Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning

Ramazan Gokberk Cinbis, Jakob Verbeek, Cordelia Schmid

▶ **To cite this version:**

Ramazan Gokberk Cinbis, Jakob Verbeek, Cordelia Schmid. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning. 2015. hal-01123482v1

HAL Id: hal-01123482

https://inria.hal.science/hal-01123482v1

Preprint submitted on 4 Mar 2015 (v1), last revised 22 Feb 2016 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning

Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, *Fellow, IEEE*

Abstract—Object category localization is a challenging problem in computer vision. Standard supervised training requires bounding box annotations of object instances. This time-consuming annotation process is sidestepped in weakly supervised learning. In this case, the supervised information is restricted to binary labels that indicate the absence/presence of object instances in the image, without their locations. We follow a multiple-instance learning approach that iteratively trains the detector and infers the object locations in the positive training images. Our main contribution is a multi-fold multiple instance learning procedure, which prevents training from prematurely locking onto erroneous object locations. This procedure is particularly important when using high-dimensional representations, such as Fisher vectors and convolutional neural network features. We also propose a window refinement method, which improves the localization accuracy by incorporating an objectness prior. We present a detailed experimental evaluation using the PASCAL VOC 2007 dataset, which verifies the effectiveness of our approach.

1 Introduction

Over the last decade significant progress has been made in object category localization, as witnessed by the PASCAL VOC challenges [18]. Training state-of-the-art object detectors, however, requires bounding box annotations of object instances, which are costly to acquire.

Weakly supervised learning (WSL) refers to methods that rely on training data with incomplete ground-truth information to learn recognition models. For object detection, WSL from image-wide labels that indicate the presence of instances of a category in images has recently been intensively studied as a way to remove the need for bounding box annotations, see *e.g.* [4], [6], [9], [12], [15], [32], [34], [35], [37], [40], [42], [43], [44], [50]. Such methods can potentially leverage the large amount of tagged images on the internet as a data source to train object detectors. We give an overview of the most relevant related work in Section 2.

Other examples of WSL include learning face recognition models from image captions [5], or subtitle and script information [17]. Another WSL example is learning semantic segmentation models from image-wide category labels [48]. Most WSL approaches are based on latent variable models to account for the missing ground-truth information. Multiple instance learning (MIL) [16] handles cases where the weak supervision indicates that at least one positive instance is present in a set of examples. More advanced inference and learning methods are used in cases where the latent variable structure is more complex, see *e.g.* [15], [37], [48]. Besides weakly supervised training, mixed fully and weakly supervised [7], active [49], and semi-supervised [37] learning methods have also been explored to reduce the amount of labeled training data for object detector training. In active learning bounding box annotations are used, but requested only for images where the annotation is expected to be most effective. Semi-supervised learning, on the other hand, leverages unlabeled images by automatically detecting objects in them, and use those to better model the object appearance variations.

• *LEAR team, Inria Grenoble Rhône-Alpes, Laboratoire Jean Kuntzmann, CNRS, Univ. Grenoble Alpes, France. E-mail: firstname.lastname@inria.fr*

In this paper we consider WSL to learn object detectors from image-wide labels. We follow an MIL approach that interleaves training of the detector with re-localization of object instances on the positive training images. Following recent state-of-the-art work in fully supervised detection [10] [20] [47], we represent (tentative) detection windows using Fisher vectors (FVs) [36] and convolutional neural network (CNN) features [27]. As we explain in Section 3, when used in an MIL framework, the high-dimensionality of the window features makes MIL quickly convergence to poor local optima after initialization. Our main contribution is a multi-fold training procedure for MIL, which avoids this rapid convergence to poor local optima. A second novelty of our approach is the use of a "contrastive" background descriptor that is defined as the difference of a descriptor of the object window and a descriptor of the remaining image area. The score for this descriptor of a linear classifier can be interpreted as the difference of scores for the foreground and background. In this manner we force the detector to learn the difference between foreground and background appearances. Finally, we propose a *window refinement* method that improves the weakly supervised localization accuracy by incorporating a category-independent objectness measure.

We present a detailed evaluation using PASCAL VOC 2007 dataset in Section 4. The experimental results show that our multi-fold MIL training improves performance for both the FV and the CNN features. We also show that WSL performance can be further improved by combining the two descriptor types and applying the window refinement method. The evaluation shows that our system obtains state-of-the-art results on VOC 2007, we also present results for VOC 2010 which was not used in previous work.

Part of the material presented here has appeared in [11]. Besides a more detailed presentation and discussion of the most recent related work, the current paper extends it in several ways. We have enhanced our WSL method by introducing a window refinement method. We have also added additional experiments using CNN features, and their combination with FV features. Finally, we included experiments when training in a mixed supervision setting, where part of the images are weakly supervised and others are labeled with full bounding-box annotations.

2 Related work

The majority of related work treats WSL for object detection as a multiple instance learning (MIL) [16] problem. Each image is considered as a "bag" of examples given by tentative object windows. Positive images are assumed to contain at least one positive object instance window, while negative images only contain negative windows. The object detector is then obtained by alternating detector training, and using the detector to select the most likely object instances in positive images.

In many MIL problems, *e.g.* such as those for weakly supervised face recognition [5], [17], the number of examples per bag is limited to a few dozen at most. In contrast, there is a vast number of examples per bag in the case of object detector training since the number of possible object bounding boxes is quadratic in the number of image pixels. Candidate window generation methods, *e.g.* [1], [22], [46], [53], can be used to make MIL approaches to WSL for object localization manageable, and make it possible to use powerful and computationally expensive object models.

Although candidate window generation methods can significantly reduce the search space per image, the selection of windows across a large number of images is inherently a challenging problem, where an iterative WSL method can typically find only a local optimum depending on the initial windows. Therefore, in the remainder of this section, we first overview the initialization methods proposed in the literature, and then summarize the iterative WSL approaches.

2.1 Initialization methods

A number of different strategies to initialize the MIL detector training have been proposed in the literature. A simple strategy, *e.g.* taken in [26], [32], [35], is to initialize by taking large windows in positive images that (nearly) cover the entire image. This strategy exploits the inclusion structure of the MIL problem for object detection. That is: although large windows may contain a significant amount of background features, they are likely to include positive object instances.

Another strategy is to utilize a class-independent saliency measure that aims to predict whether a given image region belongs to an object or not. For example, Deselaers *et al.* [15] generate candidate windows using the objectness method [2] and assign per-window weights using a saliency model trained on a small training set of non-target object classes. Siva *et al.* [41] instead estimate an unsupervised patch-level saliency map for a given image by measuring the average similarity of each patch to the other patches in a retrieved set of similar images. In each image, an initial window is found by sampling from the corresponding saliency map.

Alternatively, a class-specific initialization method can be used. For example, Chum and Zisserman [9] select the visual words that predominantly appear in the positive training images and initialize WSL by finding the bounding box of these visual words in each image. Siva and Xiang [42] propose to initially select one of the candidate windows sampled using the objectness method at each image such that an objective function based on intra-class and inter-class pairwise similarities is maximized. However, this formulation leads to a difficult combinatorial optimization problem. Siva *et al.* [40] propose a simplified approach where a candidate window is selected for a given image such that the distance from the selected window to its nearest neighbor among windows from negative images is maximal. Relying only negative windows not only avoids the difficult combinatorial optimization problem, but also has the advantage that their labels are certain, as opposed to the tentative object hypotheses, and there is a larger number of negative windows available which makes the pairwise comparisons more robust.

Shi *et al.* [37] propose to estimate a per-patch class distribution by using an extended version of the Latent Dirichlet Allocation (LDA) [8] topic model. Their approach assigns object class labels across different object categories concurrently, which allows to benefit from explaining-away effects, *i.e.* an image region cannot be identified as an instance for multiple categories. The initial windows are then localized by sampling from the saliency maps.

Song *et al.* [43] propose a graph-based initialization method. The main idea is to select a subset of the candidate windows such that the nearest neighbors of the selected windows correspond predominantly to the candidate windows in the positive images, rather than the ones in the negative images. The approach is formulated as a discriminative submodular cover problem on the similarity graph of the candidate windows. In a recent work, Song *et al.* [44] extend this approach to find multiple non-overlapping regions corresponding to object parts. The initial object windows are then generated by finding frequent part configurations and their bounding boxes.

2.2 Iterative learning methods

Once the initial windows are localized, typically an iterative learning approach is employed in order to improve the initial localizations in the training images and obtain more accurate object detectors.

One of the early examples of WSL for object detector training is proposed by Crandall and Huttenlocher [12]. In their work, object and part locations are treated as latent variables in a probabilistic model. These variables are automatically inferred and utilized during training using an Expectation Maximization (EM) algorithm. The main focus of their work, however, is on training a part-based object detector without using manual part annotations, rather than training in terms of image labels. Their approach is evaluated on datasets containing images with uncluttered backgrounds and little variance in terms of object locations, which is an unrealistic testbed for WSL of object detectors.

Several WSL methods aim to localize objects via selecting a subset of candidate windows based on pairwise similarities. For example, Kim and Torralba [26] use a link analysis based clustering approach. Chum and Zisserman [9] iteratively select windows and update the similarity measure that is used to compare windows. The window selection is done by updating one image at a time such that the average pairwise similarity across the positive images is maximized. The similarity measure, which is defined in terms of bag-of-word (BoW) descriptors [13], is updated by selecting the visual words that predominantly appear in the selected windows rather than the negative images.

Deselaers *et al.* [15] propose a CRF-based model that jointly infers the object hypotheses across all positive training images, by exploiting a fully-connected graphical model that encourages visual similarity across all selected object hypotheses. Unlike the methods of [26] and [9], the CRF-based model additionally utilize a unary potential function that scores candidate windows individually. The parameters of the pairwise and unary potential functions are updated and the positive windows are selected in an iterative fashion. Prest *et al.* [34] extend these ideas to weakly

supervised detector training from videos by extracting candidate spatio-temporal tubes based on motion cues and by defining WSL potential functions over tubes instead of windows.

Most recent work is predominantly based on iteratively selecting the highest scoring detections as the positive training examples and training the detection models. We refer to this approach as *standard MIL*. Using this approach, an off-the-shelf detector can be trained in a weakly supervised setting. For example, Nguyen *et al*. [31] and Blaschko *et al*. [7] train the branch-and-bound localization [28] based detectors over BoW descriptors in this manner. Blaschko *et al*. also investigate the use of object-center annotations as an alternative WSL setting.

The DPM model [19] has been utilized with standard MIL based training approaches by a number of other WSL approaches, see *e.g*. [32], [37], [40], [41], [42]. The majority of the works use the standard DPM training procedure and differ in terms of their initialization procedures. One exception is that Siva and Xiang [42] propose a method to detect when the iterative training procedure drifts to background regions. In addition, Pandey and Lazebnik [32] carefully study how to tune DPM training procedure details for WSL purposes. They propose to restrict each re-localization stage such that the bounding boxes between two iterations must meet a minimum overlap threshold, which avoids big fluctuations across the iterations. Moreover, they propose a heuristic to automatically crop windows with near-uniform backgrounds, where the iterative procedure may get stuck at too large object hypotheses.

Russakovsky *et al*. [35] use a similar approach based on Locality-constrained Linear Coding descriptors [51] over the candidate windows generated using the Selective Search method [46]. They allow progressively smaller windows in subsequent iterations, which avoids the method to get stuck at poor local optima. In addition, they use a background descriptor computed over features outside the window, which helps to better localize the objects as compared to only modeling the windows themselves.

Song *et al*. [43] develop a smoothed version of the standard MIL approach using Nesterov's smoothing technique [30]. The main motivation is to increase robustness against incorrectly selected windows, particularly in early iterations, by training with multiple windows per positive image. The candidate windows are generated using selective search [46] and the window descriptors are extracted using the CNN model of Krizhevsky *et al*. [27], which is pre-trained on auxiliary training images from the ImageNet dataset [14].

Bilen *et al*. [6] propose an alternative smoothed version of the standard MIL approach. Instead of selecting the top scoring window in a positive image, they propose to train over all windows that are weighted by a soft-max function over the classification scores. In addition, they utilize additional regularization terms that aim to (i) enforce that positive training windows and their horizontally mirrored images score similarly and, (ii) avoid obtaining high classification scores for multiple classes for a single window. They also utilize selective search candidate windows [46] and CNN features [27].

Recently, Wang *et al*. [50] propose a two-step method, which first groups selective search candidate windows [46] from the positive images of a class into visual clusters and then chooses the most discriminative cluster of windows. In the first step, the CNN features [27] are clustered using probabilistic latent semantic analysis (PLSA) [23]. In the second step, for each visual cluster, image descriptors are extracted from the CNN-based window descriptors of the windows associated with the cluster. Finally, one visual cluster for each class is selected based on the image classification performance of the corresponding image descriptors.

Our approach is most related to that of Russakovsky *et al*. [35]: we also rely on the selective search windows [46], and use a similar initialization strategy. A critical difference from [35] and other WSL approaches based on iterative detector training, however, is our multi-fold MIL training procedure which we describe in the next section. Our multi-fold MIL approach is also related to the work of Singh *et al*. [39] on unsupervised vocabulary learning for image classification. Starting from an unsupervised clustering of local patches, they iteratively train SVM classifiers on a subset of the data, and evaluate it on another set to update the training data from the second set.

3 Weakly Supervised Object Localization

Below, we present our multi-fold MIL approach in Section 3.2 and window refinement method in Section 3.3, but first briefly describe our FV and CNN based object appearance descriptors in in Section 3.1.

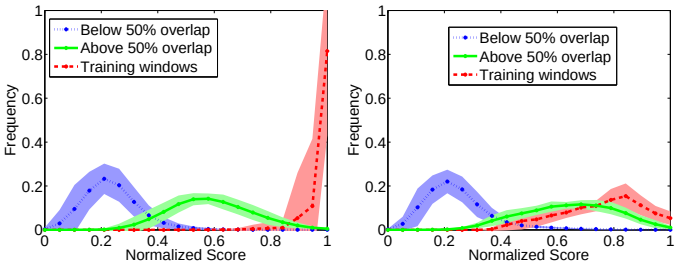
3.1 Features and detection window representation

In our experiments we rely on FV and CNN based representations. In either case, we use the selective search method of Uijlings *et al*. [46]. It generates a limited set of around 1,500 candidate windows per image. This speeds-up detector training and evaluation, while filtering out the most implausible object locations.

The FV-based representation is based on our previous work [10], which yields state-of-the-art performance for fully supervised detection. In particular, we aggregate local SIFT descriptors into an FV representation to which we apply ℓ_2 and power normalization [36]. We concatenate the FV computed over the full detection window, and 16 FVs computed over the cells in a 4×4 grid over the window, inspired by the spatial pyramid representation of Lazebnik *et al*. [29]. Using PCA to project the SIFTs to 64 dimensions, and mixture of Gaussian models (MoG) of 64 components, this yields a descriptor of 140,352 dimensions. We reduce the memory footprint, and speed up our iterative training procedure, by using the PQ and Blosc feature compression [3], [24].

Similar to Russakovsky *et al*. [35], we add contextual information from the part of the image not covered by the window. Full-image descriptors, or image classification scores, are commonly used for fully supervised object detection, see *e.g*. [10], [45]. For WSL, however, it is important to use the complement of the object window rather than the full image, to ensure that the context descriptor also depends on the window location. This prevents learning degenerate detection models, since otherwise the context descriptor can be used to perfectly separate the training images regardless of the object localization.

To enhance the effectiveness of the context descriptor we propose a "contrastive" version, defined as the difference between the background FV $\boldsymbol{x}_b$ and the 1×1 foreground FV $\boldsymbol{x}_f$. Since we use linear classifiers, the contribution to the window score of this descriptor, given by $\boldsymbol{w}^\top(\boldsymbol{x}_b - \boldsymbol{x}_f)$, can be decomposed as a sum of a foreground and a background score: $\boldsymbol{w}^\top \boldsymbol{x}_b$ and $-\boldsymbol{w}^\top \boldsymbol{x}_f$ respectively. Because the foreground and background descriptor have the same weight vector, up to a sign flip, we effectively force features to either score positively on the foreground and negatively on the background, or *vice-versa*. This prevents the detector to



(a) Fisher vectors (b) CNN features

Fig. 1. Distribution of the window scores in the positive training images after the fifth iteration of standard MIL training on VOC 2007 for FVs (left) and CNNs (right). For each figure, the right-most curve corresponds to the windows chosen in the most recent re-localization step and used for training the detector. The curve in the middle corresponds to the other windows that overlap more than 50% with the training windows. Similarly, the left-most curve corresponds to the windows that overlap less than 50%. Each curve is obtained by averaging all per-class score distributions. The surrounding regions show the standard deviation.

score the same features positively on both the foreground and the background, and to localize objects more accurately.

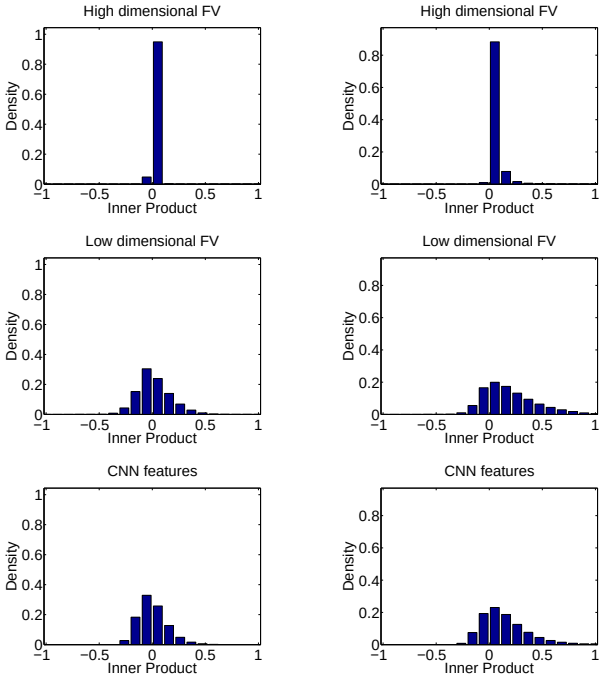
To ensure that we have enough SIFT descriptors for the background FV, we filter the detection windows to respect a margin of at least 4% from the image border, *i.e.* for a 100×100 pixel image, windows closer than 4 pixels to the image border are suppressed. This filtering step removes about half of the windows. We initialize the MIL training with the window that covers the image, up to a 4% margin, so that all instances are captured by the initial windows.

We extract the CNN features using the CNN architecture of Krizhevsky *et al*. [27]. We utilize the first seven layers of the CNN model, which consists of five convolutional and two fully-connected layers. The CNN model is pre-trained on the ImageNet ILSVRC 2012 dataset using the Caffe framework [25]. Following Girshick *et al*. [20], we crop and resize the mean-subtracted regions corresponding to the candidate windows to images of size 224×224, as required by the CNN model. Finally, we apply ℓ_2 normalization to the resulting 4096 dimensional descriptors.

An important advantage of the CNN features is that some of the feature dimensions correspond to higher level image structures, such as certain animal faces and bodies [20], which can simplify the WSL problem. On the other hand, the high-dimensional FV descriptors can possibly convey complementary information. In fact, our experimental results show that whereas the CNN features perform better than the FV features, their combination outperforms the CNN features alone.

3.2 Weakly supervised object detector training

The dominant method for weakly supervised training of object detectors is the standard MIL approach, which is based on iterating between the training and the re-localization stages, as described in Section 2.2. Note that in this approach, the detector used for re-localization in positive images is trained using positive samples that are extracted from the very same images. Therefore, there is a bias towards re-localizing on the same windows; in particular when high capacity classifiers are used which are likely to separate the detector's training data. For example, when a nearest neighbor classifier is used the re-localization will be degenerate and not move away from its initialization, since the same window will be found as its nearest neighbor.



(a) All pairs. (b) Within-image pairs.

Fig. 2. Distribution of inner products, scaled to the interval [-1 +1], of pairs of 25,000 windows sampled from 250 images using our high-dimensional FV (top), a low-dimensional FV (middle), and CNN features (bottom). (a) uses all window pairs and (b) uses only within-image pairs, which are more likely to be similar.

The same phenomenon occurs when using powerful and high-dimensional image representations to train linear classifiers. We illustrate this in Figure 1, which shows the distribution of the window scores in a typical standard MIL iteration. We observe that the windows used in SVM training score significantly higher than the other ones, including those with a significant spatial overlap with the most recent training windows, especially when the high-dimensional FV descriptors are used.

As a result, standard MIL typically results in degenerate re-localization. This problem is related to the dimensionality of the window descriptors. We illustrate this in Figure 2, where we show the distribution of inner products between the descriptors of different windows. In Figure 2a, we use random window pairs within and across images. In Figure 2b, we use only within-image pairs, which are more likely to be similar, and therefore the histograms models are shifted slightly to larger values. We show the distributions using both our 140,352 dimensional FVs, 516 dimensional FVs obtained using 4 Gaussians without spatial grid, and 4096 dimensional CNN-based descriptors.[1] Unlike in the case of low-dimensional FVs or CNN-based descriptors, almost all window descriptors are near orthogonal in the high-dimensional FV case even when we use within-image pairs only. Also, recall that the weight vector of a standard linear SVM classifier can be written as a linear combination of training samples, $\boldsymbol{w} = \sum_i \alpha_i \boldsymbol{x}_i$. Therefore, the training windows are likely to score significantly higher than the other windows in positive images in the high-dimensional case, resulting in degenerate re-localization behavior. In Section 4, we verify this hypothesis experimentally

1. To make the histograms comparable, we make all descriptors zero mean, before ℓ_2 normalization, and computing the inner products.

Algorithm 1 — Multi-fold weakly supervised training

1) Initialization: positive and negative examples are set to entire images up to a 4% border.
2) For iteration $t = 1$ to T
   a) Divide positive images randomly into K folds.
   b) For $k = 1$ to K
      i) Train using positive examples in all folds but k, and all negative examples.
      ii) Re-localize positives by selecting the top scoring window in each image of fold k using this detector.
   c) Train detector using re-localized positives and all negative examples.
   d) Add new negative windows by hard-negative mining.
3) Return final detector and object windows in train data.

by comparing the localization behavior using the low-dimensional *vs*. the high-dimensional descriptors.

We also note that it is unlikely to remedy this problem via increasing regularization weight in SVM training. The ℓ_2 regularization term with weight λ restricts the linear combination weights such that $|\alpha_i| \leq 1/\lambda$. Therefore, although we can reduce the influence of individual training samples via regularization, the resulting classifier remains biased towards the training windows since the classifier is a linear combination of the window descriptors. In Section 4, we verify this hypothesis by evaluating the regularization weight's effect on the localization performance.

To address this issue—without sacrificing the descriptor dimensionality, which would limit its descriptive power—we propose to train the detector using a multi-fold procedure, reminiscent of cross-validation, within the MIL iterations. We divide the positive training images into K disjoint folds, and re-localize the images in each fold using a detector trained using windows from positive images in the other folds. In this manner the re-localization detectors never use training windows from the images to which they are applied. Once re-localization is performed in all positive training images, we train another detector using all selected windows. This detector is used for hard-negative mining on negative training images, and returned as the final detector.

We summarize our *multi-fold MIL* training procedure in Algorithm 1. The standard MIL algorithm that does not use multi-fold training does not execute steps 2(a) and 2(b), and re-localizes based on the detector learned in step 2(c).

The number of folds used in our multi-fold MIL training procedure should be set to strike a good trade-off between two competing factors. On the one hand, using more folds increases the number of training samples per fold, and is therefore likely to improve re-localization performance. On the other hand, using more folds also requires training more detectors, which increases the computational cost. We experimentally analyze this trade-off in Section 4.

3.3 Window refinement

We now explain our window refinement method. It updates the localizations obtained by the last multi-fold MIL iteration. The final detector is, then, re-trained based on these refinements.

An inherent difficulty for weakly supervised object localization is that WSL labels only permit to determine the most repeatable and discriminative patterns for each class. Therefore, even though the windows found by WSL are likely to overlap with target object instances, it can not be ensured that they will delineate object boundaries.

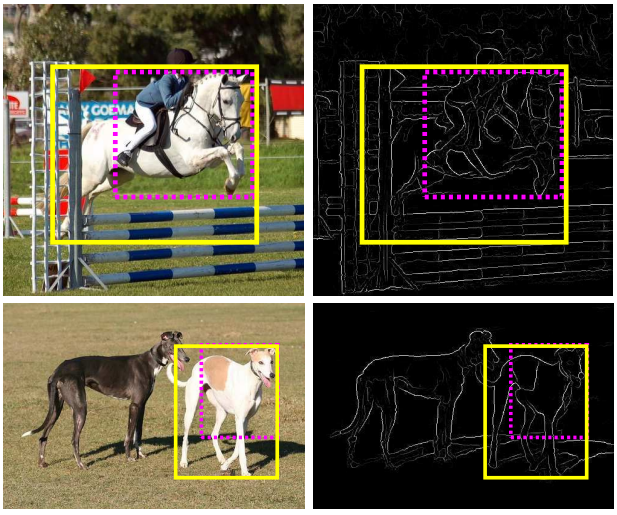


Fig. 3. Illustration of our window refinement. Dashed boxes (pink) show the localization before refinement, and the solid boxes (yellow) show the result of the window refinement method. The images on the right show the edge maps that are used to compute the objectness measure.

To better take into account object boundaries, we use the edge-driven *objectness* measure of Zitnick and Dollar [53]. The main idea in [53] is to score a given window based on the number of contours that are fully contained inside the window, with an increased weight on near-boundary edge pixels. Thus, windows that tightly enclose long contours are scored highly, whereas those with predominantly straddling contours are penalized. Additionally, in order to reduce the effect of marginal misalignments, the coordinates of a given window are updated using a greedy local search procedure that aims to increase the objectness score. In [53], the objectness measure is used for generating object proposals. For this purpose, a set of initial windows are first generated using a sliding window mechanism. The windows are, then, updated and scored using the local search procedure. The final object proposals are obtained by applying a non-maxima suppression procedure to the resulting windows and their scores.

We instead use the edge-driven objectness measure to improve WSL outputs. For this purpose, we combine the objectness measure with the classification scores given by multi-fold MIL. More specifically, we first utilize the local search procedure in order to update and score the refined candidate detection windows based on the objectness measure. Note that we do not update the classification scores. To make the classification scores and the objectness scores comparable, we scale each score channel to the range $[0, 1]$ for all windows in the positive training images. We, then, combine linearly the classification scores and the objectness scores with equal weights, and select the top detection in each image with respect to this combined score. In order to avoid selecting the windows irrelevant for the target class, but with a high objectness score, we restrict the search space to the top ten windows per image in terms of the classification score.

In Figure 3, we show two example images for the classes *horse* and *dog* in the left column, together with the corresponding edge maps in the right column. In these images, the dashed (pink) boxes show the output of multi-fold MIL training and the solid (yellow) boxes show the outputs of the window refinement procedure. Even though the initial windows are located on the object instances, they

TABLE 1
Weakly supervised learning with FV and CNN descriptors on the PASCAL VOC 2007 dataset. Evaluation in terms of the correct localization (CorLoc) measure on the training set. We compare descriptors for foreground (F), background (B) and contrastive background (C).

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
	standard MIL																				
FV: F	46.2	32.2	32.0	24.1	4.0	45.1	51.5	37.6	6.8	24.3	14.3	43.0	36.2	52.7	19.3	9.3	20.3	24.5	45.1	14.2	29.1
FV: F+B	50.3	32.2	32.4	24.8	4.0	45.1	52.2	**41.1**	6.8	25.2	14.3	**44.1**	38.2	53.7	20.5	9.3	20.3	24.5	43.4	14.2	29.8
FV: F+C	48.6	32.8	30.9	25.5	4.0	43.4	52.2	40.6	6.8	27.2	14.3	43.7	38.6	52.7	20.0	8.8	20.3	24.5	45.1	14.7	29.7
CNN	54.3	55.6	49.5	**31.7**	15.9	61.5	**72.2**	33.2	16.5	43.7	22.4	34.8	**58.5**	64.4	25.1	31.9	36.2	**34.0**	52.2	31.5	41.2
	multi-fold MIL																				
FV: F	48.0	55.6	25.8	4.1	6.3	53.3	68.3	23.3	8.8	57.3	4.1	27.6	52.7	**66.0**	33.2	15.4	55.1	14.2	49.6	**62.4**	36.5
FV: F+B	55.5	56.1	21.8	27.6	4.5	51.6	66.5	19.3	8.4	59.2	2.0	26.2	56.0	64.9	35.5	20.9	**58.0**	10.4	**56.6**	59.4	38.0
FV: F+C	**56.6**	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	10.2	29.0	58.0	64.9	**36.7**	18.7	56.5	13.2	54.9	59.4	38.8
CNN	53.2	**66.7**	**51.3**	**31.7**	**19.3**	**70.5**	72.0	23.3	**24.9**	**62.1**	**32.7**	28.0	54.6	64.9	22.1	**39.0**	55.1	33.0	54.9	40.1	**45.0**

TABLE 2
Weakly supervised learning with FV and CNN descriptors on the PASCAL VOC 2007 dataset. Evaluation in terms of the average precision (AP) measure on the test set. We compare descriptors for foreground (F), background (B) and contrastive background (C).

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
	standard MIL																				
FV: F	25.4	31.9	5.6	2.3	0.2	27.9	35.4	20.6	0.5	6.8	4.9	14.0	17.0	35.2	7.1	6.2	5.8	5.1	20.7	8.1	14.0
FV: F+B	28.8	30.7	10.5	6.6	0.3	30.1	36.2	**22.7**	0.9	7.2	3.4	16.3	22.3	35.5	7.7	9.2	7.5	3.9	26.2	6.5	15.6
FV: F+C	26.1	31.6	8.3	5.3	1.3	31.1	36.9	**22.7**	0.7	7.7	2.1	16.6	24.5	36.7	7.7	4.7	4.2	4.5	30.0	7.5	15.5
CNN	34.2	39.9	26.5	11.7	7.0	38.0	**45.6**	19.6	6.2	25.5	5.3	**18.8**	**34.2**	**42.3**	15.6	**20.0**	18.6	**23.5**	**37.0**	15.8	24.3
	multi-fold MIL																				
FV: F	29.4	37.8	7.3	0.5	1.1	33.2	41.0	14.3	1.0	21.9	9.2	9.4	29.1	37.3	15.5	9.8	27.9	4.7	29.4	40.4	20.0
FV: F+B	**36.7**	39.2	8.2	10.4	1.9	31.4	40.4	15.7	1.6	22.6	5.8	7.4	29.1	40.9	18.9	10.4	27.3	2.9	30.1	38.2	21.0
FV: F+C	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
CNN	32.1	**46.9**	**28.4**	**12.0**	**9.6**	**39.4**	45.5	16.2	**14.8**	**33.1**	**11.6**	14.0	31.2	39.3	13.1	19.7	**30.5**	23.4	**37.0**	19.6	**25.9**

are evaluated as incorrect due to the low overlap ratios with the ground-truth ones. The edge maps show that many contours, *i.e.* most object contours, straddle the initial window boundaries. In contrast, the corrected windows have higher percentages of fully contained contours, *i.e.* the contours relevant for the objects.

The refined windows are likely to be better aligned with object instances. Thus, their horizontal mirrors are more reliable and can be used as additional training examples. We evaluate the impact of window refinement and of adding flipped training examples in the next section.

4 Experimental Evaluation

In this section we present a detailed analysis and evaluation of our weakly supervised localization approach.

4.1 Dataset and evaluation criteria

We use the PASCAL VOC 2007 and 2010 datasets [18] in our experiments. Most of our experiments use the 2007 dataset, which allows us to compare to previous work. To the best of our knowledge, we are the first to report WSL performance on the VOC 2010 dataset. Following [15], [32], [37], during training we discard any images that only contain object instances marked as "difficult" or "truncated". During testing all images are included. We use linear SVM classifiers, and set the weight of the regularization term and the class weighting to fixed values based on preliminary experiments. We perform two hard-negative mining steps [19] after each re-localization phase.

Following [15], we assess performance using two measures. First, we evaluate the fraction of positive *training images* in which we obtain correct localization (CorLoc). Second, we measure the object detection performance on the *test images* using the standard protocol [18]: average precision (AP) per class, as well as the mean AP (mAP) across all classes. For both measures, we consider that a window is correct if it has an intersection-over-union ratio of at least 50% with a ground-truth object instance. Since CorLoc is not consistently measured across studies due to changes in the training settings, we use CorLoc mainly as a diagnostic measure, and use AP to measure the final detector performance and to compare to the state-of-the-art.

4.2 Multi-fold MIL training and features

In our first experiment, we compare (a) standard MIL training, and (b) our multi-fold MIL algorithm with $K = 10$ folds. Both are initialized from the full image up to the 4% boundary. We also consider the effectiveness of background features for the FV representation. We test three variants: (F) foreground only descriptor, (B) an FV computed over the window background, and (C) our contrastive background descriptor. Finally, we compare the FV representation to the CNN representation. Together, this yields eight combinations of features and training algorithms. Table 1 presents results in terms of CorLoc on the training set, and Table 2 presents results in terms of AP on the test set.

From the results we see that multi-fold MIL outperforms standard MIL in 14 out of 20 classes in terms of CorLoc and

Fig. 4. Re-localization using standard and multi-fold MIL for images of the classes *bicycle, motorbike,* and *cat* from initialization (left) to the final localization (right) and three intermediate iterations based on FV (F+C) descriptors. Correct localizations are shown in yellow, incorrect ones in pink.

12 classes in terms of AP. Furthermore, we see that the CorLoc differences across different FV descriptors are rather small when using standard MIL training. This is due to the degenerate re-localization performance with high-dimensional descriptors for standard MIL training as discussed in Section 3.2; we will come back to this point below. Finally, the CNN features give better performance overall than FV. For multi-fold training, the CNN features give better results than FV for 12 and 13 classes in terms of CorLoc and AP, respectively. They also benefit significantly from our multi-fold training procedure, although to a lesser extent than the FV. This is due to the lower dimensionality of the CNN features compared to the FV features.

Figure 4 presents examples of re-localization using standard and multi-fold MIL training. In all three cases, we observe that standard MIL gets stuck with the windows found by the first re-localization step. In contrast, multi-fold MIL is able to progressively localize down to smaller image regions. In the *bicycle* and *motorbike* examples, multi-fold MIL successfully localizes the object instances. In the *cat* example, on the other hand, while the window localized by standard MIL is correct, multi-fold MIL localizes the cat face, which has below 50% overlap with the object bounding box. The failure example in Figure 4 affirms the difficulty for weakly supervised localization that we have pointed out in Section 3.3: the WSL labels only indicate to learn a model for the most repeatable structure in the positive training images. For the cat class, due to the highly deformable body, the head can be argued to be the most distinctive and reliably detectable structure. This is what multi-fold MIL learns, but it degrades its CorLoc and AP scores. Parkhi *et al*. [33] also observed this, and proposed to localize cats and dogs based on a head detector in a fully supervised detector setting. Our window refinement method, which we evaluate below, resolves this issue to some extent.

In our next experiment, we further investigate the localization performances of the algorithms in terms of CorLoc across the training iterations. In the left panel of Figure 5 we show the results for standard MIL, and our multi-fold MIL algorithm using 2, 10, and 20 folds. The results clearly show the degenerate re-localization performance obtained with standard MIL training, of which CorLoc stays (almost) constant in the iterations following the first re-localization stage. Our multi-fold MIL approach leads

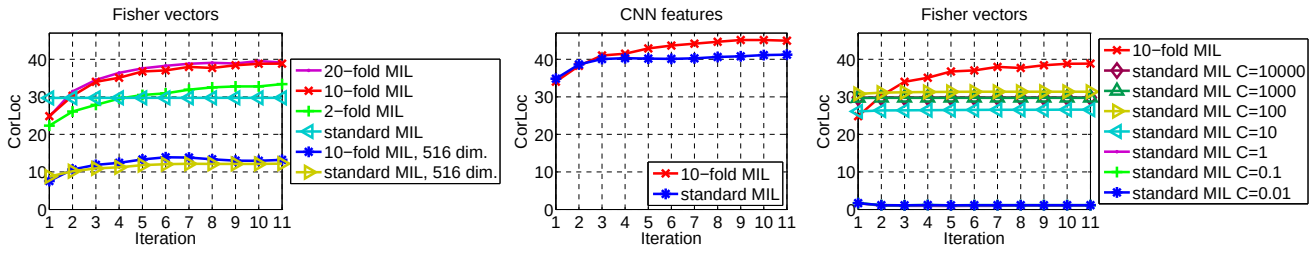


Fig. 5. Correct localization (CorLoc) performance (averaged across VOC 2007 classes) over the MIL iterations. We show results for the high and low dimensional FVs (left panel), and the CNN features (middle panel). In the right panel, we compare 10-fold training with standard MIL training using different SVM cost parameters (C) for the high-dimensional FVs.

TABLE 3
Evaluation for feature combination and the window refinement method on the PASCAL VOC 2007 dataset, in terms of the CorLoc measure.

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
FV	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	10.2	29.0	58.0	64.9	36.7	18.7	56.5	13.2	54.9	59.4	38.8
CNN	53.2	**66.7**	51.3	31.7	19.3	**70.5**	72.0	23.3	24.9	62.1	32.7	28.0	54.6	64.9	22.1	39.0	55.1	33.0	54.9	40.1	45.0
FV+CNN	57.2	62.2	50.9	37.9	**23.9**	64.8	74.4	24.8	**29.7**	64.1	40.8	37.3	55.6	68.1	25.5	38.5	65.2	35.8	56.6	33.5	47.3
	after window refinement																				
FV	62.4	62.2	40.7	35.2	5.1	67.2	76.9	33.2	12.9	63.1	16.3	39.4	62.8	67.6	37.2	22.5	63.8	22.6	**65.5**	**65.5**	46.1
CNN	**67.1**	66.1	49.8	34.5	23.3	68.9	**83.5**	**44.1**	27.7	**71.8**	**49.0**	48.0	65.2	**79.3**	**37.4**	**42.9**	65.2	**51.9**	62.8	46.2	**54.2**
FV+CNN	65.3	55.0	**52.4**	**48.3**	18.2	66.4	77.8	35.6	26.5	67.0	46.9	**48.4**	**70.5**	69.1	35.2	35.2	**69.6**	43.4	64.6	43.7	52.0

TABLE 4
Evaluation for feature combination and the window refinement method on the PASCAL VOC 2007 dataset, in terms of the AP measure.

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
FV	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
CNN	32.1	46.9	28.4	12.0	9.6	39.4	45.5	16.2	14.8	33.1	11.6	14.0	31.2	39.3	13.1	19.7	30.5	23.4	37.0	19.6	25.9
FV+CNN	38.1	**47.6**	28.2	13.9	**13.2**	45.2	**48.0**	19.3	**17.1**	27.7	17.3	19.0	30.1	45.4	13.5	17.0	28.8	24.8	38.2	15.0	27.4
	after window refinement																				
FV	36.9	38.3	11.5	11.1	1.0	39.8	45.7	16.5	1.2	26.4	4.3	17.7	31.8	44.0	13.1	11.0	31.4	9.7	38.5	36.9	23.3
CNN	**40.4**	43.5	**29.5**	11.4	9.4	42.2	47.3	**25.6**	7.6	**33.8**	15.8	27.7	37.4	46.4	**20.5**	19.9	30.2	23.5	40.6	19.6	28.6
FV+CNN	39.3	43.0	28.8	**20.4**	8.0	**45.5**	47.9	22.1	8.4	33.5	**23.6**	**29.2**	**38.5**	**47.9**	20.3	**20.0**	**35.8**	**30.8**	**41.0**	20.1	**30.2**

to substantially better performance, and ten MIL iterations suffice for the performance to stabilize. Results increase significantly by using 2-fold and 10-fold training respectively. The gain by using 20 folds is limited, however, and therefore we use 10 folds in the remaining experiments. We also include experiments with the 516 dimensional FV obtained using a 4-component MoG model, to verify the hypothesis of Section 3.2. The latter conjectured that the degenerate re-localization observed for standard MIL training is due to the trivial separability obtained for high-dimensional descriptors. Indeed, the lowest two curves in the left panel of Figure 5 show that for this descriptor we obtain non-degenerate re-localization using standard MIL similar to multi-fold MIL. The performance is poor, however, due to limited representative power of the low-dimensional FVs.

In the middle panel of Figure 5, we compare standard MIL and multi-fold MIL using the CNN features. We observe that standard MIL is less affected by degenerate re-localization problem, compared to the case for high-dimensional FVs. This is in accordance with our observations for low-dimensional FVs, as the CNN features have an intermediate dimensionality of 4,096. Nevertheless, multi-fold MIL leads to significant improvements over the iterations, which results in 43.8% CorLoc, compared to 40.3% CorLoc for standard MIL.

The degenerate re-localization of standard MIL using high-dimensional descriptors can be interpreted as over-fitting to the training data at an early stage. Therefore, the question is whether we can improve standard MIL by carefully tuning the trade-off between the regularization terms and the loss functions for SVM training. In the right panel of Figure 5, we investigate this question by evaluating the standard MIL approach at a number of different cost parameters (C) using high-dimensional FVs. The results show that, although choosing a proper C value is important, it is not possible to solve the degenerate re-localization problem of standard MIL in this manner. Whereas using a too low C value ($C \leq 1$) causes standard MIL to drift off to a poor solution, larger C values ($C \geq 10$) result in degenerate re-localization.

In Figure 6 we provide examples of the localization results on the training images using standard and multi-fold MIL for FV and CNN features. The first three examples (*car*, *chair*, and *potted plant*) are only correctly localized using multi-fold MIL with FVs. These examples demonstrate the ability of our multi-fold training procedure to handle cases with multiple instances

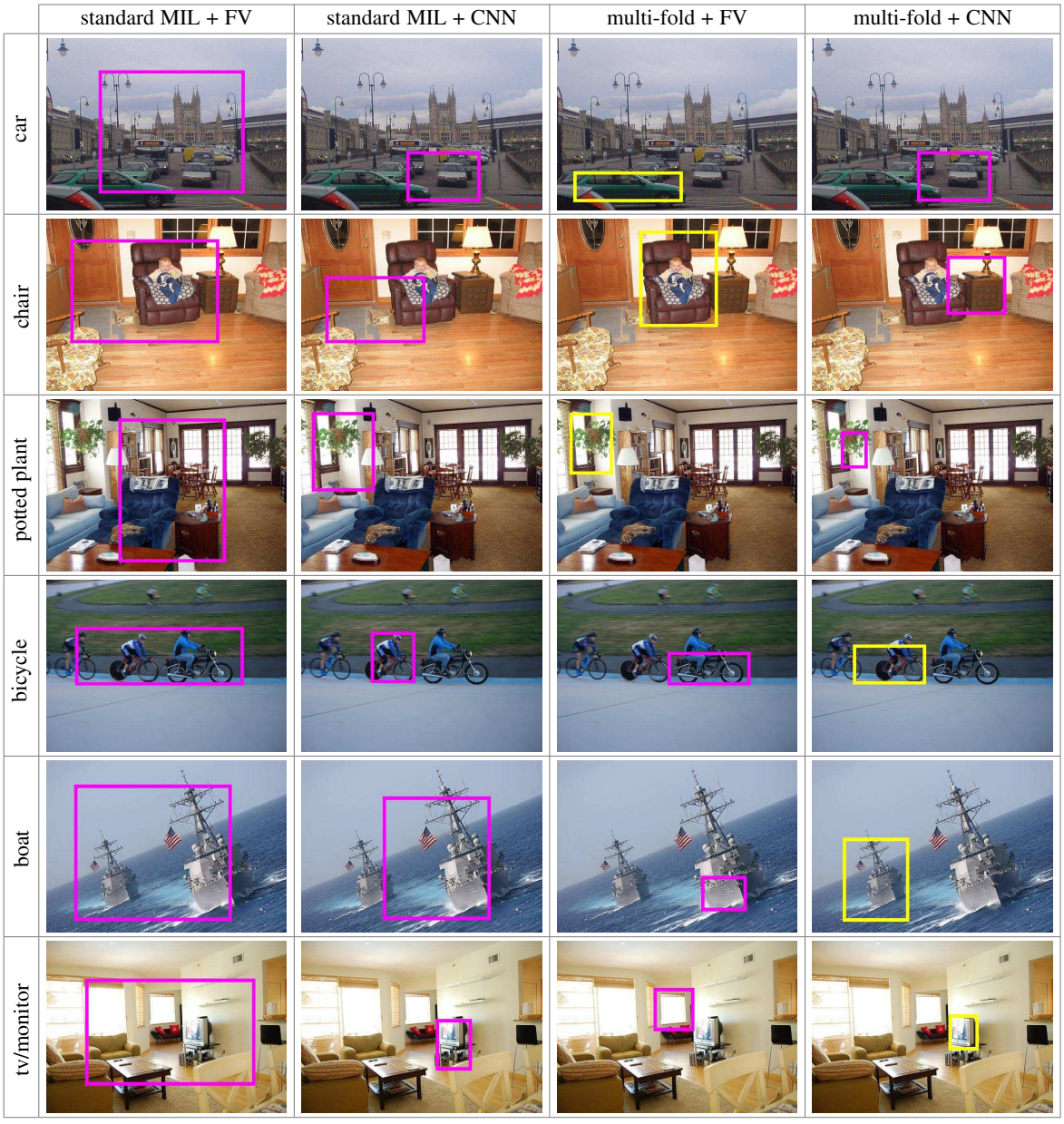


Fig. 6. Example localization results on the training images for standard MIL and multi-fold MIL algorithms with high-dimensional FV and CNN features. Correct localizations are shown in yellow, incorrect ones in pink.

that appear in near proximity and with considerable background clutter. The last three examples (*bicycle*, *boat*, and *tv/monitor*) are only correctly localized using multi-fold MIL with CNNs. In the *bicycle* example, we observe that multi-fold MIL with FVs mistakes a visually alike motorbike object for a bicycle. Likewise, in the *tv/monitor* example, multi-fold MIL over FVs localizes a window that looks similar to a bright monitor. These examples suggest that FV and CNN features can be complimentary to each other, and that some near-miss localizations might be corrected by a window refinement method.

We present the CorLoc and AP results for feature combination (by means of concatenating the descriptors) and the window refinement method in Table 3 and Table 4, respectively. All reported results are based on multi-fold training. The first parts of the tables show the results for the FV, CNN, and the combined features, without window refinement. We observe that the feature combination improves over the individual features in 13 of 20 classes in terms of both CorLoc and AP scores. We also note that standard MIL over the combined feature space performs poorly at 34.4% CorLoc and 22.0% mAP (not shown in the tables for brevity) compared to 47.3% CorLoc and 27.4% mAP for multi-fold MIL.

The second part of Table 3 and Table 4 show the results for the window refinement method. We observe that the refinement method significantly improves the average CorLoc and AP scores for all three window descriptor types. In the case of FV+CNN features, applying the window refinement method improves CorLoc and detection AP in 16 out of 20 classes, where we measure the largest three improvements in CorLoc for the classes *horse*, *dog* and *cat*. The instances of these three classes have deformable shapes, therefore, the weakly supervised localization tends to result in imprecise localizations or part localizations, some of which are corrected by the window refinement method. The four classes for which the window refinement method deteriorates CorLoc

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
Pandey and Lazebnik'11 [32]	11.5	—	—	3.0	—	—	—	—	—	—	—	—	20.3	9.1	—	—	—	—	13.2	—	—
Siva and Xiang'11 [42]	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	0.0	13.9
Russakovsky *et al*.'12 [35]	30.8	25.0	—	3.6	—	26.0	—	—	—	—	—	—	21.3	29.9	—	—	—	—	—	—	15.0
Ours (FV-only)	36.9	38.3	11.5	11.1	1.0	39.8	45.7	16.5	1.2	26.4	4.3	17.7	31.8	44.0	13.1	11.0	31.4	9.7	38.5	36.9	23.3
	methods using additional training data																				
Song *et al*.'14 [43]	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
Song *et al*.'14 [44]	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
Bilen *et al*.'14 [6]	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
Wang *et al*.'14 [50]	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
Wang *et al*.'14 [50] +context	**48.9**	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	**52.7**	19.1	17.4	**35.9**	**33.3**	34.8	**46.5**	**31.6**
Ours	39.3	43.0	**28.8**	**20.4**	8.0	**45.5**	**47.9**	22.1	8.4	33.5	**23.6**	29.2	**38.5**	47.9	**20.3**	20.0	35.8	30.8	**41.0**	20.1	30.2

TABLE 5
Comparison of WSL detectors on PASCAL VOC 2007 in terms of test-set detection AP. Results for Pandey and Lazebnik [32] are taken from [34].

are *bicycle*, *bottle*, *chair* and *potted-plant*. These classes typically have highly textured and/or small instances, where the edge-driven objectness measure can be misleading. Finally, we note that the results obtained with refinement also include the addition of horizontal flips of the positive training windows. This has only a minor effect on performance: without these the detection mAP for the FV+CNN features drops by only 0.4% to 29.8%. Overall, these results show that the FV and CNN features are complimentary, and that window refinement can improve localization performance.

4.3 Comparison to state-of-the-art WSL detection

We compare our multi-fold MIL approach to the state-of-the-art in terms of detection AP in Table 5. We separate the recent work into two groups in terms of their utilization of auxiliary training data. To the best our knowledge, only three previous studies that do not use auxiliary training data reported detection AP scores on PASCAL VOC 2007. Other work, such as *e.g.* that of Deselaers *et al*. [15], was evaluated only under simplified conditions, such as using viewpoint information and using images from a limited number of classes. Russakovsky *et al*. [35] report mAP over all 20 classes, but report separate AP values for only six classes. Multi-fold MIL over the FV-only features with window refinement, results in a detection mAP of 23.3%, which is significantly better than the 13.9% and 15.0% reported in [42] and [35].

The second half of Table 5 presents the recent work that uses CNN-based features, which involves representation learning on the ImageNet dataset. For comparison, we use our multi-fold MIL approach over the FV+CNN features with window refinement. Our detection mAP of 30.2% is significantly better than the 22.7% by Song *et al*. [43], 24.6% by Song *et al*. [44], and the 26.4% by Bilen *et al*. [6]. Wang *et al*. [50] report a detection mAP of 30.9%, and additionally an improved mAP of 31.6% mAP using the contextual rescoring method of [19]. Our detection mAP is comparable to Wang *et al*. [50] without inter-class context, and we obtain better AP scores in 11 out of 20 classes.

4.4 Discussion and analysis

To analyze the causes of difficulty of WSL for object detection, we consider the performance of our detector when used in a fully supervised training setting. For the sake of brevity, we analyze the WSL results based on the FV-only and CNN-only features, without combining the features or applying window refinement.

There are several factors that change between the WSL and fully supervised training. In order to evaluate the importance of each factor, we progressively move from the original WSL setting to the fully supervised setting. In Table 6, we report the resulting mAP values for each step using the FV features and the CNN features in the final two columns, respectively.

TABLE 6
Performance in test-set detection mAP on VOC 2007 using FV and CNN features, with varying degrees of supervision.

Supervision	Neg on Pos	Positive Set	mAP(FV)	mAP(CNN)
Image labels only	No	Non-diff/trunc	22.4	25.9
Cand box for one obj	No	Non-diff/trunc	30.8	36.5
Cand box for all obj	No	Non-diff/trunc	30.7	35.7
Cand box for all obj	Yes	Non-diff/trunc	32.0	41.2
Exact box for all obj	Yes	Non-diff/trunc	32.8	40.5
Exact box for all obj	Yes	All	35.4	42.8

In WSL we have to determine the object locations in the positive training images. If in each positive training image we fix the object hypothesis to the candidate window that best overlaps with one of the ground-truth objects, we no longer need to use MIL training. In this case, we obtain a detection mAP of 30.8% using FVs and 36.5% using the CNN features, as shown in the second row of Table 6. Even though this is a significant improvement w.r.t. WSL, there is still a gap of 4.6% and 6.3% in detection mAP respectively for the FV and the CNN features compared to the fully supervised setting.

The remaining difference in performance is due to several factors, we list them now and give the performance improvements when making the WSL training scenario progressively more similar to the supervised one. (i) WSL uses only one instance per positive training image, including all instances makes a relatively minor effect on the performance, see the third row in Table 6. (ii) In WSL hard-negative mining is based on negative images only, when positive images are used too performance rises to 32.0% and 41.2% mAP for the FV and the CNN features, respectively, as shown in the fourth row. (iii) WSL is based on the candidate windows, using the ground-truth windows instead makes a relatively small impact on the performance, see the fifth row. (iv) In WSL, we do not use positive training images marked as difficult or truncated, if these are added to the fully supervised training, performance rises to 35.4% and 42.8% mAP for the FV and the CNN features, which corresponds to the final row.[2]

2. We note that our fully-supervised mAP of 42.8% is comparable to the 44.7% of Girshick *et al*. [20], which uses similar CNN features.

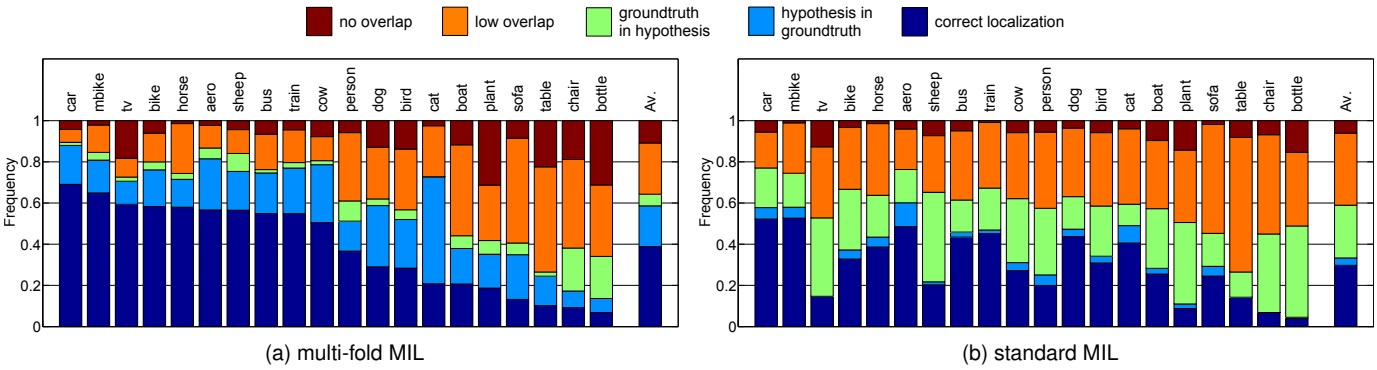


Fig. 8. Per-class frequency of error modes, and averaged across all classes using FV features with 10-fold MIL and standard MIL training.

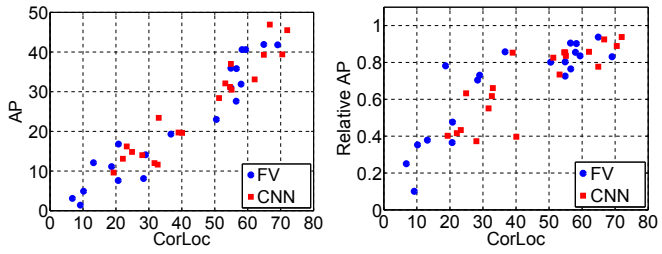


Fig. 7. AP *vs.* CorLoc for multi-fold MIL (left), and ratio of WSL over supervised AP as a function of CorLoc (right) using FV features (red squares) and CNN features (blue circles).

These results show that the most important two factors both for the FV and the CNN features are the use of correct training windows and hard-negative mining on positive training images. We also observe that the multi-fold MIL achieves 63% of the representational performance limit (22.4% out of 35.4% mAP) for the FV features and similarly 61% (25.9% out of 42.8% mAP) for the CNN features. With respect to the 30.8% mAP (FV) and 36.5% mAP (CNN) for training from ideal localization, our multi-fold MIL approach attains 73% and 71% of the WSL performance limit for the FV and the CNN features, respectively. With standard MIL training the FV and CNN features only attain 50% and 67% of this limit respectively, *c.f*. Table 2.

In Figure 7 we further analyze the results of our weakly supervised detector, and its relation to the optimally localized version. In the left panel, we visualize the close relationship between the per-class CorLoc and AP values for our multi-fold MIL detector. The three classes with lowest CorLoc are *bottle*, *chair*, and *dining table* using FVs, and *bottle*, *chair*, and *cat* using CNNs. Most instances of these classes appear in highly cluttered indoor images, and are often occluded by objects (*dining table*), or people (*chair*), or have extremely variable appearance due to transparency (*bottle*) and deformation (*cat*). In the right panel of Figure 7 we plot the ratio between our WSL detection AP (22.4% mAP with the FV features and 25.9% mAP with the CNN features) and the AP obtained with the same detector trained with optimal localization (30.8% mAP with the FV features and 36.5% mAP with the CNN features). In this case there is also a clear relation with our CorLoc values. The relation is quite different, however, below and above 40% CorLoc. Below this threshold, the amount of noisy training examples is so large that WSL essentially breaks down. Above this threshold, however, the training is able to cope with the noisy positive training examples, and the weakly supervised detector performs relatively well: on average above 80% relative to optimal localization.

In order to better understand the localization errors, we categorize each of our object hypotheses in the positive training images into one of the following five cases: (i) correct localization (overlap $\geq$ 50%), (ii) hypothesis completely inside ground-truth, (iii) reversed inclusion, (iv) none of the above, but non-zero overlap, and (v) no overlap. For the sake of brevity, we analyze only the WSL outputs for the FV features. In Figure 8a we show the frequency of these five cases for each object category and and averaged over all classes for multi-fold MIL. We observe that *hypothesis in ground-truth* category is the second largest error mode. For example, as expected from Figure 4, most localization hypotheses for the class *cat*, and similarly for the class *dog*, are fully contained within a ground-truth window. Although the instances of this mis-localization category may significantly degrade CorLoc and AP measures, they could as well be interpreted as correct localizations in certain applications where it is not necessary to localize with bounding boxes fully covering target objects. Interestingly, we observe that, with 10.8% on average, the "no overlap" case is rare. This means that 89.2% of our object hypotheses overlap to some extent with a ground-truth object. This explains the fact that detector performance is relatively resilient to frequent mis-localization in the sense of the CorLoc measure.

Figure 8b presents the error distribution corresponding to the standard MIL training. Whereas *hypothesis in ground-truth* is much more frequent than *ground-truth in hypothesis* for multi-fold MIL training, the situation is reversed for standard MIL training. This is a result of the fact that whereas multi-fold MIL is able localize most discriminative sub-regions of the object categories, standard MIL tends to get stuck after the first few iterations, resulting in too large bounding box estimates. The effect of multi-fold training on the distribution of different localization error types is similar for the CNN features.

Finally, we note that while multi-fold MIL using k-folds results in training k additional classifiers per iteration, training duration grows sublinearly with k since the number of re-localizations and hard-negative mining work does not change. In a single iteration of our implementation using the FV features, (a) all SVM optimizations take 10.5 minutes for standard MIL and 42 minutes for 10-fold MIL, (b) re-localization on positive images take 5 minutes in both cases and (c) hard-negative mining takes 20 minutes in both cases. In total, standard MIL takes 35.5 minutes per iteration and 10-fold MIL takes 67 minutes per iteration.

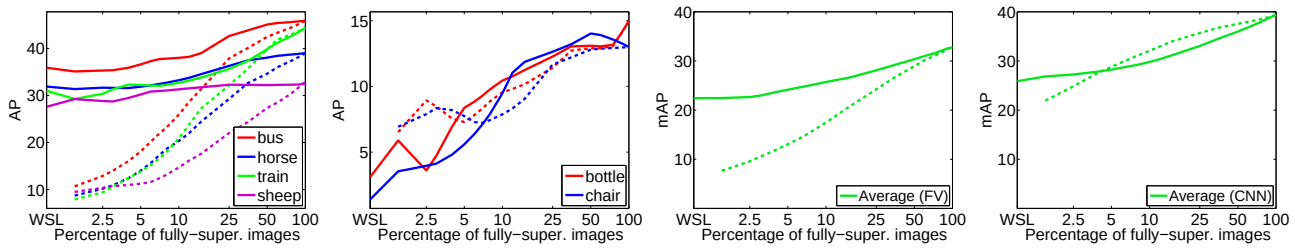


Fig. 9. Object detection results for training with mixed supervision. Each curve shows the detection AP as a function of the percentage of fully supervised positive training images. The first two plots show per-class curves for selected classes, using the FV features. The last two plots show the detection AP values averaged over all classes for the FV and the CNN features, respectively. The solid curves correspond to the mixed supervision results. The dotted curves correspond to results obtained by using only the fully supervised training examples.

4.5 Training with mixed supervision

In our experiments so far, we have considered the WSL and fully supervised scenarios, where each training image is annotated with either class labels (WSL) or object bounding boxes (fully supervised). We now consider training using a mixture of the two paradigms, which we refer to as *mixed-supervision*.

One way to combine weakly supervised and fully supervised training for object localization is to leverage an existing dataset of fully supervised training images of non-target classes during WSL of a new object category detector, also referred to as transfer learning, see *e.g.* [15], [38]. Such an approach, however, does not provide any fully supervised example for the target class and does not allow hard negative mining on the positive images, both of which are important factors as shown in our previous analysis.

We, instead, consider a setup where a subset of the positive training images for each class is fully supervised. For this purpose, we randomly sample a subset of the positive training images and add ground-truth box annotations for all objects in them. These images are then excluded from the re-localization steps in the multi-fold training procedure and instead their ground-truth windows are used as positive training examples. We also use the fully supervised positive images for hard-negative mining, in addition to the negative images.

Figure 9 presents detection AP scores as a function of the percentage of fully supervised positive training images. Each curve is obtained by evaluating the performance when the number of fully supervised images per class is limited up to 2^i, for all $i \in \{0, \ldots, 10\}$. We also evaluate the baseline detection results where only the fully supervised images are used for training. We repeat each experiment twice and average the detection AP scores. In each plot, the resulting mixed supervision and baseline curves are shown using solid and dotted lines, respectively.

The left panel in Figure 9 shows the mixed-supervision evaluation results for four classes (*bus*, *horse*, *train*, and *sheep*), where the per-class performances are similar to the average case for FVs (the latter is shown in the third panel). In these four classes, and on average, we observe significant performance gains using mixed supervision compared to conventional full supervision for FVs.

The only two classes where mixed supervision is not more effective than fully supervised training are *bottle* and *chair*, for which AP curves are presented in the second panel of Figure 9. We note that *bottle* and *chair* are also the classes with the lowest CorLoc values for multi-fold training, which explains why mixed-supervised training does not work well for these two classes.

In the third panel we observe that the fewer images are fully supervised, the more significant the benefit of additional weakly labeled images. We also observe that the benefit of adding fully supervised images into WSL is significant typically when at least 2.5% of the examples are fully supervised. Below this percentage, the mixed-supervision mAP is nearly the same as the WSL detection mAP. Therefore, in the regime where the percentage of fully supervised images is roughly between 2.5% and 50%, there is a benefit in combining the fully supervised images with weakly supervised ones, when the FV features are utilized.

The last panel in Figure 9 presents the results for the CNN descriptors. We observe that training with mixed supervision improves the detection mAP compared to training with only the fully supervised examples when up to 5% of the positive training images are fully supervised. At larger fully supervised image percentages, training over only the fully supervised images outperforms mixed-supervision based training. Regarding this result, we can interpret the CNN features as the outputs from a pre-trained classifier, and therefore, having a few training images can be sufficient for effectively learning a detection model over the CNN features. As a result, utilizing weakly supervised examples during training can sometimes deteriorate the detection performance due to the imperfect localizations provided by the WSL methods.

Overall, the results suggest that fully supervised images can be successfully integrated into multi-fold WSL training in order to improve detection rates by annotating objects only in a small number of images. This holds in particular, when auxiliary training data, such as the ImageNet images used for training the CNN model, is not available. One possible future work direction is to develop measures for weighting fully supervised examples more heavily than the weakly supervised ones during classifier training, especially in the early MIL iterations.

4.6 VOC 2010 evaluation

We now present an evaluation on the VOC 2010 dataset in order to verify the effectiveness of multi-fold training and the window refinement method on a second dataset. We are the first to present weakly supervised results on this dataset, and can therefore not compare to other weakly supervised methods. We show the resulting CorLoc values in Table 7 and detection AP results in Table 8. Overall, our results on VOC 2010 are similar to those on the 2007 dataset in the sense that multi-fold MIL significantly improves the WSL performance compared to standard MIL training, especially when high-dimensional FV descriptors are included. Using multi-fold MIL over the combined FV and CNN features results in 24.7% mAP. The window refinement method further improves the performance to 27.4% mAP.

TABLE 7
Comparison of standard MIL training vs our 10-fold MIL on VOC 2010 in terms of training set localization accuracy (CorLoc).

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
	standard MIL																				
FV	58.9	45.2	33.7	24.1	6.7	66.1	43.3	**50.6**	16.2	36.0	25.5	41.8	53.4	57.5	21.5	11.6	32.9	30.5	50.0	21.6	36.4
CNN	54.8	60.1	52.3	40.2	26.6	73.9	64.1	23.4	35.7	58.1	24.5	32.4	71.3	63.8	28.0	36.4	61.6	44.7	48.1	55.3	47.8
	multi-fold MIL																				
FV	47.3	47.1	36.2	34.8	24.9	68.9	59.8	18.9	21.3	52.9	26.6	32.2	44.1	60.7	33.7	17.3	63.9	32.6	48.1	**66.6**	41.9
CNN	53.4	59.1	52.6	39.9	27.1	73.1	65.2	18.6	40.6	68.0	33.0	30.1	71.0	63.2	27.1	**37.8**	61.6	43.3	48.1	58.9	48.6
FV+CNN	60.7	60.1	53.4	38.7	27.8	77.7	67.1	20.3	42.6	64.0	**39.4**	38.8	70.6	65.2	28.5	36.1	58.8	46.1	55.8	49.7	50.1
FV+CNN+Refinement	**61.1**	**65.0**	**59.2**	**44.3**	**28.3**	**80.6**	**69.7**	31.2	**42.8**	**73.3**	38.3	**50.2**	**74.9**	**70.9**	**37.3**	37.1	**65.3**	**55.3**	**61.7**	58.2	**55.2**

TABLE 8
Comparison of standard MIL training vs our 10-fold MIL on VOC 2010 in terms of test set AP.

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
	standard MIL																				
FV	41.9	30.4	6.9	5.2	1.6	38.6	24.8	**29.6**	1.3	8.7	2.3	18.7	22.1	40.0	9.9	0.9	9.7	6.4	18.6	11.5	16.4
CNN	35.8	38.6	21.9	10.1	8.6	39.0	33.9	20.5	8.0	22.8	7.5	17.9	33.4	46.1	15.8	**13.6**	26.7	15.5	26.8	22.2	23.2
	multi-fold MIL																				
FV	27.9	23.2	8.1	11.8	9.6	35.7	31.3	10.7	3.6	14.9	6.0	12.8	18.6	41.8	16.3	3.0	**27.6**	10.3	22.4	**34.6**	18.5
CNN	34.7	39.1	21.9	10.5	8.8	37.7	34.4	18.1	10.1	**26.4**	11.2	16.5	33.0	44.7	15.6	13.2	26.2	15.6	24.8	24.8	23.4
FV+CNN	42.2	41.5	22.5	11.3	8.6	41.7	36.1	19.4	**13.3**	24.3	**14.5**	21.3	32.7	48.3	15.2	11.3	25.0	18.0	27.9	18.4	24.7
FV+CNN+Refinement	**44.6**	**42.3**	**25.5**	**14.1**	**11.0**	**44.1**	**36.3**	23.2	12.2	26.1	14.0	**29.2**	**36.0**	**54.3**	**20.7**	12.4	26.5	**20.3**	**31.2**	23.7	**27.4**

If we train the object detectors in a fully supervised manner, we obtain 33.6% mAP using the FV features, and 37.7% mAP using the CNN features. This verifies that we have an effective object representation outperforming DPMs [21] (29.6% mAP). Highest fully supervised detection result on this dataset is 39.7% mAP [52] without using auxiliary data and 53.7% mAP [20] with using auxiliary data. We note that whereas the CNN model we use is trained on the ImageNet images only, Girshick *et al*. [20] utilize a CNN model *fine-tuned* on the VOC ground-truth boxes, which leads to a better fully-supervised detection performance. We plan to explore weakly supervised CNN fine-tuning in future work.

5 Conclusions

In this article, we have introduced a multi-fold multiple instance learning approach for weakly supervised object detection, which avoids the degenerate localization performance observed without it. Second, we have presented a contrastive background descriptor, which encourages the detection model to learn the differences between the objects and their context. Third, we have designed a window refinement method, which improves the localization accuracy by using an edge-driven objectness prior.

We have evaluated our approach and compared it to state-of-the-art methods using the VOC 2007 dataset. Our results show that multi-fold MIL effectively handles high-dimensional descriptors, which allows us to obtain state-of-the-art results by jointly using FV and CNN features. On the VOC 2010 dataset we observe similar improvements by using our multi-fold multiple instance learning method.

A detailed analysis of our results shows that, in terms of test set detection performance, multi-fold MIL attains 73% and 71% of the MIL performance upper-bound, which we measure by selecting one correct training example from each positive image, for the FV and the CNN features, respectively.

Acknowledgements. This work was supported by the European integrated project AXES and the ERC advanced grant ALLEGRO.

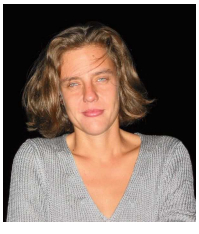
References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.
[3] F. Alted. Why modern CPUs are starving and what can be done about it. *Computing in Science & Engineering*, 12(2):68–71, 2010.
[4] S. Bagon, O. Brostovski, M. Galun, and M. Irani. Detecting and sketching the common. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[5] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
[6] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference*, 2014.
[7] M. Blaschko, A. Vedaldi, and A. Zisserman. Simultaneous object detection and ranking with weak supervision. In *Advances in Neural Information Processing Systems*, 2010.
[8] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
[9] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
[10] R. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with Fisher vectors. In *International Conference on Computer Vision*, 2013.
[11] R. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL training for weakly supervised object localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[12] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *European Conference on Computer Vision*, 2006.
[13] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Int. Workshop on Stat. Learning in Computer Vision*, 2004.
[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[15] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal on Computer Vision*, 100(3):257–293, 2012.
[16] T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.
[17] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5):545–559, 2009.
[18] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal on Computer Vision*, 88(2):303–338, 2010.
[19] P. Felzenszwalb, R. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 2010.
[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[21] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5, 2012.
[22] C. Gu, P. Arbeláez, Y. Lin, K. Yu, and Malik. Multi-component models for object detection. In *European Conference on Computer Vision*, 2012.
[23] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001.
[24] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.
[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
[26] G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *Advances in Neural Information Processing Systems*, pages 4–2, 2009.
[27] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.
[28] C. Lampert, M. Blaschko, and T. Hofmann. Efficient subwindow search: a branch and bound framework for object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2129–2142, 2009.
[29] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
[30] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.
[31] M. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *International Conference on Computer Vision*, 2009.
[32] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *International Conference on Computer Vision*, 2011.
[33] O. Parkhi, A. Vedaldi, C. Jawahar, and A. Zisserman. The truth about cats and dogs. In *International Conference on Computer Vision*, 2011.
[34] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
[35] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *European Conference on Computer Vision*, 2012.
[36] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: Theory and practice. *International Journal on Computer Vision*, 105(3):222–245, 2013.
[37] Z. Shi, T. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *International Conference on Computer Vision*, 2013.
[38] Z. Shi, P. Siva, T. Xiang, and Q. Mary. Transfer learning by ranking for weakly supervised object annotation. In *BMVC*, 2012.
[39] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision*, 2012.
[40] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *European Conference on Computer Vision*, 2012.
[41] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[42] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *International Conference on Computer Vision*, 2011.
[43] H. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *International Conference on Machine Learning*, 2014.
[44] H. Song, Y. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *Advances in Neural Information Processing Systems*, 2014.
[45] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[46] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal on Computer Vision*, 104(2):154–171, 2013.
[47] K. E. van de Sande, C. G. Snoek, and A. W. Smeulders12. Fisher and VLAD with FLAIR. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[48] J. Verbeek and B. Triggs. Region classification with Markov field aspect models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
[49] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[50] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision*, 2014.
[51] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[52] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. In *International Conference on Computer Vision*, 2013.
[53] C. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, 2014.

Ramazan Gokberk Cinbis graduated from Bilkent University, Turkey, in 2008, and received an M.A. degree in computer science from Boston University, USA, in 2010. He was a doctoral student in the LEAR team, at INRIA Grenoble, France, from 2010 until 2014, and received a PhD degree in computer science from Université de Grenoble, France, in 2014. His research interests include computer vision and machine learning.

Jakob Verbeek received a PhD degree in computer science in 2004 from the University of Amsterdam, The Netherlands. After being a post-doctoral researcher at the University of Amsterdam and at INRIA Rhône-Alpes, he has been a full-time researcher at INRIA, Grenoble, France, since 2007. His research interests include machine learning and computer vision, with special interest in applications of statistical models in computer vision.



Cordelia Schmid holds a M.S. degree in computer science from the University of Karlsruhe and a doctorate from the Institut National Polytechnique de Grenoble. She is a research director at INRIA Grenoble where she directs the LEAR team. In 2006 and 2014, she was awarded the Longuet-Higgins prize for fundamental contributions in computer vision that have withstood the test of time. In 2012, she obtained an ERC advanced grant. She is a fellow of IEEE.