Weakly Supervised Learning of Deformable Part-Based Models for Object Detection via Region Proposals

Yuxing Tang, Xiaofang Wang, Emmanuel Dellandréa, and Liming Chen, *Senior Member, IEEE*

***Abstract*—The success of deformable part-based models (DPMs) for visual object detection relies on a large number of labeled bounding boxes. With only image-level annotations, our goal is to propose a model enhancing the weakly supervised DPMs by emphasizing the importance of location and size of the initial class-specific root filter. To adaptively select a discriminative set of candidate bounding boxes as this root filter estimate, first, we explore the generic objectness measurement to combine the most salient regions and "good" region proposals. Second, we propose learning of the latent class label of each candidate window as a binary classification problem, by training category-specific classifiers used to coarsely classify a candidate window into either a target object or a nontarget class. Finally, we design a flexible enlarging-and-shrinking postprocessing procedure to modify the DPMs outputs, which can effectively match the approximative object aspect ratios and further improve final accuracy. Extensive experimental results on the challenging PASCAL Visual Object Class 2007 and the Microsoft Common Objects in Context 2014 dataset demonstrate that our proposed framework is effective for initialization of the DPM's root filter. It also shows competitive final localization performance with state-of-the-art weakly supervised object detection methods, particularly for the object categories that are relatively salient in the images and deformable in structures.**

***Index Terms*—Deformable part-based models (DPMs), object detection, region proposals, weakly supervised learning.**

I. Introduction

OBJECT detection/localization in images/videos is one of the most widely studied problems in computer vision applications [1]–[3] with the explosive growth of online images/videos today. It can also be extended to numerous applications related to the multimedia community, e.g., image and video retrieval, video surveillance [4], [5], traffic safety: self or assisted driving systems, *etc*. This task remains challenging mainly due to scale and viewpoint variation, deformation, occlusion, background clutter, intra-class variations and inter-class similarities for objects in real world images/videos. For most of the existing methods, a fully supervised learning (FSL) approach is adopted [2], [3], [6]–[8], where positive training images are manually annotated with bounding boxes encompassing the objects of interest. This manual annotation of object location for large-scale image databases is extremely laborious and unreliable though quite valuable for learning accurate object detectors. However, it is usually far easier to obtain weakly labeled data, where image-level labels (e.g., user generated image tags on Internet) are presented. For example, the recently popular ImageNet ILSVRC dataset [9] contains far fewer object-level annotations (bounding boxes) than image-level labels. As a result, in this paper, in contrast to the traditional FSL, we are concerned with weakly supervised learning (WSL) for object detection, where the exact object locations in positive training examples are not provided, giving only the binary labels indicating the presence or absence of the objects of interest.

Manuscript received February 16, 2016; revised June 26, 2016 and September 6, 2016; accepted September 21, 2016. Date of publication October 3, 2016; date of current version January 17, 2017. This work was supported by the French Research Agency, Agence Nationale de Recherche (ANR) in part through the VideoSense Project under Grant 2009 CORD 026 02 and in part through the Visen project under Grant ANR-12-CHRI-0002-04 within the framework of the ERA-Net CHIST-ERA. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

The authors are with LIRIS, CNRS UMR 5205, École Centrale de Lyon, Écully F-69134, France (e-mail: yuxing.tang@ec-lyon.fr; xiaofang.wang@ec-lyon.fr; emmanuel.dellandrea@ec-lyon.fr; liming.chen@ec-lyon.fr).

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

Digital Object Identifier 10.1109/TMM.2016.2614862

A. Related Work

In recent years, there has been a substantial amount of work on weakly supervised object detection. Based on weakly annotated examples, the common practice is to jointly learn an appearance model together with the latent object location. The majority of related work treats WSL for object detection as a multiple instance learning (MIL) [10] problem. In the MIL framework, there are some positive and some negative bags. A bag is positive when it has at least one positive instance, while it is negative if all the instances are negative. The objective of MIL is to train a classifier which can correctly classify a test instance as either positive or negative. MIL problems are usually solved by finding a local minimum of non-convex objective functions (e.g., MI-SVM [11]). Galleguillos *et al.* [12] first use the MIL model to recognize and localize objects based on multiple stable segmentations. [13] and [14] use variants of MIL to learn object detectors from weakly labeled images and videos. Cinbis *et al.* [15] use a multi-fold training procedure for MIL to avoid rapid convergence to poor local optima. Also, to get rid of bad local minima, Song *et al.* [16] initialize the object locations via a discriminative submodular covering method.

Another main strategy for WSL detection is to utilize a category-independent saliency measure to predict whether a given image region belongs to an object or not. For example, Deselaers *et al.* [17] propose a fully connected conditional random field (CRF) [18] which aims at selecting a candidate window

with the highest objectness score [19] in each positive training image.

Some works identify the WSL problem as a transfer learning (TL) problem. For example, Shi *et al.* [20] formulate a ranking based transfer learning method, which effectively transfers a model for predicting object location from an auxiliary dataset to a target dataset with completely unrelated object categories. Hoffman *et al.* [21] propose an algorithm which can learn the difference between the image classifier and the object detector, and which can transfer this knowledge to classifiers for categories without bounding box annotated data, turning them into detectors. However, for both of these methods, auxiliary object level annotations for part of the dataset are required.

In addition, Pandey and Lazebnik [22] modify the fully supervised DPMs in a weakly supervised manner without object level annotations, which learns structural object detectors based on randomly initialized windows in the positive training images. Shi *et al.* [23] propose a WSL framework based on Bayesian joint topic modeling which localizes objects across different classes concurrently. Recently, Wang *et al.* [24] propose to learn the latent categories using probabilistic latent semantic analysis (pLSA), and to select the target object category by evaluating each latent category's discrimination. Bilen [25] *et al.* propose to couple a smooth discriminative learning procedure with a convex clustering algorithm, by imposing the similarity among objects of the same class.

Tang *et al.* [26] focus on the problem of unsupervised object detection through co-localization, which further alleviates the need for annotations, requiring only a set of images each containing some common object to be localized. In object co-localization, we do not know which objects are contained in the image set, and no negative images or images known not to contain the object are provided. Co-localization outputs bounding boxes as weakly supervised localizations without strong supervision. [26] proposes a joint optimization of the prior, similarity, and discriminability of both images and boxes. The proposed formulation is capable of accounting for noisy annotations in real-word images.

B. Motivation and Contribution

Deformable part-based models (DPMs) [7] and their variants [27]–[29] have achieved remarkable success in supervised object detection on challenging PASCAL VOC datasets [30] for a long period. The DPMs represents an object with a holistic *root* filter that approximately covers an entire object and with several higher resolution *part* filters that capture smaller local appearances (parts) of the object. It also characterizes the deformations by links connecting different parts. In the standard (fully supervised) DPMs framework, the root filter is initialized with the positive ground-truth object bounding box, and is allowed to move around in its small neighborhood to maximize the filter score. The locations of object parts are always treated as latent information due to the unavailability of object part annotations upon most occasions. A *latent* SVM (LSVM) is adopted to learn object deformation, which can alternate between fixing latent values (part locations) for positive examples and optimizing its objective function.

Pandey and Lazebnik [22] modify the fully supervised DPMs in a weakly supervised manner without object level annotations: this treats the location of root filter and part filters fully latent and learns structural object detectors based on the entire image. Root filter location is initialized randomly, based on a window that has at least 40% overlap with the positive training image, while its aspect ratio is initialized roughly to the average of the aspect ratios of positive training examples. However, the specific size and location of the initial root filter, as well as their aspect ratio, are indicated to have a significant impact on the final localization result [6], [7], [22]. By random initialization, the object detector tends to learn spurious models of other classes or background regions, leading to lower accuracy during testing. To the best of our knowledge, methods for initializing the root filter based on theoretical deduction in weakly supervised DPMs, as well as the definition of the object aspect ratios, have not been properly studied in [22].

To make up the performance gap between weakly and fully supervised DPMs, in this paper, our goal in this paper is to propose a model enhancing the weakly supervised DPMs by emphasizing the importance of location and size of the initial class specific root filter. To be more precise, our goal is to discover a reliable initial set of image windows that are likely to contain the target objects in the positive training images with only category level annotations, so as to represent the object instances. Hence, our WSL framework incorporates adaptive window selection from class independent object proposals and training of deformable part-based models. In particular, we explore the "objectness" approaches [19], [31], which generate class independent object proposals with corresponding scores indicating their probabilities of being object instances. We then adaptively select a reliable set of windows from the derived object proposals for each image as initialization, by incorporating visual saliency and "objectness" scores. Two different initialization schemes are developed: *single* region and *multiple* region initilization. The former tends to select one relative larger bounding box which may contain the most salient part in the image, while the latter is far more general, which selecting a small number of object estimations that can also capture smaller and scattered objects. For multiple region initialization, the region labels are latent information. We learn the latent class label by framing it as a classification problem, which tries to coarsely classify each region into a target object class or a non-target class by some class specific classifiers. The generated object estimations are treated as the initial root filter estimates for training DPMs detectors.

The main contributions in this work are four-fold:

1) We propose a selection model based on generic "objectness" and visual saliency to adaptively select a discriminative set of candidate windows which tend to represent the object instances in each weakly labeled training image.
2) We frame the learning of the latent class label of each candidate window as a binary classification problem, by training category specific classifiers, which try to coarsely classify a candidate window into either a target object or a non-target class.

3) We propose to use a flexible enlarging-and-shrinking post-processing procedure to modify the predicted output of the DPMs detector, which can effectively generate more accurate bounding boxes by better conserving foreground and cropping out plain background regions, to approximatively match the object aspect ratios.
4) Extensive experiments are carried out on two subsets and on the entire set of the challenging PASCAL VOC 2007 database [30] with different criteria, namely annotation accuracy in terms of correct localization on training set, and detection accuracy in terms of average precision on test set. Experimental results demonstrate that our proposed framework is effective for initialization of the DPMs root filter and that it shows shows competitive final localization performance with the state-of-the-art weakly supervised object detection methods. To the best of our knowledge, we are the first to present weakly supervised results on the Microsoft COCO 2014 dataset [32].

A preliminary version of this work appeared in [33], incorporating the generic "objectness" with deformable part-based models for WSL detection. While including that work, this paper significantly extends it in the following ways. First, we explore a far more general M-WDPMs (multiple region initialization for weakly supervised deformable part-based models) model which tries to select multiple regions, and we learn the latent label information of these regions in an effective way. This model shows its superiority in discovering not only salient objects but also smaller and scattered objects in comparison with S-WDPMs (single region initialization for weakly supervised DPMs) in [33]. Second, we additionally experiment with advanced region proposals generated by Selective Search [31], as well as adopting the deep convolutional neural network (CNN) [34] features as image representation in contrast with traditional low-level handcrafted features (e.g., HOG [6]). Third, we evaluate our framework on the entire PASCAL VOC 2007 dataset, and compare it with state-of-the-arts. We also analyze the types of error that our detection framework tends to make, in order to give insights for future improvement. Finally, we report the detection results on the challenging Microsoft COCO 2014 dataset.

C. Organization of the Paper

The rest of the paper is organized as follows: we present our weakly supervised DPMs framework in detail in Section II, while in Section III we present our experimental results and the comparison with other methods on PASCAL VOC 2007 and Microsoft COCO 2014 datasets. In Section IV, we conclude our work.

II. Fusing Generic Objectness and Deformable Part-Based Models for Weakly Supervised Object Detection

In this section, we detail our approach of the weakly supervised DPMs for object detection. First, we introduce our approach to adaptively select the representative and discriminative candidate regions from the category-independent object proposals. Second, we elaborate how to learn latent class information when multiple regions are selected. We then briefly describe the weakly supervised learning procedures using the selected regions with DPMs and the detection rescoring algorithm for testing. Finally, we propose our new post-processing method to further refine the predicted object bounding box obtained by a weak DPMs detector, so as to cover the object more precisely.

A. Object Estimations: Initialization

In the weakly supervised DPMs training procedure, good initialization of the root filter is crucial. Our goal is thus to discover a reliable initial set of image windows likely to contain the target objects in the positive training images with only image-level annotations, so as to represent the object instances.

1) Region Extraction: Two general approaches have been proposed for generating class-independent object proposals in recent years: *window scoring methods* such as Objectness [19], BING [35], EdgeBoxes [36] and *grouping methods* such as Selective Search [31], Constrained Parametric Min-Cuts (CPMC) [37], Multiscale Combinatorial Grouping (MCG) [38]). We use Selective Search since it has been used as the proposal generating method by the state-of-the-art supervised R-CNN detector [3]. We also report results using the Objectness method [19] to compare with prior detection work [19], [33].

Given an input image $\boldsymbol{I}$ (shown in Fig. 1(a)), we first select top n scored windows $\boldsymbol{W} = \{w_1, w_2, \ldots, w_n\}$ and corresponding scores, denoted as $\boldsymbol{S} = \{s_1, s_2, \ldots, s_n\}$, indicating the probabilities of covering objects within them, generated by Selective Search (shown in Fig. 1(b)). To balance a high recall (i.e., covering more objects) and computation efficiency (i.e., small number of region proposals), we set $n = \min(1000, N)$ according to [39], where N is the number of proposals generated by Selective Search.

Based on the fact that the region proposal method is designed to capture all possible objects within an image, we assume that it is sufficiently reliable to provide a set of good candidate windows $\boldsymbol{W}^* \subseteq \boldsymbol{W}$ covering the objects of interest. However, windows with higher scores are not always the effective choices [20]: they usually encompass other noisy background, or they may cover only some object parts. To extract a reliable set of object estimations from the pool of n windows, we design a sequential selection scheme shown in Fig. 1(c)–(g).

2) Salient Reference Region: For weakly supervised learning of DPMs detectors, it is obvious that the initialization of the root filter is significant. The detector will be seriously damaged if it shoots on the background region. Consequently, it is an absolute necessity to start from visually meaningful regions (foreground objects). Inspired by the success of visual saliency applied in salient object recognition [1], [40], we compute the reference region $\boldsymbol{R}$ [shown in Fig. 1(d)] by taking the threshold and merging the discrete saliency map (or heat map) $\boldsymbol{M}$ into one or more connected region(s) using [41] [shown in Fig. 1(c)]. The value of saliency map $\boldsymbol{M}$ at pixel $I(i, j)$ is obtained by summing up the scores of the windows that cover this pixel

$$\boldsymbol{M}(i, j) = \sum_{k=1}^{n} \boldsymbol{M}_k(i, j) \tag{1}$$

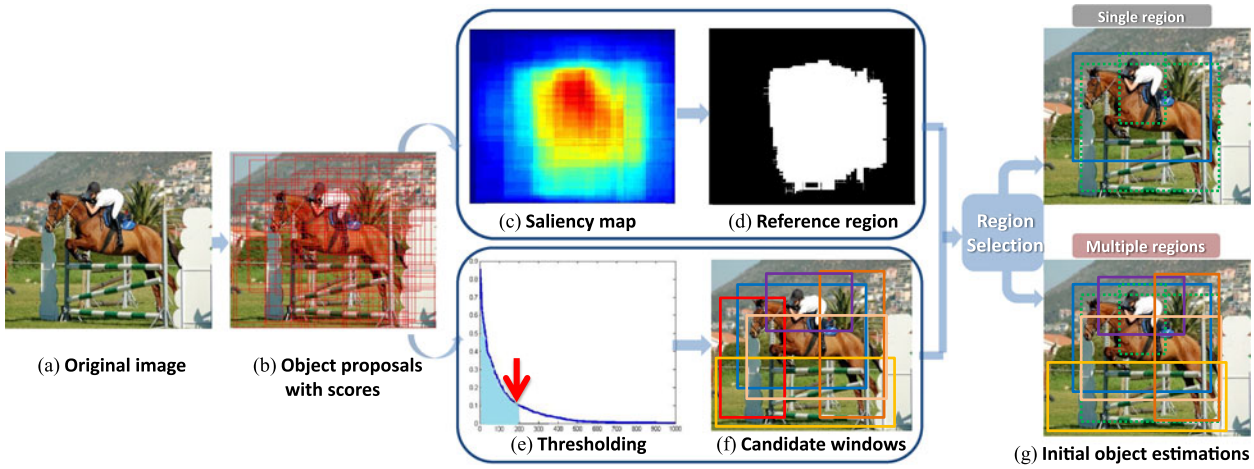


Fig. 1. Illustration of our proposed method to extract the initial object estimations: for (a) an input image, (b) object proposals and corresponding scores indicating the probabilities of containing objects are generated using the objectness [19] or selective search [31] method. (c) is the saliency map derived from (b), and (d) is the reference region obtained by thresholding (c). A coarse set of candidate windows (f) is selected based on the sorted scores of object proposals (e) after non-maximum suppression (NMS). In the top image of (g), which indicates the single region selection scheme, the blue window is our initial object estimation obtained by optimizing the overlap between (d) and (f). The bottom image of (g) indicates the multiple region selection scheme. Its color windows with solid lines are multiple finer regions which are assumed to represent the objects in the original image. For both images of (g), the green dot line windows are ground-truth bounding boxes for person and horse, respectively.

where

$$\boldsymbol{M}_k(i,j) = \begin{cases} s_k, & \text{if } I(i,j) \in w_k, \forall w_k \in \boldsymbol{W} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The reference region $\boldsymbol{R}$ can be one connected (continuous) region or several discrete regions in the image according to the score range and threshold value.

3) Coarse Candidate Window Pool: It is known that the score predicted by Selective Search (i.e., objectness score) corresponds to the probability of containing a target object to some extent. To take advantage of this auxiliary information, we concurrently select the top 200 scored windows out of n windows as candidates, [shown in Fig. 1(e)]. To avoid near duplicate candidate windows, we further perform non-maximum suppression (NMS) to obtain a finer set of candidates. Fig. 1(f) illustrates the derived smaller set of l confident candidates $\hat{\boldsymbol{W}} = \{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_l\}$ and their corresponding scores denoted as $\hat{\boldsymbol{S}} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_l\}$.

4) Object Invariant Estimations: Given the reference region $\boldsymbol{R}$ which implies the most salient region (or regions) within an image, and confident candidate windows $\hat{\boldsymbol{W}}$ with scores $\hat{\boldsymbol{S}}$, the overlap between them provides valuable information for finding the locations of target objects. We will propose two different schemes to fuse the salient region(s) with the extracted candidate windows.

a) Single region initialization: In [22], the root filter of the DPMs is randomly initialized from a *single* window which covers at least a 40% overlap with the original image. Hence, we also filter out only one *single* window w^* from the candidate pool $\hat{\boldsymbol{W}}$ in order to obtain a direct comparison with [22]. Intuitively, we expect this window estimation to have a larger overlap with the salient reference region $\boldsymbol{R}$, as well as a relatively higher objectness score. Therefore, the estimation of the initial object bounding box with objectness score (w^*, s^*) (Fig. 1(g), top image) can be determined by optimizing the following function:

$$(w^*, s^*) = \underset{\hat{w}_i \in \hat{\boldsymbol{W}}, \hat{s}_i \in \hat{\boldsymbol{S}}}{\arg\max} \left[\alpha \hat{s}_i + (1-\alpha) \frac{\text{area}(\boldsymbol{R} \cap \hat{w}_i)}{\text{area}(\boldsymbol{R} \cup \hat{w}_i)} \right], \quad i \in [1, l] \tag{3}$$

where α is a parameter used to control the influence of the objectness score s_i. In practice, $\alpha = 0.2$, was selected by a grid search over $\{0.1, 0.2, 0.3, 0.4\}$ on a validation set, for the purpose of emphasizing the priority of the intersection over union (IoU) overlap between the candidate window and the merged salient reference region.

The single region initialization scheme prefers to select a relatively large region which may contain the most salient part in the image. When very few objects are closely gathered in images, it can produce good DPMs object detectors in a weakly supervised manner. For example, by adopting the single region scheme, the blue window in Fig. 1(g) top image, is used as a positive training example (i.e. DPMs root filter initialization) for both the *horse* and the *person* categories. Moreover, the strategy of taking large windows in positive images exploits the inclusion structure of the multiple instance learning (MIL) problem for object detection: although large windows may contain a significant amount of background features, they are likely to include positive object instances and their contextual information.

b) Multiple region initialization: In fact, multiple objects (e.g., 2.5 objects on average for PASCAL VOC 2007 trainval dataset, 7.7 for MS COCO 2014) can be scattered anywhere in an image. We can therefore further improve DPMs detectors by providing more object estimations as root filter initialization, instead of training the object detectors with a single window for each image. For each image, we are motivated to select a small number of object estimations that can also capture smaller and scattered objects, better representing the original image. Meanwhile, object proposal algorithms such as Selective Search and Objectness tend to generate more overlapping bounding

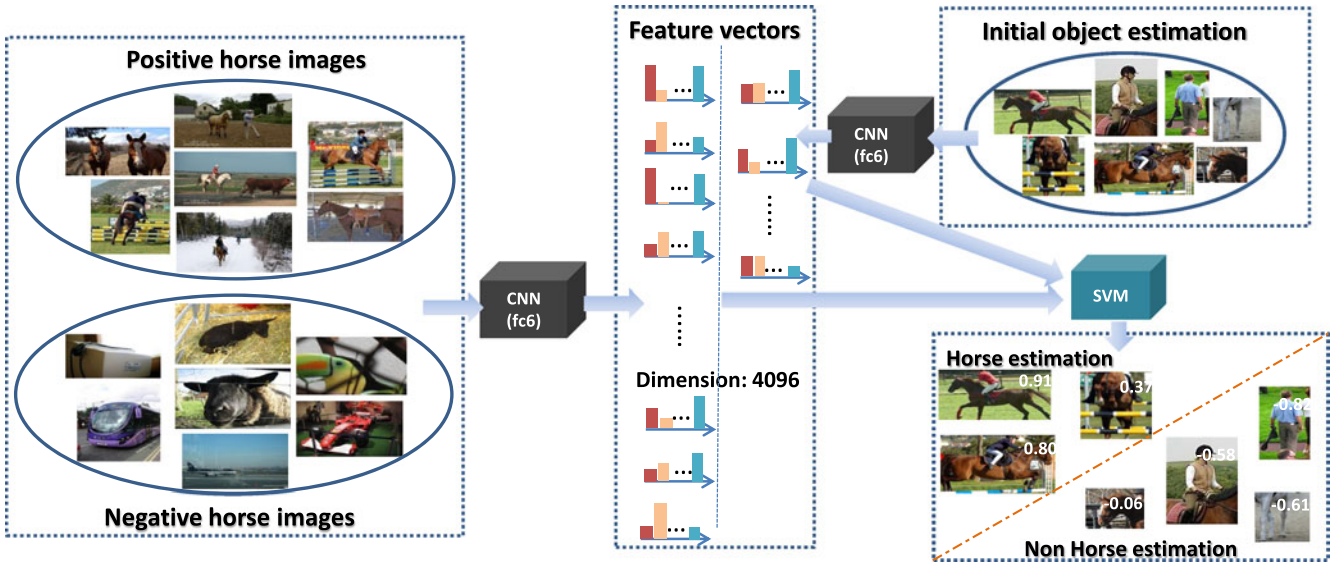


Fig. 2. Illustration of our latent class learning framework for the *horse* category. For each object category, we train a linear SVM classifier with the CNN features (output of CNN's $fc6$ layer) of image-level samples (as shown in the left part). Object estimations from the positive training images of this category are scored by its SVM. We select the regions with higher scores by thresholding as the representative objects of this category (*horse* vs. *non horse* for this example).

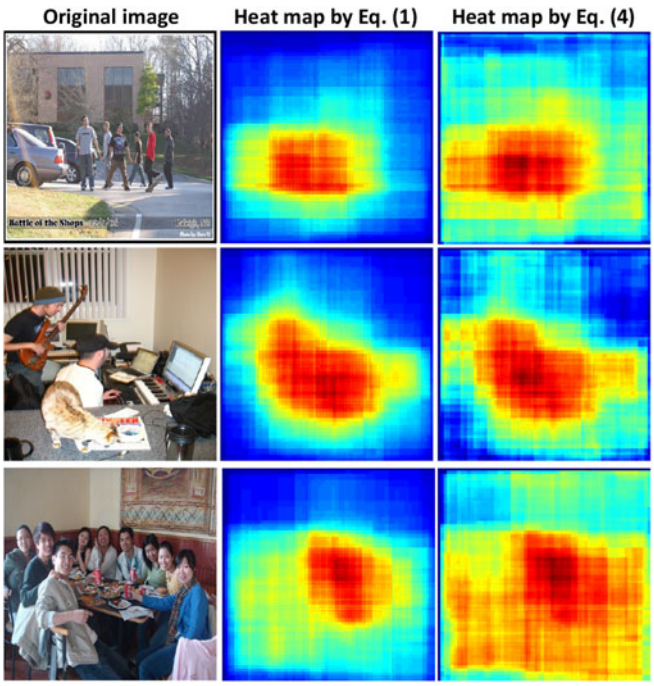


Fig. 3. Some heat map examples generated by (1) and (4).

boxes on larger objects than on smaller ones. Consequently, scattered small objects are likely to be ignored using (1). Hence, in order to fully consider these objects which were originally ignored by (1), we modified it by dividing the sum of scores by the square root of the number of windows that cover this pixel:

$$\boldsymbol{M}(i,j) = \frac{1}{\sqrt{\hat{k}}} \sum_{k=1}^{n} \boldsymbol{M}_k(i,j) \tag{4}$$

where, $M_k(i,j)$ is defined as the same in (2), and $\hat{k}$ is the number of windows that cover pixel $I(i,j)$. We show some heat map examples generated by (1) and (4) in Fig. 3.

We adopt similar criteria to the score function (3), with the best α being set to 0.4 (for both PASCAL VOC 2007 and MS COCO 2014) from a grid search over $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Instead of only selecting the maximum scoring window in (3), we pick out top Q scored windows W^* for each image. We will discuss the value of Q in the experiment part.

After generating several object estimations from each image, the next step is to approximately identify the class label of each estimation given only the labels of the whole image. For example, in Fig. 1(g) bottom image, the color windows with solid lines are associated with the *horse* and *person* labels. However, so far we have no idea which object(s) (or even background) is/are inside each bounding box. Our goal will be to solve this problem in the next subsection.

B. Learning Latent Object Classes via Region Classification

For each positive training image, we have generated Q object invariant estimations with the multiple region initialization scheme ($Q = 1$ for single region initialization, and we use the image-level labels as training annotations). Consider an object category, e.g., *horse*, which has P positive training images, we can obtain a total number of $z = P * Q$ object estimations. Obviously, some of these object estimations come from other categories (e.g., *person*, *sheep*, object parts or the background regions as well), where the class labels are latent information. For single region initialization, the unique generated window is used to initialize the DPMs root filter for any categories appearing in the image. As for multiple region initialization, in this paper we frame the latent class learning problem as a classification problem by coarsely classifying these object estimations into either the target object category or the non-target category (i.e., other classes, object parts or background).

1) Region Representation: We use the deep convolutional neural network (CNN) features to represent the regions (object

estimations). Firstly, we pre-train an eight-layer (five convolutional layers and three fully-connected layers) *Alex-Net* [34] CNN with *caffe* implementation [42] on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 classification dataset [9], which contains 1.2 million images of 1000 categories. We then warp each region into a required fixed pixel size of 227×227, and subtract it with the mean RGB image of the training set, before forward propagating it through the network. Finally, we take the output of the $fc6$ layer as R-CNN [3], which is a 4096-dimensional feature vector, to represent the input region. While this feature extraction process is similar to that of R-CNN, it is worth noticing that we do not fine-tune the pre-trained CNN on the target dataset. This is because the object level annotations are assumed not to be available in the weakly annotated data. We do not pad the region with additional image context around it either, as our region estimation is already expected to have a significant coverage of the context information due to our selection schemes in Section II-A.3.

2) Region Classification: Consider training a *horse* detector. For all the P positive training images in the *horse* category, we generate z object invariant estimations. Intuitively, only part of these z regions contains the target *horse* object, others may have *person*, *sheep*, *dog* or even background. We learn the latent categories in these regions via region classification.

We first train a *horse* linear SVM classifier [43] using the images labeled with *horse* as positive training examples and those without *horses* as negative examples. We compute the $fc6$ 4096-dimensional CNN features as in Section II-B.1 on whole images. We then run the trained *horse* classifier on the z object invariant estimations in the positive training images. By thresholding the SVM scores, finally we obtain a subset z' regions from z estimations ($z' < z$). These z' regions are assumed to represent the target *horse* category, which can be treated as positive training examples of the *horse* detector.

Suppose we have K categories that we want to detect. We train one binary SVM classifier on positive and negative images of each category, and run these K classifiers on their corresponding object estimations. We select high scoring regions for each target category so as to represent the objects of interest. Fig. 2 shows the latent class learning framework using SVM classification on the *horse* category.

C. Weakly Supervised DPMs Training and Testing Details

We design two different kinds of deformable part based models for weakly supervised object detection according to different initialization schemes in Section II-A.

1) Single Region Initialization for Weak DPMs (S-WDPMs) Detection: Similarly to [7], each root filter hypothesis in a positive training image is initialized with the corresponding derived bounding box from the single region initialization scheme. The size and aspect ratio of the DPMs root filter are decided by the average size and aspect ratio of the object estimation boxes (ground-truth bounding box and aspect ratio are used in [7]). The root filter hypothesis is allowed to move around in a small neighborhood to maximize the filter score so as to compensate for imprecise bounding box estimation from Section II-A4a. In order to obtain a direct comparison with [22], we also represent an image by a multiscale HOG feature pyramid [6] of 16 levels. For this S-WDPMs model, we use only a single component, since the multiple components are used for detecting objects with different views (S-WDPMs is trained on each view/category, e.g., *Left, Right*, etc.). We set the number of parts in this DPMs to 8 as in [7]). For negative training examples, we use random negatives from other object categories. For testing, the sliding window approach is adopted. This single region initialized weakly supervised DPMs detection model is denoted as *S-WDPMs*. We refer the reader to [7] for more details concerning the DPMs training and detection procedures.

2) Multiple Region Initialization for Weak DPMs (M-WDPMs) Detection: For the M-WDPMs (multiple region initialized weakly supervised DPMs), we make it much "*deeper*" with the *DeepPyramid* feature [44], for the reason that the HOG feature is suboptimal compared to deep features computed by CNN [3], [8], [24], [45], [46]. The feature map is computed by the fifth convolutional layer ($conv5$), which has 256 feature channels. We represent each image (or region) with a feature pyramid of 7 levels as in [44]. For training, the selected object estimations from Section II-B.2 are treated as positive training examples, and the random windows from negative images are defined as negative examples. We use a DPMs with 3 components and 8 parts per component according to [44]. The training and testing procedures are similar to S-WDPMs above, but we add a simple bounding box rescoring stage with the help of a front-to-end CNN padded with a softmax classifier as follows.

The contextual information provided by classification and detection can mutually boost the performance of the other, based on the assumption that they adopt different information [47], [48]. Classification looks at the objects and their contextual information, while detection mainly focuses on the object shape and all parts. For example, if an object is occluded or truncated, it will be difficult for the detector while the classifier could still have enough information such as context and certain parts. Inversely, the detector is able to find small objects and objects appearing in non standard context, while the classifier may fail. Hence, we are motivated to combine the classification score and the detection score. We formulate the rescoring function as a linear combination of the DPMs detection score and region classification score

$$s^i_{\rm det} = \kappa s^i_{M-{\rm WDPMs}} + (1-\kappa) s^i_{\rm cls}, \quad i \in [1, K] \qquad (5)$$

where, $0 \leqslant s^i_{M-{\rm WDPMs}} \leqslant 1$ is the normalized DPMs detection score on a sub-window of the $i^{\rm th}$ detector, and $0 \leqslant s^i_{\rm cls} \leqslant 1$ is the softmax classification score of the corresponding $i^{\rm th}$ category on this sub-window. κ is a hyper-parameter used to leverage the two scores, which ranges from 0.5 to 1.0. K is the number of object categories. The final predicted windows are obtained by thresholding the $S^i_{\rm det}$ in (5).

To train this front-to-end CNN classifier described above, we fine-tune the pre-trained CNN with image level annotations on the training set of the target dataset. We implement it by removing the last 1000-way softmax layer while keeping all the other parameters and adding a new randomly initialized K-way softmax classification layer. We then fine-tune the entire network based on the image-level labels.

In [7], contextual information is exploited to rescore the bounding boxes. However, it needs object-level annotations

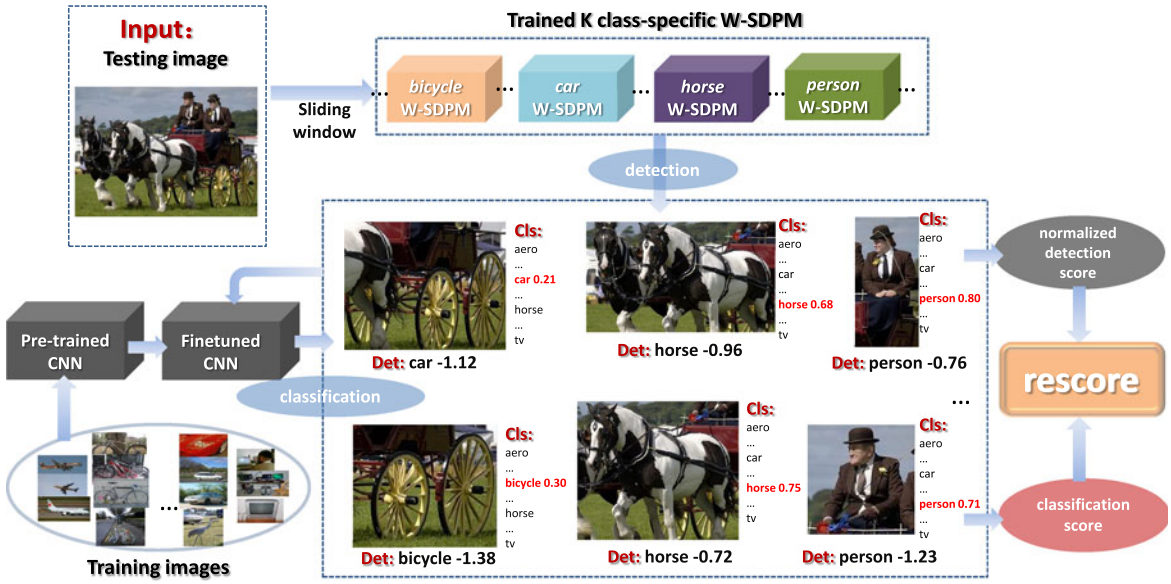


Fig. 4. Illustration of detection rescoring using an M-WDPMs and CNN softmax classifier. For a testing image, K (number of classes in the target dataset) class-specific M-WDPMs are applied on it in a sliding window manner. For each subwindow detected by M-WDPMs, the normalized detection score is rescored by the softmax classifier of the detected category. In this example, the wrongly detected car and bicycle are finally discarded by the detector after the rescoring stage.

to extract the contextual information. Our detection rescoring method does not require the object-level annotations, and leads to a remarkable improvement in average precision on several classes in the PASCAL VOC 2007 dataset (see Section III-B). In [48], the image classification scores are used as contextual features, and concatenated with the object detection scores to form a final feature vector, based on which a linear SVM is learned to refine the detection score. An example of our bounding box rescoring procedure is shown in Fig. 4.

D. Bounding Box Post-Processing

In many cases, the bounding boxes generated by DPMs detectors are too large (resp. small) when detecting very small (resp. large) objects due to the restrictions of the size of the root filter and the scale of the feature pyramid. To improve localization and obtain a more precise prediction of the bounding box aspect ratio, we post-process each bounding box by enlarging or shrinking (*ES* post-processing) it to cover the object as much as possible. This is done using an improved version of the method proposed in [49], which measures the amount of area that the edge energy occupies. In brief, we first augment the original bounding box $w = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ to 120% of the original width and height (i.e., 144% in total area, denoted as $w^{\text{aug}} = (x^{\text{aug}}_{\min}, y^{\text{aug}}_{\min}, x^{\text{aug}}_{\max}, y^{\text{aug}}_{\max})$. We expand from the centroid if applicable. Otherwise, we stop when reaching the border of the image and calculate the absolute values of the gradients $L'_{w^{\text{aug}}}$ by applying a 3×3 Laplacian filter with $\gamma = 0.2$ over the augmented bounding box. To simplify calculation of the edge spatial distribution, we then resize the gradient magnitude image size to 100×100 and normalize the image sum to 1, i.e., $L'_{w^{\text{aug}}}$. Moreover, we set the values that are less than 10% of the maximum $L_{\max}$ to 0. Finally, we expand the bounding box in four directions from the current centroid (x'_c, y'_c) and stop when it contains 98% of the total gradient magnitude (edge energy) in the augmented box. The detailed algorithms are shown in Algorithm 1.

This post-processing technique is not only able to crop out plain background regions, but can also expand to cover the foreground regions which are not encompassed by the original box. However, the cropping method in [22] can only shrink to reduce the background. Fig. 5 shows a few examples of our bounding box post-processing results. It is also worth noticing that this post-processing technique works efficiently for the objects with a unique or plain background, but is of limited help for those with cluttered or textured backgrounds.

III. Experimental Evaluation

In this section, we present the experimental results of our proposed framework with two different initialization schemes (i.e., S-WDPMs using single region initialization and M-WDPMs using multiple region initialization) on the challenging PASCAL VOC 2007 dataset [30] and the Microsoft COCO 2014 dataset [32].

A. Experiments With S-WDPMs

1) Datasets: Following the protocol used in previous works [17], [22], [26], [50], we evaluate the performance of our proposed S-WDPMs (single region initialized weak DPMs) framework on two subsets from the training and validation set (*trainval*) of the PASCAL VOC 2007 dataset (*VOC07*)[30]: *VOC07-6 × 2* and *VOC07-14*. The *VOC07-6 × 2* subset contains 6 classes (*aeroplane, bicycle, boat, bus, horse* and *motorbike*) with *Left* and *Right* views (aspects) of each class, resulting in a total of 12 separating classes. The *VOC07-14* subset (same as *PASCAL07-all* defined in [22]) consists of 42 class/view combinations covering 14 classes and 5 views (*Left, Right, Frontal,*

Algorithm 1: Bounding box post-processing pipeline.

Input:

Original bounding box:
$w = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$;
Original image width: w_o; original image height: h_o;
Maximal expanding rate: $\beta = 1.2$;
Laplacian filter shape: $\gamma = 0.2$.

Output:

Cropped bounding box:
$w' = (x'_{\min}, y'_{\min}, x'_{\max}, y'_{\max})$.

centroid: $(x_c, y_c) = (\frac{x_{\min}+x_{\max}}{2}, \frac{y_{\min}+y_{\max}}{2})$
augmented width: $a = \beta * (x_{\max} - x_{\min})$
augmented height: $b = \beta * (y_{\max} - y_{\min})$
if $x_c - \frac{a}{2} > 0$ **then**
  $x^{\text{aug}}_{\min} = ceil(x_c - \frac{a}{2})$
else
  $x^{\text{aug}}_{\min} = 1$
end if
if $x_c + \frac{a}{2} < w_o$ **then**
  $x^{\text{aug}}_{\max} = floor(x_c + \frac{a}{2})$
else
  $x^{\text{aug}}_{\max} = w_o$
end if
$y^{\text{aug}}_{\min}$ and $y^{\text{aug}}_{\max}$: process in the same way as x;
$w^{\text{aug}} = (x^{\text{aug}}_{\min}, y^{\text{aug}}_{\min}, x^{\text{aug}}_{\max}, y^{\text{aug}}_{\max})$;
$L_{w^{\text{aug}}} = filter(image(w^{\text{aug}}), 'laplacian', \gamma)$;
$L'_{w^{\text{aug}}} = norm(resize(|L_{w^{\text{aug}}}|, [100, 100]), 1)$;
$L_{\max} = max(L'_{w^{\text{aug}}})$;
for $i = 1, 2, \ldots, 100$ **do**
  for $j = 1, 2, \ldots, 100$ **do**
    if $L'_{w^{\text{aug}}}(i, j) < 0.1 * L_{\max}$ **then**
      $L'_{w^{\text{aug}}}(i, j) = 0$
    end if
  end for
end for
current centroid: $(x'_c, y'_c) \leftarrow$ average energy point of $L'_{w^{\text{aug}}}$;
while energy in $w'' < 0.98 * \sum(L'_{w^{\text{aug}}})$ **do**
  $w'' = (x''_{\min}, y''_{\min}, x''_{\max}, y''_{\max}) \leftarrow$ update by expanding bounding box in four directions $(-x, -y, x+, y+)$ from the current centroid (x'_c, y'_c).
end while
project w'' into original image:
$w' = (x'_{\min}, y'_{\min}, x'_{\max}, y'_{\max}) \leftarrow$
$w'' = (x''_{\min}, y''_{\min}, x''_{\max}, y''_{\max})$

Rear and *Unspecified*). Similar to [17], [22], [26], [50], we remove all the images annotated as *difficult* or *truncated* in both the training and the evaluation steps.

2) Evaluation Protocol: To make fair comparisons with previous works [17], [22], [23], [50], we only choose the detection window with the highest DPMs score per image, although our method can detect multiple instances appearing in the image using the sliding window approach. We also report both results

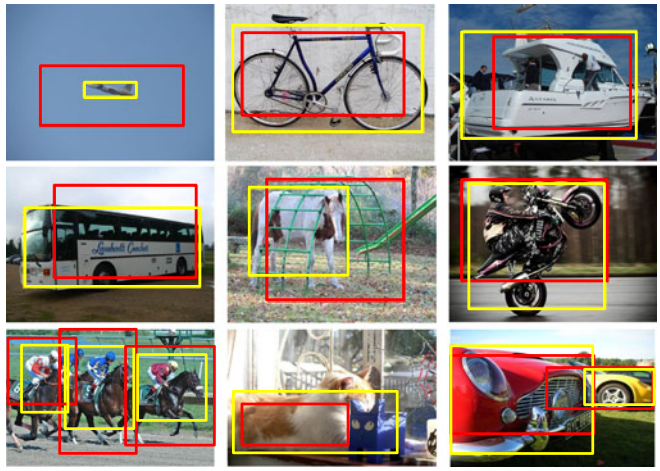


Fig. 5. Examples of bounding box enlarging and shrinking. Boxes before (resp. after) post-processing are shown in red (resp. yellow).

for initial and refined localization as [22], [50]. A refined localization is obtained by an iteratively trained DPMs detector for one/several iteration(s) to refine the initial detection using the previous annotations as ground truth. Performance is evaluated with the percentage of *training* (train + val) images in which an object is correctly covered by the window (i.e. CorLoc [17]), if the strict PASCAL-overlap criterion IoU (intersection-over-union) $\geqslant 0.5$ is satisfied.

3) Experimental Evaluation: We compare our S-WDPMs with Weak DPMs [22], Weak objectness [17] and the Joint topic model [23]. For the Weak objectness approach [17], the region proposal with the highest "Objectness" score is selected as the predicted window. As shown in Table I, our method outperforms [17] and our baseline approach [22] on both datasets. Both [22] and our S-WDPMs use the same HOG feature pyramid for the DPMs. We present our results using two kinds of object proposal generating methods: *Objectness* (*obj*) and *Selective Search* (*SS*). For *obj*, our average performance of initial detection before post-processing the bounding boxes on the *VOC07-6* $\times$ *2* and *VOC07-14* subsets is 38.74% and 21.73% respectively, versus 37.22% and 19.98% in [22]. These improvements are due to the initial object estimate of our method described in Section II-A.4a, which ensures better initialization of the root filter of DPMs detectors. We can also observe that both the post-processing method of cropping [22] (i.e., S-WDPMs(crop) in Table I) and our enlarging-or-shrinking [i.e., S-WDPMs(ES)] post-processing method steadily improves average localization accuracy. In particular, our ES method is superior to the cropping method of [22], as our cropped bounding box is able not only to shrink to crop out the background regions, but is also capable of enlarging to cover the whole foreground object resulting from incomplete coverage of the original window. An example is shown in the last row of Fig. 6, where the target object (*motorbike*) is only partially localized by the initial detector (shown in red rectangles in the middle and right images) for both [22] and our method. However, in the final detection (shown in yellow) after post-processing, our method is able to enlarge the

TABLE I
AVERAGE LOCALIZATION ACCURACY (AS A %) OF OUR S-WDPMS (SINGLE REGION INITIALIZED WEAK DPMS WITH HOG FEATURES) COMPARED WITH STATE-OF-THE-ART COMPETITORS ON THE TWO VARIATIONS OF THE PASCAL VOC 2007 DATASETS

	no post-processing			with post-processing						
		S-WDPMs			S-WDPMs(crop)		S-WDPMs(ES)		[23]	
	[22]	*obj*	*SS*	[22]-crop	*obj*	*SS*	*obj*	*SS*	*S*	*G*
Dataset	**VOC07-6 × 2**									
Initialization	37.22	38.74	41.52	44.62	47.85	48.40	48.59	51.01	50.8	**51.5**
Refinement 1	51.63	55.85	63.31	53.11	56.78	64.25	58.02	**67.13**	65.5	66.1
Refinement 2	56.99	59.82	—	59.31	63.31	—	63.91	—	—	—
Refinement 3	59.32	—	—	61.05	—	—	—	—	—	—
Result from [17]	50.00									
Dataset	**VOC07-14**									
Initialization	19.98	21.73	24.87	23.00	24.20	26.30	25.12	31.84	**32.2**	30.5
Refinement 1	25.11	27.46	31.15	26.38	28.21	33.10	28.94	**34.91**	33.8	32.5
Refinement 2	27.69	28.95	—	29.39	32.87	—	32.82	—	—	—
Refinement 3	28.98	—	—	30.31	—	—	—	—	—	—
Result from [17]	26.00									

"crop" and "ES" denote the cropping method from [22] and our enlarging & shrinking post-processing. "*obj*" and "*SS*" denote the objectness and Selective Search region proposal generating method. "*S*" and "*G*" denote the Sampling and Gaussian strategy from [23].

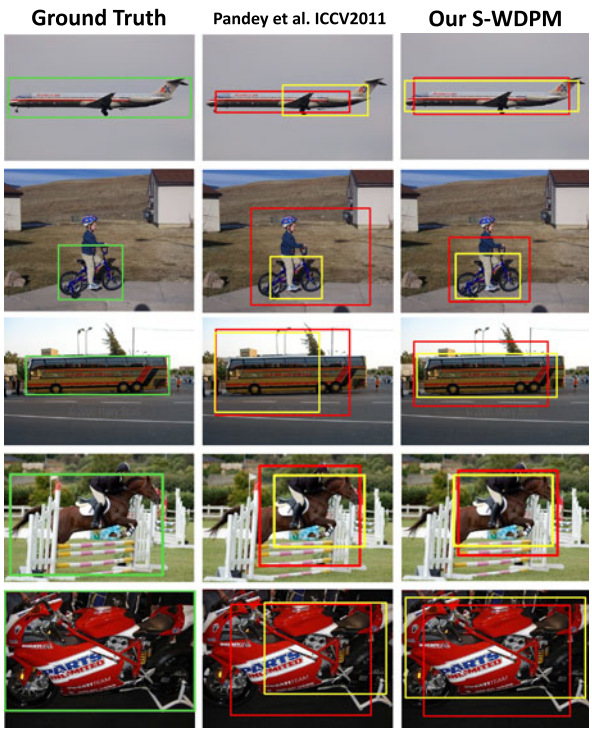


Fig. 6. Examples of localization results for our S-WDPMs on PASCAL VOC 2007 images. The left column: ground-truth bounding boxes in green rectangles. The middle and right columns are detection results with [22] and our S-WDPMs framework, respectively. Initial detections are shown in red, while detections refined by detectors are shown in yellow. Both results use the individual post-processing approach.

bounding box to approximately include the whole object, while [22] tends to crop out both foreground and background regions.

Furthermore, the rows starting with "Refinement" in Table I indicate that localization accuracy can benefit from the iterative refinement process. It is worth mentioning that with a better initialization, our models converge to a steady level of performance after one less round of costly re-training than in [22] (both using *Objectness*), and achieve slightly better results in the meantime.

TABLE II
CLASS LEVEL LOCALIZATION ACCURACY (AS A %) FOR THE *VOC07-6×2* DATASET FOR OUR S-WDPMS(ES) USING *Objectness* PROPOSALS VERSUS [17], [22], [50]

	Initialization				Refined by detector		
	ours	[22]	[50]	[26]	**ours**	[22]	[17]
aero left	**65.1**	55.8	39.1	49.1	**69.7**	65.1	58.0
aero right	**64.1**	61.5	50.0	51.3	**84.6**	82.1	59.0
bike left	31.3	**31.3**	28.4	25.0	85.4	**87.5**	46.0
bike right	42.0	**44.0**	30.6	24.0	54.0	**68.0**	40.0
boat left	9.1	4.6	**15.1**	11.4	**13.6**	2.3	9.0
boat right	9.3	9.3	**20.7**	11.6	14.0	7.0	**16.0**
bus left	23.8	23.8	**31.0**	38.1	**42.9**	28.6	38.0
bus right	**65.2**	52.2	35.1	56.5	69.6	47.8	**74.0**
horse left	**64.6**	60.4	48.5	43.5	**87.5**	83.3	58.0
horse right	**73.9**	67.4	45.2	52.2	76.1	**80.4**	52.0
mbike left	**64.1**	48.7	46.3	51.3	87.2	**92.3**	67.0
mbike right	70.6	**76.5**	55.3	64.7	82.4	**88.2**	76.0
average	**48.6**	44.6	37.1	39.3	**63.9**	61.1	50.0

The detailed comparisons for our S-WDPMs using *Objectness* with the state-of-the-arts on the *VOC07-6 × 2* dataset are listed in Table II. The results show that our method outperforms [17], [22], [50] for many of the categories. In particular, our method achieves the state-of-the-art results in the classes where the target object possesses the most salient regions in that category (e.g., *aeroplane, bus, horse*). Interestingly, even without the refinement process, the accuracy of our method in certain categories (e.g., *aeroplane left*) is superior to competitors using the time-consuming refinement procedure. Fig. 6 visually compares some of our results with those of [22]. We also list the

TABLE III
COMPARISONS OF WEAKLY SUPERVISED OBJECT DETECTORS ON PASCAL VOC 2007 TRAINVAL SET IN TERMS OF CORRECT LOCALIZATION (CORLOC [17], AS A %) ON POSITIVE TRAINING IMAGES

method / class	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mean**
our S-WDPMs-HOG	49.1	32.8	27.2	9.8	6.6	38.0	46.7	48.2	8.9	35.7	15.3	34.5	42.2	49.5	16.7	13.8	31.6	26.3	47.8	23.1	30.2
our M-WDPMs-HOG	67.9	52.4	34.4	21.9	12.1	42.0	59.9	58.4	9.9	42.0	13.5	38.9	48.1	58.6	20.4	19.5	40.8	24.9	48.9	42.7	37.9
our M-WDPMs-deep†	72.0	58.8	38.5	24.6	14.8	46.2	63.4	63.0	18.4	49.9	17.0	40.3	52.6	63.2	22.2	22.9	46.1	26.2	52.8	46.8	42.0
our M-WDPMs-rescore†	**80.3**	59.1	38.9	**26.0**	14.9	48.8	65.4	65.1	16.6	**58.5**	17.3	42.7	**58.8**	**69.6**	22.8	20.7	52.9	24.0	53.3	46.6	44.1
Joint Learning [13]	30.7	16.5	23.0	14.9	4.9	29.6	26.5	35.3	7.2	23.4	20.5	32.1	24.4	33.1	17.2	12.2	20.8	28.8	40.6	7.0	22.4
MI-SVM [51]	37.8	17.7	26.7	13.8	4.9	34.4	33.7	46.6	5.4	29.8	14.5	32.8	34.8	41.6	19.9	11.4	25.0	23.6	45.2	8.6	25.4
Model Drift [14]	42.4	46.5	18.2	8.8	2.9	40.9	73.2	44.8	5.4	30.5	19.0	34.0	48.8	65.3	8.2	10.6	16.7	32.3	54.8	5.5	30.4
MIL-Negative [50]	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7.0	29.8	**27.5**	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
Transfer Learning [20]	54.7	22.7	33.7	24.5	4.6	33.9	42.5	57.0	7.3	39.1	24.1	43.3	41.3	51.5	**25.3**	13.3	28.0	29.5	54.6	11.8	32.1
Joint Topic [23]	67.3	54.4	34.3	17.8	1.3	46.6	60.7	**68.9**	2.5	32.4	16.2	**58.9**	51.5	64.6	18.2	3.1	20.9	34.7	**63.4**	5.9	36.2
Convex Cluster.† [25]	66.4	59.3	42.7	20.4	**21.3**	**63.4**	**74.3**	59.6	21.1	58.2	14.0	38.5	49.5	60.0	19.8	**39.2**	41.7	30.1	50.2	44.1	43.7
LCL-pLSA† [24]	80.1	**63.9**	**51.5**	14.9	21.0	55.7	74.2	43.5	**26.2**	53.4	16.3	56.7	58.3	69.5	14.1	38.3	**58.8**	**47.2**	49.1	**60.9**	**48.5**

† indicates methods using auxiliary training data from ILSVRC 2012.

TABLE IV
COMPARISON OF WEAKLY SUPERVISED OBJECT DETECTORS ON PASCAL VOC 2007 IN TERMS OF AP (AVERAGE PRECISION, AS A %) IN THE TEST SET

method / class	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mean**
our S-WDPMs-HOG	26.2	25.0	8.8	9.1	6.5	37.4	40.7	22.9	5.8	19.8	10.6	20.6	27.9	35.1	8.2	6.6	15.3	14.9	27.8	12.2	19.1
our M-WDPMs-HOG	34.5	41.6	10.0	14.1	9.0	39.8	43.9	26.6	5.8	22.8	10.8	24.1	32.2	41.7	10.0	12.3	22.5	14.6	32.9	19.1	23.6
our M-WDPMs-deep	38.3	43.2	18.1	15.9	10.3	40.2	41.9	33.1	6.2	31.4	11.3	27.4	34.3	45.2	12.7	12.5	25.0	14.9	34.3	19.1	25.7
our M-WDPMs-rescore	43.3	43.5	18.6	**16.8**	10.5	**45.2**	42.3	33.8	6.6	**37.2**	12.5	32.7	**36.7**	50.8	14.1	13.8	**28.2**	14.7	**38.0**	20.6	27.7
Model Drift [14]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0	13.9
Multi-fold MIL [15]	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23.0	4.9	14.1	31.9	41.9	**19.3**	11.1	27.6	12.1	31.0	40.6	22.4
Min-Supervision [16]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Pattern Config [52]	36.3	**47.6**	23.3	12.3	11.1	36.0	46.6	25.4	0.7	23.5	12.5	23.5	27.9	40.9	14.8	19.2	24.2	17.1	37.7	11.6	24.6
Posterior Reg. [53]	42.2	43.9	23.1	9.2	**12.5**	44.9	45.1	24.9	8.3	24.0	13.9	18.6	31.6	43.6	7.6	**20.9**	26.6	20.6	35.9	29.6	26.4
Convex Cluster. [25]	46.2	46.9	**24.1**	16.4	12.2	42.2	**47.1**	35.2	7.8	28.3	12.7	21.5	30.1	42.4	7.8	20.0	26.8	20.8	35.8	29.6	27.7
LCL-pLSA [24]	**48.8**	41.0	23.6	12.1	11.1	42.7	40.9	**35.5**	**11.1**	36.6	**18.4**	**35.3**	34.8	**51.3**	17.2	17.4	26.8	**32.8**	35.1	**45.6**	**30.9**
†DPMs 5.0 [7]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
†DP-DPMs conv5 [44]	42.3	65.1	32.2	24.4	36.7	56.8	55.7	38.0	28.2	47.3	37.1	39.2	61.0	56.4	52.2	26.6	47.0	35.0	51.2	56.1	44.4
†R-CNN [3]	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5

(† supervised methods using object level annotations.)

co-localization results of [26], which does not utilize negative images.

We find that the best detection result using *Selective Search* (63.31%) is 3.49% better than *Objectness* (59.82%) within the same S-WDPMs detection model without post-processing, and is 3.22% better (67.13% *vs.* 63.91%) with post-processing, on the *VOC07-6 × 2* dataset. This tallies with the conclusion in [39], where *Selective Search* provides more reliable detection proposals than *Objectness*. Moreover, it achieves comparable or slightly better results than the sophisticated joint topic learning models in [23] when running DPMs refinement only once. As shown in Table I, *SS* also outperforms *obj* on the *VOC07-14* dataset. Consequently, we entirely adopt the *Selective Search* method ("fast" option) for our subsequent experiments.

The localization accuracy on full PASCAL VOC 2007 trainval set and detection precision on test set using S-WDPMs are shown in the first row of Tables III and IV.

B. Experiments With M-WDPMs on PASCAL VOC

1) Dataset and Settings: We evaluate our generalized model: M-WDPMs (multiple region initialized weak DPMs) on the far more challenging dataset: the whole PASCAL VOC 2007 dataset. This contains a total number of 9963 images of 20 object categories, which are split into training (2501), validation (2510) and test (4952) sets. This dataset is challenging because it has large inter-class similarities, intra-class variances, cluttered backgrounds, and scale changes. We only use the image level category labels for this task. Moreover, images labeled as "difficult" are discarded as common practice in previous studies. With respect to M-WDPMs testing, we only run the DPMs once for efficiency, although iterative detector refinement can steadily improve final performance to a certain extent. Annotation accuracy (i.e., correct localization, CorLoc) on the trainval (training + validation) set and average precision (AP) for detection on the test set are reported. For *DeepPyramid* feature extraction, we use NVIDIA GeForce GTX Titan X GPUs, each with a 12 GB memory, thus allowing us to upsample image pyramids to 1713×1713 as in [44] to facilitate detection of small objects.

2) Parameter Selection: As discussed in Section II-A.4b, we can generate Q region estimations for each image. Q is a parameter which impacting the quality of the positive training

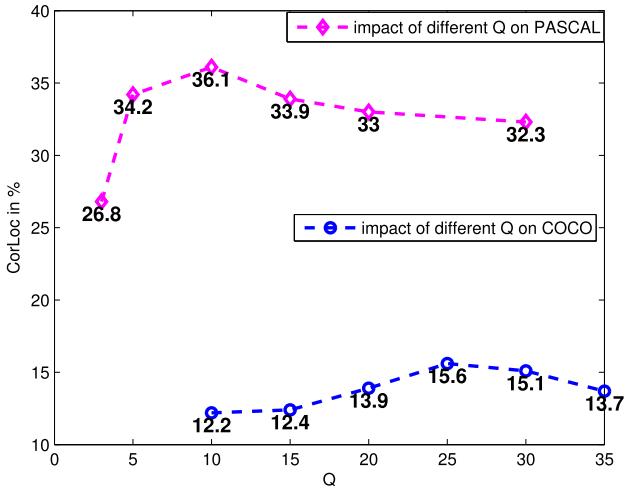


Fig. 7. Impact of parameter Q (number of selected regions for each image in the multiple region initialization scheme). The average annotation accuracy on PASCAL VOC 2007 validation (HOG feature) and MS COCO 2014 val1 (CNN feature) is evaluated with different Q.

Fig. 8. Three example images and their 10 selected regions (resized to the same squared size for regularity).

examples. If it is too large, there would be an enormous number of noisy samples for latent class learning. However, if it is set to be very small, the instances in the original image could not be comprehensively represented. Therefore, we experimentally vary $Q = \{3, 5, 10, 15, 20, 30\}$ to see which one performs best on the PASCAL VOC 2007 validation set. We implement this by directly measuring average annotation accuracy for all classes, on the generated bounding boxes (Q per image) with the PASCAL-overlap criterion. Fig. 7 shows annotation accuracy for different Q. We find that $Q = 10$ obtains the best result (36.1% average accuracy). When it is very small (e.g., 3), performance drops dramatically to 26.8%. This is because some of the "good" region proposals are not selected due to very small Q, while some selected "bad" regions may degenerate the model. When Q rises from 10 to 30, performance deteriorates progressively.. One explanation for this might be that many object parts or background regions would be included when Q is large. Hence, we set $Q = 10$ in all of our experiments on PASCAL VOC. Fig. 8 shows three example images and their 10 selected regions. The κ in (5), which leverages the classification and detection scores, is set to 0.7 according to cross-validation on a subset of the validation data.

3) Annotation Evaluation: We evaluate the same CorLoc [17] as in Section III-A.2 on the PASCAL VOC 2007 trainval set. Table III reports our experimental results compared with the state-of-the-art WSL methods for object detection.

Concerning our M-WDPMs-HOG baseline, which computes the HOG features and does not make use of auxiliary training data from the ILSVRC 2012 classification task [9] as [24], [25], it outperforms most of the previous works [13], [14], [20], [23], [50], [51] (ours: 37.9% *vs.* best from the previous works (Joint topic): 36.2%). The M-WDPMs-HOG outperforms the S-WDPMs-HOG (by 7.7%) by benefiting initialization of DPMs from multiple regions in the image. Our M-WDPMs-HOG shows modest improvement in most of the classes, thus proving that our multiple region initialization method has very discriminative power for selecting the "good" regions in the original image for training the DPMs root filters.

We also observe that, with the help of auxiliary training data and recently popular deep features, the average accuracy of our M-WDPMs-deep model increases by 4.1% in comparison with the M-WDPMs-HOG model. Moreover, our detection rescoring method (i.e., M-WDPMs-rescore) further improves performance for most of the categories. The average improvement for detection rescoring on all 20 classes is 2.1% (44.1% *vs.* 42.0%). Our M-WDPMs-rescore method is slightly better than the newly invented convex clustering approach [25], but is worse than the LCL-pLSA method [24] on average. Though [24] achieves state-of-the-art performance on many classes, it depends on more sophisticated Super-Vector (SV) coding [54] of the deep CNN features, thus unfortunately increasing feature dimensionality (e.g., 10,000 visual words). It also fails in some categories such as *boat* and *table*. However, our M-WDPMs-rescore exhibits a steady agreeable performance in all the categories with acceptable feature dimension (256 dimensional *conv 5* features for detection and 4,096 dimensional *fc6* features for classification). In particular, our M-WDPMs-rescore works well in categories where target objects are relatively salient, such as *aeroplane, boat, cow, horse* and *motorbike*. Among these categories, *cow, horse* and *motorbike* have deformable shapes, thus ensuring good detection for the DPM.

4) Detection Evaluation: Table IV shows the comparison of our M-WDPMs and other methods for object detection on the PASCAL VOC 2007 test set. Our M-WDPMs-HOG baseline method achieves an mAP of 23.6%, which outperforms [14] (13.9%) by a large margin, and is slightly better than [15] (22.4%). Both [14] and [15] represent the image windows with a SIFT [57] descriptor. [14] uses a Bag-of-Words (BOW) [58] histogram with 2000 dimensions, while [15] use Fisher Vectors (FV) encoding [59] to represent the candidate windows. [22]

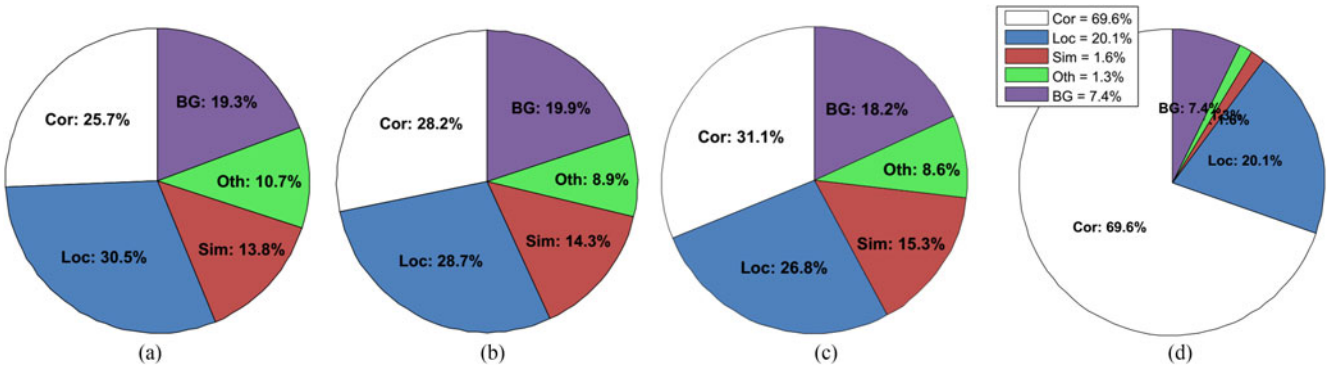


Fig. 9. Analysis of top-ranked detections on PASCAL VOC 2007 test set. Pie charts show the distributions of the true positives (TP) and false positives (FP) generated by the detection error analysis tool of [55]. Percentage of the top T detections (T is the number of ground truth objects in the whole test dataset) that are correct (Cor), or false positives due to poor localization (Loc), confusion with similar objects (Sim), confusion with other objects (Oth), or confusion with background or unlabeled objects (BG) [55]. The three charts on the left show the analysis of our methods, while the one on the right is the analysis of the state-of-the-art supervised detection results obtained by NoC [56].(a) M-WSDPM-HOG mAP = 23.6%; (b) M-WSDPM-deep mAP = 25.7%; (c) M-WSDPM-rescore mAP = 27.7%; (d) state-of-the-art supervised: NoC mAP = 68.8%.

uses the same HOG pyramid features. M-WDPMs shows consistently better performance than S-WDPMs (19.1%), except for the *sofa* category, where S-WDPMs shows trivial superiority. Among these methods that adopt low level visual features, our M-WDPMs-HOG works best. Although [16] utilizes powerful deep CNN features to represent the discovered object windows, its performance (22.7%) is more or less the same with our HOG based M-WDPMs, which proves the stronger discrimination of our window selection method. When using the deep features with additional training data from ImageNet [9], our M-WDPMs-deep can achieve an mAP of 25.7%. The boost (2.1%) is not as much as that of the annotation task (4.1%, see Section III-B.3), it is probably due to the use of distinct measuring criteria (mean average precision *v.s* percent of correct localization). Our detection rescoring method M-WDPMs-rescore continues to improve the average precision (mAP = 27.7%) for nearly all classes except for the *sofa* class (0.2% decrease). Its performance is better when compared with [52], [53], and it displays the same range of performance in comparison with [25]. The performance gap (3.2%) between our method and that of [24] might be partly caused by the use different deep feature representations as discussed in Section III-B.3. We conjecture that our detection performance could be further boosted if a complex feature encoding method such as SV [54] was adopted as [24]. We achieve the best detection results for the *boat, bus, cow, horse, sheep* and *train* classes for this dataset. We attribute the success on these categories to object saliency (e.g., *boat, bus*), deformable structures (e.g., *cow, horse, sheep*), and possibly their combination (e.g., *train*) which united by our framework. Image saliency and object structures provide good representations for these kinds of object categories. Hence, the combination of the two ensures good detection results on these categories. And our M-WDPMs yields moderately low average precision on categories such as *bird, bottle, chair* and *potted plant*. These categories typically have notably small and/or textured instances, where object proposal method such as Selective Search can be misleading, and they are hard to detect even by supervised DPMs [7], [44].

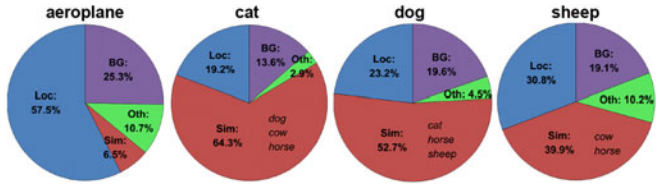


Fig. 10. Analysis of false positives for some classes on which our M-WDPMs outperforms DPMs 5.0 [7]. Each category named within "Sim" shows the category names on which detector tends towards confusion.

In addition, we provide the results obtained by popular supervised object detection methods [3], [7], [44] in the bottom lines of Table IV. It is clear that there is still a gap between the weakly supervised detection framework and supervised frameworks, although our weakly supervised DPMs yields better results for some classes (e.g., *aeroplane, cat, dog, sheep etc.*) than the supervised DPMs 5.0 [7] which uses the low level HOG feature. The state-of-the-art supervised object detection framework (i.e., Faster R-CNN [60]) achieves 78.9% mAP by adopting very deep neural networks (VGG-16 [61]).

5) Error Analysis: We present an analysis of the types of errors that our M-WDPMs make on the PASCAL VOC 2007 test set. We use the diagnosis tool of [55] and consider four types of false positive (FP) errors: *Loc* (poor localizations), *Sim* (confusion with similar objects), *Oth* (confusion with other objects, e.g., correctly localizing an object but classifying it to a wrong class) and *BG* (confusion with background or unlabeled objects). In addition, *Cor* indicates correctly located true positives (TP).

We visually show the fraction of correct detections (TP) and errors of each kind (FP) among the top ranking T windows in Fig. 9, where T is the number of ground-truth object windows in the PASCAL VOC 2007 test set.

We consider the M-WDPMs-HOG as our baseline and show the distribution of TP and each kind of FP in Fig. 9(a). We can see that the majority of errors are due to poor localizations (*Loc*) and confusion with background regions (*BG*). When adopting the deep features, our M-WDPMs-deep encounters fewer *Loc*

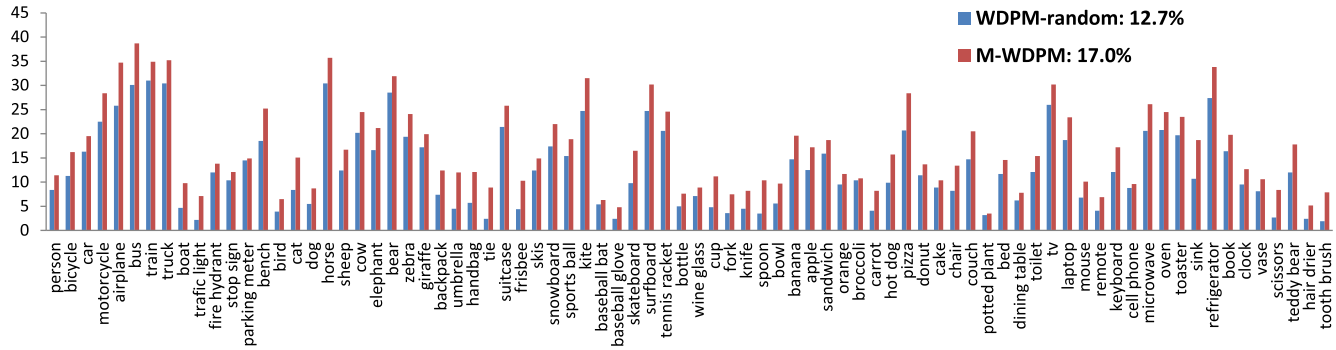


Fig. 11. Detection results of weakly supervised DPMs detectors on MS COCO 2014 val2 in terms of average precision (AP, as a %). For both methods, deep *con5* features are adopted.

and *Oth*, but continues to suffer from the *Sim* and *BG* error [as shown in Fig. 9(b)]. On the contrary, after detection rescoring, our best performing method M-WDPMs-rescore has fewer errors caused by *Loc*, *BG* and *Oth* [Fig. 9(c)], thus confirming that our rescoring approach is very efficient in excluding the background regions and avoiding misclassification. Fig. 9(d) shows the error distribution of the state-of-the-art supervised object detection framework NoC (Networks on Convolutional feature maps) [56]. NoC adopt even deeper VGG-16 nets [61] with bounding box fine-tuning on PASCAL VOC 2007 + 2012 trainval. A comparison between NoC and our M-WDPMs indicates that: 1) a deeper network helps increase correct localization (*Cor*) substantially; 2) fine-tuning deep CNN and supervised training with ground-truth bounding boxes yield far fewer *Sim* and *Oth* errors.

We also display some class specific false positive analysis in Fig. 10, on the classes where our M-WDPMs outperforms DPMs 5.0 [7].

6) Running Time: The time it takes to extract the Selective Search region proposals (can be shared among different detector learning) is 10.27 s. Reference region computation takes 778 ms, generation of initial object estimations from region proposals and reference regions takes 190 ms, while computation of CNN features is 18.97 s and *conv5* feature pyramid from CNN feature is 631 ms. Running time is averaged on 100 random PASCAL images, and is evaluated on an Intel Core i7-5960X CPU @ 3.00GHz with 32GB memory and a single NVIDIA Titan X GPU. For M-WDPMs, training binary SVM and learning latent class takes 228.20 s on the *horse* class and 196.05 s on the *motorbike* class (except for CNN pre-training and feature extraction time). Besides, training of the *horse* DPMs detector takes 84.82 min and 76.45 min for *motorbike*. Running a detector costs 9.76 s per image (including rescoring time) on average on the PASCAL dataset.

C. Preliminary Results With M-WDPMs on MS COCO

The Microsoft Common Objects in Context (MS COCO) dataset [32] involves 80 object categories. It contains considerably more object instances per image (7.7) compared to PASCAL VOC (2.5), and has 82,783 training images and 40,504 validating images in the 2014 release (COCO 2014). We split the validation set equally into val1 and val2, where val1 is used as a validation set and val2 is used as a test set. In spite of this, this subset of MS COCO is much larger and more difficult than PASCAL VOC. We evaluate the PASCAL VOC metric (mAP @IoU $= 0.5$) on val2.

We set the parameter Q to 25 regions, since there are significantly more object instances per image on MS COCO than on PASCAL VOC. The influence of Q on MS COCO is shown in Fig. 7, while the rescoring weight κ is set to 0.8 by choosing from [0.5, 1.0] on val1. The increase of κ on MS COCO probably means that there is a larger number of smaller objects in this dataset and that the detector has more influence than the classifier on the final detection score. The other training and testing settings of M-WDPMs remain as the same as on PASCAL VOC. We compare our method with the WDPM-random baseline method [22], which sets a large random window as initialization. For both of these two methods, we adopt deep *con5* feature pyramids.

Fig. 11 shows the detection results of our M-WDPMs and the WDPMs-random baseline. Overall, our M-WDPMs results in 17.0% mAP on this MS COCO val2 set, boosting the mAP by 4.3 points over the WDPMs-random. The results on 20 common categories in MS COCO are significantly lower than on PASCAL. This is because there are far more small objects on COCO, making it a fairly challenging dataset for detection. We observed that our M-WDPMs exhibits a relatively good performance on similar categories both in COCO and in PASCAL, such as *aeroplane, bus, horse, motorbike* and *train*, and has favorable performances on *truck, bear* and *oven*, etc. classes in COCO. This confirms that our M-WDPMs is capable of detecting object categories that are salient visually and/or deformable structurally.

IV. Conclusion

In this paper, we proposed a model enhancing weakly supervised learning by emphasizing the importance of location and size of the initial class specific root filter of deformable part-based models. We follow the general setup of [22] and introduce several substantial improvements to the weakly supervised deformable part-based model (DPMs). The main contributions included a new selection model based on generic

"objectness" (region proposals) and visual saliency to adaptively select a reliable set of candidate windows which tend to represent the object instances in the image, and a latent class learning process by coarsely classifying a candidate window into either a target object or a non-target class. Furthermore, we designed a flexible enlarging-and-shrinking post-processing procedure to modify the output bounding boxes of DPMs, which can effectively further improve the final accuracy. Experimental results on the PASCAL VOC 2007 database according to various criteria demonstrate that our proposed framework is efficient and competitive with the state-of-the-art, especially for the object categories which are relatively salient and deformable. We also report some preliminary weakly supervised detection results on the very challenging MS COCO 2014 dataset. Future work includes extracting more knowledge from different domains (e.g., both visual and semantic domains), using better representations, and investigating the possibility of using category-invariant properties, e.g., the difference between feature distributions of whole images and target objects, to further improve weakly supervised object detection.

ACKNOWLEDGMENT

The authors would like to thank NVIDIA for providing the Titan X GPU.

REFERENCES

[1] W. Zhang, Q. M. J. Wu, G. Wang, and H. Yin, "An adaptive computational model for salient object detection," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 300–316, Jun. 2010.

[2] Y. Zhu, J. Zhu, and R. Zhang, "Contextual object detection with spatial context prototypes," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1585–1596, Oct. 2014.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 580–587.

[4] G. L. Foresti, L. Marcenaro, and C. S. Regazzoni, "Automatic detection and indexing of video-event shots for surveillance applications," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 459–471, Dec. 2002.

[5] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 761–774, Aug. 2006.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[7] 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.

[8] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2013, pp. 2553–2561.

[9] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

[10] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 341–349.

[11] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2003, pp. 577–584.

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

[13] M. 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.

[14] 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.

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

[16] H. O. Song *et al.*, "On learning to localize objects with minimal supervision," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1611–1619.

[17] 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.

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

[19] 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.

[20] Z. Shi, P. Siva, and T. Xiang, "Transfer learning by ranking for weakly supervised object annotation," in *Proc. Brit. Mach. Vis., Conf.*, 2012, pp. 78.1–78.11.

[21] J. Hoffman *et al.*, "LSDA: Large scale detection through adaptation," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2014, pp. 3536–3544.

[22] 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.

[23] 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.

[24] C. Wang, K. Huang, W. Ren, J. Zhang, and S. Maybank, "Large-scale weakly supervised object localization via latent category learning," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1371–1385, Apr. 2015.

[25] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1081–1089.

[26] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1464–1471.

[27] R. Girshick, P. Felzenszwalb, and D. McAllester, "Object detection with grammar models," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst. Conf.*, 2011, pp. 442–450.

[28] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 836–849.

[29] X. Ren and D. Ramanan, "Histograms of sparse codes for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3246–3253.

[30] 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.

[31] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, pp. 154–171, 2013.

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

[33] Y. Tang, X. Wang, E. Dellandrea, S. Masnou, and L. Chen, "Fusing generic objectness and deformable part-based models for weakly supervised object detection," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 4072–4076.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2012, pp. 1106–1114.

[35] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 3286–3293.

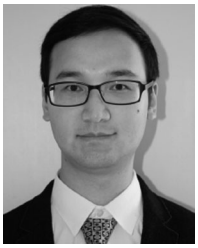
[36] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

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

[38] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 328–335.

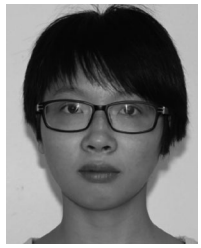
[39] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" in *Proc. Brit. Mach. Vis. Conf.*, 2014. [Online]. Available: http://dx.doi.org/10.5244/C.28.24

[40] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1896–1909, Dec. 2013.
[41] N. OTSU, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[42] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *CoRR*, 2014. [Online]: Available: http://arxiv.org/abs/1408.5093
[43] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, May 2011, Art. no. 27.
[44] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 437–446.
[45] P. Sermanet *et al.*, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2014.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[47] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1585–1592.
[48] W. Ouyang *et al.*, "DeepID-Net: Deformable deep convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 2403–2412.
[49] X. T. Y. Ke and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 1, pp. 419–426.
[50] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 594–608.
[51] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2003, pp. 577–584.
[52] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, "Weakly-supervised discovery of visual pattern configurations," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2014, pp. 1637–1645.
[53] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with posterior regularization," in *Proc. Brit. Mach. Vis. Conf.*, 2014. [Online]. Available: http://dx.doi.org/10.5244/C.28.52
[54] 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.
[55] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 340–353.
[56] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[57] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 1999, vol. 2, pp. 1150–1157.
[58] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vision/Eur. Conf. Comput. Vis.*, 2004, pp. 1–22.
[59] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
[60] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2015, pp. 91–99.
[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.



Yuxing Tang received the B.S. and M.S. degrees from the Department of Information and Telecommunication Engineering, Beijing Jiaotong University, Beijing, China, in 2009 and 2011, respectively, and is currently working toward the Ph.D. degree with the Department of Mathematics and Computer Science, École Centrale de Lyon, Écully, France.

His research interests include computer vision and machine learning, in particular models for visual category recognition and object detection.



Xiaofang Wang received the B.S. and M.S. degrees in biomedical engineering from Central South University, Changsha, China, and the Ph.D. degree in computer science from the École Centrale de Lyon, Écully, France in 2015.

She is currently a Postdoctoral Researcher with the Department of Mathematics and Computer Science, École Centrale de Lyon. Her current research interests include image/video processing, medical image segmentation and analysis, multiple object tracking, and semantic segmentation.

Emmanuel Dellandréa received the M.S. and Engineering degrees in computer science from the University of Tours, Tours, France, in 2000, and the Ph.D. degree in computer science from the University of Tours in 2003.

He then joined the École Centrale de Lyon, Écully, France, in 2004, as an Associate Professor. His research interests include multimedia analysis, image and audio understanding, and affective computing, including recognition of affect from image, audio, and video signals.

Liming Chen (A'04–M'06–SM'14) received the joint B.Sc. degree in mathematics and computer science from the University of Nantes, Nantes, France, in 1984, and the M.Sc. and Ph.D. degrees in computer science from the University of Paris 6, Paris, France, in 1986 and 1989, respectively.

He first served as an Associate Professor with the Université de Technologie de Compiègne, before joining the École Centrale de Lyon, Écully, France, as a Professor in 1998, where he leads an advanced research team on multimedia computing and pattern recognition. From 2001 to 2003, he also served as Chief Scientific Officer in a Paris-based company, Avivias, specializing in media asset management. In 2005, he served as Scientific Multimedia Expert for France Telecom R&D China, Beijing, China. He has been the Head of the Department of Mathematics and Computer Science, École Centrale de Lyon, since 2007. He has taken out three patents, authored more than 100 publications, and acted as chairman, PC member, and reviewer in a number of high-profile journals and conferences since 1995. He is a (co)-principal investigator on a number of research grants from EU FP programs, French research funding bodies, and local government departments. He has directed more than 30 Ph.D. dissertations. His current research ranges from 2D/3D face analysis and recognition, image and video analysis and categorization, to affect analysis in image, audio, and video signals.