Pattern Recognition 48 (2015) 2983–2992

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

CRF learning with CNN features for image segmentation

Fayao Liu [a], Guosheng Lin [a,b], Chunhua Shen [a,b,*]

[a] School of Computer Science, University of Adelaide, Australia
[b] ARC Center of Excellence for Robotic Vision, Australia

ARTICLE INFO

Article history:
Received 9 September 2014
Received in revised form
28 March 2015
Accepted 17 April 2015
Available online 24 April 2015

Keywords:
Conditional random field (CRF)
Convolutional neural network (CNN)
Structured support vector machine (SSVM)
Co-occurrence

ABSTRACT

Conditional Random Rields (CRF) have been widely applied in image segmentations. While most studies rely on hand-crafted features, we here propose to exploit a pre-trained large convolutional neural network (CNN) to generate deep features for CRF learning. The deep CNN is trained on the ImageNet dataset and transferred to image segmentations here for constructing potentials of superpixels. Then the CRF parameters are learnt using a structured support vector machine (SSVM). To fully exploit context information in inference, we construct spatially related co-occurrence pairwise potentials and incorporate them into the energy function. This prefers labelling of object pairs that frequently co-occur in a certain spatial layout and at the same time avoids implausible labellings during the inference. Extensive experiments on binary and multi-class segmentation benchmarks demonstrate the promise of the proposed method. We thus provide new baselines for the segmentation performance on the Weizmann horse, Graz-02, MSRC-21, Stanford Background and PASCAL VOC 2011 datasets.

1. Introduction

The task of image segmentation is to produce a pixel level labelling of different object categories, with a wide variety of applications ranging from image retrieval to object recognition. It is challenging as the objects may appear in various backgrounds and different visual conditions. CRFs [1] model the conditional distribution of labels given observations, representing the state-of-the-art in image/object segmentation [2–6]. In [2], Szummer et al. proposed to learn the coefficients of CRF potentials using structured support vector machines (SSVM) and graph cuts. Since then, SSVM has been widely applied for CRF learning in segmentation tasks.

In the pipeline of CRF learning based image segmentation, finding a good feature representation is of great significance, and can have a profound impact on the segmentation accuracy. Most previous studies rely on hand-crafted features, e.g., using color histograms, HOG or SIFT descriptors to construct bag-of-words features [7,4,5,8,9]. Recently, feature learning and especially deep learning methods have gained great popularity in machine learning and related fields. This type of methods typically takes raw images as an input and learn a (deep) representation of the images, and have found phenomenal success in various tasks such as speech recognition [10], image classification [11,12], object detection [13] etc. See Bengio et al. [14] for a detailed review. Deep learning methods attempt to model high-level abstractions in data at multiple layers, inspired from the cognitive processes of human brains, which generally starts from simpler concepts to more abstract ones. The learning is achieved by using deep architectures, e.g., deep belief networks (DBNs) [10], stacked autoassociator networks [15], deep convolutional neural networks (CNNs) [16,11], etc. Among them, CNNs are high-capacity machine learning models with a very large number of (typically a few million) parameters that are optimized from labelled training examples. The success of CNNs in various vision tasks [16,11] is mainly due to their ability to learn rich mid-level features that accommodate within-class variance and at the same time possess discriminative information. This is in contrast to low-level hand-crafted features.

On the other hand, prior work [17–19] has demonstrated that holistic reasoning about the occurrences of all classes helps to improve segmentation performance. These are based on the considerations that neighbouring image regions may be occupied by frequently co-occurring objects, and object pairs of mutual exclusion are less likely to appear together. For example, a cow is more likely to show up together with grass rather than a monitor, and grass is less likely to appear above sky. Therefore, we here propose to construct spatially related co-occurrence pairwise potentials to exploit the context information during inference.

In summary, we highlight the main contributions of this work as follows:

- We show that cross-domain image features learned by CNNs with labelled data from ImageNet[1] can be successfully transferred for

* Corresponding author.
E-mail address: chhshen@gmail.com (C. Shen).

[1] http://image-net.org

http://dx.doi.org/10.1016/j.patcog.2015.04.019
0031-3203/

segmentation purpose. By thoroughly evaluating the performance of the CNN features of different depths and comparing with the traditional bag-of-words and unsupervised feature learning methods, we demonstrate the power of CNN features in image segmentation.
- We illustrate that SSVM based CRF learning with CNN features yields astounding results and thus provide new baselines for segmentation performance on the Weizmann horse, Graz02, MSRC-21, Stanford Background and PASCAL VOC 2011 datasets.
- We incorporate spatially related co-occurrence pairwise potentials into the inference and gain further performance boost.

2. Related work

We briefly review some work that is relevant to ours. The first work on using convolutional networks for scene parsing is [20]. In [20], they train a deep CNN using a supervised greedy learning strategy taking pixels as an input to yield a pixel-wise labelling of an image. While somewhat preliminary, they achieved marginal improvement over CRF learning based segmentation methods. We show in this paper that deep CNN features transferred from ImageNet (ImageNet is an image dataset organized according to the WordNet hierarchy, containing millions of labelled images.) combined with SSVM based CRF learning outperforms most state-of-the-art methods. Schulz et al. [21] propose to predict the segmentation mask by adding a pairwise class location filter to the conventional CNN architecture of [16]. In the work of [22], the authors use a multiscale convolutional network trained from raw pixels to extract dense feature vectors that encode regions of multiple sizes centered on each pixel and present impressive results on several datasets. Our work differs from [22] in two aspects. First, we transfer a deep CNN trained on the ImageNet [11] dataset to segmentation while [22] trains a 3-stage convolutional network [16] on the current training data of the segmentation

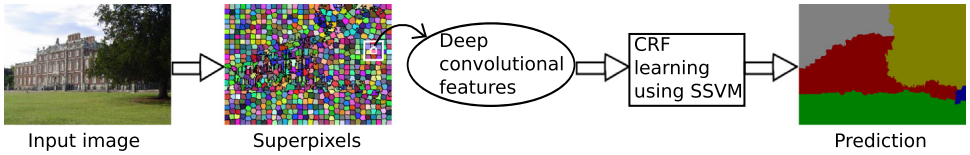

Fig. 1. An illustration of the proposed segmentation pipeline. We first over-segment the image into superpixels and then compute deep convolutional features of the patch around each superpixel centroid using a pre-trained deep CNN. The learned features are then used to learn a CRF for segmentation.

Table 1
Performance of different methods on the Weizmann horse dataset. CNN features perform significantly better than the traditional BoW feature and the unsupervised feature learning method, with features of the 6th layer performing marginally better than other compared layers. SSVM based CRF learning performs far better than SVM.

Metric	SVM					SSVM				
	BoW	UFL	L5	L6	L7	BoW	UFL	L5	L6	L7
Sa	87.5	89.3	90.1	**92.7**	91.1	92.3	94.6	95.2	**95.7**	95.1
So	58.7	63.6	68.9	**74.6**	72.9	72.5	80.1	82.4	**84.0**	82.3

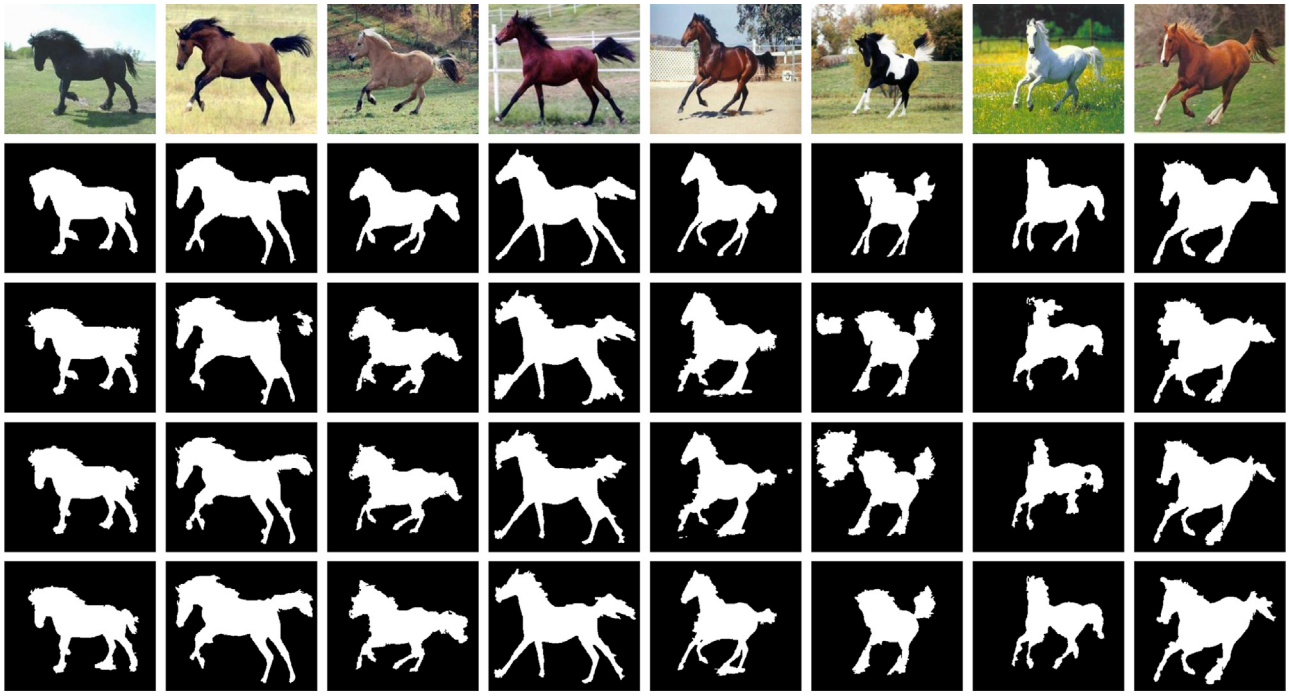

Fig. 2. Segmentation examples on Weizmann horse. 1st row: test images; 2nd row: ground truth; 3rd row: segmentation results produced by SSVM based CRF learning with bag-of-words feature; 4th row: segmentation results produced by SSVM based CRF learning with unsupervised feature learning; 5th row: segmentation results produced by SSVM based CRF learning with the 6th layer CNN features.

Table 2
Compared results of the average intersection-over-union score and average pixel accuracy on the Graz-02 dataset. We include the foreground and background results in the brackets. CNN features perform significantly better than the traditional BoW feature and the unsupervised feature learning, with features of the 6th layer performing the best among the compared layers in both SVM and SSVM. SSVM based CRF learning performs far better than SVM.

Category		Intersection/union(foreground, background) (%)			pixel Accuracy (foreground, background) (%)		
		bike	car	people	bike	car	people
SVM	BoW	66.5 (50.4, 82.7)	66.8 (42.2, 91.5)	64.0 (41.9, 86.2)	79.0 (67.9, 90.2)	75.8 (55.2, 96.3)	74.5 (55.4, 93.7)
	UFL	69.7 (55.0, 84.5)	73.1 (52.7, 93.4)	61.4 (37.2, 85.6)	81.7 (72.4, 91.1)	80.9 (64.4, 97.4)	71.2 (48.2, 94.3)
	L5	74.6 (62.4, 86.8)	76.0 (58.4, 93.7)	65.9 (47.0, 84.9)	86.3 (81.2, 91.4)	86.3 (76.2, 96.4)	80.9 (72.4, 89.4)
	L6	**77.7** (66.7, 88.6)	**78.1** (61.8, 94.5)	**68.9** (51.1, 86.6)	**88.4** (84.4, 92.5)	**87.2** (77.3, 97.0)	**83.0** (75.2, 90.8)
	L7	77.1 (66.0, 88.2)	77.6 (60.8, 94.3)	68.4 (50.5, 86.3)	88.2(84.1, 92.2)	86.6 (76.3, 97.0)	82.8 (75.1, 90.5)
SSVM	BoW	70.9 (56.6, 85.2)	75.7 (57.2, 94.1)	71.3 (53.5, 89.1)	82.5 (73.5, 91.6)	83.2 (68.9, 97.6)	81.4 (68.2, 94.7)
	UFL	74.2 (61.5, 86.9)	77.9 (60.9, 94.9)	70.9 (53.0, 88.8)	85.4 (78.6, 92.1)	83.8 (69.3, 98.4)	81.5 (68.9, 94.2)
	L5	81.6 (72.3, 90.8)	84.5 (72.6, 96.4)	75.4 (61.1, 89.7)	91.0 (88.0, 93.9)	90.6 (82.8, 98.3)	88.8 (85.3, 92.3)
	L6	**82.0** (73.1, 91.0)	**85.6** (74.5, 96.6)	**79.6** (67.2, 92.1)	**91.6** (89.5, 93.7)	**91.4** (84.4, 98.4)	**90.0** (85.1, 94.8)
	L7	81.7 (72.6, 90.8)	85.1 (73.7, 96.5)	76.0 (62.0, 90.0)	91.3 (89.0, 93.6)	91.2 (84.0, 98.4)	89.3 (86.1, 92.4)

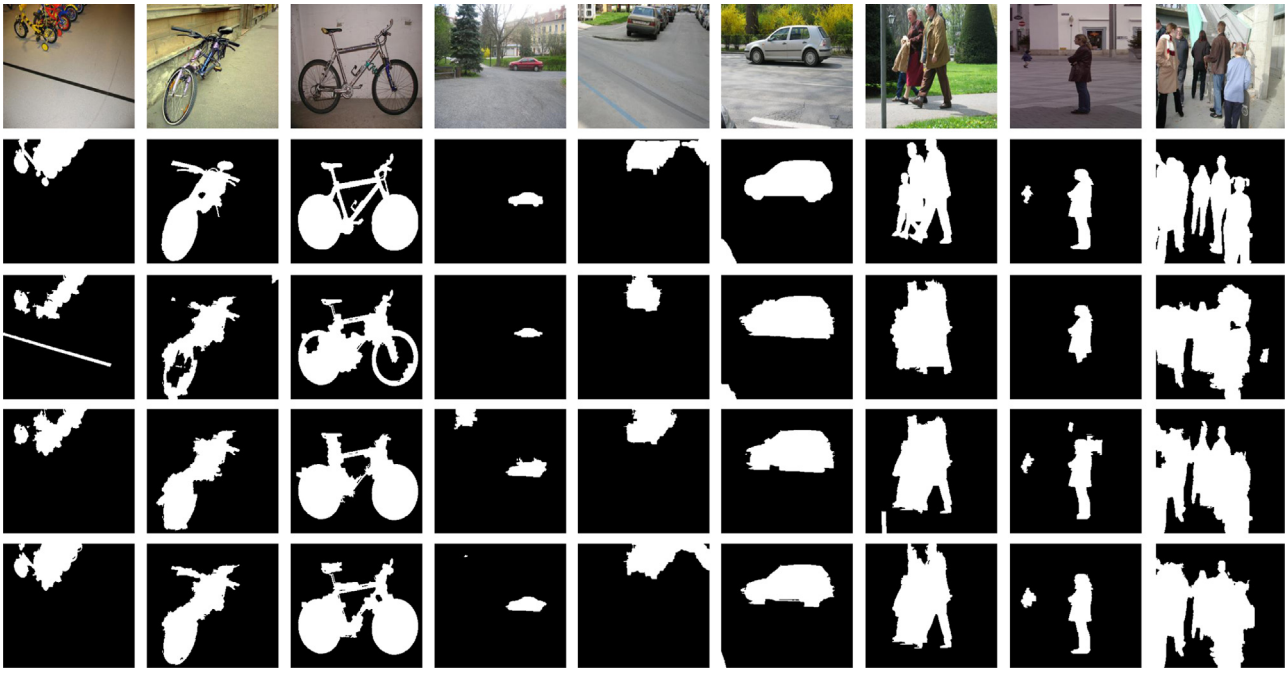

Fig. 3. Segmentation examples on the Graz-02 dataset. 1st row: test images; 2nd row: ground truth; 3rd row: segmentation results produced by SSVM based CRF learning with bag-of-words feature; 4th row: segmentation results produced by SSVM based CRF learning with unsupervised feature learning; 5th row: segmentation results produced by SSVM based CRF learning with the 6th layer CNN features.

dataset, and we demonstrate experimentally that better performance can be achieved by our method. Secondly, our method uses SSVM to learn CRF potentials while no learning is involved in [22]. Fig. 1 shows a sketch of our segmentation pipeline.

Most recently, Girshick et al. [13] demonstrate that a deep CNN trained on ImageNet can be successfully transferred to object detection and great performance boost is achieved on the PASCAL VOC 2012 dataset. As an extension of their statement, they also conduct a scene labelling experiment on the PASCAL VOC segmentation dataset to validate the power of deep CNN features on the segmentation task. Our work is mainly inspired from theirs, but differs in that we combine deep CNN features with SSVM based CRF learning in contrast to their region proposals and support the vector regression based method. Furthermore, we thoroughly evaluate the performance of deep CNN features compared to the bag-of-words features and unsupervised learned features, and provide new baselines for labelling performance on various segmentation benchmarks.

Co-occurrence statistics have been exploited and demonstrated its strength in the community. In [17], the authors incorporate semantic object context as a post-processing step by considering the co-occurrence counts of label pairs. Ladicky et al. [18] explore the inference methods for CRF with co-occurrence statistics by considering a class of global potentials. Different from their methods that ignore spatial relations of the co-occurrences, we propose to construct spatially related co-occurrence pairwise potentials, which favor labellings of object pairs that frequently co-occur in a certain spatial layout while at the same time prevents unreasonable labellings. Our method is inspired from [19] but differs in that they incorporate the mutex information by adding a mutex constraint to the inference problem while we simply construct co-occurrence pairwise potentials, and most

Table 3
Segmentation results on the MSRC-21 dataset. We report the pixel-wise accuracy for each category as well as the average per-category scores and the global pixel-wise accuracy (%). Deep learning performs significantly better than the BoW feature and the unsupervised feature learning, with SSVM based CRF learning using features of the 7th layer of the deep CNN achieving the best results. SSVM based CRF learning performs far better **than SVM.

Method	Feature	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**Average**	**Global**
SVM	BoW	61	87	60	29	47	83	56	66	60	54	66	53	68	7	61	33	51	27	35	19	29	50.1	62.7
	UFL	57	95	77	55	59	96	56	70	61	41	67	65	31	17	67	30	75	52	26	32	6	54.1	69.5
	L5	77	91	86	79	83	95	80	85	81	76	84	81	52	55	82	64	83	81	63	68	25	74.8	82.1
	L6	78	95	88	81	87	95	83	88	86	75	86	83	55	58	86	69	85	84	67	72	28	77.6	84.9
	L7	80	98	89	82	91	96	86	87	89	76	86	86	58	59	87	68	87	85	67	74	31	**79.0**	**86.0**
SSVM	BoW	65	89	87	64	74	90	58	75	78	56	85	54	55	6	60	14	66	50	35	38	8	57.4	70.7
	UFL	70	97	87	69	77	98	45	75	77	49	86	82	26	12	81	40	79	49	14	47	1	60.1	76.1
	L5	71	97	92	86	95	98	94	82	93	80	95	92	76	65	94	72	89	87	71	78	51	83.9	86.9
	L6	71	94	93	89	96	96	95	85	92	85	95	90	71	68	94	77	92	93	75	81	54	85.8	87.3
	L7	71	95	92	87	98	97	97	89	95	85	96	94	75	76	89	84	88	97	77	87	52	**86.7**	**88.5**

importantly, we explore CNN features combined with SSVM based CRF learning.

3. Method

3.1. Segmentation using CRF models

Given an image instance $\mathbf{x}$ and its corresponding labelling $\mathbf{y}$, CRF [1] models the conditional distribution of the form

$$P(\mathbf{y}|\mathbf{x};\mathbf{w}) = \frac{1}{Z}\exp(-E(\mathbf{y},\mathbf{x};\mathbf{w})). \tag{1}$$

where $\mathbf{w}$ are parameters and Z is the normalization term. The energy E of an image $\mathbf{x}$ with segmentation labels $\mathbf{y}$ over the nodes (superpixels) $\mathcal{N}$ and edges $\mathcal{S}$ takes the following form:

$$E(\mathbf{y},\mathbf{x};\mathbf{w}) = \sum_{p\in\mathcal{N}} \Phi^{(1)}(y^p,\mathbf{x};\mathbf{w}) + \sum_{(p,q)\in\mathcal{S}} \Phi^{(2)}(y^p,y^q,\mathbf{x};\mathbf{w}). \tag{2}$$

Here $\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}$; $\Phi^{(1)}$ and $\Phi^{(2)}$ are the unary and pairwise potentials, both of which depend on the observations as well as the parameter $\mathbf{w}$. CRF seeks an optimal labelling that achieves maximum a posterior (MAP), which mainly involves a two-step process [2]: (1) learning the model parameters from the training data; (2) inferring a most likely label for the test data given the learned parameters. The segmentation problem thus reduced to minimizing the energy (or cost) over $\mathbf{y}$ by the learned parameters $\mathbf{w}$, which is

$$\mathbf{y}^* = \operatorname*{argmin}_{\mathbf{y}\in\mathcal{Y}} E(\mathbf{y},\mathbf{x};\mathbf{w}). \tag{3}$$

3.2. Learning CRF in the large-margin framework

Applying the large-margin based CRF learning is to solve the following optimization:

$$\begin{aligned} \min_{\mathbf{w},\xi\geq 0}\quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{m}\sum_i \xi_i \\ \text{s.t.}:\quad & E(\mathbf{y},\mathbf{x}_i;\mathbf{w}) - E(\mathbf{y}_i,\mathbf{x}_i;\mathbf{w}) \geq \Delta(\mathbf{y}_i,\mathbf{y}) - \xi_i, \\ & \forall i = 1,\ldots,m \text{ and } \forall \mathbf{y}\in\mathcal{Y}; \end{aligned} \tag{4}$$

where $\Delta : \mathcal{Y}\times\mathcal{Y}\mapsto\mathbb{R}$ is a loss function associated with the prediction and the true label mask. In general, we have $\Delta(\mathbf{y},\mathbf{y}) = 0$ and $\Delta(\mathbf{y},\mathbf{y}') > 0$ for any $\mathbf{y}'\neq\mathbf{y}$. Intuitively, the optimization in (4) is to encourage the energy of the ground truth label $E(\mathbf{y}_i,\mathbf{x}_i;\mathbf{w})$ to be lower than any other *incorrect* labels $E(\mathbf{y},\mathbf{x}_i;\mathbf{w})$ by at least a margin $\Delta(\mathbf{y}_i,\mathbf{y})$. The SSVM solves (4) by iteratively finding the most violated constraint for each example i

$$\mathbf{y}_i^* = \operatorname*{argmin}_{\mathbf{y}\in\mathcal{Y}} E(\mathbf{y},\mathbf{x};\mathbf{w}) - \Delta(\mathbf{y}_i,\mathbf{y}). \tag{5}$$

To learn CRF in the large margin framework, we consider energy functions that are linear in the parameter $\mathbf{w}$, which indicates that the unary and the pairwise potentials in (2) can be written as

$$\Phi^{(1)}(y^p,\mathbf{x};\mathbf{w}) = \langle \mathbf{w}^{(1)}, \phi^{(1)}(y^p,\mathbf{x})\rangle, \tag{6}$$

and

$$\Phi^{(2)}(y^p,y^q,\mathbf{x};\mathbf{w}) = \langle \mathbf{w}^{(2)}, \phi^{(2)}(y^p,y^q,\mathbf{x})\rangle, \tag{7}$$

where $\phi^{(1)},\phi^{(2)}$ are the unary and pairwise feature mappings respectively and $\langle\cdot,\cdot\rangle$ denotes inner products. Clearly we have $\mathbf{w} = \mathbf{w}^{(1)}\otimes\mathbf{w}^{(2)}$ ($\otimes$ stacks two vectors). We will show how to construct the feature mappings over the learned deep features in the following.

Implementation details: After obtaining the learned deep features, we define feature mappings upon them to construct the energy function. Consider the image $\mathbf{x}$ with label $\mathbf{y}$, let $\mathbf{x}^p$ be the

Fig. 4. Segmentation examples on MSRC. 1st row: test images; 2nd row: ground truth; 3rd row: segmentation results produced by SSVM based CRF learning with bag-of-words feature; 4th row: segmentation results produced by SSVM based CRF learning with unsupervised feature learning; 5th row: segmentation results produced by our method with co-occurrence pairwise potentials.

Table 4
State-of-the-art comparison of segmentation performance (%) on the Weizmann horse (top) and Graz-02 (bottom) datasets.

Method	Sa	So
Levin & Weiss [32]	95.5	–
Cosegmentation [33]	80.1	–
Bertelli et al. [26]	94.6	80.1
Kuttel et al. [27]	94.7	–
Ours	**95.7**	**84.0**

Method	bike	car	people	average
Marszalek & Schimid [28]	61.8	53.8	44.1	53.2
Fulkerson et al. [7]	66.4	54.7	51.4	57.5
Aldaverft et al. [34]	71.9	62.9	58.6	64.5
Kuttel et al. [27]	63.2	74.8	66.4	68.1
Ours	**84.5**	**85.4**	**80.4**	**83.4**

feature vector associated with the pth superpixel, and K is the number of classes (possible labels). Then we define the unary feature mappings as

$$\phi^{(1)}(y^p, \mathbf{x}) = \left[I(y^p = 1)\mathbf{x}^{p\top}, \ldots, I(y^p = K)\mathbf{x}^{p\top} \right]^\top, \tag{8}$$

where $I(\cdot)$ is an indicator function which equals 1 if the input is true and 0 otherwise. In the case of multi-class, the dimension of $\phi^{(1)}(y^p, \mathbf{x})$ can be too large when $\mathbf{x}^p$ is high dimensional. To address this issue, we first train an one-vs.-all multi-class linear SVM over the features of superpixels, and then use the output confidence scores of the pth superpixel as $\mathbf{x}^p$ to construct the unary potential. A similar strategy is used in [4,5]. Accordingly, the pairwise feature mapping is constructed as

$$\phi^{(2)}(y^p, y^q, \mathbf{x}) = L_{pq} \cdot I(y^p \neq y^q), \tag{9}$$

where L_{pq} can be the shared boundary length or inversed color difference between neighbouring superpixels.

The energy function in (2) can then be written as

$$E(\mathbf{y}, \mathbf{x}; \mathbf{w}) = \left\langle \mathbf{w}^{(1)}, \sum_{p \in \mathcal{N}} \phi^{(1)}(y^p, \mathbf{x}) \right\rangle + \left\langle \mathbf{w}^{(2)}, \sum_{(p,q) \in \mathcal{S}} \phi^{(2)}(y^p, y^q, \mathbf{x}) \right\rangle. \tag{10}$$

To deal with the unbalanced appearance of different categories in the dataset, we define $\Delta(\mathbf{y}_i, \mathbf{y})$ as the weighted Hamming loss, which weighs errors for a given class inversely proportional to the frequency it appears in the training data, similar to [5]. We use the method of [23] to solve the inference in (5).

3.3. Inference with co-occurrence pairwise potentials

To fully exploit context information, we consider the frequency of co-occurred object pairs in different spatial layouts during the inference. On one hand, this prefers labelling of frequently co-occurred label pairs in a certain spatial relation; while on the other hand, it excludes unreasonable labellings of co-occurrences (mutex constraint, similar as [19]), such as grass, water or road appearing

Table 5
State-of-the-art comparison of global and average per-category pixel accuracy on the MSRC-21 (top) and the Stanford Background (bottom) datasets.

Method	Global (%)	Average(%)
Shotton et al. [3]	72	67
Ladicky et al. [36]	86	75
Munoz et al. [37]	78	71
Gonfaus et al. [38]	77	75
Lucchi et al. [5]	73	70
Yao et al. [8]	86.2	79.3
Lucchi et al. [9]	83.7	78.9
Ladicky et al.[18]	87	77
Roy et al. [19]	**91.5**	–
Ours	88.5	86.7
Ours (mutex)	90.3	89.2
Ours (co-occur)	91.1	**90.5**

Method	Global (%)	Average (%)
Gould et al.[29]	76.4	–
Munoz et al. [37]	76.9	66.2
Lempitsky et al. [39]	81.9	72.4
Farabet et al. [22]	81.4	76.0
Roy et al. [19]	81.1	–
Ours	82.6	76.2
Ours (mutex)	82.6	76.3
Ours (co-occur)	**83.5**	**76.9**

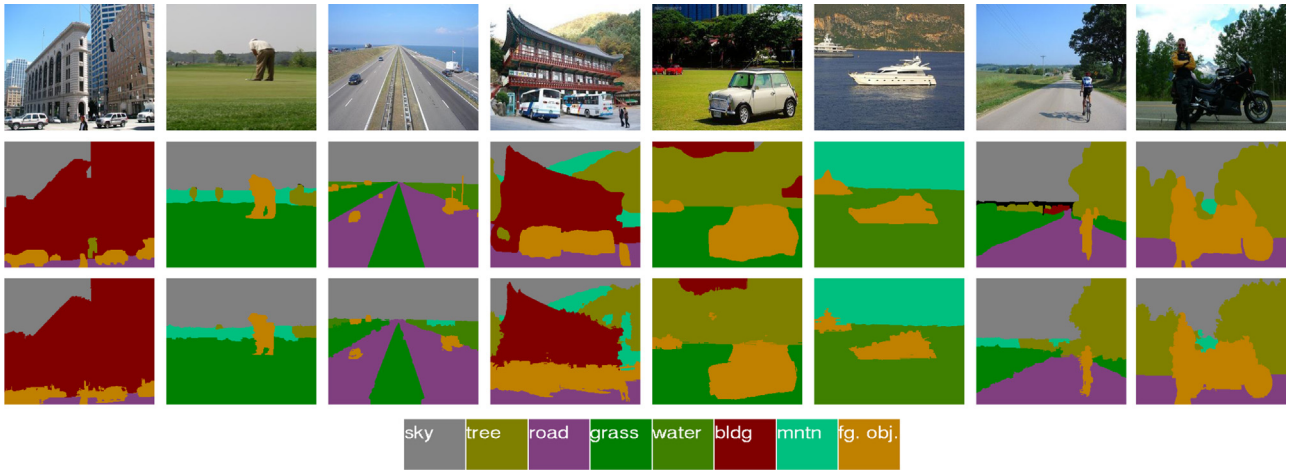

Fig. 5. Segmentation examples on the Stanford Background dataset. 1st row: test images; 2nd row: ground truth; 3rd row: segmentation results produced by our method with co-occurrence pairwise potentials.

above sky. Different from the mutex constraint used in [19], we incorporate the co-occurrence constraint into the pairwise term by devising spatially related co-occurrence pairwise potentials. We consider four spatial relations of the adjacent superpixel pairs: p is above q, p is below q, p is left to q and p is right to q. Then the feature mapping for the pairwise potential in 10 is written as

$$\sum_{(p,q)\in\mathcal{S}}\phi^{(2)}(y^p,y^q,\mathbf{x})=\sum_{(p,q)\in\mathcal{S}_1}\phi_1^{(2)}(y^p,y^q,\mathbf{x})+\sum_{(p,q)\in\mathcal{S}_2}\phi_2^{(2)}(y^p,y^q,\mathbf{x})+\sum_{(p,q)\in\mathcal{S}_3}\phi_3^{(2)}(y^p,y^q,\mathbf{x})+\sum_{(p,q)\in\mathcal{S}_4}\phi_4^{(2)}(y^p,y^q,\mathbf{x}),\tag{11}$$

where $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$ are the sets of edges where p and q are in the spatial relations "above", "below", "left" and "right" respectively, and $\mathcal{S}=\mathcal{S}_1\cup\mathcal{S}_2\cup\mathcal{S}_3\cup\mathcal{S}_4$, and $\mathcal{S}_i\cap\mathcal{S}_j=\emptyset$ for $i\neq j, i,j=1,2,3,4$.

To construct the co-occurrence pairwise potentials, we assume that the training data is sufficiently large. The pairwise potentials in (11) can then be written as

$$\phi_i^{(2)}(y^p,y^q,\mathbf{x})=L_{pq}\cdot I(y^p\neq y^q)\cdot g_i(y^p,y^q),\quad i=1,2,3,4.\tag{12}$$

where $g_i(y^p,y^q)=\frac{1}{f^i_{co-occur}(y^p,y^q)}$ with $f^i_{co-occur}(y^p,y^q)=\frac{N^i_{pq}}{N_{pq}}$. Here, N_{pq} is the number of training images in which y^p and y^q co-exist, and N^i_{pq} $(i=1,2,3,4)$ are the numbers of training images in which y^p and y^q appear in the four spatially related neighbouring superpixels respectively. If $N^i_{pq}=0$, meaning that y^p and y^q never appear in the ith spatial relation, then $g_i(y^p,y^q)=inf$, preventing the inference to yield such pair labellings. Intuitively, this would prefer labellings that frequently co-occurred in certain spatial relations in the training data, and avoid those mutual exclusion labellings, such as grass appear above sky.

Note that the mutex constraint used in [19] can be seen a special case of our co-occurrence pairwise potentials, as it is equivalent to ours when we set $g_i(y^p,y^q)=inf$ for $f^i_{co-occur}(y^p,y^q)=0$ and $g_i(y^p,y^q)=1$ for $f^i_{co-occur}(y^p,y^q)\neq 0$. We will provide experimental

Table 6
Results of per-category and mean segmentation accuracy (%) on the PASCAL VOC 2011 validation dataset. Best results are bold faced.

VOC 2011 val	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	house	mbike	person	plant	sheep	sofa	train	tv	mean
Ours	78.3	43.9	20.4	23.2	22.7	24.6	42.2	41.0	36.1	12.6	24.9	19.8	25.0	23.8	38.6	53.3	20.0	36.6	20.2	38.1	24.6	31.9
Ours (mutex)	79.8	53.1	**23.8**	**26.4**	**28.8**	**28.6**	51.6	48.2	**37.8**	13.1	29.7	22.3	**28.4**	29.6	45.2	52.7	21.0	46.2	20.9	46.2	29.6	36.3
Ours (co-occur)	**81.5**	**55.7**	23.6	24.0	27.7	27.3	**52.8**	**54.1**	37.1	**14.9**	**37.1**	**28.6**	22.9	**33.1**	**49.7**	**54.2**	**27.4**	**49.3**	**22.3**	**49.3**	**30.9**	**38.3**

Table 7
Comparison of the mean segmentation accuracy (%) on the PASCAL VOC 2011 validation dataset.

Method	Mean (%)
HOG [35]	14.1
SIFT-PCA-FISHER [35]	31.9
O_2P [35]	**38.3**
Ours (co-occur)	**38.3**

comparison with this case in Section 4.3. After learning the CRF using SSVM, we construct co-occurrence pairwise potentials for prediction. We add a trade-off parameter α multiplied to the pairwise term and tune it from 0.5 to 2 based on validation sets.

4. Experiments

To demonstrate the effectiveness of the proposed method, we first compare the CNN features with the traditional bag-of-words feature and an unsupervised feature learning method [24] as well as evaluate the impact of depths to the performance of the CNN features in Section 4.2. We then compare with state-of-the-art methods on several image segmentation datasets in Section 4.3.

4.1. Experimental setup

For the CNN features, we use the model trained on ImageNet provided by Caffe [25]. The network follows the famous AlexNet [11], and is composed of 5 convolutional layers and 2 fully connected layers together with a soft-max layer.

We evaluate the performance of the proposed method on Weizmann horse, Graz-02, MSRC-21, Standford Background and PASCAL VOC 2011 segmentation challenge dataset. The Weizmann horse dataset[2] consists of 328 horse images from various backgrounds, with groundtruth masks available for each image. We use the same data split as in [26,27], and we simply resize the images to 256×256. The Graz-02 dataset[3] contains 3 categories (bike, car and people). This dataset is considered challenging as the objects appear at various backgrounds and with different poses. We follow the evaluation protocol in [28] to use 150 for training and 150 for testing for each category.

The MSRC-21 dataset [3] is a popular multi-class segmentation benchmark with 591 images containing objects from 21 categories. We follow the standard split to divide the dataset into training/validation/test subsets. The Standford Background dataset [29] is a collection of outdoor scene images from several publicly available datasets, which consists of 715 images coming from 8 categories. Each image is approximately 320×240 pixels and contains at least one foreground object. We use the same evaluation protocol as in [29] to report 5-fold cross validation accuracy (global and per-category). The VOC 2011 dataset consists of images from 20 objects and background. We train on the training set and test on the validation images. The performance are quantified by the standard VOC measure [30].

We start with over-segmenting the images into superpixels using SLIC [31] (~ 700 superpixels per image) and then compute features within regions around each superpixel centroid with different block sizes (36×36, 48×48, 64×64, 72×72). We construct four types of pairwise features also using different block sizes to enforce spatial smoothness, which are color difference in LUV space, color histogram difference, texture difference in terms of LBP operators as well as

[2] http://www.msri.org/people/members/eranb/
[3] http://www.emt.tugraz.at/~pinz/

Fig. 6. Segmentation examples on the VOC 2011 dataset. 1st row: test images; 2nd row: ground truth; 3rd row: segmentation results produced by our method with co-occurrence pairwise potentials.

Fig. 7. Failure examples on the VOC 2011 dataset. 1st row: test images; 2nd row: ground truth; 3rd row: segmentation results produced by our method with co-occurrence pairwise potentials.

shared boundary length [4]. Training our model on the MSRC-21 dataset takes around 2 h. During prediction, the inference is rather efficient (less than 1 s per image).

4.2. Baseline comparison

To show the superiority of the deep CNN over the unsupervised feature learning, we compare with the traditional bag-of-word (BoW) feature and features learned from a popular unsupervised feature learning method [24]. Specifically, we first extract dense SIFT descriptors within each superpixel block and then quantize them into BoW feature using nearest neighbour search with a codebook size of 400. For the unsupervised feature learning, we first learn a dictionary of size 400 and patch size 6×6 based on the evaluated image dataset using Kmeans, and then use the soft threshold coding [24] to encode patches extracted from each superpixel block. The final feature vectors are obtained by performing a three-level max pooling over the superpixel block.

To investigate the roles of different layers in the proposed segmentation method, we evaluate the performance of features from the last three layers of the CNN model (5th, 6th and 7th layers). The 5th layer (with dimension 9216) is the last convolutional layer of the CNN. The 6th layer (with dimension 4096) is a fully connected layer and follows the 5th layer and the 7th (with dimension 4096) is the final layer of the feature learning pipeline. Using the two types of learned features, we compare the SSVM based CRF learning with a baseline method, namely linear SVM, which classifies each superpixel independently without CRF learning. The datasets used in this section are Weizmann horse, Graz-02 and MSRC-21. We use BoW to denote the bag-of-words feature, UFL represents the unsupervised feature learning method, and L5, L6, and L7 are CNN features of the 5th, 6th and 7th layers respectively.

Weizmann horse: We first test on the Weizmann horse dataset. The performance are quantified by the global pixel-wise accuracy S_a and the foreground intersection over union score S_o, similar as in [26]. S_a measures the percentage of pixels correctly classified while S_o directly reflects the segmentation quality of the foreground. The compared results are reported in Table 1. We can observe that the CNN features perform consistently better than the bag-of-words feature and the unsupervised learned feature in both SVM and SSVM. By enforcing smoothness term, SSVM based CRF learning obtains far better segmentations than simple binary model as SVM. Furthermore, features of different depths exhibit almost similar performance with the 6th layer performing marginally better than the other compared layers in both SVM and SSVM. In Fig. 2, we show some examples of qualitative evaluation, which yields conclusions that are in accordance with those from Table 1.

Graz-02: For a comprehensive evaluation, we use two measurements to quantify the performance of our method on the Graz-02 dataset, which are intersection over union score and the pixel accuracy (including foreground and background). We report the results in Table 2. It can be observed that feature learning methods generally outperform the traditional bag-of-words feature, with CNN features standing as the best. As for different depths, feature

of the 6th layer consistently outperforms all the other compared layers in both SVM and SSVM, which is in accordance with the conclusion of [13]. We show some segmentation examples in Fig. 3, from which we can see that SSVM based CRF learning with CNN features produces segmentation similar to ground truth.

MSRC-21: The compared results with features of different layers are summarized in Table 3. Different from the binary cases as Weizmann horse and Graz-02, features of the 7th layer perform the best, which may results from the fact that MSRC is much more difficult due to the many categories. Fig. 4 shows some qualitative results of SSVM based CRF learning with different features, from which similar conclusions can be drawn.

4.3. State-of-the-art comparison

Based on the above evaluation, we choose the best performed 6th layer for the binary (Weizmann horse and Graz-02) and 7th layer features for the multi-class datasets (MSRC-21, Stanford Background and VOC 2011) to learn CRF and compare with state-of-the-art results in this section. For the three multi-class datasets, we add the results of incorporating the mutex and co-occurrence pairwise potentials introduced in Section 3.3.

Binary datasets: Table 4 shows the compared segmentation results on the Weizmann horse and the Graz-02 datasets. We use a different evaluation metric for comparison on the Graz-02 dataset, which is the *F*-score ($F=2pr/(p+r)$, where p is the precision and r is the recall) for each class and the average over classes. In both cases, our method outperforms all the compared methods.

Multi-class datasets: The compared global and average per-category pixel accuracies on the MSRC-21 and the Stanford Background datasets are summarized in Table 5. On the MSRC dataset, our method outperforms all the methods except [19]. When incorporated with mutex or co-occurrence pairwise potentials in inference, we obtain further improvements. As expected, the co-occurrence potentials outperform the mutex potentials. Ref. [19] performs slightly better than ours in terms of global accuracy (they did not report average per-category accuracy), which may results from the fact that they use a fully connected CRF while ours are not.

As for the Stanford Background dataset, we can see that our method performs better than [22] and outperforming all the others. The work of [22] trains a 3-stage multiscale convolutional network on the training images while we directly transfer the deep CNN trained on the ImageNet to here sparing the effort of network training. Adding mutex potentials to our method does not bring any performance boost. On further investigations, we found that this is because there is only eight categories (one of which is the ambiguous foreground category) in this dataset, which leads to the fact that the only mutex information obtained is that grass, water and road cannot appear above sky. Instead, our co-occurrence potentials perform much better, leading to further performance boost. We show some segmentation examples in Fig. 5.

The segmentation results on the PASCAL VOC 2011 validation dataset are reported in Table 6. In [13], Girshick et al. achieved an average accuracy of 47.9 by using augmented training data and extra annotation set. Here we did not use any extra dataset but only the VOC training set. By introducing mutex or co-occurrence pairwise potentials, constant improvements are observed on most of the categories. As expected, our co-occurrence potential again outperforms the mutex potential. In Table 7, we compare with the recent work of Carreira et al. [35], which performed evaluations with the same settings as ours (using the train/val set). Our method achieves the same accuracy as [35]. Note that the dimension of the feature descriptors used in [35] is tens of thousands of (33,589) while ours is 4096. Qualitative examples and some failure cases are shown in Figs. 6 and 7.

5. Conclusion

We propose to learn CRF using SSVM based on features learned from a pre-trained deep convolutional neural network for image segmentation. The deep CNN is trained on ImageNet and proved to perform exceptionally well when transferred to object segmentation. We learn the CRF in the large margin framework by SSVM, and then conduct inference with co-occurrence pairwise potentials incorporated. Extensive experimental evaluations on the Weizmann horse, Graz-02, MSRC-21, Stanford Background and the PASCAL VOC 2011 dataset demonstrate the advantages of our method and provide new baselines for further research.

Conflict of interest

None declared.

References

[1] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the International Conference on Machine Learning, 2001.

[2] M. Szummer, P. Kohli, D. Hoiem, Learning CRFs using graph cuts, in: Proceedings of the European Conference on Computer Vision, 2008.

[3] J. Shotton, M. Johnson, R. Cipolla, Semantic text on forests for image categorization and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[4] B. Fulkerson, A. Vedaldi, S. Soatto, Class segmentation and object localization with superpixel neighborhoods, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.

[5] A. Lucchi, Y. Li, K. Smith, P. Fua, Structured image segmentation using kernelized features, in: Proceedings of the European Conference on Computer Vision, 2012.

[6] S. Nowozin, P. Gehler, C.H. Lampert, On parameter learning in CRF-based approaches to object class image segmentation, in: Proceedings of the European Conference on Computer Vision, 2010.

[7] B. Fulkerson, A. Vedaldi, S. Soatto, Localizing objects with smart dictionaries, in: Proceedings of the European Conference on Computer Vision, 2008.

[8] J. Yao, S. Fidler, R. Urtasun, Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[9] A. Lucchi, Y. Li, P. Fua, Learning for structured prediction using approximate subgradient descent with working sets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[10] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.

[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012.

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: a deep convolutional activation feature for generic visual recognition, in: Proceedings of the International Conference on Machine Learning, 2014.

[13] R.B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[14] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[15] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2007.

[16] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86(11) (1998) 2278–2324.

[17] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie, Objects in context, in: Proceedings of the IEEE International Conference on Computer Vision, 2007.

[18] L. Ladicky, C. Russell, P. Kohli, P.H.S. Torr, Inference methods for crfs with co-occurrence statistics, Int. J. Comput. Vis. 103 (2) (2013) 213–225.

[19] A. Roy, S. Todorovic, Scene labeling using beam search under mutex constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[20] D. Grangier, L. Bottou, R. Collobert, Deep convolutional networks for scene parsing, in: ICML Deep Learning Workshop, 2009.

[21] H. Schulz, S. Behnke, Learning object-class segmentation with convolutional neural networks, in: Proceedings of the European Symposium on Artificial Neural Networks, 2012.

[22] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.

[23] Z. Zhang, Q. Shi, Y. Zhang, C. Shen, A. van den Hengel, Constraint Reduction Using Marginal Polytope Diagrams for Map LP Relaxations, URL ⟨http://arxiv.org/abs/1312.4637⟩.
[24] A. Coates, A.Y. Ng, The importance of encoding versus training with sparse coding and vector quantization, in: Proceedings of the International Conference on Machine Learning, 2011.
[25] Y. Jia, Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding, ⟨http://caffe.berkeleyvision.org/⟩, 2013.
[26] L. Bertelli, T. Yu, D. Vu, B. Gokturk, Kernelized structural SVM learning for supervised object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[27] D. Kuettel, V. Ferrari, Figure-ground segmentation by transferring window masks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[28] M. Marszalek, C. Schmid, Accurate object localization with shape masks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[29] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.
[30] M. Everingham, L.J.V. Gool, C.K.I. Williams, J.M. Winn, A. Zisserman, The Pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[31] R. Achanta, K. Smith, A. Lucchi, P. Fua, S. Ssstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.
[32] A. Levin, Y. Weiss, Learning to combine bottom-up and top-down segmentation, in: Proceedings of the European Conference on Computer Vision, 2006.
[33] A. Joulin, F.R. Bach, J. Ponce, Discriminative clustering for image co-segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[34] D. Aldavert, A. Ramisa, R.L. de Mntaras, R. Toledo, Fast and robust object segmentation with the integral linear classifier, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[35] J. Carreira, R. Caseiro, J. Batista, C. Sminchisescu, Free-form region description with second-order pooling, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2014) 1.
[36] L. Ladicky, C. Russell, P. Kohli, P.H. S. Torr, Associative hierarchical crfs for object class image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.
[37] D. Munoz, J.A. Bagnell, M. Hebert, Stacked hierarchical labeling, in: Proceedings of the European Conference on Computer Vision, 2010.
[38] J.M. Gonfaus, X.B. Bosch, J. van de Weijer, A.D. Bagdanov, J.S. Gual, J.G. Sabaté, Harmony potentials for joint classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[39] V.S. Lempitsky, A. Vedaldi, A. Zisserman, Pylon model for semantic segmentation, in: Proceedings of the Advances in Neural Information Processing Systems, 2011.

Fayao Liu received the B.Eng. and M.Eng. degrees from the School of Computer Science, National University of Defense Technology, Hunan, China, in 2008 and 2010, respectively. She is currently pursuing the Ph.D. degree with the University of Adelaide, Adelaide, Australia. Her current research interests include machine learning and computer vision.

Guosheng Lin is a Research Fellow at School of Computer Science, The University of Adelaide. He completed his Ph.D. degree at the same university in 2014. His research interests are on computer vision and machine learning. He received a Bachelor degree and a Master degree from the South China University of Technology in computer science in 2007 and 2010 respectively.

Chunhua Shen received the Ph.D. degree from the University of Adelaide, Adelaide, Australia, in 2006. He has been a Faculty Member with the School of Computer Science, University of Adelaide, since 2011. He was with the Computer Vision Program, National ICT Australia, Canberra Research Laboratory, Canberra, Australia. His current research interests include the intersection of computer vision and statistical machine learning. His recent work has been on real-time object detection, large-scale image retrieval and classification, and scalable nonlinear optimization. Prof. Shen received the Australian Research Council Future Fellowship in 2012.