Joint CRF and Locality-Consistent Dictionary Learning for Semantic Segmentation

Yi Li, Yanqing Guo, Jun Guo, Zhuang Ma, Xiangwei Kong, and Qian Liu

***Abstract*—Semantic image segmentation can be accomplished by assigning a proper object category label to each meaningful region of an image. Beyond the original bottom-up models, the use of top-down categorization information has been applied to semantic segmentation to improve performance. An excellent example of such a top-down scheme is to integrate a Conditional Random Field (CRF) model with sparse dictionary learning. However, the existing solutions merely consider the discrimination of dictionaries to obtain better sparse codes, without considering the inherent data locality characteristics. In this paper, we explore such characteristics and propose a novel semantic segmentation framework based on an innovative CRF model with locality-consistent dictionary learning. In particular, we propose two new locality-consistent dictionary learning strategies by capturing the local consistencies in the feature space and the label space. In addition, we develop a joint dictionary and a CRF model parameter learning algorithm to seamlessly integrate the proposed locality-consistent dictionary learning strategies into the CRF model. Extensive experiments are conducted with two popular databases of different traits (i.e., Graz-02 and PASCAL-CONTEXT). The simulation results confirm the efficiency of the proposed scheme, especially when training data are limited.**

***Index Terms*—Image semantic segmentation, conditional random field, dictionary learning, locality consistency.**

Manuscript received January 14, 2017; revised November 16, 2017 and February 20, 2018; accepted August 1, 2018. Date of publication August 29, 2018; date of current version March 22, 2019. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants U1636219 and 61379151, 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, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT18JC06. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wenwu Zhu. This paper was presented in part at the 3rd IAPR Asian Conference on Pattern Recognition, Kuala Lumpur, Malaysia, November 2015. *(Corresponding author: Yanqing Guo.)*

Y. Li was with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China, and now with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, CASIA, Beijing 100190, China (e-mail: liyi@mail.dlut.edu.cn).

Y. Guo, Z. Ma, and X. Kong are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: guoyq@dlut.edu.cn; mz_dip@mail.dlut.edu.cn; kongxw@dlut.edu.cn).

J. Guo is with the Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China (e-mail: guoj16@mails.tsinghua.edu.cn).

Q. Liu is with the Department of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China, and also with the Chair of Media Technology and the Chair of Communication Networks, Technical University of Munich, Munich 80333, Germany (e-mail: qianliu@dlut.edu.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/TMM.2018.2867720

I. Introduction

SEMANTIC image segmentation can be considered a "pixel labeling" task that assigns predefined object labels to groups of meaningful pixels (or objects) in an image. It has raised tremendous interest in computer vision research, with various applications ranging from video analysis [1] to clothing retrieval [2]. The major challenge of semantic segmentation is that the predefined objects may have a significant amount of variation in their appearance, viewpoint and background. As a result, a single visual element (e.g., pixel, patch or superpixel) is too small to maintain sufficient information for object categorization [3].

Over the past several decades, the Conditional Random Field (CRF) [4] has been considered one of the most classical models in segmentation by combining the benefits of graphical modeling and classification. The CRF is able to leverage low-level contextual information and is particularly suitable for structural prediction, e.g., for image segmentation. Early works, e.g., [5], [6], modeled an image as a second-order CRF with unary and pairwise potentials. Such approaches generally obtained an efficient Maximum a Posteriori (MAP) solution via graph cuts [7], [8] or other approximate graph inference algorithms. However, due to the weak expressive ability of the second-order CRF, these algorithms cannot handle long-range interactions among objects or miscellaneous object deformations.

To address this issue, Kohli *et al.* [9] showed the superior performance of higher-order potentials in the form of a Robust P^n model to that of the traditional second-order CRF formulation. Nevertheless, the higher order incurred higher computational complexity and was generally more time-consuming. Another promising approach was to simultaneously leverage both high-level object class priors and low-level contextual information. The related solutions explored shape priors, scene information and other top-down clues of objects, integrating them into CRF potentials. Based on the Bag of Features (BoF) model, [10] and [11] augmented the second-order CRF with a global top-down categorization potential added to the energy. Compared with the previous algorithms, this scheme could obtain category-level priors for effective dictionary-based representation. However, a drawback of this scheme was that the visual words in [11] were determined by classifiers independent of the CRF. As a result, [12] proposed extending the framework by learning a discriminative dictionary jointly with CRF parameters. Nonetheless, the extended method was restricted to the BoF representation. Since recently, the sparse representation has begun to demonstrate its superiority over the basic BoF model for image classification.

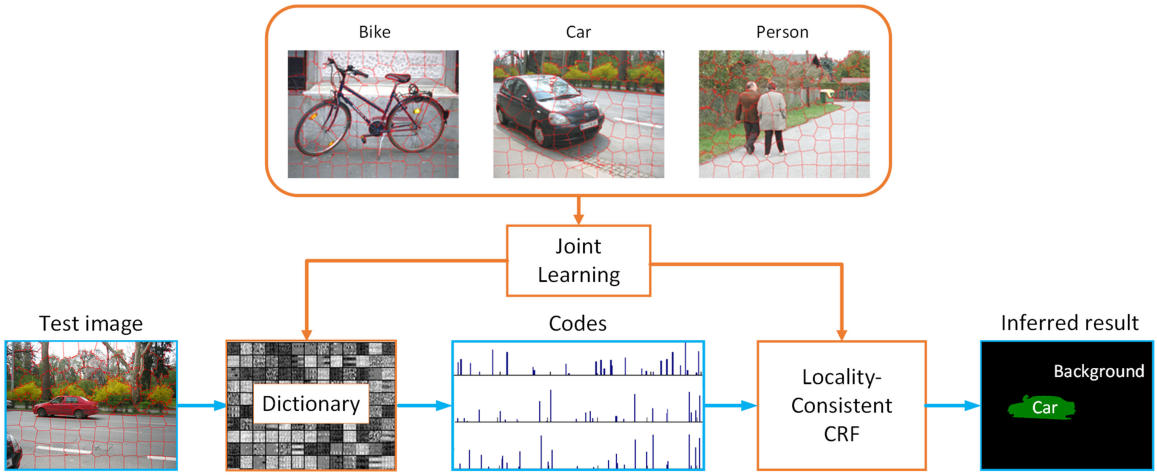


Fig. 1. An overview of the system framework with a joint locality-consistent dictionary and CRF model parameter learning (see Sec. IV for details. The orange lines represent the learning phase, and cyan denotes the testing phase. Given a test image, we first encode all superpixels in the trained dictionary and then utilize the locality-consistent CRF to infer the segmentation. The data locality is well preserved by obtaining similar codes for similar superpixels.

Hence, [13] further improved the framework proposed in [11] by incorporating a discriminative sparse dictionary learning-based top-down potential into the CRF.

It is noted that the above solutions can be considered amendments to the second-order CRF approach because they add an extra top-down potential to the basic CRF infrastructure. In addition, the above algorithms only considered the discrimination of dictionaries for sparse code design, where the underlying local consistency of the original input data was completely ignored.

Recently, [14] and [15] revealed an important phenomenon in image sparse representation, that similar inputs (e.g., neighbors in the feature space) result in similar codes due to local-consistency (or locality) of input data. However, in the existing CRF-based semantic segmentation solutions, the sparse representation only encodes each data point as a linear combination of dictionaries without considering the locality of input data. A simple linear combination of dictionaries uses dictionary items from different subspaces and therefore leads to a scattered distribution and ultimately degrades the discrimination of sparse codes. Hence, we can conclude that ignoring data locality will impair the effect of top-down categorization information in CRF and hinder the segmentation performance.

In this paper, we propose a novel semantic segmentation scheme to overcome the shortcomings of existing solutions. The proposed scheme is based on an innovative CRF model that can preferentially preserve the locality of original data when utilizing top-down category priors. In contrast to the existing approaches (which only introduce an extra potential into the second-order CRF), in the proposed scheme, we consider the sparse codes with local consistency to be latent variables in the CRF model. Thus, there is no need to develop a customized inference algorithm for our system. In the preliminary stage of this research [16], we developed a novel locality-consistent dictionary learning strategy in the feature space. In this paper, we move one step further by efficiently exploiting locality in both feature and label spaces. From a top-down schematic perspective, the CRF model in the proposed scheme is to capture the low-level contexture information in the location, and the dictionary learning is to structure the high-level relationship in the feature/label space. In particular, the first locality-consistent dictionary learning strategy is developed from the input perspective, aiming to generate similar sparse codes for similar superpixels of an image. It preserves the locality in the feature space and is unsupervised. In contrast, the second locality-consistent dictionary learning strategy is developed from the output perspective, focusing on obtaining similar codes for superpixels with the same labels. It can be considered a supervised strategy. It is worth noting that the dictionary and CRF parameters should be learned simultaneously in the proposed scheme. To this end, we develop a joint dictionary and model parameter learning algorithm, enabling a seamless integration of a locality-consistent dictionary learning strategy with the CRF model. An illustration of the proposed scheme is exhibited in Fig. 1.

The main contributions of this research can be summarized as follows.

1. We propose a novel semantic segmentation scheme based on an innovative CRF model with locality-consistent dictionary learning. The proposed scheme effectively leverages the inherent locality characteristics of the input data into semantic segmentation.
2. Two novel dictionary learning strategies are presented, exploiting the locality-consistency of an image in the feature space and the label space.
3. The tight integration of the dictionary learning and CRF parameters learning makes the proposed scheme profoundly distinct from the existing solutions, so it can be considered a new reference infrastructure for sematic-oriented segmentation designs.

II. Related Works

A. Semantic Segmentation

As a problem that developed from image segmentation, semantic segmentation has been widely studied. Some works formulated the problem as CRFs. He *et al.* [5] proposed a multiscale framework for incorporating contextual features into the CRF

model. TextonBoost, combining appearance, shape and context information, was developed and integrated in the CRF model for object recognition and segmentation [6]. The work in [17] improved the performance of a pixel-based CRF using cues from object detectors. [18] extended [17] with shape and scene classification priors and proposed a region-based method for holistic scene parsing. Tao *et al.* [13] further proposed a semantic segmentation scheme, appending a discriminative sparse dictionary learning-based potential to the CRF model. We highlight the locality of the input data to encourage the discrimination of the dictionary, which can bring a promising improvement to performance.

Recently, network-based semantic segmentation and object parsing approaches [19]–[24] have attracted extensive attention with the rise of deep learning. Grangier *et al.* [19] applied convolutional networks to scene parsing by training a deep convolutional neural network (CNN) with a supervised greedy learning strategy. Liu *et al.* [20] utilized a deep CNN that was pretrained on ImageNet [25] to extract deep features and further used it as the input of CRF with spatially related co-occurrence potentials. [21] used the response of deep CNN as the input of a fully connected CRF for segmentation refinement, which successfully resolved the poor localization property of deep networks. Long *et al.* [22] proposed to build fully convolutional networks (FCN) that make the output have the same size as the input. They adapted contemporary classification networks to FCN and produced detailed segmentations. A novel deep Local-Global Long Short-Term Memory (LG-LSTM) architecture was developed in [24] for semantic object parsing. The system was able to capture both short- and long-distance spatial dependencies simultaneously for the feature learning. However, the deep learning approaches usually require tremendous amount of images for training, and their performance will degrade dramatically when limited resources are available.

B. Conditional Random Field

CRF was proposed by Lafferty *et al.* [4] and first utilized in the field of natural language processing. It has recently been widely adopted in image segmentation. Initially, a second-order CRF was directly built on an image based on its pixel-wise (or patch-wise) observation and the corresponding labels. Kumar *et al.* [26] first imported the 1-D CRF to 2-D images and presented Discriminative Random Fields (DRFs) for image understanding. After years of development, CRF in semantic segmentation is often utilized to harmonize different cues [27], [28] or refine the results from other systems [21], [29]. In contrast, we integrate top-down objective cues from locality-consistent sparse coding into CRF into our solution. Our CRF parameters are basically node classifier coefficients on sparse codes instead of coefficients of different cues, as in [27]. Additionally, a dictionary for sparse coding is jointly learned with the CRF parameters, which means that the codes are actually latent variables in the algorithm. Although they share similar characters, the hidden CRF (HCRF) is a discriminative latent variable model for structured classification and recognition [30]–[32] that has inherent differences from our approach. A vector of latent variables was included in the HCRF to indicate unobserved/intermediate labels of image parts and was used for the single label decision of each image. However, our approach regards sparse codes as latent variables of CRF to perform discriminative dictionary learning.

C. Dictionary Learning

Olshausen and Field [33] first introduced the idea of learning a dictionary from data, improving conventional signal reconstruction significantly with predefined dictionaries. Then, unsupervised dictionary learning was applied to image classification [34]. Aharon *et al.* [35] proposed a K-SVD algorithm to find the dictionary that best represented the training signals under strict sparsity constraints. Nevertheless, the supervised dictionary learning has been extensively studied [36]–[38] because the performance of sparse coding depends strongly on the quality of its dictionary. [36] presented a general framework called task-driven dictionary learning, which was able to adapt different representation coefficients to different tasks. The label-consistent K-SVD (LC-KSVD) algorithm, demonstrated in [37], enforced dictionary discrimination by associating labels with dictionary items. A structured dictionary was learned based on the Fisher discrimination criterion in [38]. Instead of classification or recognition, as mentioned above, we focus on the problem of semantic segmentation and present both unsupervised and supervised strategies for locality-consistent dictionary learning and how to incorporate them into the CRF.

III. Conditional Random Field

The CRF directly builds the posterior distribution of the label field conditioned on the observation field. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m \times n}$ be an image with n superpixels (each represented by an m-dimensional descriptor), $\boldsymbol{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$ be the corresponding labels, where $y_i \in \{1, \ldots, L\}$ and L denote the total number of categories. Each superpixel $\boldsymbol{x}_i$ is considered a node, with two nodes connected by an edge if they are adjacent in a certain space, e.g., the feature space. Then, a graph $G = \{V, E\}$ is built on the image, where V refers to the nodes and E the edges. The CRF models the posterior distribution with a Gibbs distribution in the form of

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

where θ is the set of model parameters and Z is the normalization factor. In a second-order CRF, the energy can be written as

$$E(\boldsymbol{y}, \boldsymbol{X}, \theta) = \sum_{i \in V} \phi_1(y_i, \boldsymbol{X}, \theta) + \sum_{(i,j) \in E} \phi_2(y_i, y_j, \boldsymbol{X}, \theta). \tag{2}$$

The unary potential ϕ_1 formulates the cost of assigning a label y_i to a node $\boldsymbol{x}_i$, and the pairwise potential ϕ_2 models the cost of assigning a pair of labels (y_i, y_j) to a pair of adjacent nodes $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ [13]. There are various CRF potentials, and the adopted functions are described below.

The objective is to obtain an optimal labeling that maximizes the conditional probability, i.e., MAP, which is equivalent to

minimizing the energy E:

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

E can be efficiently minimized by an α expansion, an $\alpha-\beta$ swap, or another approximate inference algorithm. There is an intuitive meaning of CRF: image regions that are alike or close to each other tend to share the same category label. Conversely, regions that are dissimilar or far from each other tend to have different labels.

IV. CRF with Locality-Consistent Dictionary Learning

When integrating sparse coding into a CRF model, we intend to efficiently leverage global top-down categorization priors and preserve the locality relationship of original data synchronously to achieve greater discrimination and a higher segmentation quality. As a result, we exploit the local-consistency of image superpixels by embodying the similarities of superpixels in terms of both location and feature/label space in our algorithm. Two strategies that incorporate locality-consistent dictionary learning are to present from the input or the output views. As illustrated in Fig. 1, the detailed description of our proposed model is as follows.

A. Basic Model

We structure the segmentation problem as classifying superpixels in an image. As the same classification scheme will be applied to all superpixels, they are expected to have a similar size or area. Therefore, we first over-segment each training image into superpixels following [39]. Each superpixel is denoted by an m-dimensional descriptor $\boldsymbol{x}_i$ and is considered a node when building the graph of an image. Assume $\boldsymbol{D} \in \mathbb{R}^{m\times K}$ is a dictionary to be learned, where K denotes the dictionary size. Then, each superpixel descriptor $\boldsymbol{x}_i$ can be encoded over $\boldsymbol{D}$ via the LASSO model [40] to produce a sparse representation. The optimization is formulated as

$$\boldsymbol{a}_i = \arg\min_{\boldsymbol{a}} \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}\|^2 + \lambda\|\boldsymbol{a}\|_1 \quad (4)$$

where λ controls the sparse penalty.

Note that $\boldsymbol{a}_i$ represents an optimization problem instead of a variable. Accordingly, $\boldsymbol{A} = [\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n] \in \mathbb{R}^{K\times n}$ is incorporated in Eq. (1) as a latent variable, with the energy function in Eq. (2) becoming

$$E(\boldsymbol{y},\boldsymbol{A},\theta) = \sum_{i\in V}\phi_1(y_i,\boldsymbol{A},\theta) + \sum_{(i,j)\in E}\phi_2(y_i,y_j,\boldsymbol{A},\theta) \quad (5)$$

where $\boldsymbol{A}$ depends nonlinearly on $\boldsymbol{D}$. Define $\theta \triangleq \boldsymbol{W} = [\boldsymbol{w}_1,\ldots,\boldsymbol{w}_L]^{\mathrm{T}} \in \mathbb{R}^{L\times K}$ where $\boldsymbol{w}_l$ are the coefficients of a linear classifier for the l-th category. Then, the unary and pairwise potentials can be written as

$$\phi_1(y_i,\boldsymbol{A},\boldsymbol{W}) = \langle \boldsymbol{w}_{l=y_i}, -\boldsymbol{a}_i\rangle = -\boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{W}\boldsymbol{a}_i \quad (6)$$

and

$$\phi_2(y_i,y_j,\boldsymbol{A},\boldsymbol{W}) = \omega I(y_i \neq y_j) \quad (7)$$

where $\langle\cdot,\cdot\rangle$ represents the inner product, $I(\cdot)$ is equal to 1 if the input is true, and ω controls the contribution of the pairwise potential. For the convenience of deriving the learning algorithm, we also provide the matrix form in Eq. (6), where $\boldsymbol{l}_i \in \mathbb{R}^L$ is the label vector of y_i. The l-th entry of $\boldsymbol{l}_i$ is set to 1 if $y_i = l$.

B. Locality-Consistent Sparse Coding

In this research, locality-consistency can be colloquially defined as encouraging similar inputs to have similar codes. In the following, we describe two locality-consistent sparse coding schemes to extend the basic model. For a semantic segmentation system, the input is an image, and the output is the corresponding semantic labels. The first scheme is considered from the system input view without using the label information of each superpixel, and the second one is considered from the output view with supervision from superpixel labels. Additionally, using both schemes is optional in our algorithm.

1) Unsupervised Strategy From Input View: An input image is described by a set of superpixel feature vectors $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$. To preserve the locality relationships of the input in the feature space, we allow visually similar superpixels in the image to have similar sparse codes, as shown in Fig. 2(b). In contrast to Eq. (4), a sparse representation $\boldsymbol{a}_i^I$ of $\boldsymbol{x}_i$ with locality consistency in the feature space is obtained by optimizing the following problem:

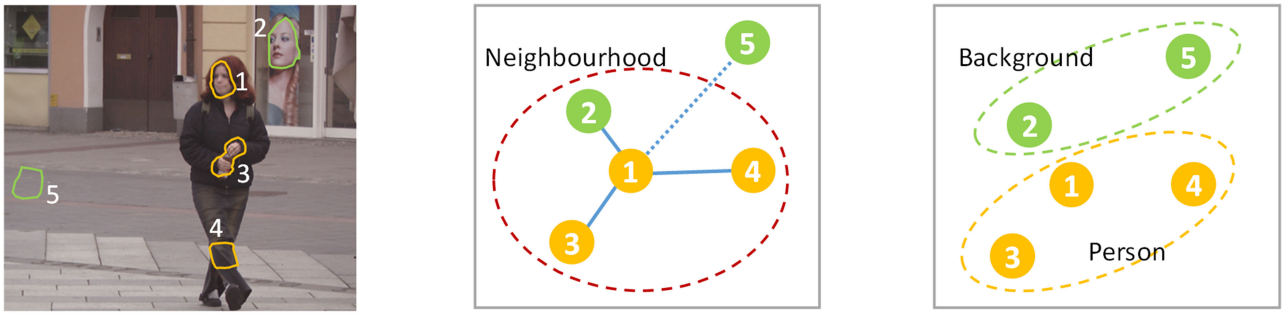
$$\boldsymbol{a}_i^I = \arg\min_{\boldsymbol{a}} \left\{ \frac{1-\gamma}{2}\|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}\|^2 + \frac{\gamma}{2}\sum_{\boldsymbol{z}_i\in\mathcal{N}(\boldsymbol{x}_i)} v_i\|\boldsymbol{z}_i - \boldsymbol{D}\boldsymbol{a}\|^2 + \lambda\|\boldsymbol{a}\|_1 \right\} \quad (8)$$

where $0 \le \gamma < 1$ weights the contribution of nearby points denoted by $\boldsymbol{z}_i$. $\mathcal{N}(\boldsymbol{x}_i)$ is the neighborhood set of $\boldsymbol{x}_i$ in the feature space, obtained by searching for k superpixels with the minimum Euclidean distance in image $\boldsymbol{X}$. v_i refers to the adapted weights for different $\boldsymbol{z}_i$ and $v_i = \exp\{-dist(\boldsymbol{x},\boldsymbol{z}_i)\}$ before normalization. Hence, the medial term is the locality constraint in the formula that enforces the reconstruction of $\boldsymbol{x}_i$, depending on its local neighborhood. It models the locality property: original inputs with uniform appearance lead to similar codes. Thus, the locality consistency in the feature space is preserved.

2) Supervised Strategy From Output View: The output of a semantic segmentation system is the inferred collection of labels for the input image. Based on the observation of system outputs, we propose preserving the label locality relationship in this section. In other words, superpixels with the same label in an image are encouraged to have similar sparse codes, as shown in Fig. 2(c). Acting as another substitute for Eq. (4) in the basic model, we propose to formulate the statement above as

$$\boldsymbol{a}_i^O = \arg\min_{\boldsymbol{a}} \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}\|^2 + \frac{\mu}{2}\sum_{j\in V\setminus i}\|\boldsymbol{a}-\boldsymbol{a}_j\|^2\boldsymbol{L}_{ij} + \lambda\|\boldsymbol{a}\|_1 \quad (9)$$

where μ is the weight of the label locality-consistent term and $\boldsymbol{L}$ is the label relationship matrix of the image. Its concrete form



(a) (b) (c)

Fig. 2. The proposed locality-consistent sparse coding strategies. (a) is a training image with 5 sample superpixels. Different colors indicate different labels, with orange for "person" and light green for "background". (b) The unsupervised strategy from input view allows visually alike superpixels (in the neighborhood circle) to have similar codes, preserving the locality in the feature space. (c) The supervised strategy from output view encourages superpxels with the same label to have similar codes, preserving the locality in the label space.

depends on the specific problem; we define it as

$$\boldsymbol{L}_{ij} = \begin{cases} +1, & \text{if } y_i = y_j \\ -1, & \text{if } y_i \neq y_j \\ 0, & \text{unassigned} \end{cases}. \tag{10}$$

In particular, the third branch refers to the situation in which y_i or y_j is unassigned (e.g., in the testing process), so their relation remains unclear. Similar to Eq. (8), the middle term ensures that the codes of superpixels with the same label in the image are similar but that codes of superpixels with different labels diverge. Thus, the label priors are well preserved when incorporating sparse coding into CRF.

C. Inference

Given the optimal CRF parameters $\boldsymbol{w}^*$ and a normalized dictionary $\boldsymbol{D}^*$, the proposed model is settled and can be utilized to infer the segmentation of a test image. Each test image is oversegmented into superpixels, and the corresponding graph G is built. In the following sections, we use the notation $\boldsymbol{a}_i$ to unify sparse coding optimization in Eq. (8) and Eq. (9) for concision. Note that the energy function in Eq. (5) does not employ any complex calculation of the latent variables, so it is plausible to decompose the inference into two steps.

For a test image $\boldsymbol{X}$,

Step 1: evaluate the sparse codes $\boldsymbol{A}$ over $\boldsymbol{D}^*$ by optimizing Eq. (8) or Eq. (4), and

Step 2: infer the semantic label vector $\boldsymbol{y}$ by minimizing the energy E in Eq. (5).

The pair relation of (y_i, y_j) remains unclear if $i \neq j$ and the corresponding $\boldsymbol{L}_{ij}$ is set to 0 because the superpixel labels of a test image are unknown. Therefore, the optimization in Eq. (9) reduces to Eq. (4) in the inference. The CRF in this paper is a multi-class CRF whose global optimum has been proved to be NP-hard. Thus, we employ an α expansion algorithm that can be efficiently solved by graph cuts for an approximate inference [7], [8], [41].

D. Joint Learning of Parameters and Dictionary

Inspired by [13], [42], we develop a solution for joint learning of CRF parameters and dictionary. Note that we follow the mathematical approach and the solution procedure of [13], [42]. However, the concrete formulae and the derivation are tailored to our proposed model. The detailed derivation proceeds as follows. Suppose that $\{\boldsymbol{X}^i\}_{i=1}^N$ is the training image set and $\{\boldsymbol{y}^i\}_{i=1}^N$ are the corresponding segmentation labels. The classifier coefficients $\boldsymbol{W}$ and the dictionary $\boldsymbol{D}$ are to be learned. Our model is clearly linear in $\boldsymbol{W}$. However, it is nonlinear in $\boldsymbol{D}$, and the dependency is implicit, which makes the learning procedure challenging. Nonetheless, if $\boldsymbol{D}$ is known, the proposed model is similar to most CRF models, and the large-margin framework can be adopted to learn the parameters, which is to solve the following optimization

$$\begin{aligned} \min_{\boldsymbol{W}, \{\xi_i\}} \quad & \frac{\beta}{2}||\boldsymbol{W}||_F^2 + \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & E(\boldsymbol{y}, \boldsymbol{A}^i, \boldsymbol{W}) - E(\boldsymbol{y}^i, \boldsymbol{A}^i, \boldsymbol{W}) \geq \Delta(\boldsymbol{y}, \boldsymbol{y}^i) - \xi_i, \\ & \forall i \in \{1, \ldots, N\}, \end{aligned} \tag{11}$$

where $\{\xi_i\}$ are the slack variables of the constraints, $\Delta(\cdot, \cdot)$ is a loss function, $\boldsymbol{y}^i$ is the ground truth label and $\boldsymbol{y}$ denotes the prediction. In general, $\Delta(\boldsymbol{y}^i, \boldsymbol{y}^i) = 0$, but $\Delta(\boldsymbol{y}, \boldsymbol{y}^i) > 0$ for any $\boldsymbol{y} \neq \boldsymbol{y}^i$. We define $\Delta(\boldsymbol{y}, \boldsymbol{y}^i) = \sum_{j=1}^{n} I(y_j, y_j^i)$ in this paper. The intuitive meaning of Eq. (11) is to determine the value of $\boldsymbol{W}$ with a small norm to make the energy of the ground truth label $E(\boldsymbol{y}^i, \boldsymbol{A}^i, \boldsymbol{W})$ smaller than that of any other incorrect label $E(\boldsymbol{y}, \boldsymbol{A}^i, \boldsymbol{W})$ by at least a margin of $\Delta(\boldsymbol{y}, \boldsymbol{y}^i)$ [20]. We first search for the most violated constraint by solving

$$\tilde{\boldsymbol{y}}^i = \arg\min_{\boldsymbol{y}} E(\boldsymbol{y}, \boldsymbol{A}^i, \boldsymbol{W}) - \Delta(\boldsymbol{y}, \boldsymbol{y}^i). \tag{12}$$

Then, the objective optimization in Eq. (11) can be rewritten as

$$f(\boldsymbol{W}, \boldsymbol{D}) = \frac{\beta}{2}||\boldsymbol{W}||_F^2 + \sum_{i=1}^{N} E(\boldsymbol{y}^i, \boldsymbol{A}^i, \boldsymbol{W}) - E(\tilde{\boldsymbol{y}}^i, \boldsymbol{A}^i, \boldsymbol{W}). \tag{13}$$

We adopt stochastic gradient descent to solve the optimization. The matrix form of Eq. (6) is used to simplify the derivation, as mentioned above. Given the initial values $\boldsymbol{D}_0$ and $\boldsymbol{W}_0$, we randomly select an instance $(\boldsymbol{X}, \boldsymbol{y})$ (i.e., $N = 1$) from the training set to update $\boldsymbol{D}$ and $\boldsymbol{W}$ according to their partial derivatives, respectively, at each iteration. Let $\boldsymbol{D}_t$ and $\boldsymbol{W}_t$ be the results of the t-th iteration, then the partial derivative of f with respect to $\boldsymbol{W}$ of the selected instance can be easily calculated by

$$\frac{\partial f}{\partial \boldsymbol{W}} = \beta \boldsymbol{W}_t + \sum_{i \in V} (\tilde{\boldsymbol{l}}_i - \boldsymbol{l}_i) \boldsymbol{a}_i^{\mathrm{T}}. \tag{14}$$

For the partial derivative of f with respect to $\boldsymbol{D}$, we utilize the chain rule of differentiation in the form of

$$\frac{\partial f}{\partial \boldsymbol{D}} = \sum_{i \in V} \left(\frac{\partial f}{\partial \boldsymbol{a}_i} \right)^{\mathrm{T}} \frac{\partial \boldsymbol{a}_i}{\partial \boldsymbol{D}} \tag{15}$$

since the energy depends implicitly on the dictionary. In the following, we omit the subscripts i and t for simplicity. Now, we calculate $\frac{\partial \boldsymbol{a}_i}{\partial \boldsymbol{D}}$. Under certain conditions, the sparse representation $\boldsymbol{a}$ in Eq. (8) and Eq. (9) must separately satisfy [36]

$$\boldsymbol{D}^T \left\{ \boldsymbol{D}\boldsymbol{a} - \left((1-\gamma)\boldsymbol{x} + \gamma \frac{1}{k} \sum_{\boldsymbol{z} \in \mathcal{N}(\boldsymbol{x})} \boldsymbol{z} \right) \right\} = -\lambda sign(\boldsymbol{a}) \tag{16}$$

and

$$\boldsymbol{D}^T (\boldsymbol{D}\boldsymbol{a} - \boldsymbol{x}) = -\lambda sign(\boldsymbol{a}) - \mu \sum_{j \in V \setminus i} \boldsymbol{L}_{ij} (\boldsymbol{a} - \boldsymbol{a}_j). \tag{17}$$

We now consider the unsupervised strategy of Sec. IV-B as an example. Suppose the active set in $\boldsymbol{a}$ (denoted by $\boldsymbol{a}_\Lambda$) does not change when $\boldsymbol{D}$ experiences a small perturbation. $\boldsymbol{D}_\Lambda$ is composed of columns in $\boldsymbol{D}$, corresponding to $\boldsymbol{a}_\Lambda$. Let $\boldsymbol{M}_I \triangleq (\boldsymbol{D}_\Lambda^T \boldsymbol{D}_\Lambda)^{-1}$ and $\boldsymbol{x}_{\boldsymbol{z}} \triangleq (1-\gamma)\boldsymbol{x} + \frac{\gamma}{k} \sum_{\boldsymbol{z} \in \mathcal{N}(\boldsymbol{x})} \boldsymbol{z}$. We then calculate the derivative of $\boldsymbol{D}$ in Eq. (16) and

$$\frac{\partial \boldsymbol{a}_{(k)}}{\partial \boldsymbol{D}} = (\boldsymbol{x}_{\boldsymbol{z}} - \boldsymbol{D}\boldsymbol{a})\, \boldsymbol{M}_{I[k]} - \left(\boldsymbol{D}\boldsymbol{M}_I{}^{\mathrm{T}} \right)_{\{k\}} \boldsymbol{a}^{\mathrm{T}}, \quad \forall k \in \Lambda \tag{18}$$

where (k) indicates the k-th entry of $\boldsymbol{a}$, $[k]$ denotes the k-th row of $\boldsymbol{M}_I$ and $\{k\}$ the k-th column of $\boldsymbol{D}\boldsymbol{M}_I{}^{\mathrm{T}}$. Hence, for the selected training instance $(\boldsymbol{X}, \boldsymbol{y})$, the partial derivative of f with respect to $\boldsymbol{D}$ can be computed by

$$\begin{aligned} \frac{\partial f}{\partial \boldsymbol{D}} &= \sum_{i \in V} \left(\frac{\partial f}{\partial \boldsymbol{a}_i} \right)^{\mathrm{T}} \frac{\partial \boldsymbol{a}_i}{\partial \boldsymbol{D}} = \sum_{i \in V} \sum_{k \in \Lambda_i} \frac{\partial f}{\partial \boldsymbol{a}_{i(k)}} \frac{\partial \boldsymbol{a}_{i(k)}}{\partial \boldsymbol{D}} \\ &= \sum_{i \in V} \sum_{k \in \Lambda_i} \boldsymbol{g}_{i(k)} \left(-\left(\boldsymbol{D}\boldsymbol{M}_I{}^{\mathrm{T}} \right)_{\{k\}} \boldsymbol{a}_i^{\mathrm{T}} + (\boldsymbol{x}_{\boldsymbol{z}} - \boldsymbol{D}\boldsymbol{a}_i)\, \boldsymbol{M}_{I[k]} \right) \\ &= \sum_{i \in V} (\boldsymbol{x}_{\boldsymbol{z}} - \boldsymbol{D}\boldsymbol{a}_i) \left(\boldsymbol{M}_I \boldsymbol{g}_i \right)^{\mathrm{T}} - \boldsymbol{D}\boldsymbol{M}_I{}^{\mathrm{T}} \boldsymbol{g}_i \boldsymbol{a}_i^{\mathrm{T}}. \end{aligned} \tag{19}$$

The supervised strategy can be derived similarly. Let $\boldsymbol{M}_O \triangleq (\boldsymbol{D}_\Lambda^T \boldsymbol{D}_\Lambda + (\mu \sum_{j \in V \setminus i} \boldsymbol{L}_{ij}) \boldsymbol{I})^{-1}$, then we obtain

$$\frac{\partial f}{\partial \boldsymbol{D}} = \sum_{i \in V} (\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}_i) \left(\boldsymbol{M}_O \boldsymbol{g}_i \right)^{\mathrm{T}} - \boldsymbol{D}\boldsymbol{M}_O{}^{\mathrm{T}} \boldsymbol{g}_i \boldsymbol{a}_i^{\mathrm{T}} \tag{20}$$

Algorithm 1: Joint Learning of Parameters and Dictionary

Input: N images $\{\boldsymbol{X}^i\}_{i=1}^N$ and their ground truth label vectors $\{\boldsymbol{y}^i\}_{i=1}^N$, initiation $\boldsymbol{W}_0$ and $\boldsymbol{D}_0$, iteration T, and initial learning rate ρ_0

Output: classifier coefficient matrix $\boldsymbol{W}$ and dictionary $\boldsymbol{D}$

1) $i = 1, t = 0$
2) **while** $i \leq T$ **do**
3) permute training data randomly
4) **for** $j = 1, \ldots, N$ **do**
5) $t++$
6) calculate sparse codes $\boldsymbol{A}$ by Eq. (8) or Eq. (9)
7) find the most violated constraint $\tilde{\boldsymbol{y}}$ by Eq. (12)
8) update $\boldsymbol{W}$ using Eq. (14):

$$\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \rho_{t-1} \frac{\partial f}{\partial \boldsymbol{W}}$$

9) update $\boldsymbol{D}$ using Eq. (19) or Eq. (20):

$$\boldsymbol{D}_t = \boldsymbol{W}_{t-1} - \rho_{t-1} \frac{\partial f}{\partial \boldsymbol{D}}$$

10) normalize $\boldsymbol{D}_t$
11) update learning rate: $\rho_t = \rho_0 / i$
12) **end for**
13) $i++$
14) **end while**

where $\boldsymbol{g}_i = \frac{\partial f}{\partial \boldsymbol{a}_i} = \boldsymbol{W}^{\mathrm{T}} (\tilde{\boldsymbol{l}}_i - \boldsymbol{l}_i)$, as in Eq. (19). The detailed learning process is summarized in Algorithm 1.

E. Superpixel Descriptor: Zoomed-Out SIFT

The essential intention of incorporating the CRF and dictionary learning is to effectively leverage low-level contexture information and high-level object class priors simultaneously. Inspired by [43], we propose a novel architecture based on the basic SIFT descriptor [44] to catch object-level priors, called the zoom-out SIFT. In this section, we explain how superpixel descriptors are extracted in our algorithm. Fig. 3 illustrates the zoomed-out regions and the proposed structure. "A" and "B" stand for two superpixels in the image, and the nested rectangles around them are the zoomed-out regions. We first extract the dense-SIFT at four levels for each image in a fine-to-coarse manner. Specifically, the grid size of the first level closely matches the superpixel area to capture the local features of each superpixel. The remaining three levels are designed to acquire proximal, distant and scene information centered at a superpixel (as indicated in [43]). As for the choice of scales in the zoomed-out SIFT, we expect the feature areas at different levels to cover about $frac132$, $frac116$ and $frac14$ of the image, respectively. For instance, the side lengths of each dense SIFT square at the three levels are set to 80, 160 and 320 for Graz-02 images. Consequently, a 128-dimensional SIFT map is generated for every level of an image, with all map resolutions being lower than that of the original image. We then upsample the feature maps to the image resolution sing bilinear interpolation. A vector is obtained by pooling the feature map over the superpixel area

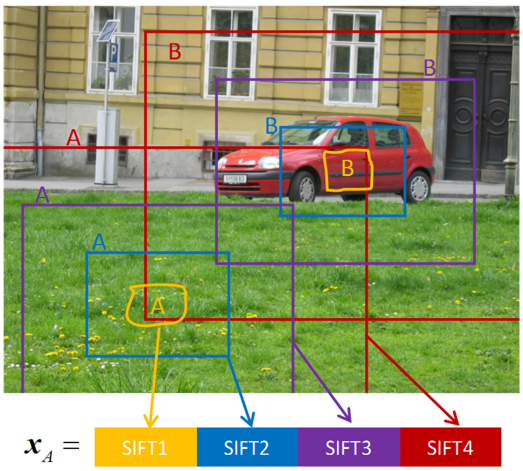


Fig. 3. An illustration of the zoomed-out SIFT.

for each level. Finally, vectors from four levels pooled over the same superpixel are concatenated to form a 512-dimensional superpixel descriptor, denoted $\boldsymbol{x}_A$, as shown in Fig. 3.

Our solution is implemented on image superpixels instead of pixels. However, the ground truth label of every superpixel is unavailable in the databases. We propagate the original ground truth from the pixel to the superpixel level as follows: a label y_i is set to l if the majority of the area of superpixel $\boldsymbol{x}_i$ belongs to the l-th category.

V. Experimental Studies

A. Parameter Settings and Metrics

Since the joint $\boldsymbol{D}$ and $\boldsymbol{w}$ learning is non-convex, a good initialization is vital to the proposed algorithm. We employ the dictionary learning algorithm of [37] to initialize the dictionary $\boldsymbol{D}$ and evaluate the sparse coding of the training data with the initial dictionary. A linear SVM is trained to obtain the initial value of $\boldsymbol{W}$. All images are empirically over-segmented into about 300 superpixels. The sparse penalty is set at $\lambda = 0.15$ and the weight penalty at $\beta = 1e-5$ in Eq. (11), based on the results of [42]. In the experiments, the initial learning rate ρ_0 is set to $1e-4$, and the iteration number T is set to 200. For the other key parameters, there are detailed studies in the following subsections, including the dictionary size, the weight ω in Eq. (7), the weight γ and the neighborhood size in Eq. (8), and the weight μ in Eq. (9).

Two typical segmentation performance metrics are applied in our experiments: accuracy and the intersection-over-union score (*i.e.*, IoU) over all classes, at the pixel level. Specifically, accuracy $= \frac{\#TP}{\#TP+\#FN}$ and IoU $= \frac{\#TP}{\#TP+\#FN+\#FP}$ where $\#TP$, $\#FN$ and $\#FP$ stand for the number of true positives, false negatives and false positives, respectively. We also calculate the global accuracy, defined as the percentage of all correctly classified pixels of all classes.

B. Experiments on Graz-02

The Graz-02 database is composed of four categories—bicycles, cars, persons and background. Each category contains 300 images of size 480×640. Following [42], we utilize the 150 odd-numbered images in each category for training and the others for testing. The database is considered to be challenging for its various object poses, different foreground scales and the presence of busy background. It is worth noticing that the ground truth merely differentiates foreground from background, which is essentially a binary classification problem. However, we consider all four categories together and formulate our scheme in a multi-class way, which is widely believed to be more difficult than the binary one.

Comparison with State-of-the-art Methods: The comparison results are shown in Table I, with the comparison metrics of pixel accuracy and IoU. "I" and "O" stand for the two proposed locality-consistent dictionary learning strategies from the input and output views, respectively. "I+O" is the result of applying both strategies simultaneously in the learning phase. Two results for each item are shown in bold. Global BoF [11], VDL [12] and SD [13] are the leading BoF-based segmentation approaches. The proposed solution achieves the best results for most terms. All three schemes of our approach outperform the compared solutions and bring obvious improvements to the global accuracy and mean IoU score in particular. CNN-CRF [20] considers the acknowledged powerful CNN feature to be the input of a second-order CRF and carries out state-of-the-art results on Graz-02. Our results are comparable to those of CNN-CRF, as there is increase in the accuracy and a decrease in the IoU score. This is understandable because the network in [20] is trained on the ImageNet database and fine-tuned on Graz-02, which means that it has many more materials for learning and can thus capture more semantic information than our model. Therefore, our solution is more plausible than are deep learning-based approaches for a scenario with less training data.

Ablation Studies: To further investigate the contribution of each component of the overall system, we conduct experiments on several degenerated versions, and the results are exhibited in the Table II. "U+P" refers to the traditional CRF model in Eq. (2) but adopting the same unary and pairwise potentials as our scheme. The baselines are obtained by the basic model described in Sec. IV-A but employing different feature to describe superpixels. Note that the basic model does not consider locality-consistency in dictionary learning. The effectiveness of the proposed zoomed-out SIFT feature is illustrated by the comparison of two baseline results. The zoomed-out SIFT feature incurs improvements of 2% and 5% in the mean accuracy and mean IoU score, respectively. We augment the basic model with locality-consistency, which improves the baseline by more than 2% in the global accuracy. For the comparison of different strategies, we can see that the locality-consistency produces a general amelioration for the bike category. Moreover, Proposed-I performs better on the car category, whereas Proposed-O achieves better results on the person category. Further, the Proposed-I+O obtains the best performance by integrating the advantages of the two strategies.

Parameter Evaluation: In Fig. 4, we show the performance variation over key parameters in the algorithm, including dictionary size, ω in Eq. (7), γ in Eq .(8), and neighborhood size

TABLE I
ACCURACY (%) AND IOU (%) ON GRAZ-02. THE MEAN WITHOUT BACKGROUND IS MARKED WITH†

Method	Accuracy						IoU				
	bg	bike	car	person	mean/mean†	global	bg	bike	car	person	mean/mean†
Global BoF [11]	86.4	73.0	68.7	71.3	74.9/71.0	–	82.3	46.2	36.5	39.0	51.0/40.6
VDL [12]	75.9	**84.9**	**76.7**	**79.8**	**79.3/80.5**	–	78.0	55.6	41.5	37.3	53.1/44.8
SD [13]	**90.6**	77.8	66.3	66.7	75.4/70.3	87.6	**86.4**	52.8	44.1	41.2	**56.1**/46.0
CNN-CRF [20]	–	84.4	**77.3**	75.2	– /79.0	–	–	**66.7**	**61.8**	**51.1**	– /**59.9**
Proposed-I+O	**89.7**	**87.0**	73.3	**79.4**	**82.4/79.9**	**90.1**	**86.0**	**58.3**	**49.9**	**48.0**	**60.6/52.1**

TABLE II
ABLATION STUDIES ON GRAZ-02. THE MEAN WITHOUT BACKGROUND IS MARKED WITH†

Method	Accuracy						IoU				
	bg	bike	car	person	mean/mean†	global	bg	bike	car	person	mean/mean†
U+P	64.6	85.0	76.4	74.0	75.0/78.5	66.4	63.6	36.7	30.1	33.4	41.0/33.4
Baseline (SIFT)	84.7	80.9	73.5	71.3	77.6/75.2	81.1	80.3	53.9	38.3	35.1	51.9/42.4
Baseline (Zoomed-out SIFT)	87.6	82.7	76.6	72.6	79.9/77.3	86.5	82.0	54.1	45.3	43.7	56.3/47.7
Proposed-I	88.4	86.3	**79.3**	75.1	81.6/79.3	88.7	83.9	55.1	47.4	45.0	57.9/49.2
Proposed-O	88.5	85.1	72.7	**80.4**	81.4/79.1	89.2	83.6	57.6	45.2	**48.7**	58.8/50.5
Proposed-I+O	**89.7**	**87.0**	73.3	79.4	**82.4/79.9**	**90.1**	**86.0**	**58.3**	**49.9**	48.0	**60.6/52.1**

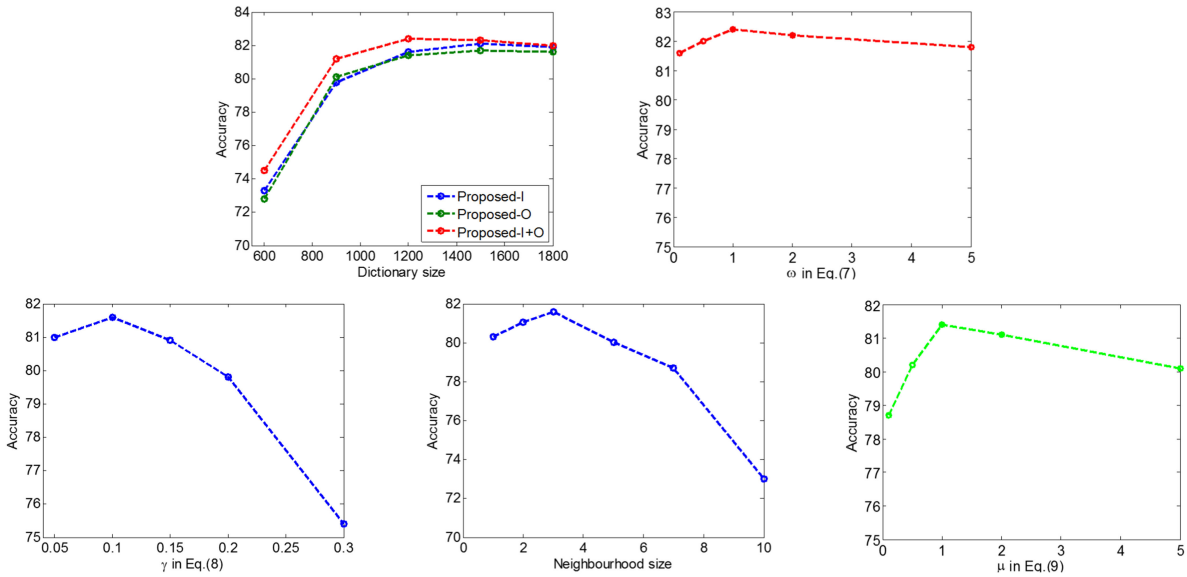


Fig. 4. Performance trend of the proposed solution against different parameters, experiments conducted on Graz-02 database.

and μ in Eq. (9). As has been concluded, Proposed-I+O has the highest mean accuracy for the advantage integration. The performance of the proposed solution remains stable when the dictionary size exceeds 1200 for the Graz-02 database. We set the dictionary size as 1200 for the other experiments on Graz-02, considering the time consumption. ω is the weight parameter when balancing the contributions of unary and pairwise potentials in CRF. To explore the effect derived by ω, we take the Proposed-I+O scheme as an instance and conduct experiments. Since ω mainly controls the "sharpness" of boundaries, it has relatively less influence on the final results. γ and neighborhood size are the parameters in the first locality-consistent dictionary learning strategy, so their curves are results of Proposed-I. Correspondingly, the last curve is obtained by experiments on the second strategy, Proposed-O. As expected, values of these three parameters that are excessively large or small will lead to performance degradation. In particular, a large γ and neighborhood size will mix too much information from other superpixels into the sparse codes and make the codes ambiguous. For parameter μ, an extremely large value relies too strongly on label supervision and thus forces the codes to neglect the characteristics of different superpixels.

Visualized segmentation results are presented in Fig. 5. The contribution of each part of the system is clearly illustrated. We construct a single dictionary for multiple categories in our approach. Our zoom-out SIFT helps recognize what categories are in the image compared with the SIFT feature. Additionally, the subsequent learning process will further advance the segmentation quality. The step-by-step refinement in Fig. 5 demonstrates that by considering locality in both location and feature/label spaces, our method is robust to varying viewpoints and complex backgrounds.

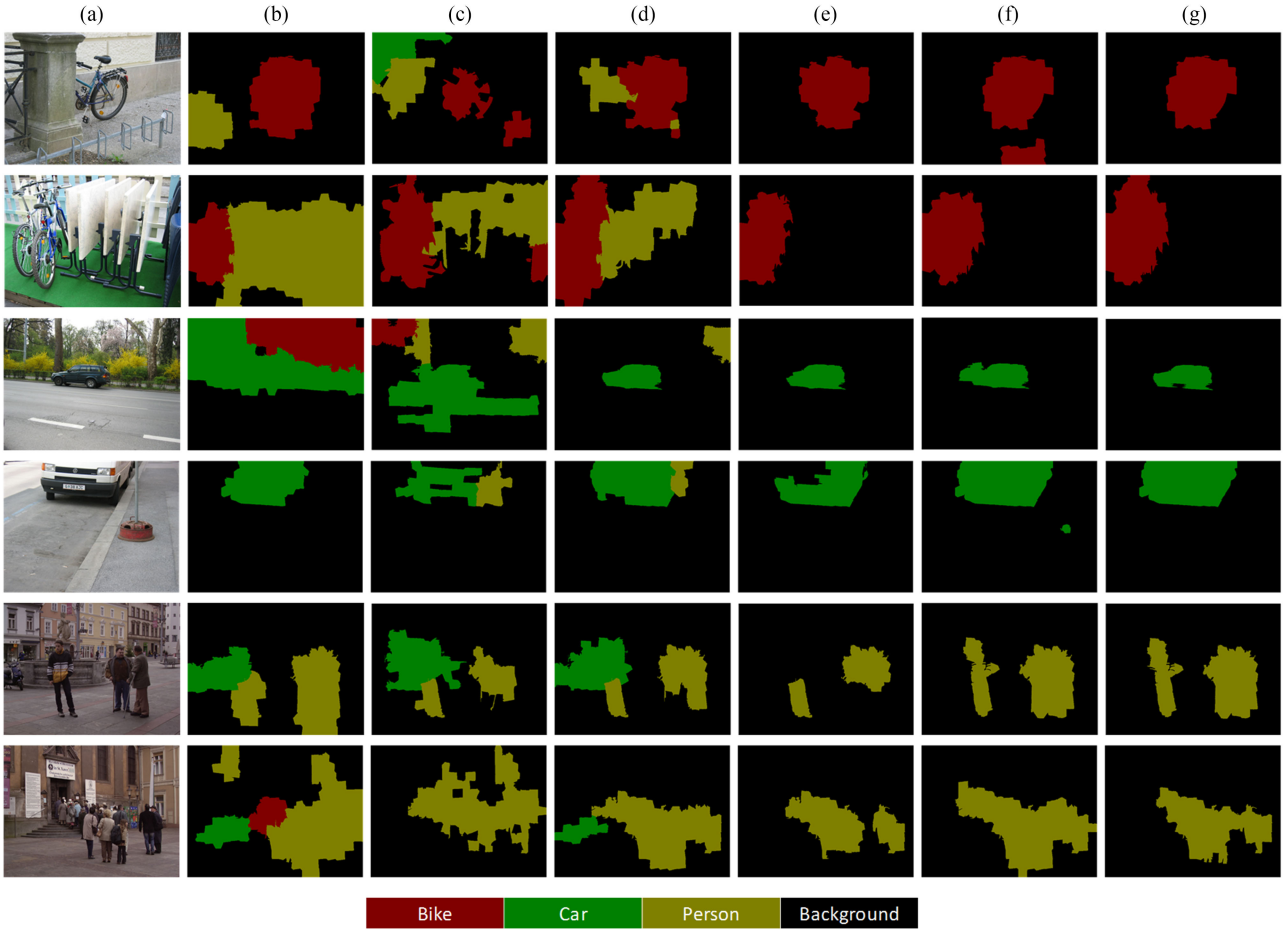


Fig. 5. Visualized samples on Graz-02. (a) Original images. (b) U+P results without sparse coding. (c) Baseline with SIFT descriptors. (d) Baseline with zoomed-out SIFT descriptors. (e) Results of our solution using the unsupervised strategy. (f) Our results using the supervised strategy. (g) Results of model learning simultaneously with both strategies.

C. Experiments on PASCAL-CONTEXT

Compared with Graz-02, the data scales of PASCAL-CONTEXT are much larger, and the involved categories are more complicated. We utilize it to evaluate the large-scale scenario. This is a newly labeled database that enriches PASCAL VOC 2010 with annotations for the whole scene. The ground truth is also on the pixel level, and the semantic label set includes objects (e.g., chair, table, book) and stuff (e.g., ground, grass, sky). There are 4,998 images in the training set and 5,105 images in the validation set. As official ground truth of the test data is unavailable, we adopt the training set for learning and the validation set for testing, as done by [23], [47], [49]. Following the protocol of [49], performance is evaluated on the most frequent 59 categories. The results for a subset of 33 easier categories are also calculated.

A comparison of the mean IoU scores is presented in Table III. Note that BoxSup is marked with * because it utilizes extra bounding box annotations from Microsoft's COCO [50] database in the training phase. Similar to Graz-02, our scheme yields the best performance when applying the two locality-consistent strategies simultaneously. SuperParsing [45] and O_2 P [46] (results reported in [49]) are state-of-the-art methods without deep networks, while CFM [47], FCN [22], BoxSup [23] and [48] are deep learning solutions. Obviously, all three of our schemes yield significant improvements over the other non-deep approaches of more than 3.5% on the easier 33 categories and more than 4.8% overall. This finding further demonstrates that preserving data locality is indispensable in semantic segmentation, even for large-scale cases. However, it is also the case that our methods perform slightly worse than the state-of-the-art deep learning approaches. In Table IV, we further compare the equipment used in experiments and per image processing time of our method to those of CFM. The per image time is computed by averaging the total training/testing time over all

TABLE III
MEAN IOU SCORES ON PASCAL-CONTEXT VALIDATION SET

Method	mean on 33	mean on 59
SuperParsing [45]	15.2	–
O_2P [46]	29.2	18.1
CFM [47]	49.5	34.4
FCN [22]	–	35.1
BoxSup* [23]	–	40.5
End-to-end [48]	–	32.5
Proposed-I	32.7	22.9
Proposed-O	33.4	23.2
Proposed-I+O	36.2	25.2

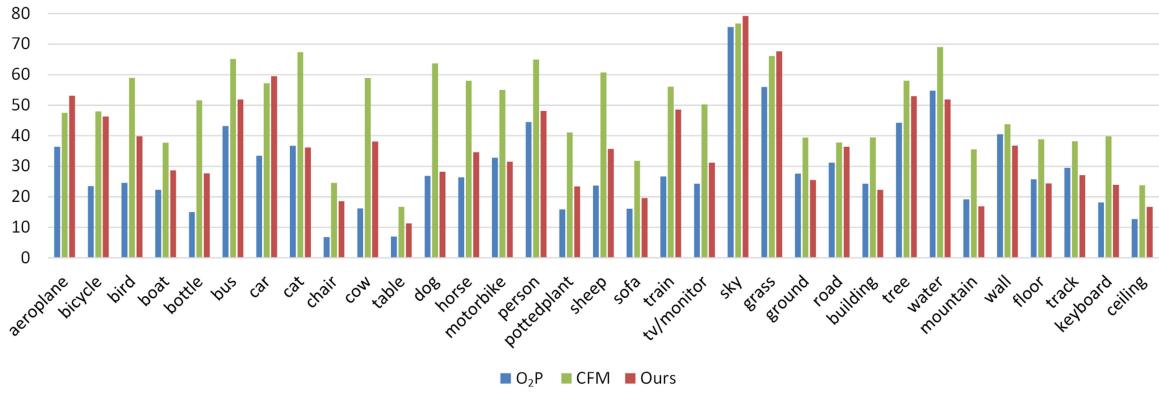


Fig. 6. Category IoU score comparison with PASCAL-CONTEXT validation set. Best viewed in color.

TABLE IV
EQUIPMENT AND PER IMAGE TIME COMPARISON ON PASCAL-CONTEXT

Method	Equipment	train	test
CFM [47]	Nvidia Titan GPU	1.32s	1.18s
Proposed-I+O	Intel i7 CPU	0.89s	0.96s

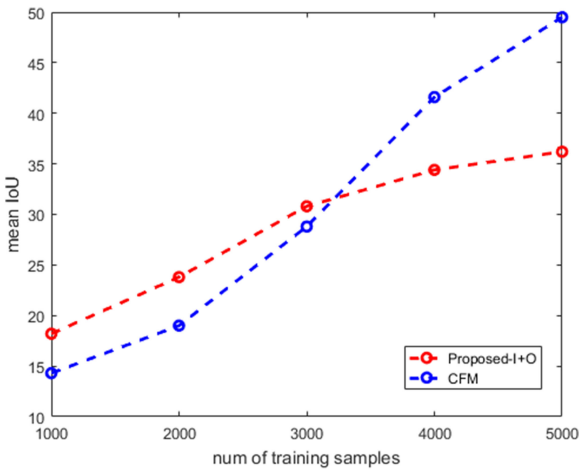


Fig. 7. Mean IoU comparison with different numbers of training samples.

involved images, excluding all the pre-processes in both algorithms. Even run on CPU, our scheme consumes less time than CFM does on GPU for per image computation in both the training and testing procedures. We summarize the reasons as follows.

In practice, the performance of deep networks usually depends significantly on kernels in each layer and link patterns between layers. It will generally import millions of parameters, so the training must be accomplished on a GPU nowadays, and the process may last for days. In contrast, our solution requires quite common hardware and is able to operate on PCs, achieving more convenience.

We further compare our method with another two leading approaches by exhibiting the per category IoU score on the new PASCAL-CONTEXT validation set in Fig. 6. Thus, O_2 P [46] and CFM [47] are selected to represent the non-deep and deep schemes, respectively, because they yield relatively closer mean IoU scores to the proposed ones. Note that the per category IoU scores are not available in End-to-end [48]. Our performance surpasses that of O_2 P for 24 of the 33 total categories. Among them, the scores of 10 categories (including "bicycle", "cow" and "train") are increased by over 10%. Our system can achieve higher scores in some categories, such as "aeroplane", "car" and "grass", compared with the deep network-based method CFM. For the other categories, most of the gaps are less than 14%, which is astonishing for a non-deep approach. For CFM, we conduct experiments with different numbers of training samples. The samples are randomly selected, and the training set is exactly the same in each branch for both methods. It shows in Fig. 7 that when there are fewer than 3000 training samples, our scheme achieves better results than CFM does. The above findings indicate that the proposed framework can outperform the state-of-the-art segmentation methods, especially when the training dataset is small.

VI. CONCLUSION

In this paper, we proposed a novel semantic segmentation scheme based on a new CRF model with locality-consistent dictionary learning. The framework explored the role of data locality in the synthesis of the CRF and dictionary learning. To this end, we proposed two locality-consistent dictionary learning strategies to be combined in the CRF model. The two strategies modeled the high-level local-consistency in feature and label space, respectively, in addition to capturing low-level context-ture in location by CRF. We also developed an algorithm to jointly learn the dictionary and parameters. The experimental results demonstrated that the proposed algorithm outperforms the leading non-deep segmentation approaches and is superior to the prevailing network-based methods for small training sets, showing that locality considered by the proposed scheme helps improve performance.

REFERENCES

[1] J. Huang, Z. Liu, and Y. Wang, "Joint scene classification and segmentation based on hidden Markov model," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 538–550, Jun. 2005.
[2] X. Liang *et al.*, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1175–1186, Jun. 2016.
[3] J. Yang, Y.-H. Tsai, and M.-H. Yang, "Exemplar cut," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 857–864.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, Jun. 2001, vol. 1, pp. 282–289.

[5] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán, "Multiscale conditional random fields for image labeling," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, USA, Jun. 2004, vol. 2, pp. II–695C–II–702.

[6] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 1–15.

[7] 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.

[8] 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.

[9] P. Kohli *et al.*, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 302–324, May 2009.

[10] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statistical Learn. Comput. Vis.*, Prague, Czech, May 2004, vol. 1, no. 1–22, pp. 1–2.

[11] D. Singaraju and R. Vidal, "Using global bag of features models in random fields for joint categorization and segmentation of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2313–2319.

[12] A. Jain, L. Zappella, P. McClure, and R. Vidal, "Visual dictionary learning for joint object categorization and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Firenze, Italy, Oct. 2012, pp. 718–731.

[13] L. Tao, F. Porikli, and R. Vidal, "Sparse dictionaries for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 549–564.

[14] K. Ohki, S. Chung, Y. H. Ch'ng, P. Kara, and R. C. Reid, "Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex," *Nature*, vol. 433, no. 7026, pp. 597–603, Feb. 2005.

[15] X. Peng, L. Zhang, Z. Yi, and K. K. Tan, "Learning locality-constrained collaborative representation for robust face recognition," *Pattern Recognit.*, vol. 47, no. 9, pp. 2794–2806, Sep. 2014.

[16] 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, Nov. 2015, pp. 509–513.

[17] L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. H. Torr, "What, where and how many? combining object detectors and CRFs," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 424–437.

[18] 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, Jun. 2012, pp. 702–709.

[19] D. Grangier, L. Bottou, and R. Collobert, "Deep convolutional networks for scene parsing," in *Proc. Int. Conf. Mach. Learn. Deep Learn. Workshop*, Montreal, QC, Canada, Jun. 2009, vol. 3, no. 6, p. 109.

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

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," arXiv preprint arXiv:1412.7062, 2014.

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

[23] 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, Dec. 2015, pp. 1635–1643.

[24] X. Liang *et al.*, "Semantic object parsing with local-global long short-term memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, Nevada, USA, Jun. 2016, pp. 3185–3193.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, Nevada, USA, Dec. 2012, pp. 1097–1105.

[26] S. Kumar and M. Hebert, "Discriminative random fields," *Int. J. Comput. Vis.*, vol. 68, no. 2, pp. 179–201, Jun. 2006.

[27] 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," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, Jan. 2009.

[28] H. Zhu *et al.*, "Multiple human identification and cosegmentation: A human-oriented CRF approach with poselets," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1516–1530, Aug. 2016.

[29] B. Fulkerson *et al.*, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep. 2009, vol. 9, pp. 670–677.

[30] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1852, Oct. 2007.

[31] Y. Wang and G. Mori, "Max-margin hidden conditional random fields for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 872–879.

[32] K. Bousmalis, S. Zafeiriou, L.-P. Morency, and M. Pantic, "Infinite hidden conditional random fields for human behavior analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 170–177, Jan. 2013.

[33] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, Dec. 1997.

[34] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1794–1801.

[35] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[36] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[37] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.

[38] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, Sep. 2014.

[39] R. Achanta *et al.*, "Slic superpixels," School Comput. Commun. Sci., École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Tech. Rep. 149300, Jun. 2010.

[40] T. Hastie, R. Tibshirani, and J. Friedman, "Linear methods for regression," *Elements Statistical Learn.*, pp. 43–99, 2009. [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-84858-7_3#citeas

[41] 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.

[42] J. Yang and M.-H. Yang, "Top-down visual saliency via joint CRF and dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2296–2303.

[43] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3376–3385.

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

[45] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 352–365.

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

[47] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3992–4000.

[48] H. Caesar, J. Uijlings, and V. Ferrari, "Region-based semantic segmentation with end-to-end training," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 381–397.

[49] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 891–898.

[50] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 740–755.

Yi Li received the B.E. degree in electronic and information engineering from the Dalian University of Technology, Dalian, China, in 2014, and the M.S. degree in information and communication engineering from the Dalian University of Technology, in 2017. She is currently working toward the Ph.D. degree with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, CASIA, Beijing, China. Her current research interests include computer vision and pattern recognition.

Yanqing Guo received the B.S. degree and Ph.D. degree 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 Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. His research interests focus on machine learning, computer vision, and multimedia security.

Jun Guo received the B.S. degree in electronic and information engineering and the M.S. degree in information and communication engineering from Dalian University of Technology, Dalian, China, in 2013 and 2016, respectively. He is currently working toward the Ph.D. degree with Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Beijing, China. His research interests include pattern recognition and machine learning. In particular, he focuses on sparse learning, matrix factorization and their applications on multimedia and data processing.

Zhuang Ma received the B.E. degree in electronic and information engineering from the Dalian University of Technology, Dalian, China, in 2017, where he is currently working toward the Postgraduate degree with the School of Information and Communication Engineering. His research interests include image sematic segmentation and image quality assessment.

Xiangwei Kong received the Ph.D. degree in management science and engineering from the Dalian University of Technology, Dalian, China, in 2003. From 2006 to 2007, she was a Visiting Researcher with the Department of Computer Science, Purdue University, West Lafayette, IN, USA. She is currently a Professor with the School of Information and Communication Engineering, Dalian University of Technology, China. Her research interests include digital image processing and recognition, multimedia information security, digital media forensics, image retrieval and mining, multisource information fusion, knowledge management and business intelligence.

Qian Liu received the B.S. and M.S. degrees from the Dalian University of Technology, Dalian, China, in 2006 and 2009, respectively, and the Ph.D. degree from The State University of New York at Buffalo, Buffalo, NY, USA (SUNY-Buffalo), in 2013. She has been working as an Associate Professor with the Department of Computer Science and Technology, Dalian University of Technology, China, since Dec. 2015. She was a Postdoctoral Fellow with the Ubiquitous Multimedia Laboratory, SUNY-Buffalo from 2013 to 2015. She was an Alexander von Humboldt Fellow at the Chair of Media Technology and the Chair of Communication Networks, Technical University of Munich from 2016 to 2017. Her current research interests include wireless multimedia communications, energy-aware multimedia delivery, haptic communications and teleoperation systems. Prof. Liu provides services to the IEEE Haptic Codec Task Group as a secretary for standardizing haptic codecs in the Tactile Internet. She also served as the Technical Program Co-Chair of 2017 IEEE Haptic Audio Visual Environments and Games HAVE'17 and HAVE'18.