for training, validation and test, respectively. The augmented ground truth of additional 9118 images provided by Hariharan *et al.* [40] are also generally used for deep network training. But for our framework, the training phase is merely conducted on the standard segmentation subset and it finally yields the competitive performance. The database is usually used for, respectively, large scale examination, even with cluttering background, illumination variation and occlusion.

A. Implementation Details

Many region-based segmentation approaches (see [12], [41]) utilize the output of multiscale combinatorial grouping (MSG) [42] directly for convenience. However, since MSG proposes overlapped regions which disagrees the puzzle thought, we merge superpixels from simple linear iterative clustering (SLIC) [43] to obtain image pieces. Seeing that images from different databases have different size, we over-segment images from MSRC-21 and PASCAL VOC into about 100 and 200 superpixels (the exact number varies for every image), respectively. Concretely, the compactness of SLIC is set to 20. Then each superpixel is represented by concatenating its CNN and LAB features. For the CNN feature, we engage the publicly available Caffe [44] and extract a 4096-dimensional feature vector, using the AlexNet [45] pretrained on ImageNet without fine tuning on the used databases. To remove redundancy and alleviate computational burden, dimensionality reduction by principal component analysis [46] is conducted on the CNN feature, retaining about 80% of the original energy. Finally, each superpixel is described by a 128 (CNN) +3 (LAB) = 131-dimension feature vector. As for how the key parameters influence the performance, there are detailed studies later in this paper.

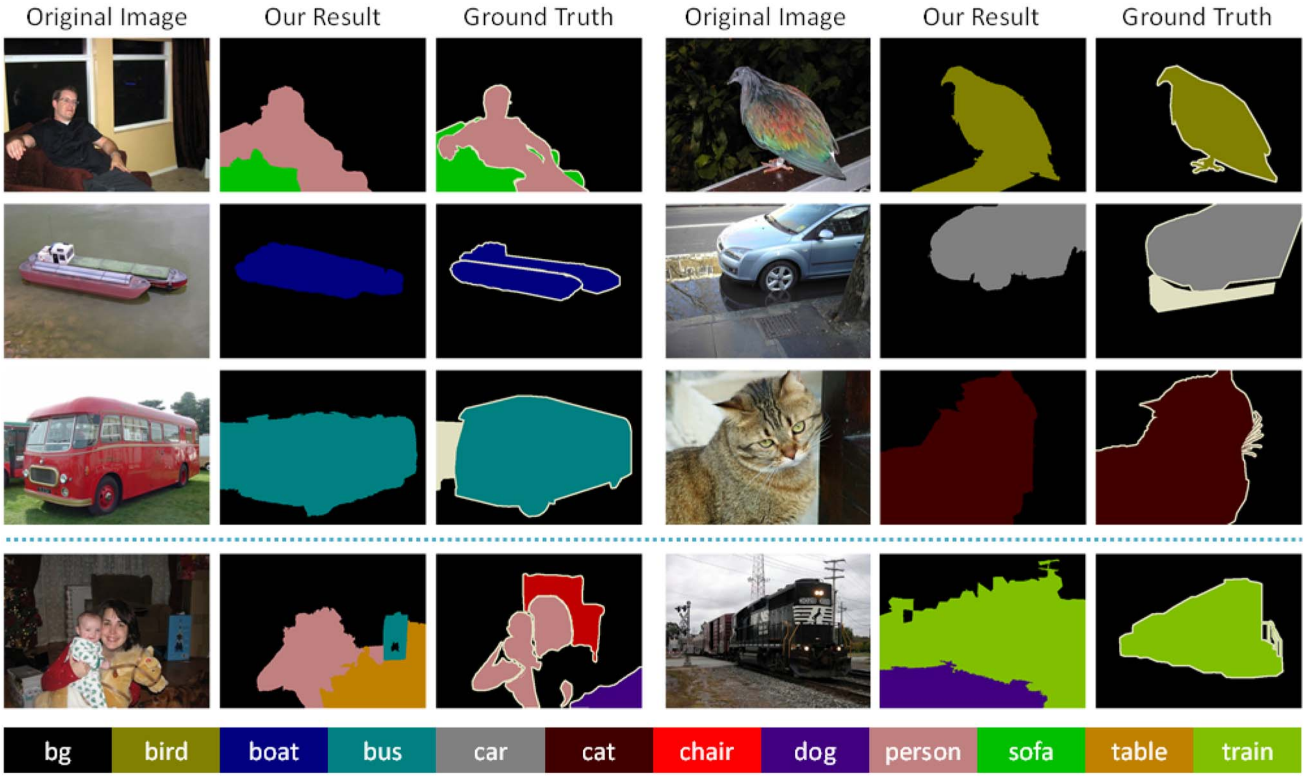

Fig. 5. Qualitative results on PASCAL VOC 2007.

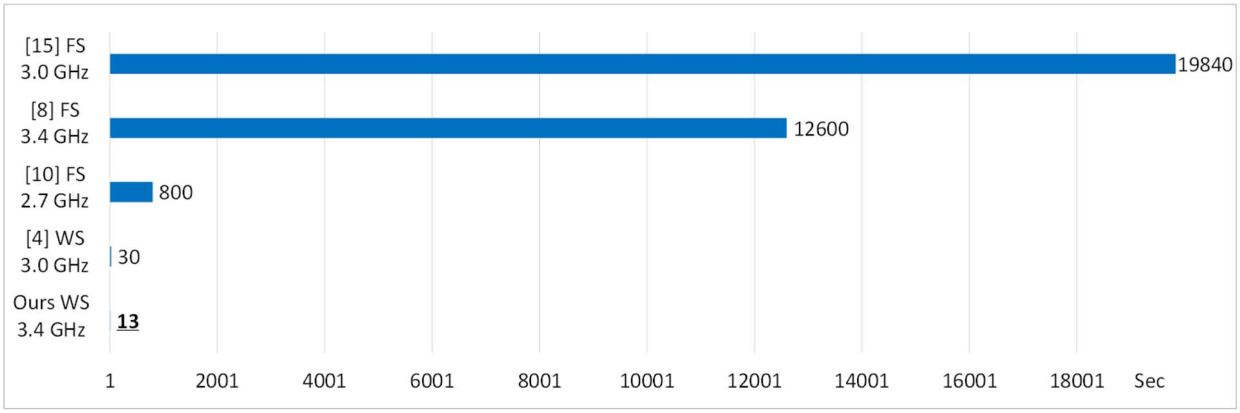

Fig. 6. Comparison of training time(s) on MSRC-21.

The CRF model used in this paper is multiclass CRF which can be converted to an energy minimization problem. Although the global optimum of the problem has been turned out to be NP-hard, with the form of summing unary and pairwise potentials, the mentioned optimization can be efficiently solved by existing approximate inference algorithms, e.g., α expansion and $\alpha - \beta$ swap. In our implementation, we adopt α expansion for inference, which can be efficiently solved by graph cuts algorithm [47]–[49]. And it has been proved that the energy obtained by α expansion is within a known factor of the global optimum [47]. As for the weights of the pairwise potentials, we empirically set $\lambda_1 = 0.05$ in (5) and $\lambda_2 = 0.1$ in (9).

B. Results and Analysis

1) PASCAL VOC 2007: The existing weakly supervised segmentation approaches [3], [26] evaluate their models by per-class accuracy. For fair comparison, we also use accuracy as the metric on VOC 07 dataset. In Table I, we compare the performance of our method with other approaches, including both fully and weakly supervised approaches. Particularly, the compared fully supervised methods are baselines, while the weakly supervised ones are state-of-the-art results. And the best result of each item is shown in boldface. Since K-means brings random factor to the results, we run the system with the same parameters five times and calculate the mean of per-class

TABLE II
ACCURACY (%) ON VOC 07 WITH DIFFERENT FEATURES

Feature	SIFT	PCA-CNN	LAB+SIFT	LAB+CNN
Dimension	128	128	131	131
Accuracy	40.9	44.2	45.2	47.7

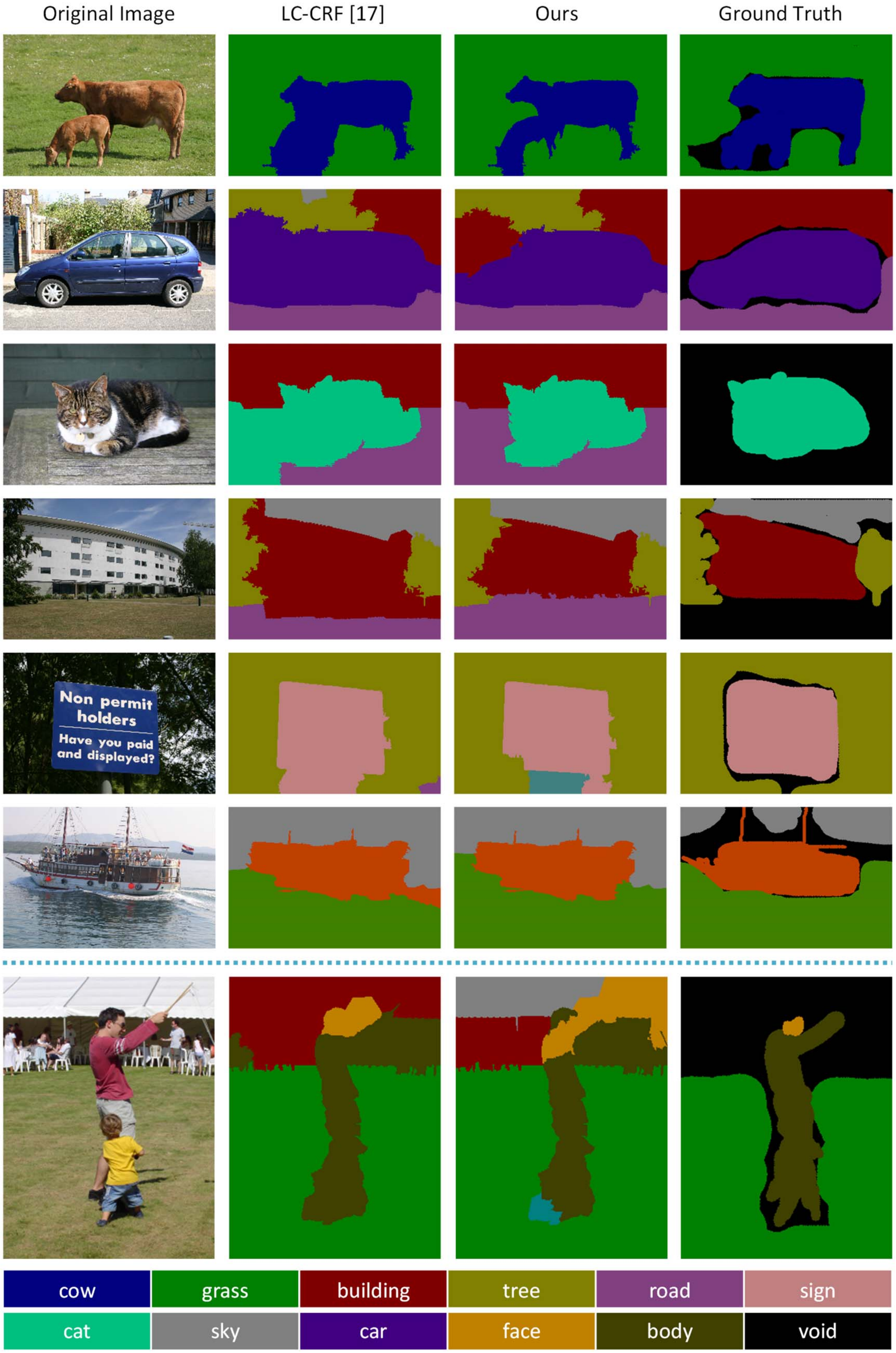

Fig. 7. Sample results on MSRC-21.

accuracy. For average per-class accuracy, our approach achieves 47.7%, outperforming both state-of-the-art weakly supervised methods and baseline fully supervised methods. It demonstrates that we take advantage of the limited priors efficiently. However, our accuracy of the "background" is much lower than that of [3]. This is understandable since our method

TABLE III
AVERAGE PER-CLASS ACCURACY (%) ON MSRC-21

	Method	building	grass	tree	cow	sheep	sky	plane	water	face	car	bike	flower	sign	bird	book	chair	road	cat	dog	body	boat	mean
FS	[50]	**75**	93	78	70	79	88	66	63	75	76	81	74	44	25	75	24	79	54	55	43	18	63.6
	[8]	71	**98**	90	79	86	93	88	86	90	84	94	**98**	76	53	**97**	71	**89**	83	55	68	17	79.3
	Cpmc [7]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**94.5**
	HP [51]	66	87	84	81	83	93	81	82	78	**86**	94	96	**87**	48	90	81	82	82	75	70	52	79.9
	MRP [10]	70	**98**	87	76	79	96	81	75	86	74	88	96	72	36	90	79	87	74	60	54	35	75.9
	LC-CRF [15]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	79.7
	CNN-CRF [13]	71	95	**92**	**87**	**98**	**97**	**97**	**89**	**95**	85	**96**	94	75	**76**	89	**84**	88	**97**	**77**	**87**	**52**	86.7
WS	MIML [52]	7	96	18	32	6	**99**	0	46	**97**	54	74	54	14	9	82	1	28	47	5	0	0	36.6
	MIM [23]	12	83	70	81	**93**	84	**91**	55	**97**	87	92	82	69	51	61	59	66	53	44	9	58	66.5
	SR [26]	63	93	**92**	62	75	78	79	64	95	79	93	62	76	32	95	48	**83**	63	38	68	15	69.2
	PAM [27]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	80
	LC [4]	**89**	**97**	89	**94**	92	96	89	**87**	88	**96**	89	87	90	**82**	89	79	77	87	**89**	**88**	**96**	**89.0**
	Ours	56	92	75	80	88	93	89	76	80	83	**94**	**99**	**95**	51	**97**	**88**	58	**98**	83	58	86	81.9

is designed to capture the most representative information of each category. The clutter background leads to the dispersion in the feature space. For the reasons above, our model tends to assign nonbackground labels to image regions in the test. This helps us recognize and segment more tangible objects.

To further inspect whether the performance benefit comes from the novel framework or the state-of-the-art deep feature, we conduct experiments of our framework with different features. The acknowledged SIFT feature [53] is considered to be one of the most powerful manually designed features in image segmentation. For fair comparison, a same 128-dimension SIFT vector is extracted to describe each superpixel. We compare the mean accuracy in Table II. Although the CNN feature brings higher accuracies to our framework than the SIFT feature, the "LAB+SIFT" combination achieves an accuracy of 45.2% which still outperforms state-of-the-art weakly supervised methods. We present performance variation curve in Fig. 4, key parameters containing image piece number K in every training image, the confidence ϵ in (6) and the neighborhood threshold η in the testing. Note that $\text{norm}_0(\boldsymbol{l}) = \|\boldsymbol{l}\|_0$ in the figure. It is apparent that the change of K exerts the greatest influence among the three parameters. The reason lies in that all the following system steps are on the basis of image pieces, which forces us to ensure its high quality by controlling the piece number in a proper range. In Fig. 5, qualitative segmentation samples on partially labeled database PASCAL VOC 2007 are presented, the last row showing the failing cases. Analyzing the failures we can conclude that:

1) the strong illumination contrast would impact the performance of our framework, mainly due to the scarce priors for training;
2) the approximation between object and background would fuzz up the boundary which is rather difficult even for human eyes to tell apart, and thus deteriorate the process of piece merging.

2) MSRC-21: The per-class accuracy of our method is compared with fully supervised models and weakly supervised models on MSRC-21 in Table III. It can be concluded that our result is better than or comparable to state-of-the-art weakly supervised results. It is also worth noticing that the performance of our model is comparable to the fully supervised approaches while much less information is required by our system. The average accuracy achieves 89% in [4], which is remarkable and surpass most fully supervised approaches.

TABLE IV
MEAN IOU SCORE ON VOC 12 *Test* SET

	Method	mean IoU
FS	O_2P [54]	47.8
	SDS [41]	51.6
	FCN [14]	62.2
	DeepLab-CRF [56]	66.4
	CRF-RNN [17]	72.0
WS	MIL-FCN [55]	25.7
	EM-Adapt [57]	39.6
	MIL-seg [28]	40.6
	Ours	33.4

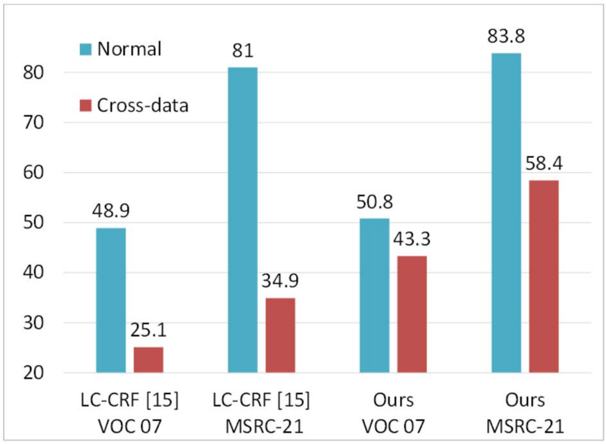

Fig. 8. Comparison of average accuracy between normal and cross-dataset experiments.

However, since the training in [4] put all the images regions together to learn the model, it is reported to take about 30 s to train on MSRC-21, while ours only takes less than 15 s on computers with similar computational capability, as is shown in Fig. 6. The computational cost reduction benefits a lot from the first step of our framework: merging superpixels into larger pieces, greatly reducing node amount in the latter graph. We merge superpixels in a training image with $\|\boldsymbol{l}\|_0$ labels into $2\|\boldsymbol{l}\|_0$ pieces. And in MSRC-21, the label number of a training image ranges from 2 to 5, which is close to the real scenario. The merging is conducted within each image, leading the number of the piece centers in (4) to no more than 10. Even for the larger set VOC 07, there are in total 2902 pieces from 422 training images in the pool.

TABLE V
ACCURACY (%) OF THE CROSS-DATASET EXPERIMENT

Test set	Train set	plane	bike	bird	boat	car	cat	chair	cow	dog	person	sheep	mean
VOC 07	VOC 07	46	35	42	45	38	65	60	23	67	61	77	50.8
	MSRC-21	31	74	29	25	31	34	49	46	56	58	43	43.3
MSRC-21	MSRC-21	89	91	63	86	83	92	87	79	83	81	88	83.8
	VOC 07	74	87	32	65	76	37	45	51	34	73	68	58.4

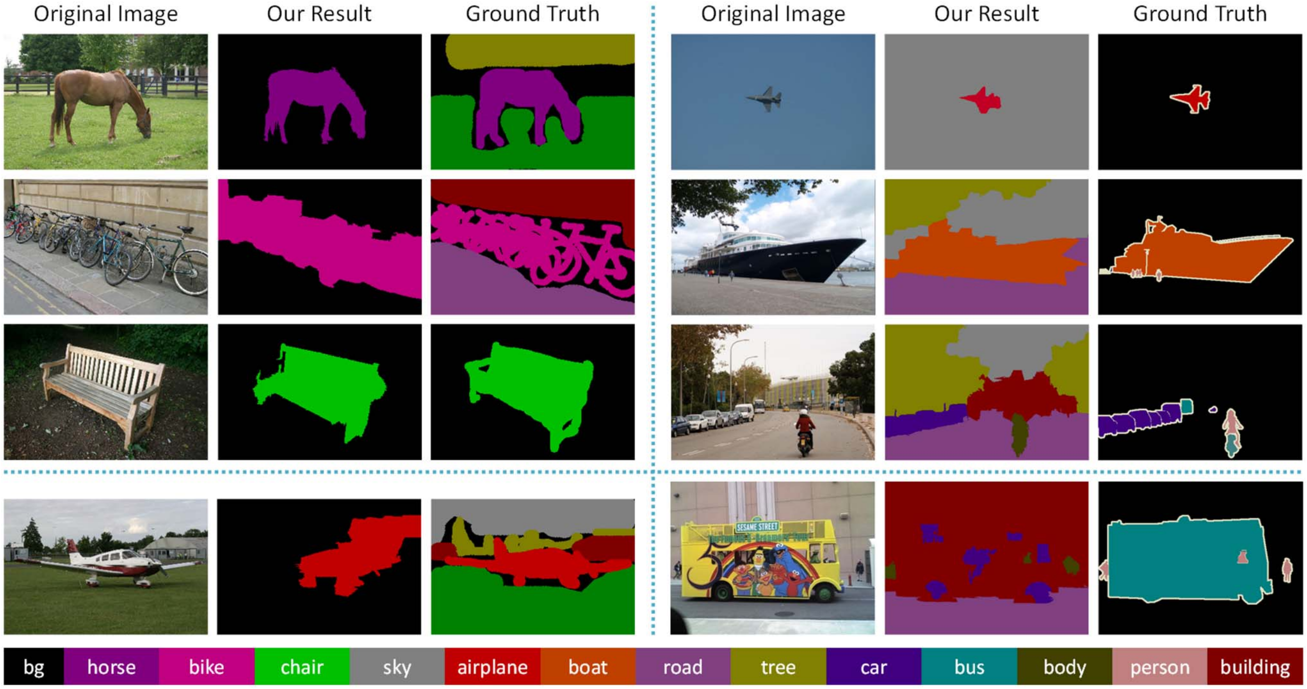

Fig. 9. Example segmentation results of the cross-dataset experiment.

Compared with the original superpixels, the merging process brings remarkable data compression. Some visualized results are exhibited in Fig. 7. Our results are compared with results from fully supervised method in [15] with the last row exhibiting a failing sample. For fair comparison, we use the same superpixel setting as our method to rerun the algorithm in [15]. It can be observed that our model achieves qualitatively better segmentation.

3) PASCAL VOC 2012: To examine the scalability of the proposed framework, we also conduct experiments on PASCAL VOC 2012. As has been mentioned above, this database is much larger than the former two. In stead of accuracy, we adopt IoU score as the metric and the results are compared in Table IV. Among these methods, O_2P [54] is a leading non-CNN-based approach, while the others are all based on deep networks. Technically, our framework is also nondeep since we establish it on the basis of CRF, though we use the CNN feature in our experiments. The result of our framework outperforms multiple instance learning fully convolutional network (MIL-FCN) [55] by nearly 8% and is also comparable with the scores of state-of-the-art methods. For the reason that there are roughly millions of parameters in a deep network, these deep learning methods require extremely more data than the traditional ones for training. It leads to the fact that they have to be trained for days even on a powerful graphics processing unit, whereas our method has the advantage of achieving more time efficiency. However, the large training data scale results in the macro library, and thus indeed hold back the system running speed. For every superpixel in a test image, our system has to compare it with every piece in the library to find its close neighbors. And we put all the merged pieces in the library without valid sieving. It would finally lead to the library data dispersity especially in the face of large data. The designed testing unary potential helps to alleviate the problem, but it remains improvable. In the future, we will consider to integrate better retrieval mechanism and piece eliminating strategy into our framework.

C. Cross-Data Performance

After analyzing the labels in PASCAL VOC 2007 and MSRC-21, we find that 11 out of 21 categories are shared by the two databases. Among the shared 11 categories, the "person" from MSRC-21 is the combination of "face" and "body." There are both indoor and outdoor images in PASCAL VOC 2007. Its label set emphasizes more on object categories, while summarize the others as the "background." Whereas the labels in MSRC-21 divide the background into more detailed classes: "grass," "sky," "water," etc.

In this section, we conduct cross-dataset experiments by training on one database while testing on the other one, inspired by [58]. The parameters in this section remain the same as the mentioned experiments. It is unsurprising that the cross-dataset test brings a drop to the accuracy, considering the great gap between the databases. The comparison of average accuracy over the 11 classes is shown in Fig. 8. Our framework succeeds in controlling the drop within a smaller range than [15]. Table V presents the per-class accuracy of the shared 11 categories. It is observed that the "bicycle" in VOC 2007 achieves a higher accuracy when trained on MSRC-21. The reason is that, the bicycles in MSRC-21 are easier to recognize in the training phase than VOC 2007. In other words, our framework is promising when trained on naive data but tested on more complex ones, which is of great practicality.

Another advantage of our framework is the ability of segmenting extra categories that are not defined in the original database. There are some qualitative results of the cross-dataset experiment in Fig. 9. The left column is the result of our system trained on PASCAL VOC 2007 while tested on MSRC-21, and the right column is the inverse case. The horse on the left and the road, tree and building on the right are all extra classes recognized and segmented by our system. In the cross-dataset experiment, the different label characteristics facilitate the segmentation of the classes that are not defined in the original database, demonstrating the flexibility and universality of our approach.

V. Conclusion

In this paper, we propose a novel weakly supervised semantic segmentation framework based on CRF, to incorporate cues from multiple structural priors. Our aim is to capture the most representative information of each semantic category. To that end, we first merge superpixels into larger pieces to attach more semantic priors. It also contributes much to computational cost reduction. Then all pieces are gathered and associated with appropriate semantic labels by CRF, based on their image-level annotation. In the associating process, the framework effectively utilizes label correlation, including co-occurrence statistics and label similarity. The piece library constructed in our framework has the advantages of universality and flexibility. Extensive experiments are conducted on PASCAL VOC 2007, MSRC-21, and VOC 2012 databases. The results demonstrate that our approach outperforms or is comparable to state-of-the-art segmentation methods, both fully and weakly supervised. Furthermore, we conduct cross-dataset experiment to verify the robustness of our method and a promising performance is achieved.

Acknowledgment

The authors would like to thank the associate editor and the reviewers for their valuable comments and suggestions.

References

[1] J. Lei, X. You, and M. Abdel-Mottaleb, "Automatic ear landmark localization, segmentation, and pose classification in range images," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 2, pp. 165–176, Feb. 2016.

[2] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 447–456.

[3] W. Zhang, S. Zeng, D. Wang, and X. Xue, "Weakly supervised semantic segmentation for social images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 2718–2726.

[4] N. Pourian, S. Karthikeyan, and B. S. Manjunath, "Weakly supervised graph based semantic segmentation by learning communities of image-parts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1359–1367.

[5] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Graph cut based inference with co-occurrence statistics," in *Proc. Eur. Conf. Comput. Vis.*, Heraklion, Greece, 2010, pp. 239–253.

[6] H. Lu, G. Fang, X. Shao, and X. Li, "Segmenting human from photo images based on a coarse-to-fine scheme," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 889–899, Jun. 2012.

[7] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1312–1328, Jul. 2012.

[8] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 702–709.

[9] Y.-L. Hou and G. K. H. Pang, "Multicue-based crowd segmentation using appearance and motion," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 2, pp. 356–369, Mar. 2013.

[10] D. Pei, Z. Li, R. Ji, and F. Sun, "Efficient semantic image segmentation with multi-class ranking prior," *Comput. Vis. Image Understand.*, vol. 120, pp. 81–90, Mar. 2014.

[11] X. Yuan, J. Guo, X. Hao, and H. Chen, "Traffic sign detection via graph-based ranking and segmentation algorithms," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 12, pp. 1509–1521, Dec. 2015.

[12] Y.-T. Chen, X. Liu, and M.-H. Yang, "Multi-instance object segmentation with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3470–3478.

[13] F. Liu, G. Lin, and C. Shen, "CRF learning with CNN features for image segmentation," *Pattern Recognit.*, vol. 48, no. 10, pp. 2983–2992, 2015.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3431–3440.

[15] Y. Li, Y. Guo, J. Guo, M. Li, and X. Kong, "CRF with locality-consistent dictionary learning for semantic segmentation," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit.*, Kuala Lumpur, Malaysia, 2015, pp. 509–513.

[16] J. Wang and A. L. Yuille, "Semantic part segmentation using compositional model combining shape and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1788–1797.

[17] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1529–1537.

[18] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Comput. Vis. Image Understand.*, vol. 110, no. 2, pp. 260–280, 2008.

[19] F. Wang, Q. Huang, M. Ovsjanikov, and L. J. Guibas, "Unsupervised multi-class joint image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 3142–3149.

[20] L. Khelifi and M. Mignotte, "A novel fusion approach based on the global consistency criterion to fusing multiple segmentations," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[21] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[22] G. Csurka and F. Perronnin, "A simple high performance approach to semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, Leeds, U.K., 2008, pp. 1–10.

[23] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 643–650.

[24] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, "Weakly-supervised dual clustering for image semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 2075–2082.

[25] L. Zhang *et al.*, "Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 1908–1915.

[26] K. Zhang, W. Zhang, Y. Zheng, and X. Xue, "Sparse reconstruction for weakly supervised semantic segmentation," in *Proc. Int. Joint Conf. Artif. Intell.*, Beijing, China, 2013, pp. 1889–1895.

[27] L. Zhang *et al.*, "A probabilistic associative model for segmenting weakly supervised images," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4150–4159, Sep. 2014.

[28] P. O. Pinheiro and R. Collobert, “From image-level to pixel-level labeling with convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1713–1721.

[29] J. Dai, K. He, and J. Sun, “Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1635–1643.

[30] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proc. 18th Int. Conf. Mach. Learn.*, vol. 1. Williamstown, MA, USA, 2001, pp. 282–289.

[31] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[32] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani, “Self-taught object localization with deep networks,” presented at the 2016 IEEE Winter Conf. Applications of Computer Vision (WACV), Mar. 7–10, 2016. doi: 10.1109/WACV.2016.7477688

[33] Z. Li, J. Liu, J. Tang, and H. Lu, “Robust structured subspace learning for data representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.

[34] J. Tang, Z. Li, M. Wang, and R. Zhao, “Neighborhood discriminant hashing for large-scale image retrieval,” *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, Sep. 2015.

[35] J. Shotton, M. Johnson, and R. Cipolla, “Semantic texton forests for image categorization and segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.

[36] C. H. Lampert, M. B. Blaschko, and T. Hofmann, “Beyond sliding windows: Object localization by efficient subwindow search,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.

[37] V. Viitaniemi and J. Laaksonen, “Techniques for image classification, object detection and object segmentation,” in *Proc. Int. Conf. Adv. Vis. Inf. Syst.*, Salerno, Italy, 2008, pp. 231–234.

[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[39] M. Everingham *et al.*, “The Pascal visual object classes challenge: A retrospective,” *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[40] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 991–998.

[41] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, “Simultaneous detection and segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, 2014, pp. 297–312.

[42] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, “Multiscale combinatorial grouping,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 328–335.

[43] R. Achanta *et al.*, “Slic superpixels,” School Comput. Commun. Sci., École Polytech. Fédrale de Lausanne, Lausanne, Switzerland, Tech. Rep. 149300, 2010.

[44] Y. Jia *et al.*, “Caffe: Convolutional architecture for fast feature embedding,” in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 675–678.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[46] I. T. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2002.

[47] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[48] V. Kolmogorov and R. Zabin, “What energy functions can be minimized via graph cuts?” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[49] Y. Boykov and V. Kolmogorov, “An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[50] G. Csurka and F. Perronnin, “An efficient approach to semantic segmentation,” *Int. J. Comput. Vis.*, vol. 95, no. 2, pp. 198–212, 2011.

[51] X. Boix *et al.*, “Harmony potentials,” *Int. J. Comput. Vis.*, vol. 96, no. 1, pp. 83–102, 2012.

[52] A. Vezhnevets and J. M. Buhmann, “Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, 2010, pp. 3249–3256.

[53] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[54] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, “Semantic segmentation with second-order pooling,” in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, 2012, pp. 430–443.

[55] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional multi-class multiple instance learning,” in *Proc. Int. Conf. Learning Representations*, San Diego, CA, USA, May 2015.

[56] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” in *Proc. Int. Conf. Learning Representations*, San Diego, CA, USA, May 2015.

[57] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, “Weakly- and semi-supervised learning of a DCNN for semantic image segmentation,” in *Proc. IEEE Int. Conf. Computer Vision*, Santiago, Chile, Dec. 2015, pp. 1742–1750.

[58] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 1521–1528.

Yi Li received the B.E. degree in electronic and information engineering from the Dalian University of Technology, Dalian, China, in 2014, where she is currently pursuing the postgraduate degree with the School of Information and Communication Engineering.

Her research interests include computer vision and machine learning.

Yanqing Guo (M’13) received the B.S. and Ph.D. degrees in electronic engineering from the Dalian University of Technology of China, Dalian, China, in 2002 and 2009, respectively.

He is currently an Associate Professor with the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. His research interests include machine learning, computer vision, and multimedia security.

Yueying Kao received the B.E. degree from Xidian University, Xi’an, China. She is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Her research interests include computer vision and machine learning.

Ran He (M’09–SM’15) received the B.E. and M.S. degrees in computer science from the Dalian University of Technology, Dalian, China, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2001, 2004, and 2009, respectively.

Since 2010, he has been with the Institute of Automation, Chinese Academy of Sciences, where he is currently a Professor. His research interests include information theoretic learning and computer vision.