Multi-fold MIL Training for Weakly Supervised Object Localization

Ramazan Gokberk Cinbis Jakob Verbeek Cordelia Schmid
Inria*

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 high-dimensional representations, such as the Fisher vectors, are used. We present a detailed experimental evaluation using the PASCAL VOC 2007 dataset. Compared to state-of-the-art weakly supervised detectors, our approach better localizes objects in the training images, which translates into improved detection performance.

1. Introduction

Object category localization is an important computer vision problem where the goal is to identify all instances of a given object category in an image, and to report these by means of bounding boxes around the objects. As compared to image classification, localization is significantly more challenging since precise estimates of the object locations need to be produced. Over the last decade significant progress has been made in this area, as witnessed by the PASCAL VOC challenges [14]. Training such detectors, however, requires bounding box annotations of object instances, which are more error prone and costly to acquire as compared to the labels required for image classification.

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 indicating the absence or presence of instances of the category in images has recently been intensively studied as a way to remove the requirement of bounding box annotations, see *e.g.* [2, 8, 10, 11, 20, 22, 23, 25, 27, 29]. Such methods can potentially leverage the large amount of tagged images on the internet as a source of data to train object detection models.

Other examples of WSL include learning face recognition models from image captions [3], or subtitle and script information [13]. Another WSL example is learning semantic segmentation models from image-wide category labels [32]. Most WSL approaches are based on latent variable models to account for the missing ground-truth information. Multiple instance learning (**MIL**) [12] 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.* [11, 25, 32]. Besides weakly supervised training, mixed fully and weakly supervised [4], active [33], and semi-supervised [25] 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 uses those to better model the object appearance variations. We review the most relevant related work in more detail in Section 2.

In this paper we consider the WSL problem 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. To represent (tentative) detection windows, we use the high-dimensional Fisher vector (**FV**) image representation [24], which has shown to yield state-of-the-art performance for image classification and object detection [6, 7, 9]. As we explain in Section 3, when used in an MIL framework, the high-dimensionality of the FV representation 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

*LEAR team, Inria Grenoble Rhône-Alpes, Laboratoire Jean Kuntzmann, CNRS, Univ. Grenoble Alpes, France

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. In Section 3 we present our multi-fold training procedure and object representation in full detail.

We present a detailed evaluation using the PASCAL VOC 2007 dataset in Section 4, and also report results on the VOC 2010 dataset. A comparison to the current state of the art in WSL shows that our approach leads to better localization on the training images, which translates into a substantial improvement in detection performance. Finally, we summarize our conclusions in Section 5.

2. Related work

The majority of related work treats WSL for object detection as an MIL [12] 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 single most likely object instance in each positive image. In the case of object detector training there is a vast number of examples per bag since the number of possible object bounding boxes is quadratic in the number of image pixels. This is different from many other MIL problems, *e.g.* such as those for weakly supervised face recognition [3, 13], where the number of examples per bag is limited to a few dozen at most. In order to make MIL approaches to WSL more manageable, tentative object windows can instead be obtained using recent window proposal methods, which effectively reduce the number of candidate windows for object detection to several hundreds or thousands by exploiting low-level segmentation-based cues [1, 17, 31].

A number of different strategies to initialize the MIL detector training have been proposed in the literature. The first strategy, *e.g.* taken in [20, 23], 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: although large windows may contain a significant amount of background features, they are likely to include a positive object instance.

The second 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*. [11] initialize the training using the candidate window with the highest objectness score [1]. Siva *et al*. [28] 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. An initial window for each image is found by sampling from the corresponding salieny map.

The third strategy is to use a class-specific initialization method. For example, Siva and Xiang [29] propose to initially select one of the candidate windows sampled using objectness [1] for 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*. [27] propose a simplified approach where a candidate window is selected for a given positive image such that the distance to its nearest neighbor among windows from negative images is maximal. Relying only on 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*. [25] recently used an extended version of the LDA topic model [5] to obtain object hypotheses in positive images. Their approach localizes object categories across different categories concurrently, which allows to benefit from explaining-away effects, *i.e.* an image region cannot be identified as an instance for multiple categories. Their approach, however, associates image patches with object categories, rather than complete object hypotheses.

The majority of related work relies on off-the-shelf detectors for MIL training. They iteratively select the maximum scoring detections as the positive training examples and train the detection models. For example, Nguyen *et al*. [19] use a branch-and-bound localization based detector [18]. A number of other works [20, 25, 27, 28, 29] build upon deformable part-based model (**DPM**) detectors [15].

Our approach is most related to that of Russakovsky *et al*. [23]: we also rely on selective search windows [31], and use a similar initialization strategy. A critical difference from [23] and other related work, 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*. [26] 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.

3. Weakly supervised object localization

We present our multi-fold MIL approach in Section 3.2, but first briefly describe our FV object model in Section 3.1.

3.1. Features and detection window representation

To represent the detection windows, we rely on methods that have shown to yield state-of-the-art performance for image classification and fully-supervised detection [6, 7, 9]. In particular, we aggregate local SIFT descriptors into a FV representation to which we apply ℓ_2 and power normalization [24]. 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. Using PCA to project the SIFTs to 64 dimensions, and Gaussian mixture models (**GMM**) of 64 components, this yields a descriptor of 140,352 dimensions. We reduce the memory footprint by using feature compression [24] in combination with the selective search method of Uijlings *et al*. [31]. The latter, 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.

Similar to Russakovsky *et al*. [23], we also 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*. [9, 30]. 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 degenerate object localization on the training images, 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 score the same features positively on both the foreground and the background, and localizes 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.

3.2. Weakly supervised object detector training

The dominant method in the literature for weakly supervised object detector training is the iterative training and re-localization MIL approach described above, which we will henceforth refer to as *Standard MIL*. In this approach, the detector used for re-localization in positive images is trained using samples that are extracted from the same images. This generates a bias towards re-localizing on the same windows; in particular when high capacity classifiers are used which easily separate the detector's training data.

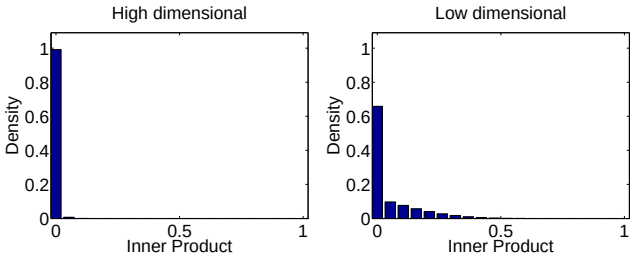

Figure 1: Distribution of inner products, scaled to the unit interval, of random window pairs using our high-dimensional FV (left), and a low-dimensional FV (right).

To understand this effect, we consider in Figure 1 the distribution of inner products between FVs of different windows. We show the distribution using both our 140,352 dimensional FVs and 516 dimensional FVs obtained using 4 Gaussians without spatial grid. Unlike in the low-dimensional case, almost all FVs are near orthogonal in the high-dimensional case. Recall that the weight vector of a linear (SVM) classifier can be written as a linear combination of training samples, $\boldsymbol{w} = \sum_i \alpha_i \boldsymbol{x}_i$. Therefore, due to the near orthogonality, the training windows will score significantly higher than the other windows in positive images in the high-dimensional case, resulting in degenerate re-localization behaviour. Although this may appear as a classic overfitting problem, increasing the weight of the regularization term in SVM training will not solve the problem, since weight vector will remain a linear combination of the training samples. We verified this experimentally, but did not include these experimental results due to lack of space.

Instead, to address this issue—without sacrificing the FV 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

Algorithm 1 — Multi-fold weakly supervised training

1. Initialization: positive and negative windows 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 positives in all folds but k.
      ii. Re-localize positives in fold k using this detector.
   (c) Train detector using positive windows from all folds.
   (d) Perform hard-negative mining using this detector.
3. Return final detector and object windows in train data.

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 will analyze this trade-off in our experiments below.

We note that while multi-fold MIL using K folds results in training K additional classifiers per iteration, the 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, (a) all SVM optimizations take 10.5 minutes for standard MIL and 42 minutes for 10-fold MIL, (b) relocalization on positive images take 5 minutes in both cases and (c) hard-negative mining takes 20 minutes in both cases. In total, for a single class, standard MIL takes 35.5 minutes per iteration and 10-fold MIL takes 67 minutes per iteration.

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 [14] 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 2010 dataset. Following [11, 20, 25], 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, see [15], after each re-localization phase.

	MIL			Multi-fold MIL, K=10		
Descriptors	F	F+B	F+C	F	F+B	F+C
CorLoc	29.1	29.8	29.7	36.5	38.0	**38.8**
Detection mAP	14.0	15.6	15.5	20.0	21.0	**22.4**

Table 1: Localization performance on the train set (CorLoc) and detection performance on the test set for VOC 2007, averaged over classes, using different training methods and features: foreground (F), background (B), contrastive (C).

Following [11], 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: 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.

4.2. Multi-fold MIL training and context 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 (near) full image window. We also consider the effectiveness of background features. We test three variants: (i) no background descriptor, (ii) an FV computed over the window background, and (iii) our contrastive background descriptor. Together, this yields six combinations of features and training algorithms. We present the performance for each of these in Table 1, in terms of CorLoc and detection AP, where both are averaged over the 20 VOC 2007 classes.

From the results we see that the CorLoc differences across different 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. Using multi-fold training instead of standard MIL training leads to a substantial improvement of the results. Best performance is obtained using our contrastive descriptor, in which case multifold training improves CorLoc from 29.7% to 38.8% and detection mAP from 15.5% to 22.4%. In the experiments below we always use our contrastive descriptor.

In our next experiment, we consider the performance in terms of CorLoc across the training iterations. In Figure 2 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. Our multi-fold MIL approach leads to substantially better per-

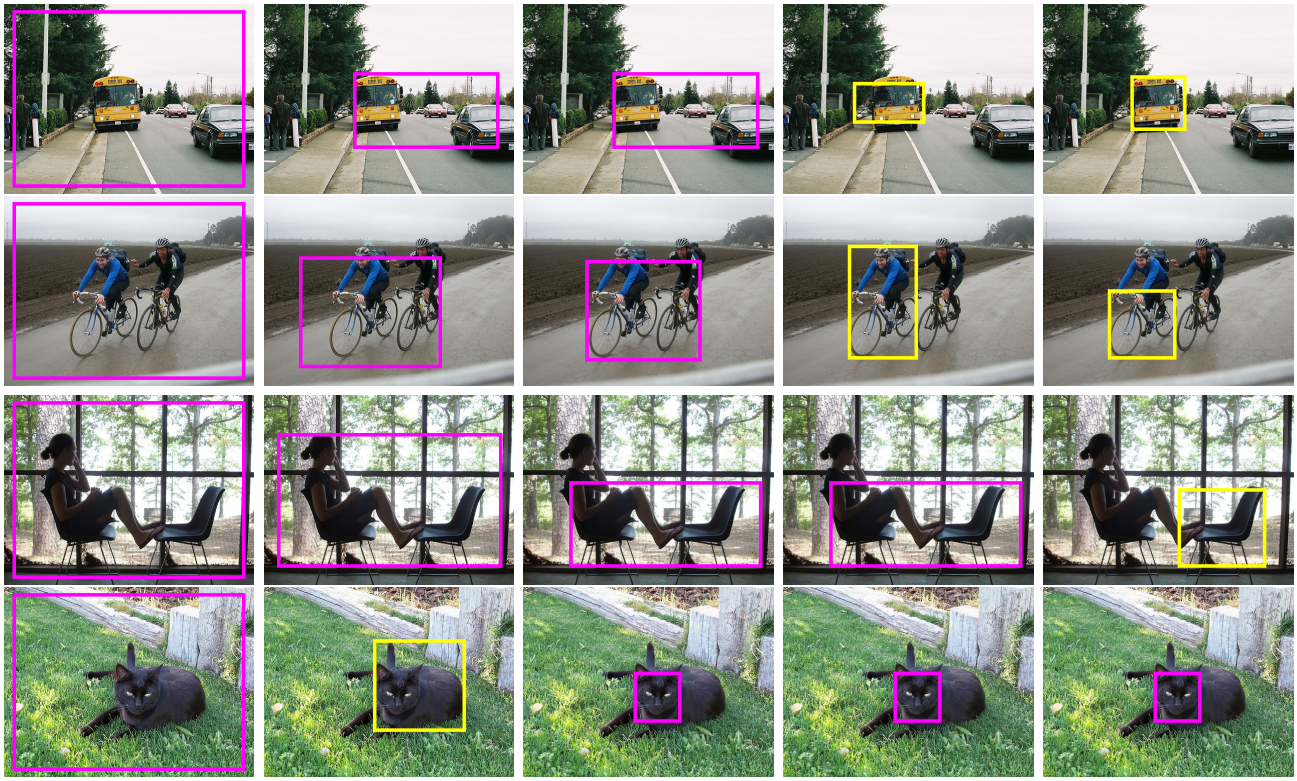

Figure 3: Examples of the re-localization process for images of four classes from initialization (left) to the final localization (right) and three intermediate iterations. The top three rows show examples of successful re-localization, the last one shows a failure case. Correct localizations are shown in yellow, incorrect ones in pink. This figure is best viewed in color.

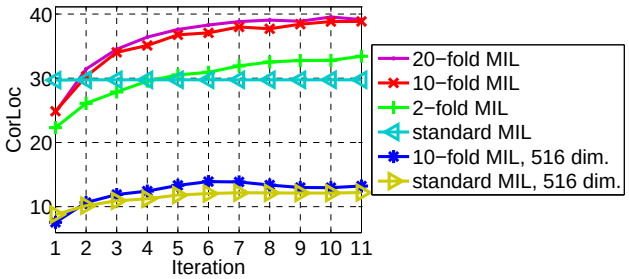

Figure 2: Correct localization (CorLoc) over the MIL iterations, averaged across classes. We show results for standard MIL training and our multi-fold training algorithm.

formance, 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.

In Figure 2, we also include experimental results with the 516 dimensional FV obtained using a 4-component GMM, 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 Figure 2 show that for this descriptor we obtain non-degenerate re-localization. The performance is, however, poor since the low dimensionality necessarily limits the capacity of the classifier. Our multi-fold MIL approach, instead, allows the use of high-dimensional features without suffering from degenerate re-localizations. In the low-dimensional case multi-fold training still helps, but to a much smaller extent since standard MIL is already non-degenerate in this case.

In Figure 3 we illustrate the re-localization performance for our multi-fold MIL algorithm with high-dimensional FVs. The success cases demonstrate the progressive improvement of the models over the MIL iterations, and the ability to correctly handle cases with multiple instances that appear in near proximity. The failure case for the cat image shows an inherent difficulty of WSL for object detection: 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 large deformability of the body, the face turns out to be the most distinctive and reliably detected structure, and this is what the detector learns. Parkhi *et al*. [21] recognized this issue, and addressed it using a segmentation-based method in a fully supervised object detection setting. Potentially, their method applies to WSL for object detection too; we plan to explore this in the future.

	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 [20]	50.9	56.7	—	10.6	0	**56.6**	—	—	2.5	—	14.3	—	50.0	53.5	11.2	5.0	—	**34.9**	33.0	40.6	—
Siva *et al*.'12 [27]	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
Siva and Xiang'11 [29]	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	9.4	16.7	32.3	54.8	5.5	30.4
Siva *et al*.'13 [28]	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	32.0
Shi *et al*.'13 [25]	**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
Ours: multi-fold MIL, 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**

Table 2: Comparison against state-of-the-art weakly-supervised detectors on PASCAL VOC 2007 in terms of correct localization on postive training images (CorLoc). The results for [20] were obtained through personal communication.

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	mAP
Pandey and Lazebnik'11 [20]	11.5	—	—	3.0	—	—	—	—	—	—	—	—	20.3	9.1	—	—	—	—	13.2	—	—
Prest *et al*.'12 [22]	17.4	—	—	**9.2**	—	—	—	—	—	—	—	—	16.2	27.3	—	—	—	—	15.0	—	—
Russakovsky *et al*.'12 [23]	30.8	25.0	—	3.6	—	26.0	—	—	—	—	—	—	21.3	29.9	—	—	—	—	—	—	15.0
Siva and Xiang'11 [29]	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
Ours: multi-fold MIL, 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**

Table 3: Comparison of weakly-supervised object detectors on PASCAL VOC 2007 in terms of test-set detection AP. The results of [22] are based on external video data for training. The results for [20] are taken from [22].

Our results on VOC'10 follow those for VOC'07. Compared to standard MIL training, multi-fold MIL training increases average CorLoc from 36.4% to 41.6%, and the detection AP from 16.4% to 18.5%.

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

We now compare the results of our multi-fold MIL approach to the state of the art. The evaluation results in Table 2 show that our multi-fold MIL training procedure leads to the best CorLoc value of 38.8% on average, as well as on 10 of the 20 classes. Compared to the 36.2% by Shi *et al*. [25], we improve by 2.6% to 38.8%, and improve over their results on 13 of the 20 classes. Pandey and Lazebnik [20] reported results on only 14 classes; for 11 of those our CorLoc values are higher than theirs. Our baseline result of 29.7% CorLoc in Table 1 for standard MIL training, is comparable to the results of Siva *et al*. [27, 28, 29].

In Table 3 we compare to the state of the art in terms of detection AP on the test set. Only two recent weakly supervised methods [23, 29] were evaluated on the VOC 2007 test set. Russakovsky *et al*. [23] provides mAP over all 20 classes, but reports separate AP values for only six classes. Other related work, *e.g*. [11], was evaluated only under simplified conditions, such as using viewpoint information and using images from a limited number of classes. Our multi-fold MIL detection mAP of 22.4% is significantly better than the 13.9% by Siva *et al*. [29], and the 15.0% by Russakovsky *et al*. [23]. Our result of 15.5% from Table 1 obtained with standard MIL training is close to the result of 15.0% by Russakovsky *et al*. [23]. For per-class comparison we included results for five classes provided by Prest *et al*. [22] based on WSL from external videos, and their evaluation of models provided by Pandey and Lazebnik [20].

4.4. Discussion and analysis

In our first analysis, we consider the performance of our detector when progressively using more supervised information, in order to quantify the performance gap between weakly and fully supervised learning.

The most notable difficulty in WSL is that we have to determine the object locations in the positive training images. If, instead, 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, and we obtain a detection mAP of 30.8%. This is an improvement of 8.4 mAP points w.r.t. the 22.4% obtained with WSL.

We now consider the remaining factors that change between the fully supervised scenario and WSL. (i) WSL uses only one instance per positive training image. If we use the optimal candidate window for all instances performance does not change significantly. (ii) In WSL hard-negative mining is based on negative images only. If we also use positive images performance rises to 32.0% mAP. (iii) WSL is based on the candidate windows which might not align well with ground-truth objects. If the ground-truth windows are used instead, performance rises to 32.8%. (iv) WSL does not use positive training images marked as difficult or truncated. If these are added to the fully supervised training, performance rises to 35.4%. Using difficult and truncated images in the original WSL setting, however, it is detrimental; since these instances are hard to recover automatically.

Our fully-supervised detection result of 35.4% mAP, compares favorably to the 33.7% of DPMs [16].This shows that our representation is reasonable, and that our WSL mAP of 22.4% achieves 63% compared to the fully-

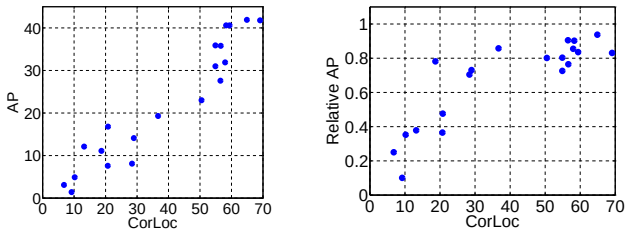

Figure 4: AP *vs*. CorLoc for multi-fold MIL (left), and ratio of WSL over supervised AP as a function of CorLoc (right).

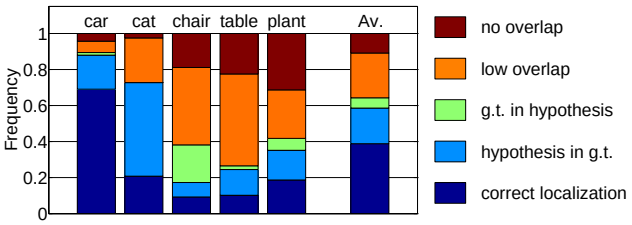

Figure 5: Distribution of localization error types for five example classes and averaged across all 20 VOC'07 classes for the positive training images.

supervised performance of 35.4% mAP.

In Figure 4 we analyze results of our weakly supervised detector, and the relation to those obtained with optimal localization. In the left panel, we visualize the relationship between the per-class CorLoc and AP values for our multi-fold MIL detector. The three classes with lowest CorLoc values are *bottle*, *chair*, and *dining table*. All of these 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*). In the right panel of Figure 4 we plot the ratio between our WSL detection AP (22.4% mAP) and the AP obtained with the same detector trained with optimal localization (30.8% mAP). In this case there is also a clear relation with our CorLoc values. The relation is quite different below and above 30% 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 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. In Figure 5 we show the frequency of these five cases, averaged over all classes, and for the five object categories that have maximum frequency for each of the five cases. As expected from Figure 3, for *cat* most localization hypotheses are fully contained within a ground-truth window. 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 mislocalization in the sense of the CorLoc measure.

4.5. Application to image classification

Since WSL requires image-wide labels only, the resulting object detectors can be used within a standard image classification paradigm. We consider two approaches for this purpose. The first one is *classification-by-detection*, where we use the maximum detection score as the image classification score. The second approach is *detection-driven* classification, where we use the top-scoring window as a data-driven and class-specific feature pooling region. Our detection-driven approach is easily integrated in most image classification methods. Below, we report the classification performance averaged over all classes in VOC 2007.

Using classification-by-detection, we obtain a mAP of 57.7%. Russakovsky *et al*. [23] obtained a similar classification-by-detection result of 57.2%. When we instead use a strong baseline image classification system with full image FV descriptors over SIFT and local color descriptors [24] and GMM vocabularies with 1,000 components, we get a mAP of 63.3%. The fact that the baseline full image descriptors performs significantly better than classification-by-detection underlines the importance of using contextual information and rich high-dimensional descriptors for image classification.

A common way of incorporating spatial information into image descriptors is adding a spatial pyramid (SPM), see [6] for a recent review. However, this may not always be an effective technique, especially for high dimensional image descriptors as in our case. Concatenating descriptors for cells in 3×1 and 2×2 grids with the full image descriptors improves performance only slightly to 63.4% mAP.

On the other hand, concatenating our detection-driven descriptors with full image descriptors significantly improves the performance to 65.6% mAP. This is a gain of 2.3 points over the baseline by adding one pooling region, where the seven rigid pooling regions of SPM only lead to a marginal improvement of 0.1 point. This shows that data-driven and class-specific pooling strategies have a larger potential than rigid pooling regions.

5. Conclusions

We presented a multi-fold multiple instance learning approach for weakly supervised object detection, which avoids the degenerate localization performance observed without it. Second, we presented a contrastive background descriptor, which forces the detection model to learn the differences between the objects and their context.

We evaluated our approach and compared it to state-of-the-art methods using the VOC 2007 dataset, a challenging benchmark for weakly-supervised detection. In terms of correct localization on the positive training images, we improve over the state of the art on 13 of the 20 classes, from 36.2% to 38.8% on average. Our results also improve the test set detection performance of state-of-the-art weakly-supervised methods. 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 train set localization performance, our approach attains 73% of the best performance that multiple instance learning could achieve using our image representation.

When using our weakly supervised detector for feature pooling in an FV-based image classification system, we obtain 65.6% mAP, which improves the baseline performance of 63.4% obtained using pooling across eight rigid regions.

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 *CVPR*, 2010.

[2] S. Bagon, O. Brostovski, M. Galun, and M. Irani. Detecting and sketching the common. In *CVPR*, 2010.

[3] 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 *CVPR*, 2004.

[4] M. Blaschko, A. Vedaldi, and A. Zisserman. Simultaneous object detection and ranking with weak supervision. In *NIPS*, 2010.

[5] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[6] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[7] Q. Chen, Z. Song, R. Feris, A. Datta, L. Cao, Z. Huang, and S. Yan. Efficient maximum appearance search for large-scale object detection. In *CVPR*, 2013.

[8] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007.

[9] R. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with Fisher vectors. In *ICCV*, 2013.

[10] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV*, 2006.

[11] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):257–293, 2012.

[12] 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.

[13] 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.

[14] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, June 2010.

[15] P. Felzenszwalb, R. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010.

[16] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5.

[17] C. Gu, P. Arbeláez, Y. Lin, K. Yu, and Malik. Multi-component models for object detection. In *ECCV*, 2012.

[18] C. Lampert, M. Blaschko, and T. Hofmann. Efficient subwindow search: a branch and bound framework for object localization. *PAMI*, 31(12):2129–2142, 2009.

[19] M. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009.

[20] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[21] O. Parkhi, A. Vedaldi, C. Jawahar, and A. Zisserman. The truth about cats and dogs. In *ICCV*, 2011.

[22] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[23] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

[24] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[25] Z. Shi, T. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *ICCV*, 2013.

[26] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[27] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*, 2012.

[28] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013.

[29] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV*, 2011.

[30] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2011.

[31] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

[32] J. Verbeek and B. Triggs. Region classification with Markov field aspect models. In *CVPR*, 2007.

[33] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, 2011.