Large-Scale Weakly Supervised Object Localization via Latent Category Learning

Chong Wang, *Student Member, IEEE*, Kaiqi Huang, *Senior Member, IEEE*, Weiqiang Ren, *Member, IEEE*, Junge Zhang, *Member, IEEE*, and Steve Maybank, *Fellow, IEEE*

***Abstract*— Localizing objects in cluttered backgrounds is challenging under large-scale weakly supervised conditions. Due to the cluttered image condition, objects usually have large ambiguity with backgrounds. Besides, there is also a lack of effective algorithm for large-scale weakly supervised localization in cluttered backgrounds. However, backgrounds contain useful latent information, e.g., the sky in the aeroplane class. If this latent information can be learned, object-background ambiguity can be largely reduced and background can be suppressed effectively. In this paper, we propose the latent category learning (LCL) in large-scale cluttered conditions. LCL is an unsupervised learning method which requires only image-level class labels. First, we use the latent semantic analysis with semantic object representation to learn the latent categories, which represent objects, object parts or backgrounds. Second, to determine which category contains the target object, we propose a category selection strategy by evaluating each category's discrimination. Finally, we propose the online LCL for use in large-scale conditions. Evaluation on the challenging PASCAL Visual Object Class (VOC) 2007 and the large-scale imagenet large-scale visual recognition challenge 2013 detection data sets shows that the method can improve the annotation precision by 10% over previous methods. More importantly, we achieve the detection precision which outperforms previous results by a large margin and can be competitive to the supervised deformable part model 5.0 baseline on both data sets.**

***Index Terms*— Weakly supervised learning, object localization, latent semantic analysis, large-scale.**

I. Introduction

OBJECT localization is a fundamental problem in computer vision. Most studies adopt a fully-supervised approach, which requires manually annotating both object categories and locations. However, the annotation of object location, which is specified by a bounding box around the object of interest, is usually tedious, laborious and ambiguous, especially in the large-scale localization task such as those found in the ImageNet database. Therefore, learning to annotate object locations automatically has great practical value, which leads to the problem of weakly supervised localization. Though the idea sounds attractive, this task is challenging because objects usually appear in cluttered backgrounds, and there is also a lack of effective algorithm for large-scale weakly supervised localization. In this paper, we focus on the large-scale weakly supervised localization in cluttered images.

In recent years, many studies on weakly supervised localization have been proposed and most of them adopt a similar framework, as shown in Fig. 1(a). Firstly, region proposals are used to extract candidate detection regions, which are represented by some feature such as the histogram of gradients or bag-of-words. Then, the object regions (correct localizations) are selected from these candidate regions by some region mining strategy, *e.g.,* exhaustive search [1], [2], multiple instance learning [3]–[5], inter-intra-class modeling [3], [6]–[9] and topic model [10], [11]. These strategies have achieved promising results when objects occupy a large portion of the image [12]. However, on the highly cluttered and large-scale conditions such as the PASCAL Visual Object Class (VOC) challenge [13] and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [14], weakly supervised methods are far from being competitive with supervised methods [15], and an effective algorithm for large-scale weakly supervised localization is required [16], [17].

In cluttered conditions, objects may not be salient and usually have large ambiguity with backgrounds. Besides, in the weakly supervised task, only image-level class labels are available, *e.g.,* the image contains the class of aeroplane in Fig. 1(a). However, a large quantity of candidate detection regions have large background area. With such little supervision, discovering object regions (correct localizations) with large object-background ambiguity is very challenging, *e.g.,* the localization in Fig. 1(a) contains too much background and it is a wrong localization.

Though we only know the image-level class label, there is a vast area of background regions to be explored. *Is it possible to explore the background to reduce the object-background ambiguity?* Backgrounds contain some latent information, *e.g.,* there is also sky, grass and mountain in the image of Fig. 1(a). This latent information can be very beneficial because if we can learn these latent categories, object-background ambiguity

Manuscript received June 18, 2014; revised October 18, 2014; accepted January 15, 2015. Date of publication January 26, 2015; date of current version March 3, 2015. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316302, in part by the National Natural Science Foundation of China under Grant 61322209 and Grant 61175007, and in part by the National Key Technology Research and Development Program under Grant 2012BAH07B01. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nilanjan Ray.

C. Wang, K. Huang, W. Ren, and J. Zhang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: cwang@nlpr.ia.ac.cn; kqhuang@nlpr.ia.ac.cn; wqren@nlpr.ia.ac.cn; jgzhang@nlpr.ia.ac.cn).

S. Maybank is with the Department of Computer Science and Information Systems, Birkbeck College, University of London, London WC1E 7HU, U.K. (e-mail: sjmaybank@dcs.bbk.ac.uk).

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/TIP.2015.2396361

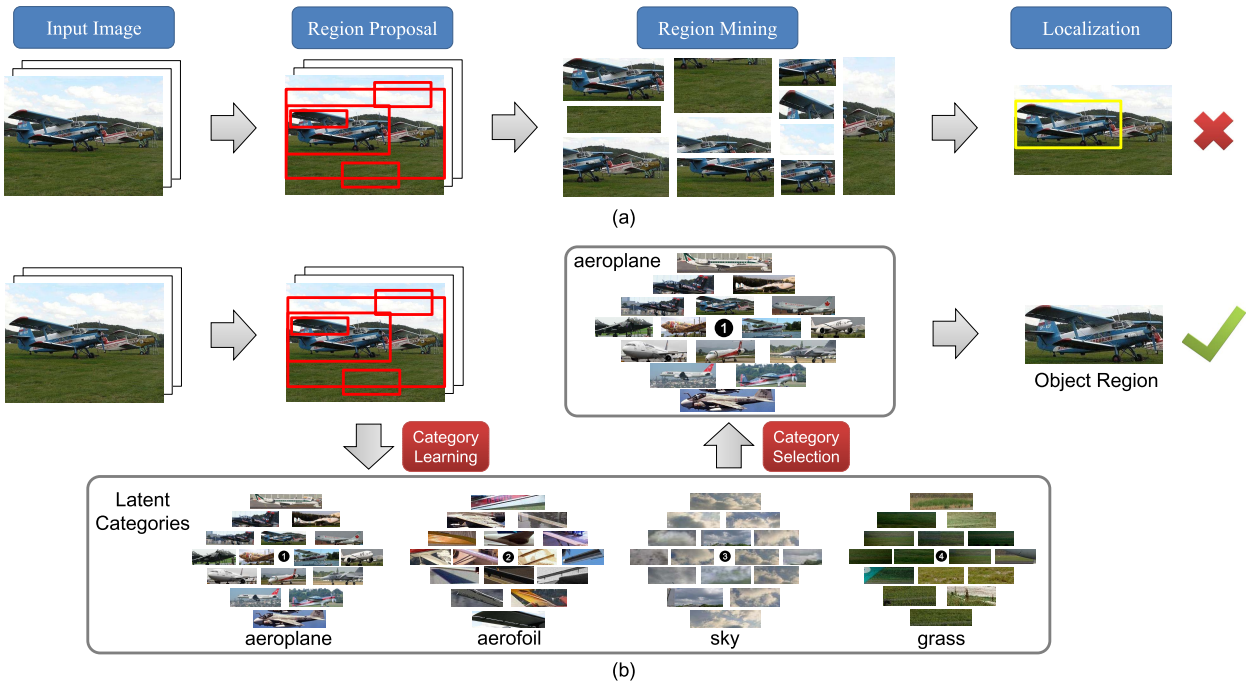

Fig. 1. The framework of the proposed latent category learning (LCL) and most previous studies for weakly supervised object localization. (a) The framework of most previous studies. (b) The framework of the proposed latent category learning (LCL).

can be largely reduced to suppress the background area in the candidate regions. For example in Fig. 1(b), the regions containing too much sky, grass and mountain can be suppressed effectively to obtain correct localizations.

Due to the unknown label of background regions, learning these latent categories is an unsupervised learning problem. In recent years, methods have been developed in finding latent categories in object-centered conditions. These categories can represent objects, object parts, backgrounds and the relations between objects and object parts [10], [16], [18]–[22]. However, in highly cluttered conditions, these methods are challenged by high object-background ambiguity. Inspired by them, we learn the latent categories in cluttered backgrounds.

In this paper, we propose the latent category learning (LCL) for large-scale weakly supervised object localization. The framework of the LCL is shown in Fig. 1(b). There are three main differences from previous studies:

1) *Category Learning: Is it possible to learn meaningful latent categories in backgrounds?* We show that the typical unsupervised semantic analysis can successfully learn the latent categories to represent objects, object parts and backgrounds, as shown in Fig. 1(b).

2) *Category Selection:* After learning these categories, *which category contains the target object class?* We propose a category selection method by evaluating the discrimination of each category and select the most discriminative one. In this paper, we denote by "class" the given image-level object class and by "category" the latent category in an object class.

3) *Online Learning:* We also propose online latent category learning (online-LCL) in this paper. We show that the category learning and category selection can be easily modified into online algorithms, which are more appropriate in practical applications and can be used in the large-scale weakly supervised localization effectively.

In the evaluation, we use the challenging PASCAL VOC 2007 database [13] and the large-scale ILSVRC 2013 detection competition. Both databases have large image variations and cluttered backgrounds. We use the complete dataset with only image-level class labels for fair comparison with the supervised method. On PASCAL VOC 2007, results show that the proposed method obtains an annotation accuracy of 48%, which is 10% higher than the previous results [3], [8], [11]. More importantly, it achieves the detection performance of 31.6%, which outperforms previous results [8], [9] by 10% and can be competitive to the supervised deformable part model 5.0 baseline 33.7% [23]. On the large-scale ILSVRC 2013 detection competition, the online LCL yields the mAP of 6.0%, which is competitive to the DPM 5.0 baseline 8.8%.

There are three contributions in this paper:

- We propose to discover the latent categories in cluttered backgrounds to reduce the object-background ambiguity. The method can effectively suppress the background area to enhance object localization.
- We achieve detection performance competitive with the supervised deformable part model 5.0. To our best knowledge, this is the first time the weakly supervised method can be competitive to the supervised approaches in cluttered conditions.
- We propose the online latent category learning to make our method applicable to large-scale localization tasks. The method can largely reduce the manual effort in annotating large-scale image datasets.

The rest of this paper is organized as follows. In Sec.II, we first review the related studies on weakly supervised

object localization. Then, we present the proposed method in Sec.III and its online algorithm in Sec.IV. To evaluate the method, we give detailed experimental results in Sec.V. Finally, Sec.VI summarizes the paper.

II. Related Work

In recent years, many studies have been proposed in weakly supervised localization, *e.g.,* the exhaustive search [1], [2], [6], [24], multiple instance learning [3]–[5], [25], [26], inter-intra-class modeling [7]–[9], [27] and topic model [10], [11], [28]–[30]. Most of them adopt a similar framework, which has three main steps: (1) *Region Extraction*: candidate regions are extracted for object detection in each image; (2) *Region Representation*: each candidate region is represented by a feature vector with semantic meaning; (3) *Region Mining*: object regions (correct localizations) are discovered among the candidate regions by region mining strategies. Though these methods achieve promising results on object-centered images, they do not work well in cluttered backgrounds because of large object-background ambiguity. More importantly, with the rapid increase in the amount of image data, there is still a lack of effective algorithms for large-scale weakly supervised object localization. In this part, we first review the main studies based on the above three steps, then we present some studies on large-scale weakly supervised object localization.

A. Region Extraction

In [1] and [2], Pandy *et al.* and Nguyen *et al.* extract dense image regions in an initial bounding box as candidate regions. However, the size and shape of these regions are fixed, which makes it difficult to take account of large object variations. As a result, not enough object regions are generated. To improve the candidate regions, many proposals for extracting more reliable detection regions based on object saliency and image segmentation have been put forward [15], [31]–[36]. Among these methods, the one popularly used in many weakly supervised localization methods [3], [7], [8] is the one proposed by Alexe *et al.* [37], who present a generic objectness measure by combining multiple image cues in a Bayesian framework. [3], [7], [8], [37]. Though high recalls have been obtained, the Maximum Average Best Overlap (MABO) [38], which measures the overlap between the candidate regions and the ground truth bounding box, is still low. A recent segmentation based region proposal, named Selective Search [38], can generate candidate regions with better quality for hierarchical segmentation and grouping strategies [38]–[40]. In addition, it yields a much higher MABO with only the comparable number of regions. In this paper, we use the selective search in region extraction.

B. Region Representation

In [1], each candidate region is represented by the histogram of oriented gradients (HOG) descriptor [15]. With additional viewpoint annotation, promising results are obtained on the subset of the PASCAL VOC 2007 challenge [13]. However, this gradient based low-level descriptor is sensitive to cluttered backgrounds and large object-background ambiguity will exist. Therefore, many studies use higher level object representation, for example the Bag-of-Words (BoW) is popularly used for its mid-level object representation [3], [7]–[9], [41], [42]. Due to the feature clustering in BoW, it can effectively remove noise, while the feature encoding and pooling can suppress the background response [41]–[47]. Furthermore, some researchers combine the multiple low-level feature representation and the mid-level BoW for better discrimination [7]. In recent years, with the great progress in theoretical achievements and parallel computing, the deep neural networks have achieved great success in many large-scale visual tasks [48]–[50]. In particular, the Convolutional Neural Network (CNN), which has achieved great success in the large-scale object recognition, can generate highly semantic object representation [39], [51]. In this paper, we use the CNN for region representation.

C. Region Mining

In exhaustive search, a detector classifier is applied to each candidate region and the one with the highest score is considered as the probable object region [1], [2]. However, the number of regions is often large, which reduces the efficiency of the exhaustive search. To improve the efficiency and discover more object regions, multiple instance learning considers inter-class relations to reduce object-background ambiguity by organizing the candidate regions as positive and negative bags [2]–[5]. Then, classification on these bags gives the object regions. To further suppress the background area and improve the quality of the object regions, researchers model intra-class relations to improve the similarity of the regions of within an object class [7]–[9], [27]. Though these methods have yield some improvements, they only consider the target objects but neglect the backgrounds, which result in large object-background ambiguity. To reduce the ambiguity, Shi *et al.* [11] propose to model objects and backgrounds in a joint Bayesian topic model, which yields considerable improvements in the annotation accuracy. Inspired by their work, we realize that backgrounds contain useful latent categories, which can represent objects, object parts or backgrounds. These latent categories can be beneficial to reduce the object-background ambiguity and suppress the background area. Given only the image-level class label, learning latent categories from backgrounds is an unsupervised learning problem. Some related studies have attempted to learn these categories from large quantities of images in object-centered conditions [10], [16], [18]–[22]. Motivated by their studies, in this paper, we propose to learn the latent categories in cluttered conditions.

D. Large-Scale Weakly Supervised Localization

In recent years, several studies have explored the weakly supervised localization in large-scale conditions [11], [16]–[18]. Shi *et al.* [11] propose the Bayesian joint topic model which can be learned with a mixture of weakly labelled and unlabelled images, allowing the large volume of unlabelled images on the Internet to be exploited for learning [11]. To learn more object categories, Chen *et al.* [16] propose the Never Ending Image

Learner (NEIL), which continuously learns and updates the object categories and locations from Internet images. They have achieved impressive results that the NEIL system can successfully learn objects, scenes, attributes and the relations between objects. However, starting the system requires a large set of seed images which are annotated with both object categories and locations. To reduce the supervision, Divvala *et al.* [17] propose to learn the visual classes in a weakly supervised way, and object detectors are learned from internet images given only object categories. However, these studies mainly focus on object-centered image conditions, in which objects usually occupy a large portion of the image. In this paper, we deal with large-scale weakly supervised localization in cluttered backgrounds.

III. Latent Category Learning

In this section, we present the offline latent category learning (LCL). We first introduce the extraction of the semantic candidate regions. Then we elaborate how to learn the latent categories, and discover object regions by category selection. Finally, we give a short summary and analyze the influence of different modeling and parameter choices. In this paper, we refer to object regions as correct localizations.

A. Region Extraction

Region proposal is an important step for generating candidate regions for object locations. It reduces the number of regions, thus making the learning more efficient. In this paper, we use a recent segmentation based region proposal named Selective Search [38], which uses multiple low-level cues, hierarchical segmentation and various grouping strategies to generate regions in which objects are likely to be found [39], [40]. Compared to other proposals [37], it is reported to have a higher Maximum Average Best Overlap (MABO) [38] and recall but only with a comparable number of regions [38]. The selective search is category independent, thus it can find the possible locations of all objects. Fig. 3(b) shows some extracted regions on the training set of the PASCAL VOC 2007 database. It is observed that although objects vary a lot in size, illumination and occlusion in cluttered backgrounds, selective search can always extract reliable regions.

After generating the candidate regions, the next step is to construct feature representation for them. In this paper, we use Convolutional Neural Network (CNN) to represent the regions. CNN has made a great breakthrough in many object recognition tasks [40], [51]. It can construct semantic object representation for its deep hierarchical structure. As demonstrated in [40], the classification results on ImageNet [14] can generalize well to the detection task in PASCAL VOC challenge. We train a CNN classification model on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012. With the same setup to [40], which uses a CNN architecture with five convolutional layers and three fully-connected layers, we represent each candidate region by the CNN output from the *fc6* layer of the classification model. The *fc6* layer is the first fully-connected layer and it contains 4096 neurons. Therefore, the feature representation of each region has the dimension of 4096.

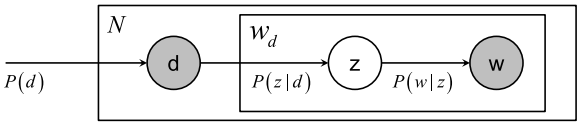

Fig. 2. An illustration of the graphical model of the probabilistic Latent Semantic Analysis (pLSA) [10].

B. Category Learning

With the candidate regions extracted, in this part, we learn the latent categories from them. Due to the unknown object class label of these regions, learning the latent category is an unsupervised learning problem. Popular methods for unsupervised learning include k-means, probabilistic Latent Semantic Analysis (pLSA) [52], [53] and Latent Dirichlet Allocation (LDA) [54], [55]. pLSA and LDA are more powerful than k-means, while pLSA is more efficient than LDA. In this paper, we use the typical pLSA for latent category learning.

We use positive images in an object class for category learning. Suppose we have N candidate regions in positive images, and the CNN representation of each region is d_j. As introduced in Sec.III-A, d_j is obtained from the *fc6* layer and has a dimension of 4096. In document analysis, the pLSA usually takes the histogram of occurrence frequency on visual words as input, while the CNN region representation satisfies this histogram input for two reasons. Firstly, due to the Rectified Linear Units in the deep network [40], all the region representation is non-negative. Secondly, we consider each neuron in the *fc6* layer as a visual word, and the CNN representation is the occurrence confidence on these words. Due to the high accuracy [37] of the extracted regions and high semantics of CNN representation, these neurons (words) can represent high-level visual patterns. More importantly, the larger confidence leads to the larger occurrence probability of a pattern. If a hard threshold function ($d_j >$T 1;else 0) is used on the CNN representation, it will turn into the 0,1 value, thus the representation is the same to the histogram of occurrence frequency on the visual patterns; while if the threshold function is not used, the CNN representation is not the strict frequency but the soft version. Therefore, this CNN region representation can fit well in the framework of topic modeling.

We denote each word (neuron) as w_i, thus the occurrence frequency of region d_j on w_i is the i-th dimension of d_j. In addition, there is a hidden topic variable z_k associated with all the visual words. We treat each topic as a latent category in an object class. The pLSA optimizes the joint probability $P\left(w_i, d_j, z_k\right)$, which has the form of the graphical model shown in Fig. 2 [10]. Marginalizing over the latent category z_k determines the conditional probability $P\left(w_i|d_j\right)$:

$$P\left(w_i|d_j\right) = \sum_{k=1}^{K} P\left(z_k|d_j\right) P\left(w_i|z_k\right), \tag{1}$$

where $P\left(z_k|d_j\right)$ is the probability of category z_k occurring in region d_j. Based on this term, each region has K probabilities

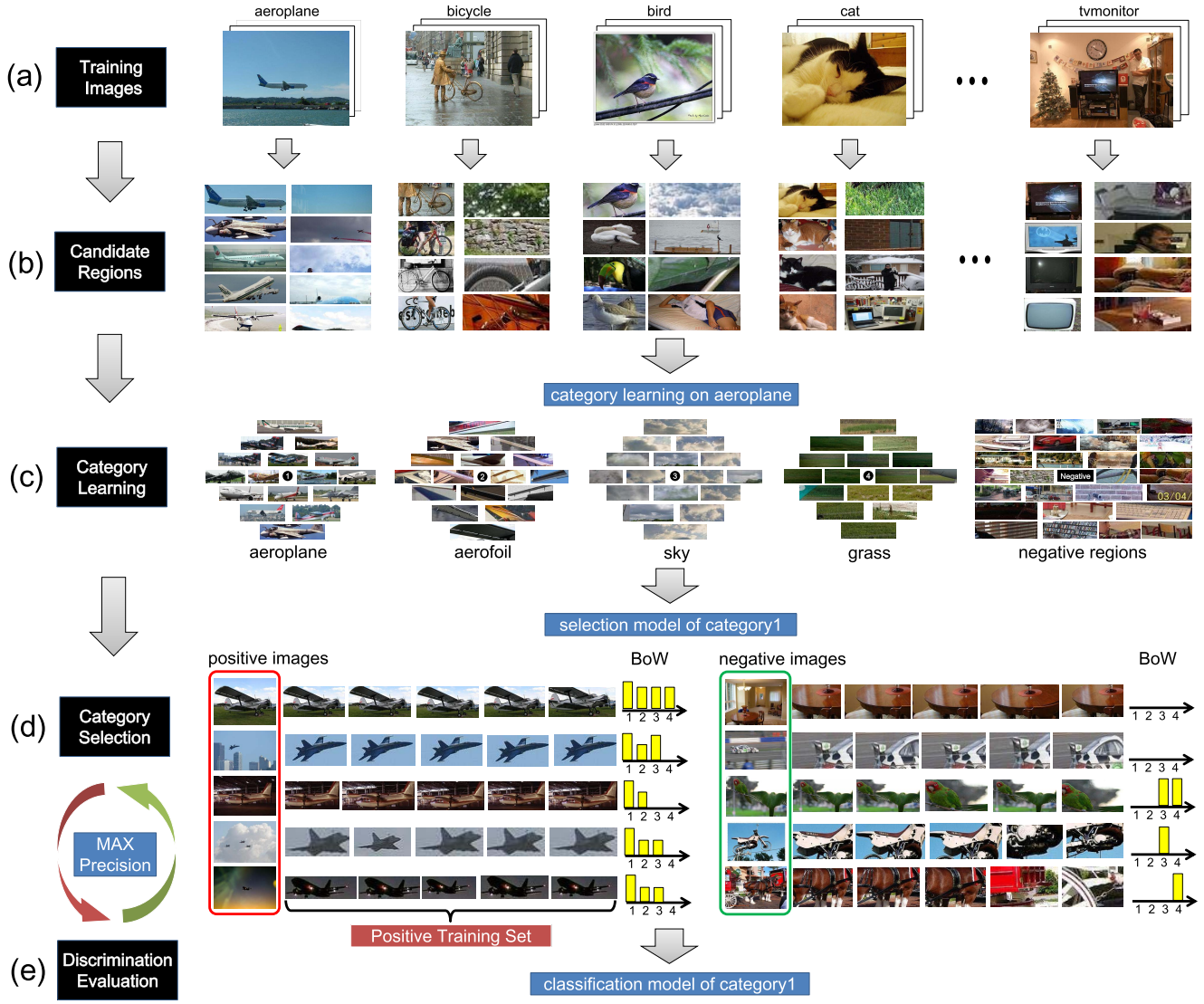

Fig. 3. The flowchart of the latent category learning (LCL) for large-scale weakly supervised object localization. (a) Original images on the PASCAL VOC 2007 training set. (b) Selective search and Convolutional Neural Network (CNN) are used to extract semantic candidate regions. (c) Probabilistic Latent Semantic Analysis (pLSA) learns latent categories. (d) A category selection model is trained for each latent category. (e) The discrimination of each category is evaluated by the classification model constructed in the manner of Bag-of-Words (BoW), and the most discriminative category is selected for training.

for K latent categories. We consider that if region d_j has the maximum probability on category z_k, then d_j only belongs to z_k. In this way, all candidate regions are divided into K sets, each of which contains the regions with a similar semantic meaning. Fig. 3(c) shows some learned latent categories of the aeroplane class. These categories have strong semantic meanings, *e.g.,* category 1 represents the aeroplane, category 2 is the aerofoil, while others contain backgrounds such as sky and grass. The latent categories in each object class are learned separately to avoid a large memory cost.

C. Category Selection

After learning the latent categories, a problem is to decide *which one contains the object regions of the target object class?* In this part, we propose a category selection strategy to discover the object regions. The idea is that the latent categories have different semantic meanings, thus they have different discrimination to the target object class. For example in Fig. 3(c), category 1 is more discriminative for describing the aeroplane than others. We exploit the different discrimination to find out the correct category. To evaluate the discrimination, it is observed that in each latent category, the regions of positive and negative images have different occurrence frequencies on all the categories. For example, in category 1, regions of positive images have a high occurrence frequency on aeroplane but much lower frequency on others, while it is the opposite for the regions of negative images, as shown in Fig. 3(d). Combined with the image-level class label, we select the category with the frequency which best differentiates the target object class and backgrounds. The detailed implementations are as follows.

Fig. 3 is used for an illustration. To construct the frequency for each category, we first have to select the regions which can represent the category. We train a selection model to select them. For any target category (category 1), we consider the regions in it as positive regions, while the negative

regions consist of two parts: the ones in other categories (category 2-4) and the ones from negative images (negative). Therefore, a selection model of the target category can be trained (category 1), and the top T scored regions in each positive and negative image are selected. Secondly, we observe that the occurrence frequencies of the T selected regions is the BoW representation on all the categories, as shown in Fig. 3(d). Based on these regions, we construct the BoW image representation for each positive and negative image. Finally, with the BoW representation, a classification model of the target latent category (category 1) is trained on the training set with the image-level class label, and the discrimination is evaluated by the classification performance on the validation set. By evaluating all categories, the one with the highest classification precision is selected, and its corresponding top T regions in positive images constitute the positive training set. Fig. 3(d) shows the selection process and the positive training set on the aeroplane class.

In constructing the BoW representation, we use three typical steps: (1) *Codebook Generation*. In our method, we quantify each latent category by averaging the regions in it. Let $\mathbf{Z} = [\mathbf{z_1}, \ldots, \mathbf{z_K}]^T \in \Re^{M \times K}$ denote the codebook with K categories. We use the average to quantify the category for two reasons: one is that the regions in a category look very similar. From the viewpoint of clustering, it is reasonable to use the center; another is that the regions in the correct category overlap heavily with the target object, thus averaging them is beneficial to suppress the background influence. (2) *Feature Encoding*. In each image, suppose the T selected regions are denoted as $[\mathbf{d_1}, \ldots, \mathbf{d_T}]^T \in \Re^{M \times T}$, we encode each region by the Super Vector Coding [46]:

$$\left[\underbrace{0, \ldots, 0}_{(j-1)*M\,dim.}, \overbrace{\mathbf{d_i} - \mathbf{z_j}}^{M \text{ dim.}}, \underbrace{0, \ldots, 0}_{(K-j)*M\,dim.}\right]. \quad (2)$$

$$s.t.\ \mathbf{z_j} = \arg\min_{\mathbf{z_k}} \|\mathbf{d_i} - \mathbf{z_k}\|_2$$

(3) *Feature Pooling*. After the encoding, average pooling [46] is used on the encoding of all the T regions to construct the BoW image representation, as shown in Fig. 3(d).

D. Algorithm Pipeline

In this part, as a short summary of the proposed idea, we give the pseudo-code of the whole algorithm in Alg.1. The pipeline summarizes the algorithm from initialization, training and testing. First, in the initialization, we set some parameters (the number of latent categories K and top selected regions T), pre-train a CNN classification model and generate CNN representation for the candidate regions in all training images. Then, in the training phase, we learn K latent categories by pLSA for each individual class, and use a BoW selection strategy to select the best category based on the classification performance on the validation set. With the best category selected, the final object detector is learned for each class. Finally, in the testing phase, we generate CNN representation for all the regions in test images, and the learned object detectors are applied on the test regions to evaluate localization performance for each object class.

Algorithm 1 The Pipeline of Latent Category Learning

Initilization:
1: Set the number of object classes C;
2: Set the number of latent categories K;
3: Set the number of the top selected region T;
4: Pre-train a CNN classification model on ILSVRC 2012;
5: Extract candidate regions by selective search for all training images;
6: Generate representation $\mathbf{d_j}$ by CNN with $fc6$ layer for all regions;
Training:
7: **for** $c = 1, ..., C$ **do**
8: Learn K latent categories with pLSA: $\mathbf{z_1}, ..., \mathbf{z_K}$;
9: **for** $k = 1, ..., K$ **do**
10: Learn a selection model for $\mathbf{z_k}$: regions in $\mathbf{z_k}$ as positive samples, the ones in other topics as negative samples;
11: Apply the selection model to select the top T regions in each positive and negative image of class c;
12: Use the T regions to generate the BoW representation for each positive and negative image of class c;
13: Learn a classification model based on the BoW representation with the classification label of class c;
14: Apply the classification model on the validation set to get classification performance;
15: **end for**
16: Select the best topic $\mathbf{z}^*$ which yields the highest classification performance on the validation set;
17: Use $\mathbf{z}^*$'s corresponding top T regions in positive images as positive samples;
18: Train the final object detector for class c;
19: **end for**
Testing:
20: Extract the candidate regions by selective search for all test images;
21: Generate representation $\mathbf{d_j}$ by CNN with $fc6$ layer for all regions;
22: Apply all the learned object detectors on all the candidate regions;
23: After processing NMS with the threshold of 0.5, preserve the regions whose score> -1;

This category selection is efficient for two reasons. (1) Both the selection and classification models are trained by the stochastic dual coordinate ascent algorithm [56], which can handle millions of samples efficently; (2) The number of the categories (K) and the top regions (T) is usually small, *e.g.,* K is around 30 and T is set to be 10, thus constructing the BoW image representation is fast. All the experiments are implemented on several computer servers with 24 cores and 128G memory, and the whole selection for an object class takes about only 1 hour.

E. Modeling and Parameter Influence

As can be seen from Alg.1, the proposed method has three important factors: category learning, category selection and their parameters (K and T). In fact, there are many choices in modeling and selecting them, *e.g.,* K-means and pLSA in category learning. These different choices will affect differently on the final quality of the model. In the following, we give some analysis to the difference. The experimental evaluation of the analysis will be given in Sec.V-E and Sec.V-F.

1) Category Learning: This step can be understood as clustering, and any clustering methods can be used, but only with the image-level class label. Here we consider two types of clustering methods, one is the typical method such as

K-means and Gaussian Mixture Model (GMM), and the other one is the topic model such as pLSA and Latent Dirichlet Allocation (LDA). In cluttered conditions, there is large object-background ambiguity. For the typical methods, this ambiguity will result in synonymous and polysemous latent categories, *e.g.,* there may be two categories for the whole aeroplane regions and the aerofoil may also appear in the boat class. As a result, the target category will not be discriminative enough, which will lead to the failure of category selection. However, the topic model can combine these ambiguous clusters to generate highly semantic topics, which are more discriminative to represent the target object class.

2) Category Selection: This step selects the category which contains clean target objects. Due to the unknown object location in the training set, the only way for selection is to use the image-level class label. The basic idea is to train a classification model for each category and test it on the validation set. Then, the category with the highest classification performance is selected as the target one. Here we compare three possible selection methods:

- *Selection Model as Classification Model:* In training the selection model, we observe that there is the case many positive samples coming from a few images. This unbalanced distribution in the positive set will hurt the performance badly. Besides, the target latent category may also contain many backgrounds regions because of the large object-background ambiguity, which will reduce the discrimination of the model.
- *Final Object Detector as Classification Model:* In testing the classification model on each validation image, we cannot guarantee the region with the highest classification score is the correct localization, *e.g.,* aerofoil may be detected by the model of aeroplane. Therefore, the ambiguity between similar latent categories such as aeroplane and aerofoil is difficult to deal with.
- *BoW Based Model as Classification Model:* Based on our observation, most of the top T regions are correct localizations. Instead of considering the top 1 in testing, BoW considers the top T to obtain an average prediction, which can reduce the ambiguity influence.

3) Hyper-Parameters: The number of latent categories K and the top region T are two important parameters. We use the best K and T by parameter selection. For K, if it is too small, the target category will contain many background regions, then the category will not be clean enough to train a good model; if K is too large, the object regions will be split into multiple categories, which will also hurt the model. For T, if it is too small, we cannot guarantee the top T regions are absolutely correct, thus several wrong localizations will hurt the model badly; while if T is too large, there will be many background regions which reduces the discrimination.

IV. Online Algorithm

In Sec.III, we have introduced the offline latent category learning. As the big data becoming important, a question arises: *"Can it be used in the large-scale condition such as the ImageNet?"* In this section, we first analyze the challenges of the offline learning in large-scale conditions, then we give an online algorithm for latent category learning.

TABLE I
THE CHANGES OF THE PIPELINE IN THE LARGE-SCALE CONDITION

	Category Learning	***Model Training***
Offline	pLSA	Liblinear
Online	Online K-means	Stochastic Gradient Descent (SGD)

Large-scale conditions have many challenges such as the large diversity of object classes, large number of object classes and large number of images. Among these factors, the large number of images is the most important factor for two main concerns: (1) In the offline category learning, we use all the regions of positive images in an object class, and we load all of them into the memory. However, on large-scale conditions, there is not enough memory to store all the regions. (2) In the offline training of the selection and classification models, due to the same reason, there is no way to train such models in an offline way. Therefore, two challenges in the offline method is the large memory cost in category learning and model training.

Based on these two limitations, we propose to replace the category learning and model training by online algorithms, which operate data in batches in limited memory. For category learning, there are many available methods such as the online k-means, online pLSA and online LDA [55]. We prefer to use the online k-means which has higher efficiency than the others. In training the selection and classification models, the Stochastic Gradient Descent (SGD) [57] is preferred for its high efficiency and comparable performance to the offline training. We summarize the difference of the pipeline for the online-LCL algorithm in Table I.

V. Experimental Evaluation

In this section, we give the experimental evaluation of the proposed method. We evaluate the method on the challenging PASCAL VOC 2007 dataset and the large-scale ILSVRC 2013 detetion competition. We first give the detailed settings, then we present the main results.

A. Experimental Settings

1) Datasets: We use two popular datasets for evaluation: PASCAL VOC 2007 dataset and ILSVRC 2013 detection competition. The PASCAL VOC 2007 dataset contains 9963 images, which are divided into three subsets: 2501 for training, 2510 for validation and 4952 for testing. There are 20 object classes, and due to cluttered backgrounds and large object variations, this dataset is very challenging. The ILSVRC 2013 detection dataset contains 464278 images, which are also divided into three subsets: 404005 for training, 20121 for validation and 40152 for testing. There are 200 object classes with large object variations and diversity, which make object detection more difficult. In both cases, we use the complete dataset with only image-level class labels for fair comparison with supervised approaches.

TABLE II
THE COMPARISON OF ANNOTATION ACCURACY BETWEEN THE PROPOSED METHOD AND PREVIOUS STUDIES ON PASCAL VOC 2007 TRAINVAL SET

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	
Joint Learning [2]	30.7	16.5	23	14.9	4.9	29.6	26.5	35.3	7.2	23.4	
MIL-SVM [60]	37.8	17.7	26.7	13.8	4.9	34.4	33.7	46.6	5.4	29.8	
Drift Detect [8]	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7	29.8	
MIL-Negative [3]	42.4	46.5	18.2	8.8	2.9	40.9	73.2	44.8	5.4	29.8	
Transfer Learning [61]	54.7	22.7	33.7	24.5	4.6	33.9	42.5	57	7.3	39.1	
Beyasian Topic [11]	67.3	54.4	34.3	17.8	1.3	46.6	60.7	68.9	2.5	32.4	
Multifold MIL [26]	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	
LCL-kmeans	74.9	61.7	49.6	13.5	17.0	57.4	73.3	44.0	27.5	70.0	
LCL-pLSA	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	

Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	Accuracy
Joint Learning [2]	20.5	32.1	24.4	33.1	17.2	12.2	20.8	28.8	40.6	7	22.4
MIL-SVM [60]	14.5	32.8	34.8	41.6	19.9	11.4	25	23.6	45.2	8.6	25.4
Drift Detect [8]	27.5	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
MIL-Negative [3]	14.5	32.8	34.8	41.6	19.9	11.4	25	23.6	45.2	8.6	30.4
Transfer Learning [61]	24.1	43.3	41.3	51.5	25.3	13.3	28	29.5	54.6	11.8	32.1
Beyasian Topic [11]	16.2	58.9	51.5	64.6	18.2	3.1	20.9	34.7	63.4	5.9	36.2
Multifold MIL [26]	10.2	29.0	58.0	64.9	36.7	18.7	56.5	13.2	54.9	59.4	38.8
LCL-kmeans	16.3	56.3	55.3	69.5	13.6	40.0	60.3	46.2	45.5	61.9	**47.7**
LCL-pLSA	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	**48.5**

2) Region Extraction: In extracting the region proposals by selective search, we use the source code released by Uijlings *et al.* [38]. The "fast" option is used for high efficiency and the minimum width of the regions is 20. About 2000 candidate regions are generated for each image. Then, to represent the regions with CNN, we train a CNN model on the ILSVRC 2012 dataset with five convolutional layers and three fully-connected layers, which is the same architecture as in [41] and [52]. We do not use any fine-tuning in the experiments. All the regions are warped to the same size of 224×224 and represented by the fc6 layer with the dimension of 4096.

3) Category Learning: For each object class, the latent categories are learned separately and all the regions from positive images in the class are used for learning. On the PASCAL VOC 2007 dataset, we use the pLSA to learn the latent categories, and the number of categories (K) is determined by the highest classification precision on the validation set based on different K. In fact, we observe that the best K is around 30 for most object classes. Therefore, on the large-scale ILSVRC 2013 detection dataset, we fix K to be 30 and use the online k-means to improve efficiency. We will give the selection of K in Sec.V-F.

4) Category Selection: One parameter in category selection is the number of the top scored regions (T), which influence the quality of the positive training set. We observe that the best performance is achieved when T is around 10, thus we set T to be 10 on both datasets. Besides, in constructing the BoW image representation, we quantify each latent category by the average representation of the regions in it, and the super-vector coding [46] and average pooling [41], [42] are used to generate the BoW representation.

5) Training and Evaluation: On PASCAL VOC 2007, we use the Liblinear SVM to train all the models, which include the selection model, the classification model and the final object detector. The Stochastic Dual Coordinate Ascent (SDCA) algorithm [56] in VLFeat [60] is used for its high efficiency in handling millions of samples. On the ILSVRC 2013 detection competition, due to the memory and efficiency problem, all these models are trained by the Stochastic Gradient Descent (SGD) algorithm [57]. In both cases, the penalty term of the classifier is determined by cross-validation. In the testing phase, we first use the trained object detector to select the regions with the score larger than -1, then the Non Maximum Suppression (NMS) [15] with the threshold of 0.5 is used to obtain the final localizations. We report the annotation precision on the trainval set and the mean average detection precision on the validation/testing set.

B. Annotation Results

Table II shows the annotation accuracy of the proposed LCL and the previous studies on the trainval set. The accuracy is measured by the percentage of training images in which an instance is correctly localized according to the PASCAL criterion, which requires the overlap of larger than 0.5 between the object region and the ground truth. We also use k-means in category learning as a baseline for comparison with pLSA. It is observed that LCL yields an annotation accuracy of 48.5%, which outperforms the previous best result by 10%. LCL improves most classes, and the improvement is quite promising on some difficult ones, *e.g.,* 18% on chair and 22% on plant. Besides, LCL-pLSA outperforms LCL-kmeans by a small margin, which shows that pLSA is slightly better in learning latent category, but it is much better than LCL-kmeans in the detection results, as shown below in Sec. V-C. Fig. 4 shows some successful and failed difficult localizations by LCL on the trainval set. Although objects vary a lot in size, occlusion and illumination in cluttered backgrounds, LCL correctly localizes most difficult samples.

Though LCL shows promising improvements, it fails on some classes such as boat and table. Based on our observation, there are two main reasons for this: (1) Too much

Fig. 4. Some successful and failed localizations on the VOC 2007 trainval set. The last column of each class shows the failed localizations.

TABLE III
THE COMPARISON OF DETECTION MAP BETWEEN THE PROPOSED METHOD AND PREVIOUS STUDIES ON PASCAL VOC 2007 TEST SET

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
Drift-Detect [8]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3
Object-Centric [9]	-	-	-	-	-	-	-	-	-	-
Multifold MIL [26]	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23.0
Latent SVM [27]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9
LCL-kmeans	41.5	29.7	24.9	12.0	10.7	30.3	40.9	31.8	10.5	21.8
LCL-pLSA	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6
DPM 5.0 [15]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1
CNN Supervise [41]	61.8	62.0	38.8	35.7	29.4	52.5	61.9	53.9	22.6	49.7

Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mAP**
Drift-Detect [8]	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
Object-Centric [9]	-	-	-	-	-	-	-	-	-	-	15.0
Multifold MIL [26]	4.9	14.1	31.9	41.9	19.3	11.1	27.6	12.1	31.0	40.6	22.4
Latent SVM [27]	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
LCL-kmeans	15.4	29.4	24.3	37.8	19.1	14.7	33.1	24.1	36.2	43.0	**26.6**
LCL-pLSA	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	**30.9**
DPM 5.0 [15]	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
CNN Supervise [41]	40.5	48.8	49.9	57.3	44.5	28.5	50.4	40.2	54.3	61.2	47.6

object variation. For example in boat, the size and appearance vary too much. Some images have small sailboats while some have large ships, which makes it difficult to learn meaningful latent categories under the limited number of positive images. (2) Similar co-occurrent classes. For example the table, it always co-exists with chairs. They look very similar in most cases, *e.g.,* both the table and chair have a flat area with several legs, which makes it difficult to learn two different latent categories. Therefore, under the cases of too much variation and similar co-occurrent classes, it is challenging for LCL to generate good localizations.

C. Detection Results

Table III shows the detection mean average precision (mAP) of the proposed LCL, the previous studies and the supervised approaches on the PASCAL VOC 2007 test set. It is observed that LCL-pLSA yields a detection mAP of 30.9%, which improves the previous best result by 8% and improves most classes by a large margin, *e.g.,* 21% on aeroplane, 13% on cow, 10% on motorbike and 15% on sofa. We also make a breakthrough on the classes which are almost zero in previous results, *e.g.,* the improvement is about 11% on chair. More importantly, compared to the supervised approach, the 30.9% obtained by LCL-pLSA can be competitive to the deformable part model 5.0 released baseline 33.7%. The precision on most classes is comparable to DPM 5.0, and some classes show better precision, *e.g.,* the improvement is about 15% on aeroplane, 12% on bird, cat and cow, and 23% on dog. This result is very encouraging because without the tedious and ambiguous annotation of object locations, the weakly supervised localization yields the comparable detection precision to the supervised methods in cluttered image conditions. Some successful and failed difficult detections on the test set are shown in Fig. 5, in which LCL correctly localizes most objects under large variations of size, occlusion and illumination.

Though LCL has achieved comparable performance to DPM 5.0, the precision on some classes is relatively low, *e.g.,* bicycle, car, horse and person. We observe that for the classes which DPM beats LCL, most of them are the classes of rigid objects, *e.g.,* bicycle, boat, bottle, chair and table.

Fig. 5. Some successful and failed localizations on the VOC 2007 test set. The last column of each class shows the failed localizations.

TABLE IV
THE DETECTION mAP OF THE PROPOSED LCL BY INCORPORATING OBJECT STRUCTURE AND INTER-CLASS RELATION

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	
Drift-Detect [8]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	
LCL-pLSA	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6	
LCL+DPM	30.2	46.9	10.4	4.6	11.1	47.0	44.9	14.7	5.6	17.4	
LCL+Context	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	

Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Drift-Detect [8]	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
LCL-pLSA	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
LCL+DPM	4.6	15.0	38.6	41.8	13.9	10.6	19.3	31.8	16.3	37.9	23.1
LCL+Context	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	**31.6**

Under this condition, object structures provide good representations because rigid objects do not change much. Combined with the HOG representation, the DPM achieves better results.

D. DPM and Context Embedding

To incorporate object structure and inter-class relations, we consider DPM and context in LCL for further enhancement. In DPM, we use the LCL annotations as ground truth, and the same setup to [15] is used, *i.e.,* 8 object parts and 3 object components. In the context, similar to the contextual operation in [15], we concatenate the region score, region location and the detection score of each class to the CNN region representation, thus the dimension of the feature vector for each candidate region is $4096 + 25 = 4121$.

Table. IV shows the detection mAP of LCL by considering DPM and context in the framework. It is observed that the LCL+DPM obtains a mAP of 23.1%, which is 9% higher than the Drift-Detect [8] which also trains DPM. However, compared to the LCL-pLSA baseline, it decreases by 7% due to the inaccurate annotations of LCL, and the precision on most classes decreases a lot. But we see some promising improvements in detecting rigid objects, *e.g.,* the improvement over LCL-pLSA is about 6% in bicycle, 5% on bus and car, and 4% in horse. Fig. 6 shows the detection model with three components on the classes of bicycle and horse. The top two components describe the side views of the objects based on the different size, and the bottom component is more like the frontal or the rear view. These results show that object structures can be beneficial to represent rigid objects.

We see that by considering inter-class relations in LCL, performance can be further improved. LCL+Context achieves the mAP of 31.6%, which outperforms the LCL-pLSA baseline by 0.7%. The improvements on some classes are promising, *e.g.,* 9% on sheep, 3% on bird and 2% on person, but this improvement is too small. The reason may be that the detection results are not accurate, *i.e.,* the locations and scores of the detections are not accurate enough to provide meaningful co-occurrence information. As a result, this will hurt the detection precision, *e.g.,* the precision decreases about $1 \sim 2\%$ on boat, bus, cow and dog.

E. Modeling Influence

In this part, we evaluate the influence of different modeling choices in category learning and selection.

In category learning, we evaluate two methods: *K-means* and *pLSA*. Table.V shows their Maximum Average Best Overlap (MABO) [40] with ground truth on the PASCAL VOC 2007 training set. It is observed that pLSA yields a 5% higher mean MABO over k-means, and the improvement on some classes is quite impressive, *e.g.,* 11% on car, 23% on table

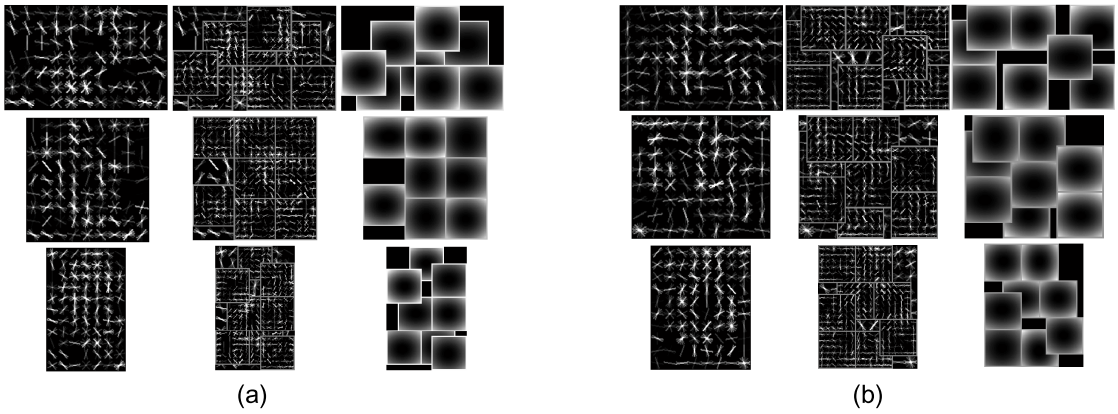

Fig. 6. The detection model trained on the LCL localizations on the PASCAL VOC 2007 trainval set. Each model is trained with three components. (a) Bicycle. (b) Horse.

TABLE V

THE MAXIMUM AVERAGE BEST OVERLAP (MABO) [40] WITH GROUND TRUTH ON THE PASCAL VOC 2007 TRAINING SET BY K-MEANS AND pLSA

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	
K-means	71.2	61.15	54.52	50.81	43.43	71.55	58.24	78.48	50.68	67.45	
pLSA	72.09	70.97	58.65	50.94	47.17	76.94	69.84	79.43	45.57	69.83	
Selective Search	85.63	80.13	78.86	76.89	67.70	83.64	79.22	86.65	78.78	80.02	
Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean MABO
K-means	55.31	76.96	64.33	71.10	46.92	35.03	48.57	50.57	71.16	60.31	59.38
pLSA	78.50	80.08	73.84	79.34	50.92	48.70	48.37	57.48	71.08	66.44	**64.80**
Selective Search	83.32	84.32	79.30	81.60	73.97	72.14	79.42	86.74	82.36	85.20	80.30

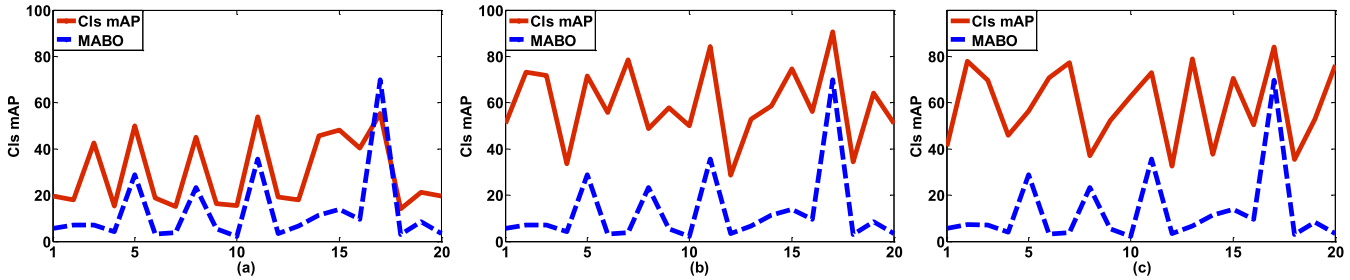

Fig. 7. The three category selection methods on the aeroplane class. (a) Selection Model; (b) Final Object Detector; (c) BoW based Selection Model. Their number of latent categories is set to be 20.

and 9% on horse. In addition, compared to the MABO given by selective search, the pLSA does not lose too much object regions. On some classes such as the dog and motorbike, it can preserve most object regions, *e.g.,* there is only a 2% difference of MABO on dog. Although there are some lost in category learning, the selection model will recall some object regions in the top selected T regions. These results imply that the latent category generated by pLSA contain more object regions than the one by K-means, which demonstrates that the topic model is more powerful in learning semantic categories.

In category selection, we evaluate three methods for the classification model: *selection model*, *final object detector* and *BoW based model*. Fig. 7(a-c) show the category selection on the aeroplane class by these three methods respectively. The number of latent categories K is set to be 20, and the Cls-mAP denotes the classification performance by the classification model on the validation set. It can be observed that for the selection model, the best Cls-mAP is much lower than the other two. Though the category selection is correct, but there is background influence and unbalanced distribution of positive samples in training the selection model. For the final object detector and BoW based model, they also give the correct selection and they have a similar and much higher Cls-mAP. However, we observe that the highest Cls-mAP does not have a large margin over the ones of other categories, *e.g.,* the $13th$ category also has a high precision. The reason is that

TABLE VI

THE PARAMETER INFLUENCE ON THE BICYCLE CLASS: THE NUMBER OF LATENT CATEGORIES K AND THE NUMBER OF TOP SELECTED REGIONS T

The number of latent categories K						The number of top selected regions T						
K	20	30	40	50	60	**T**	5	10	20	30	40	50
Cls-mAP	66.7	67.6	65.2	68.4	**69.6**	**Cls-mAP**	65.3	**69.6**	68.5	68.4	67.9	67.3
MABO	57.2	68.2	67.3	62.2	**70.0**	**Det-mAP**	39.40	**48.85**	44.24	45.52	45.26	43.32

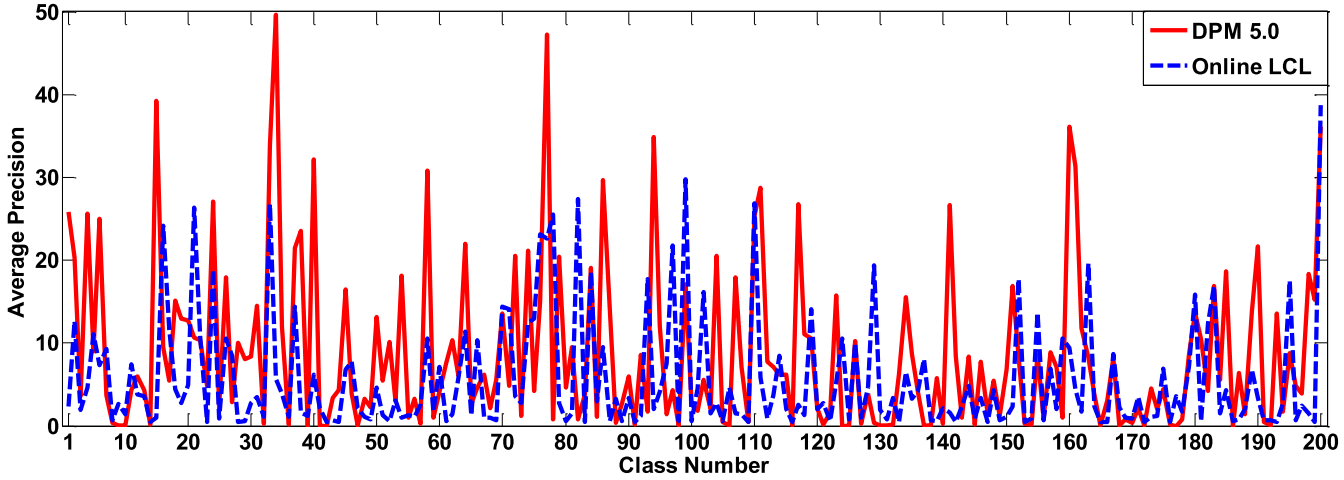

Fig. 8. The Average Precision (AP) of the deformable part model 5.0 and online LCL on the validation set of the ImageNet 2013 detection competition.

from the viewpoint of classification, the categories such as aerofoil and sky also contribute a lot to the aeroplane class. In spite of the small margin, we can guarantee that the best category always has the highest classification performance. As a result, we prefer to use the final object detector and BoW based selection model in category selection.

F. Parameter Influence

In this part, we evaluate the parameter influence in category learning and selection. In category learning, the number of latent categories K is an important factor. Table.VI shows the highest Cls-mAP and MABO of the bicycle class under different K. We initially set K to be $20 \sim 60$, then we use the above BoW based selection method to obtain the most discriminative category for each K. Finally, K with the highest Cls-mAP is used. We see the $K = 60$ is the best, which yields the improvement of 1.2% on Cls-mAP and 1.8% on MABO over the other values of K. We also test K on other classes, in which $K = 30$ yields the best performance for most classes. If K is too small, the discriminative category will contain many background regions; while if K is too large, object regions will be assigned to different latent categories which may not be discriminative to the target object class.

In category selection, the important parameter is the number of the top selected regions T. Table.VI shows the Cls-mAP on the validation set and Detection mAP (Det-mAP) on the test set of the bicycle class under different T. It is observed that $T = 10$ obtains the Cls-mAP of 69.6 and the Det-mAP of 48.85, which is the best among all the test values and has a large improvement over the other values of T. One observation is that the result of $T = 5$ is not that good. This is because if T is too small, it cannot be guaranteed that the top T regions are absolutely correct, thus several wrong localizations will hurt the model badly. However, if T is too large, there will be many background regions which reduces the discrimination, *e.g.*, the mAP of $T > 10$ begins to decrease.

G. Online-LCL Results

In this part, we validate the online LCL on the large-scale ImageNet 2013 detection competition. Based on the experiments on PASCAL VOC 2007, we observe that most classes yield the best performance when K is around 30 and T is about 10, thus we fix them to be 30 and 10. In the evaluation, we compare the DPM 5.0 baseline and online LCL on the validation set. The Average Precision (AP) of these two methods on the 200 object classes is shown in Fig. 8. The online LCL gives a mAP of 6.0% on all these classes. Given the fact the online-LCL does not select the best K for each class as did in PASCAL VOC 2007, this result can be competitive with the DPM 5.0 baseline. In addition, among the 200 object classes, the online-LCL yields a higher precision on 91 classes, which is encouraging in handling large-scale situation. Some classes improves DPM by a large margin, and 11 classes have an improvement over 10%, *e.g.,* 27% on the 82th class (hamburger), 25% on the 78th class (guacamole) and 19% on the 129th class (pizza).

However, the results of online LCL on some classes are much lower than the DPM 5.0 baseline, and there is also a

3% difference between the online-LCL and DPM 5.0. Based on our observation, there are two main reasons:

- Online k-means is not as powerful as pLSA. As discussed before, pLSA can generate more semantic category by reducing the synonymous and polysemous categories in k-means. This is shown in Table.III and Table.V, in which pLSA yields an large improvement of Det-mAP and preserves much more object regions after category learning. In addition, the online k-means, which operates data in batches, is not robust enough in clustering. Due to the limited batch size and large image variations, online-kmeans is more easily affected by image noise and the clustering centers are more easily shifted to some data which is not discriminative.
- The fixed number of latent categories is not flexible to generate semantic categories. As demonstrated in Sec.V-F, the different K will cause a large difference. Given the fixed K, we observe that in some object classes, the object regions may be assigned to multiple latent categories; while there is also the case that the category containing most object regions has many background regions. Therefore, the fixed K cannot always give discriminative latent categories, and how to determine the best K for each class is a challenging problem for using LCL in large-scale applications.

VI. Conclusion

In this paper, we have proposed the latent category learning (LCL) for weakly supervised object localization. We first use a segmentation based region proposal to generate semantic candidate regions, each of which is represented by the Convolutional Neural Network trained on ILSVRC 2012. Then, based on the large number of candidate regions, the probabilistic Latent Semantic Analysis (pLSA) is used to learn the latent categories, from which the category containing target object class is selected by evaluating each latent category's discrimination. Evaluation on the challenging PASCAL VOC 2007 dataset and the large-scale ILSVRC 2013 detection competition shows encouraging results achieved by LCL, with state-of-the-art annotation and detection performance among the weakly supervised localization methods. More importantly, the results are competitive with the supervised deformable part model 5.0 released baseline. In the future, we will design a category learning algorithm which automatically determine the number of latent categories for use in large-scale conditions.

References

[1] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1307–1314.

[2] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: A joint learning process," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1925–1932.

[3] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 594–608.

[4] S. Vijayanarasimhan and K. Grauman, "Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[5] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie, "Weakly supervised object localization with stable segmentations," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 193–207.

[6] O. Chum and A. Zisserman, "An exemplar model for learning object classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[7] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 275–293, 2012.

[8] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 343–350.

[9] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 1–15.

[10] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 370–377.

[11] Z. Shi, T. M. Hospedales, and T. Xiang, "Bayesian joint topic modelling for weakly supervised object localisation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2984–2991.

[12] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. II-264–II-271.

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

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 248–255.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[16] X. Chen, A. Shrivastava, and A. Gupta, "NEIL: Extracting visual knowledge from web data," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1409–1416.

[17] S. K. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3270–3277.

[18] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 851–858.

[19] J. Philbin, J. Sivic, and A. Zisserman, "Geometric LDA: A generative model for particular object discovery," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–8.

[20] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[21] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.

[22] H. Kang, M. Hebert, A. A. Efros, and T. Kanade, "Connecting missing links: Object discovery from sparse observations using 5 million product images," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 794–807.

[23] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. *Discriminatively Trained Deformable Part Models, Release 5*. [Online]. Available: http://people.cs.uchicago.edu/~rbg/latent-release5/, accessed 2014.

[24] Y. Zhang and T. Chen, "Weakly supervised object recognition and localization with invariant high order features," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 1–11.

[25] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2409–2416.

[26] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell, "On learning to localize objects with minimal supervision," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1611–1619.

[27] S. Wang, J. Joo, Y. Wang, and S.-C. Zhu, "Weakly supervised learning for attribute localization in outdoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3111–3118.

[28] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 431–445.

[29] C. Wang, W. Ren, and K. Huang, "Window mining by clustering mid-level representation for weakly supervised object detection," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 4067–4071.

[30] C. Wang, J. Zhang, P. Yang, and K. Huang, "Robust object recognition via visual pathway feedback," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 604–607.
[31] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 222–234, Feb. 2014.
[32] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[33] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3241–3248.
[34] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[35] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 237–244.
[36] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 606–613.
[37] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 73–80.
[38] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. N. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
[39] R. G. Cinbis, J. Verbeek, and C. Schmid, "Segmentation driven object detection with Fisher vectors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2968–2975.
[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik. (2013). "Rich feature hierarchies for accurate object detection and semantic segmentation." [Online]. Available: http://arxiv.org/abs/1311.2524
[41] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2559–2566.
[42] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 111–118.
[43] 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.*, Jun. 2009, pp. 1794–1801.
[44] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.
[45] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.
[46] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 141–154.
[47] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
[48] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Feb. 2010.
[49] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[50] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits Syst.*, May/Jun. 2010, pp. 253–256.
[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Red Hook, NY, USA: Curran Associates, 2012.
[52] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. Conf. SIGIR*, 1999, pp. 50–57.
[53] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, 2001.
[54] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 2, pp. 993–1022, Mar. 2003.
[55] M. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent Dirichlet allocation," in *Advances in Neural Information Processing Systems 23*. Red Hook, NY, USA: Curran Associates, 2010.
[56] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 567–599, 2013.
[57] Y. Lin *et al.*, "Large-scale image classification: Fast feature extraction and SVM training," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1689–1696.
[58] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems 15*. Cambridge, MA, USA: MIT Press, 2003.
[59] Z. Shi, P. Siva, and T. Xiang, "Transfer learning by ranking for weakly supervised object annotation," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 15–23.
[60] A. Vedaldi and B. Fulkerson. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. [Online]. Available: http://www.vlfeat.org/

Chong Wang received the B.Sc. degree from the Beijing University of Posts and Telecommunications, Beijing, China. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include pattern recognition, computer vision, and machine learning. He has authored several top and international conference and journal papers in ICIP, ICPR, ECCV, and the IEEE TRANSACTIONS ON IMAGE PROCESSING. In 2010 and 2011, he has participated in the famous PASCAL VOC challenge and won prizes in both years. Besides, he also won the championship of the classification task with additional data in ILSVRC 2014.

Kaiqi Huang (SM'09) received the B.Sc. and M.Sc. degrees from the Nanjing University of Science Technology, Nanjing, China, and the Ph.D. degree from Southeast University. He has worked with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he is currently a Professor. His current research interests include visual surveillance, digital image processing, pattern recognition, and biological-based vision. He has authored over 80 papers in the important international journals and conference, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Pattern Recognition*, *Computer Vision and Image Understanding*, ECCV, CVPR, ICIP, and ICPR. He was a recipient of the best student paper awards from ACPR10, the winner prizes of the detection task in both PASCAL VOC10 and PASCAL VOC11, the honorable mention prize of the classification task in PASCAL VOC11, and the winner prize of the classification task with additional data in ILSVRC 2014. He was the Deputy General Secretary of the IEEE Beijing Section (2006–2008).

Weiqiang Ren received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2009. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, image classification, object detection, and deep learning.

Junge Zhang (M'14) received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2013. In 2013, he joined the Center for Research on Intelligent Perception and Computing, Chinese Academy of Sciences, as an Assistant Professor. His major research interests include computer vision and pattern recognition. He is a Committee Member of CCF YOCSEF. In 2010 and 2011, he and his group members won the champion of PASCAL VOC challenge on object detection and ranked the second on object classification. He served as the Publicity Chair and a Technical Program Committee Member of several conferences, and the peer reviewer for over 10 international journals and conferences.

Steve Maybank received the B.A. degree in mathematics from Kings College, Cambridge, U.K., in 1976, and the Ph.D. degree in computer science from Birkbeck College, London University, London, U.K., in 1988. He joined the Pattern Recognition Group, Marconi Command and Control Systems, Frimley, U.K., in 1980, and moved to the GEC Hirst Research Center, Wembley, U.K., in 1989. From 1993 to 1995, he was a Royal Society and Engineering and Physical Sciences Research Council Industrial Fellow with the Department of Engineering Science, University of Oxford, Oxford, U.K. In 1995, he joined the Department of Computer Science, University of Reading, Reading, U.K., as a Lecturer. In 2004, he joined the School of Computer Science and Information Systems, Birkbeck College, London University, where he is currently a Professor. His research interests include the geometry of multiple images, camera calibration, visual surveillance, information geometry, and the applications of statistics to computer vision. He is a fellow of the Royal Statistical Society and the Institute of Mathematics and its Applications, and a member of the British Machine Vision Association and the Société Mathématique de France.