Min-Entropy Latent Model for Weakly Supervised Object Detection

Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han and Qixiang Ye†

University of Chinese Academy of Sciences, Beijing, China

{wanfang13,weipengxu11}@mails.ucas.ac.cn, {jiaojb,hanzhj,qxye}@ucas.ac.cn

Abstract

Weakly supervised object detection is a challenging task when provided with image category supervision but required to learn, at the same time, object locations and object detectors. The inconsistency between the weak supervision and learning objectives introduces randomness to object locations and ambiguity to detectors. In this paper, a min-entropy latent model (MELM) is proposed for weakly supervised object detection. Min-entropy is used as a metric to measure the randomness of object localization during learning, as well as serving as a model to learn object locations. It aims to principally reduce the variance of positive instances and alleviate the ambiguity of detectors. MELM is deployed as two sub-models, which respectively discovers and localizes objects by minimizing the global and local entropy. MELM is unified with feature learning and optimized with a recurrent learning algorithm, which progressively transfers the weak supervision to object locations. Experiments demonstrate that MELM significantly improves the performance of weakly supervised detection, weakly supervised localization, and image classification, against the state-of-the-art approaches.

1. Introduction

Weakly supervised object detection (WSOD) solely requires image category annotations indicating the presence or absence of a class of objects in images, which significantly reduces human efforts when preparing training samples. Despite supervised object detection having become more reliable [10, 14, 15, 25, 26, 29, 30], WSOD remains an open problem, as often indicated by low detection rates of less than 50 percent for state-of-the-art approaches [12, 18, 33, 38]. Due to the lack of location annotations, WSOD approaches require learning latent objects from thousands of proposals in each image, as well as learning detectors that compromise the appearance of various objects in training images.

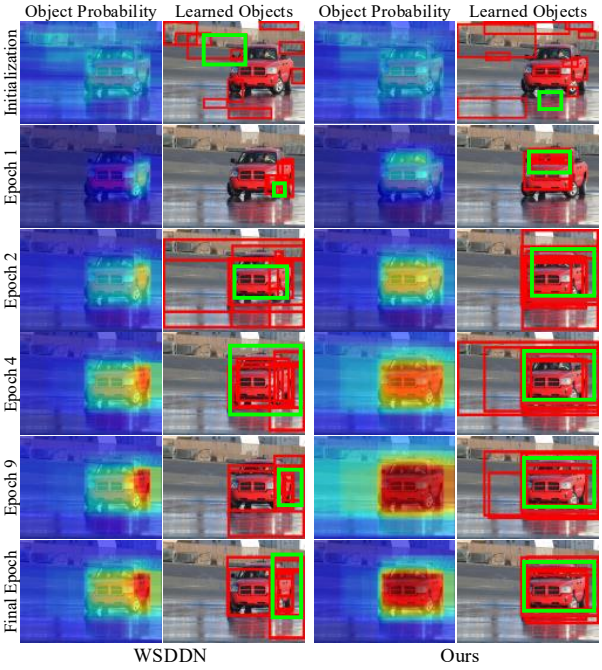

Figure 1: Evolution of object locations during learning (from top to bottom). Red boxes denote proposals of high object probability, and green ones detected objects. It shows that our approach reduces localization randomness and improves localization accuracy. Best viewed in color.

In the learning procedure of weakly supervised deep detection networks (WSDDN) [6], a representative WSOD approach, object locations evolved with great randomness, $e.g.$, switching among different object parts, Fig. 1. Various object parts were capable of optimizing the learning objective, $i.e.$, minimizing image classification loss, but experienced difficulty in optimizing object detectors due to their appearance ambiguity. The phenomenon resulted from the inconsistency between data annotations and learning objectives, $i.e.$, image-level annotations and object-level models. It typically requires introducing latent variables and solving non-convex optimization in vast solution spaces, $e.g.$, thousands of images and thousands of object proposals for

†Corresponding author.

each image, which might introduce sub-optimal solutions of considerable randomness. Recent approaches have used image segmentation [12, 24], context information [19], and instance classifier refinement [38] to empirically regularize the learning procedure. However, the issue of quantifying sub-optimal solutions and principally reducing localization randomness remains unsolved.

In this paper, we propose a min-entropy latent model (MELM) for weakly supervised object detection, motivated by a classical thermodynamic principle: *Minimizing entropy results in minimum randomness of a system.* Min-entropy is used as a metric to measure the randomness of object localization during learning, as well as serving as a model to learn object locations. To define the entropy, object proposals in an image are spatially separated into cliques, where spatial distributions and the probability of objects are jointly modeled. During the learning procedure, minimizing global entropy around all cliques discovers sparse proposals of high object probability, and minimizing local entropy for high-scored cliques identifies accurate object locations with minimum randomness. MELM is deployed as network branches concerning object discovery and object localization, Fig. 2, and is optimized with a recurrent stochastic gradient descent (SGD) algorithm, which progressively transfers the weak supervision, $i.e.$, image category annotations, to object locations. By accumulating multiple iterations, MELM discovers multiple object regions, if such exist, from a single image. The contributions of this paper include:

(1) A min-entropy latent model that effectively discovers latent objects and principally minimizes the localization randomness during weakly supervised learning.

(2) A recurrent learning algorithm that jointly optimizes image classifiers, object detectors, and deep features in a progressive manner.

(3) State-of-the-art performance of weakly supervised detection, localization, and image classification.

2. Related Work

WSOD problems are often solved with a pipelined approach, $i.e.$, an object proposal method is first applied to decompose images into object proposals, with which latent variable learning [4, 5, 36, 37, 41, 45] or multiple instance learning [2, 8, 9, 16, 42] is used to iteratively perform proposal selection and classifier estimation. With the widespread acceptance of deep learning, pipelined approaches have been evolving into multiple instance learning networks [6, 12, 17–19, 23, 28, 31, 33, 34, 38, 43, 46].

Latent Variable Learning. Latent SVM [44, 45] learns object locations and object detectors using an EM-like optimization algorithm. Probabilistic Latent Semantic Analysis (pLSA) [40, 41] learns object locations in a semantic clustering space. Clustering methods [5, 37] identify latent objects by discovering the most discriminative clusters. Entropy is employed in the latent variable methods [7, 27], but not considering the spatial relations among locations and the network fine-tuning for object detection. Various latent variable methods are required to solve the non-convex optimization problem. They often become stuck in a poor local minimum during learning, $e.g.$, falsely localizing object parts or backgrounds. To pursue a stronger minimum, object symmetry and class mutual exclusion information [4], Nesterov's smoothing [36], and convex clustering [5] have been introduced to the optimization function. These approaches can be regarded as regularization which enforces the appearance similarity among objects and reduces the ambiguity of detectors.

Multiple Instance Learning (MIL). A major approach for tackling WSOD is to formulate it as an MIL problem [2], which treats each training image as a "bag" and iteratively selects high-scored instances from each bag when learning detectors. When facing large-scale datasets, however, MIL remains puzzled by random poor solutions. The multi-fold MIL [8, 9] uses division of a training set and cross validation to reduce the randomness and thereby prevents training from prematurely locking onto erroneous solutions. Hoffman $et\ al.$ [16] train detectors with weakly annotations while transferring representations from extra object classes using full supervision (bounding-box annotation) and joint optimization. To reduce the randomness of positive instances, a bag splitting strategy has been used during the optimization procedure of MILinear [31].

Deep Multiple Instance Learning Networks. MIL has been updated to deep multiple instance learning networks [6, 38], where the convolutional filters behave as detectors to activate regions of interest on the deep feature maps [20, 22, 32]. The beam search [3] has been used to detect and localize objects by leveraging spatial distributions and informative patterns captured in the convolutional layers. To alleviate the non-convexity problem, Li $et\ al.$ [23] have adopted progressive optimization as regularized loss functions. Tang $et\ al.$ [38] propose to refine instance classifiers online by propagating instance labels to spatially overlapped instances. Diba $et\ al.$ [12] propose weakly supervised cascaded convolutional networks (WCCN) with multiple learning stages. It learns to produce a class activation map and candidate object locations based on image-level supervision, and then selects the best object locations among the candidates by minimizing the segmentation loss.

Deep multiple instance learning networks [12, 18, 38] report state-of-the-art WSOD performance, but are misled by the problem of inconsistency between data annotations (image-level) and learning objectives (object-level). With image-level annotations, such networks are capable of learning effective image representations for image classification. Without object bounding-box annotations, how-

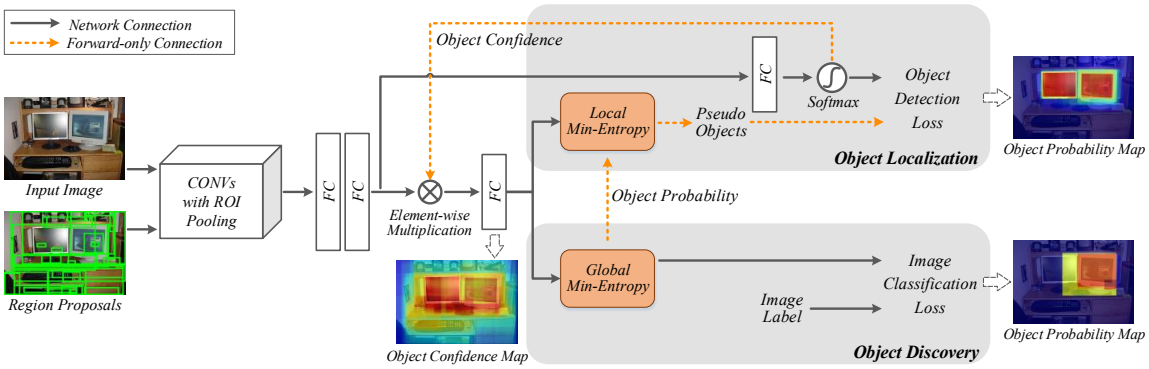

Figure 2: The proposed min-entropy latent model (MELM) is deployed as object discovery and object localization branches, which are unified with deep feature learning and optimized with a recurrent learning algorithm.

ever, their localization ability is very limited. The convolutional filters learned with image-level supervision incorporate redundant patterns, $e.g.$, object parts and backgrounds, which cause localization randomness and model ambiguity. Recent methods have empirically used object segmentation [12] and spatial label propagation [38] to solve these issues. In this paper, we provide a more effective and principled way by introducing global and local entropy as a randomness metric.

3. Methodology

Given image-level annotations, $i.e.$, the presence or absence of a class of objects in images, the learning objective of our proposed MELM is to find a solution that disentangles object samples from noisy object proposals with minimum localization randomness. Accordingly, an overview of the proposed approach is presented, followed by formulation of the min-entropy latent model. We finally elaborate the recurrent learning algorithm for model optimization.

3.1. Overview

The proposed approach is implemented with an end-to-end deep convolutional neural network, with two network branches added on top of the fully-connected (FC) layers, Fig. 2. The first network branch, designated as the *object discovery* branch, has a global min-entropy layer, which defines the distribution of object probability and targets at finding candidate object cliques by optimizing the global entropy and the image classification loss. The second branch, designated as the *object localization* branch, has a local min-entropy layer and a soft-max layer. The local min-entropy layer classifies the object candidates in a clique into pseudo objects and hard negatives by optimizing the local entropy and pseudo object detection loss.

In the learning phase, object proposals are generated with the Selective Search method [39] for each image. An ROI-pooling layer atop the last convolutional layer is used for efficient feature extraction for these proposals. The min-entropy latent models are optimized with a recurrently learning algorithm, which uses forward propagation to select sparse proposals as object instances, and back-propagation to optimize the parameters in the object localization branches. The object probability of each proposal is recurrently aggregated by being multiplied with the object probability learned in the preceding iteration. In the detection phase, the learned object detectors, $i.e.$, the parameters for the soft-max and FC layers, are used to classify proposals and localize objects.

3.2. Min-Entropy Latent Model

Modeling. Let $x \in \mathcal{X}$ denote an image, $y \in \mathcal{Y}$ denote the label indicating whether x contains an object or not, where $\mathcal{Y} = \{1, 0\}$. $y = 1$ indicates that there is at least one object in the image (positive image) while $y = 0$ indicates an image without any object (negative image). h denoting object locations is a latent variable and $\mathcal{H}$ denoting object proposals in an image is the solution space. θ denotes the network parameters. The min-entropy latent model, with object locations h^* and network parameters θ^* to be learned, is defined as

$$\begin{aligned} \{h^*, \theta^*\} &= \arg\min_{h,\theta} E_{(\mathcal{X},\mathcal{Y})}(h, \theta) \\ &= \arg\min_{h,\theta} E_d(h, \theta) + E_l(h, \theta) \\ &\Leftrightarrow \arg\min_{h,\theta} L_d + L_l, \end{aligned} \tag{1}$$

where $E_d(h, \theta)$ and $E_l(h, \theta)$ are the global and local entropy models.[1] They are respectively optimized by the loss function L_d and L_l in the object discovery and the object localization branch, Fig. 2. (h, θ) and $(\mathcal{X}, \mathcal{Y})$ in Eq. 1 are omitted for short.

Object Discovery. The object discovery procedure is implemented by selecting those object proposals which best discriminate positive images from negative ones. Accord-

[1]The entropy here is Aczél and Daróczy (AD) entropy [1].

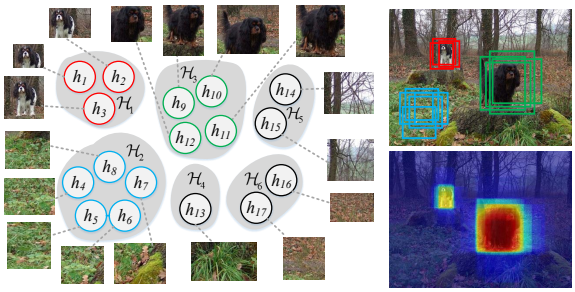

Figure 3: Object proposal cliques. The left column is an exemplar clique partition. The right-top shows some of the corresponding cliques in the image. The right-bottom shows the object confidence map of cliques.

ingly, a global min-entropy latent model $E_d(h,\theta)$, is defined to model the probability and the spatial distribution of object probability, as

$$\begin{aligned} E_d(h,\theta) &= -\log \sum_c w_{\mathcal{H}_c} p_{\mathcal{H}_c} \\ &= -\log \sum_c w_{\mathcal{H}_c} \sum_{h\in\mathcal{H}_c} p(y,h;\theta), \end{aligned} \tag{2}$$

where $p(y,h;\theta)$ is the joint probability of class y and latent variable h, given network parameters θ. It is calculated on the object confidences $s(y,\phi_h;\theta)$ with a soft-max operation, as

$$p(y,h;\theta) = \frac{\exp\left(s\left(y,\phi_h;\theta\right)\right)}{\sum_{y,h}\exp\left(s\left(y,\phi_h;\theta\right)\right)}, \tag{3}$$

where ϕ_h is the feature of object proposal h and $s(\cdot)$ denotes object confidence for a proposal computed by the last FC layer in the object discovery branch. $w_{\mathcal{H}_c}$, defined as

$$w_{\mathcal{H}_c} = 1/\left|\mathcal{H}_c\right| \sum\nolimits_{h\in\mathcal{H}_c} \left(p(y,h;\theta) / \sum\nolimits_y p(y,h;\theta)\right), \tag{4}$$

measures the probability distribution of objects to all image classes in a spatial clique $\mathcal{H}_c$, Fig. 3. $|\cdot|$ calculates the number of elements in a clique.

The spatial cliques $c, c' \in \{1,...,C\}$ are the minimum sufficient cover to an image, *i.e.*, $\bigcup_{c=1}^{C} \mathcal{H}_c = \mathcal{H}$ and $\forall c \neq c', \mathcal{H}_c \cap \mathcal{H}_{c'} = \emptyset$. To construct the cliques, the proposals are sorted by their object confidences and the following two steps are iteratively performed: 1) Construct a clique using the proposal of highest object confidence but not belonging to any clique. 2) Find the proposals that overlap with a proposal in the clique larger than a threshold (0.7 in this work) and merge them into the clique.

Eq. 2 and Eq. 3 show that minimizing the entropy $E_d(h,\theta)$ for the positive images maximizes $p(y)$, which means that the learning procedure selects the proposals of largest object probability to minimize image classification loss. For the negative images, all of the proposals are background and are simply modeled via a fully supervised way. Eq. 4 shows that $w_{\mathcal{H}_c} \in [0,1]$ is positively correlated to object confidences of the positive class in a clique, but negatively correlated to confidences of all other classes. According to the property of entropy, minimizing Eq. 2 produces a sparse selection of cliques in which proposals have significant high probability to the positive class. This sparsity of cliques with high object class confidence $w_{\mathcal{H}_c}$ shows the reduction of the randomness of selected proposals.

In the learning procedure, $E_d(h,\theta)$ is minimized by optimizing both the parameters in the object discovery branch and the parameters in the convolutional layers in an end-to-end manner. To implement this, an SGD algorithm is used, and the loss function is defined as

$$L_d = yE_d(h,\theta) - (1-y)\sum_h \log\left(1 - p(y,h;\theta)\right). \tag{5}$$

For positive images, $y = 1$, the second term of Eq. 5 is zero and only E_d is optimized. For negative images, $y = 0$, the first term of Eq. 5 is zero and the second term, image classification loss, is optimized.

Object Localization. The proposals selected by the global min-entropy model constitute good initialization for object localization, but nonetheless incorporate random false positives, *e.g.*, objects or partial objects with backgrounds. That is a consequence of the learning objective of the object discovery branch selecting those object proposals which best discriminate positive images from negative ones, but ignoring the localization of objects. A local min-entropy model is therefore defined for accurate object localization, as

$$E_l(h,\theta) = -\log \max_{h\in\mathcal{H}_c^*} w_h \cdot p(y,h;\theta), \tag{6}$$

where $\mathcal{H}_c^*$ denotes the clique of the highest average object confidence. $w_h = p(y,h;\theta)/\sum_y p(y,h;\theta)$ measures the distribution of object confidences to all image classes. Optimizing Eq. 6 produces maximum w_h and sparse object proposals of high object probability $p(y,h;\theta)$, and depresses negative proposals in a clique. With optimization results, the object proposals in a clique are classified into either pseudo objects h^* or hard negatives by a thresholding method, as

$$p(y,h^*;\theta) = \begin{cases} 1 & if \quad p(y,h^*;\theta) > \tau \\ 0 & otherwise \end{cases}, \tag{7}$$

where $\tau = 0.6$ is an empirically set threshold.

With pseudo objects and hard negatives, a object detector is learned by using the loss function defined as

$$L_l = \sum_{h^*} -\log f(h^*,\theta_l), \tag{8}$$

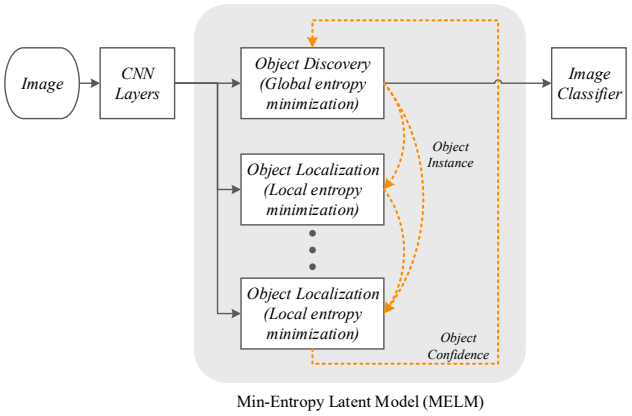

Figure 4: The flowchart of the proposed recurrent learning algorithm. The black solid lines denote network connections and orange dotted lines denote forward-only connections.

where $f(\cdot)$ denotes the object detectors with the parameters θ_l of the FC layer and soft-max layer in the object localization branch, Fig. 2.

3.3. Model Learning

In MELM, the object discovery branch learns potential objects by optimizing a min-entropy latent model using image category supervision, while the object localization branch learns object classifiers using estimated pseudo objects. The objective of model learning is to transfer the image category to object locations with min-entropy constraints, $i.e.$, minimum localization randomness.

Recurrent Learning. A recurrent learning algorithm is implemented to transfer the image-level (weak) supervision using an end-to-end forward- and back-propagation procedure. In a feed-forward procedure, the min-entropy latent models discover and localize objects which are used as pseudo-annotations for object detector learning with a back-propagation. With the learned detectors the object localization branch assigns all proposals new object probability, which is used to aggregate the object confidences with an element-wise multiply operator in the next learning iteration, Fig. 2. In the back-propagation procedure, the object discovery and object localization branches are jointly optimized with an SGD algorithm, which propagates gradients generated with image classification loss and pseudo-object detection loss. With forward- and back-propagation procedures, the network parameters are updated and the classification models and object detectors are mutually enforced. The recurrent learning algorithm is described in Alg. 1.

Accumulated Recurrent Learning. According to Eq. 6, the object localization model also performs object discovery, which may find objects different from those discovered by the object discovery model. This work extends recurrent learning to accumulated recurrent learning, Fig. 4, which accumulates different objects from both the object discovery and object localization branches, and uses them to learn object classifiers. Doing so endows this approach with the capability to localize multiple objects in a single image but also provides the robustness to process object appearance diversity by using multiple detectors.

Algorithm 1 Recurrent Learning

Input: Image $x \in \mathcal{X}$, image label $y \in \mathcal{Y}$, and region proposals $h \in \mathcal{H}$
Output: Network parameters θ and object detectors θ_l

Initialize object confidence $s(h) = s(y, \phi_h; \theta) = 1$ for all h
for i = 1 **to** $MaxIter$ **do**
  $\phi_h \leftarrow$ Compute deep features for all h through forward confidence
  $\phi_h^{'} \leftarrow \phi_h * s(h)$, aggregate features by object confidence
  Object discovery:
    $\mathcal{H}_c^* \leftarrow$ Optimize E_d using Eq. 2
    $L_d \leftarrow$ Compute classification loss using Eq. 5
  Object localization:
    $h^* \leftarrow$ Optimize E_l using Eq. 6
    $L_l \leftarrow$ Compute detection loss using Eq. 8
  Network parameter update:
    $\theta, \theta_l \leftarrow$ Back-propagation by using loss L_d and L_l
  $s(h) \leftarrow$ Update object confidence using detectors θ_l
end for

4. Experiments

The proposed MELM was evaluated on the PASCAL VOC 2007 and 2012 datasets using mean average precision (mAP) [13]. Following is a description of the experimental settings, and the evaluation of the effect of min-entropy models with randomness analysis and ablation experiments. The proposed MELM is then compared with the state-of-the-art approaches.

4.1. Experimental Settings

MELM was implemented based on the widely used VGG16 CNN model [35] pre-trained on the ILSVRC 2012 dataset [21]. As the conventional object detection task [18,31], we used Selective Search [39] to extract about 2000 object proposals for each image, removing those whose width or height was less than 20 pixels.

The input images were re-sized into 5 scales {480, 576, 688, 864, 1200} with respect to the larger side, height or width. The scale of a training image was randomly selected and the image was randomly horizontal flipped. In this way, each test image was augmented into a total of 10 images [6, 12, 38]. For recurrent learning, we employed the SGD algorithm with momentum 0.9, weight decay 5e-4, and batch size 1. The model iterated 20 epochs where the learning rate was 5e-3 for the first 15 epochs and 5e-4 for

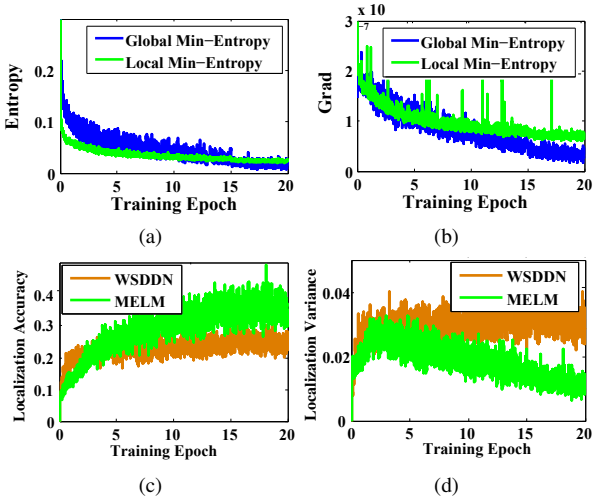

Figure 5: Localization, gradient, and entropy on the VOC 2007 dataset. (a) the evolution of entropy and (b) gradient. (c) localization accuracy and (d) localization variance.

the last 5 epochs. The output scores of each proposal from the 10 augmented images were averaged.

4.2. Randomness Analysis

Fig. 5a shows the evolution of global and local entropy, suggesting that our approach optimizes the min-entropy objective during learning. Fig. 5b provides the gradient evolution of the FC layers. In the early learning epochs, the gradient of the global min-entropy module was slightly larger than that of the local min-entropy module, suggesting that the network focused on optimizing the image classifiers. As learning proceeded, the gradient of the global min-entropy module decreased such that the local min-entropy module dominated the training of the network, indicating that the object detectors were being optimized.

To evaluate the effect of min-entropy, the randomness of object locations was evaluated with localization accuracy and localization variance. Localization accuracy was calculated by weighted averaging the overlaps between the ground-truth object boxes and the learned object boxes, by using $p(y, h; \theta)$ as the weight. Localization variance was also defined as the weighted variance of the overlaps by using $p(y, h; \theta)$ as the weight. Fig. 5c and Fig. 5d show that the proposed MELM had significantly greater localization accuracy and lower localization variance than WSDDN. This strongly indicates that our approach effectively reduces localization randomness during weakly supervised learning. Such an effect is further illustrated in Fig. 6, where the object locations learned by our approach were more accurate and less variant than those of WSDDN.

4.3. Ablation Experiments

Ablation experiments were used to evaluate the respective effects of the proposal cliques, the min-entropy model, and the recurrent learning algorithm.

Baseline. The baseline approach was derived by simplifying Eq. 2 to solely model the global entropy $E_d(h, \theta)$. This is similar to WSDDN without the spatial regulariser [6] where the only learning objective is to minimize the image classification loss. This baseline, referred to as "LOD-" in Tab. 1, achieved 24.7% mAP.

Clique Effect. By dividing the object proposals into cliques, the "LOD-" approach was promoted to "LOD". Tab. 1 shows that the introduction of spatial cliques improved the detection performance by 4.8% (from 24.7% to 29.5%). That occurred because using multiple cliques reduced the solution spaces of the latent variable learning, thus readily facilitating a better solution.

Multi-Entropy Latent Model. We denoted the multi-entropy model by "MELM-D" and "MELM-L" in Table 1, which respectively corresponded to object discovery and object localization. We trained the min-entropy latent model by simply cascading the object discovery and object localization branches, without using the recurrent optimization. Tab. 1 shows that MELM-L significantly improved the baseline LOD from 29.5% to 40.1%, with a 10.6% margin at most. This fully demonstrated that the min-entropy latent model and the implementation of object discovery and object localization branches were pillars of our approach.

Recurrent Learning. In Tab. 1, the proposed recurrent learning algorithm, "MELM-D+RL" and "MELM-L+RL", respectively achieves 34.5% and 42.6% mAP, improving the "MELM-L" (without recurrent learning) by 1.9% and 2.5%. This improvement showed that with recurrent learning and the object confidence accumulation, Fig. 2, the object discovery and object localization branches benefited from each other and thus were mutually enforced.

Accumulated Recurrent Learning. When using two accumulated object localization modules, the MELM, referred to as "MELM-L2-ARL", significantly improved the mAP of the "MELM-L-RL" from 42.6% to 46.4% (+3.8%). It further improved the mAP from 46.4% to 47.3% (+0.9%) when using three accumulated detectors, but did not significantly improve when using four detectors.

4.4. Performance and Comparison

Weakly Supervised Object Detection. Table 2 shows the detection results of our MELM approach and the state-of-the-art approaches on the PASCAL VOC 2007 dataset. MELM improved the state-of-the-art to 47.3% and respectively outperformed the OICR [38][2], Self-Taught [18], and

[2]This work reported a higher performance (47.0%) with multiple networks ensemble and Fast-RCNN re-training. For a fair comparison, the performance of OICR using a single VGG16 model is used.

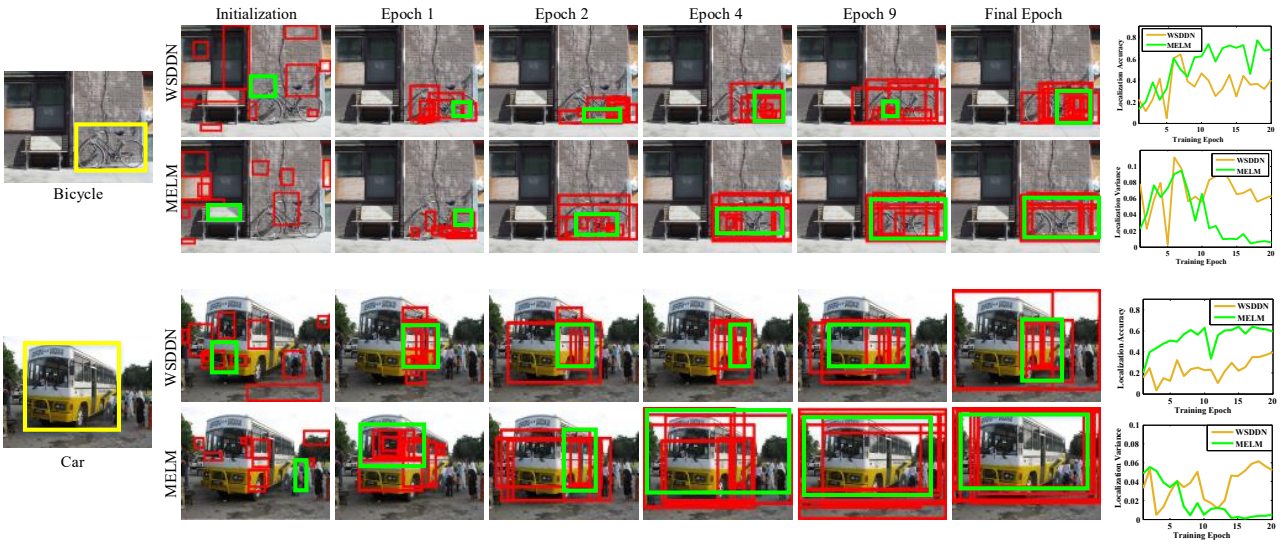

Figure 6: Comparison about object localization of our MELM to WSDDN [6]. The red solid boxes denote objects of high probability and the green solid boxes denote the detected objects. The yellow boxes in the first column denote ground-truth locations. It can be seen that the objects learned by MELM are more accurate and have less randomness, which is quantified by the localization accuracy and localization variance in the last column. Best viewed in color.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
LOD-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.7
LOD	32.2	49.6	15.9	8.1	5.0	51.1	44.8	22.3	16.6	35.3	24.0	20.4	31.0	57.1	9.8	15.3	30.9	31.7	50.1	37.8	29.5
MELM-D	36.3	47.1	19.7	13.4	3.1	61.4	52.6	12.8	13.9	40.5	33.3	12.6	29.6	62.1	**10.1**	17.5	35.0	48.7	60.4	41.3	32.6
MELM-L	49.5	54.4	26.2	19.7	12.9	59.4	63.0	39.2	22.3	46.9	39.1	36.2	43.2	64.2	2.6	21.3	40.1	48.9	57.9	54.4	40.1
MELM-D+RL	37.4	56.8	27.4	13.1	4.4	59.2	52.0	25.8	20.3	41.5	33.1	21.3	32.8	60.0	10.0	11.6	35.7	43.6	57.2	47.3	34.5
MELM-L+RL	50.4	57.6	**37.7**	23.2	13.9	60.2	63.1	**44.4**	24.3	52.0	42.3	42.7	43.7	66.6	2.9	21.4	45.1	45.2	59.1	56.2	42.6
MELM-D+ARL	42.1	61.2	26.5	17.3	7.8	61.4	55.6	20.2	21.3	46.3	35.3	36.7	37.0	63.1	1.2	18.7	38.9	52.0	57.8	48.0	37.4
MELM-L1+ARL	51.3	**66.9**	36.1	28.1	15.5	68.6	67.1	37.3	24.8	65.2	45.1	50.7	46.9	67.5	2.1	25.3	51.3	56.4	**62.9**	**59.0**	46.4
MELM-L2+ARL	**55.6**	**66.9**	34.2	**29.1**	**16.4**	**68.8**	**68.1**	43.0	**25.0**	**65.6**	**45.3**	**53.2**	**49.6**	**68.6**	2.0	**25.4**	**52.5**	**56.8**	62.1	57.1	**47.3**

Table 1: Detection average precision (%) on the PASCAL VOC 2007 test set. Ablation experimental results of MELM.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
MILinear [31]	41.3	39.7	22.1	9.5	3.9	41.0	45.0	19.1	1.0	34.0	16.0	21.3	32.5	43.4	**21.9**	19.7	21.5	22.3	36.0	18.0	25.4
Multi-fold MIL [9]	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
LCL+Context [40]	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
WSDDN [6]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
PDA [23]	54.5	47.4	**41.3**	20.8	**17.7**	51.9	63.5	46.1	21.8	57.1	22.1	34.4	**50.5**	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
OICR [38][2]	**58.0**	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	**64.3**	**62.6**	41.2
Self-Taught [18]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	**54.3**	3.0	53.6	24.7	43.6	48.4	65.8	6.6	18.8	51.9	43.6	53.6	62.4	41.7
WCCN [12]	49.5	60.6	38.6	**29.2**	16.2	**70.8**	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	42.8
MELM	55.6	**66.9**	34.2	29.1	16.4	68.8	**68.1**	43.0	**25.0**	**65.6**	**45.3**	**53.2**	49.6	**68.6**	2.0	25.4	**52.5**	**56.8**	62.1	57.1	**47.3**

Table 2: Detection average precision (%) on the PASCAL VOC 2007 test set. Comparison of MELM to the state-of-the-arts.

WCCN [12] by 6.1%, 5.6%, and 4.5%. In Table 3, the detection comparison results on the PASCAL VOC 2012 datasets are provided. MELM respectively outperformed the OICR [38], Self-Taught [18], and WCCN [12] by 4.5%, 4.1%, and 4.1%, which were significant margins for the challenging WSOD task.

Specifically, on the "bike", "car", "chair", and "cow" classes, MELM outperformed the state-of-the-art WCCN approach up to 6∼21%. Despite of the average good performance, our approach failed on the "person" class, as shown in the last image of Fig. 7. This may have been because with a large appearance variation existing in person instances, it is difficult to learn a common appearance model. The "faces" that represent the person class with minimum randomness were falsely localized.

Fig. 7 shows some of the detection results of the MELM

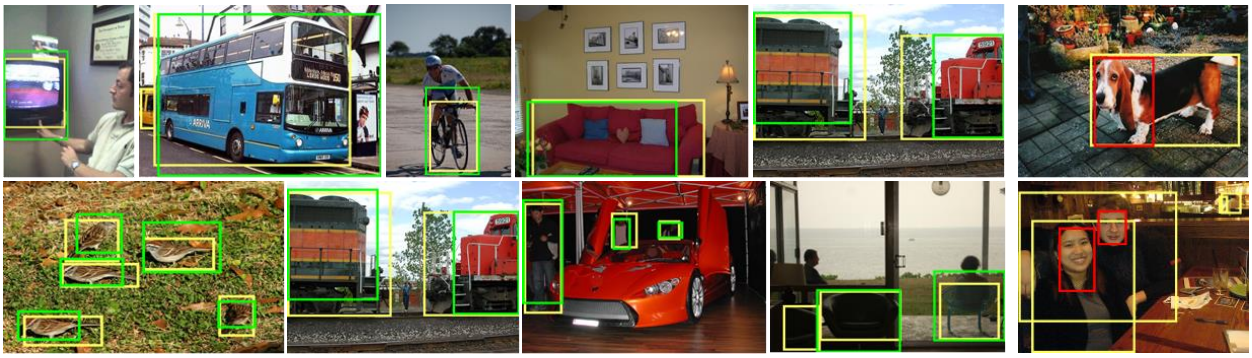

Figure 7: Examples of our object detection results. Yellow bounding boxes are ground-truth annotations. Green boxes and red boxes are positive and negative detection results respectively. Images are sampled from PASCAL VOC 2012 test set.

Method	Dataset Splitting	mAP
MILinear [31]	train/val	23.8
PDA [23]	train/val	29.1
Self-Taught [18]	train/val	39.0
ContextNet [19]	trainval/test	35.3
WCCN [12]	trainval/test	37.9
OICR [38]	trainval/test	37.9
Self-Taught [18]	trainval/test	38.3
MELM	train/val	**40.2**
MELM	trainval/test	**42.4**

Table 3: Detection average precision (%) on the VOC 2012 test set. Comparison of MELM to the state-of-the-arts.

Method	Localization (mAP)	Classification (mAP)
MILinear [31]	43.9	72.0
LCL+Context [40]	48.5	-
PDA [23]	52.4	-
VGG16 [35]	-	89.3
WSDDN [6]	53.5	89.7
Multi-fold MIL [9]	54.2	-
ContextNet [19]	55.1	-
WCCN [12]	56.7	90.9
MELM	**61.4**	**93.1**

Table 4: Correct localization rate (%) and image classification average precision (%) on PASCAL VOC 2007. Comparison of MELM to the state-of-the-arts.

approach. By accumulating proposals of high confidences, MELM localized multiple object regions and therefore learned more discriminative detectors.

Weakly Supervised Object Localization. The Correct Localization (CorLoc) metric [11] was employed to evaluate the localization accuracy. CorLoc is the percentage of images for which the region of highest object confidence has at least 0.5 interaction-over-union (IoU) with one of the ground-truth object regions. This experiment was done on the $trainval$ set because the region selection exclusively worked in the training process. Tab. 4 shows that the mean CorLoc of MELM outperformed the state-of-the-art WCCN [12] by 4.7% (61.4% vs. 56.7%). This shows that the min-entropy strategy used in our approach was more effective for object localization than the image segmentation strategy used in WCCN.

Image Classification. The object discovery and object localization functionality of MELM highlights informative regions and suppresses disturbing backgrounds, which also benefits the image classification task. As shown in Tab. 4, with the VGG16 model, MELM achieved 93.1% mAP, which respectively outperformed WSDDN [6] and WCCN [12] up to 3.4% and 2.2%. It is noteworthy that MELM outperforms the VGG16 network, which was specifically trained for image classification, by 3.8% mAP (93.1% vs. 89.3%). This shows that the min-entropy latent model learned more representative feature representations by reducing the localization randomness of informative regions.

5. Conclusion

In this paper, we proposed a simple but effective min-entropy latent model (MELM) for weakly supervised object detection. MELM was deployed as two submodels of object discovery and object localization, and was unified with the deep learning framework in an end-to-end manner. Our approach, by leveraging the sparsity produced with a min-entropy model, provides a new way to learn latent object regions. With the well-designed recurrent learning algorithm, MELM significantly improves the performance of weakly supervised detection, weakly supervised localization, and image classification, in striking contrast with state-of-the-art approaches. The underlying reality is that min-entropy results in minimum randomness of an information system, which provides fresh insights for weakly supervised learning problems.

Acknowledgements: This work is partially supported by the NSFC under Grant 61671427, 61771447, 61601466, and Beijing Municipal Science and Technology Commission.

References

[1] J. Aczél and Z. Daróczy. Charakterisierung der entropien positiver ordnung und der shannonschen entropie. *Acta Mathematica Academiae Scientiarum Hungarica*, 14(1-2):95–121, 1963.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Adv. in Neural Inf. Process. Syst. (NIPS)*, pages 561–568, 2002.

[3] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. Manjunath. Weakly supervised localization using deep feature maps. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 714–731. Springer, 2016.

[4] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *Brit. Mach. Vis. Conf. (BMVC)*, pages 1997–2005, 2014.

[5] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1081–1089, 2015.

[6] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2846–2854, 2016.

[7] D. Bouchacourt, S. Nowozin, and M. Pawan Kumar. Entropy-based latent structured output prediction. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 2920–2928, 2015.

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object lcalization. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshop*, pages 2409–2416, 2014.

[9] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(1):189–203, 2016.

[10] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Adv. in Neural Inf. Process. Syst. (NIPS)*, pages 379–387, 2016.

[11] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *Int. J. Comput. Vis*, 100(3):275–293, 2012.

[12] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool. Weakly supervised cascaded convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5131–5139, 2017.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis*, 88(2):303–338, 2010.

[14] R. Girshick. Fast r-cnn. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1440–1448, 2015.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 580–587, 2014.

[16] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, page 797823, 2015.

[17] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 340–353. Springer, 2012.

[18] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4294–4302, 2017.

[19] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 350–365. Springer, 2016.

[20] D. Kim, D. Cho, D. Yoo, and I. S. Kweon. Two-phase learning for weakly supervised object localization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. in Neural Inf. Process. Syst. (NIPS)*, pages 1097–1105, 2012.

[22] K. Kumar Singh and Y. Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[23] D. Li, J. B. Huang, Y. Li, S. Wang, and M. H. Yang. Weakly supervised object localization with progressive domain adaptation. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3512–3520, 2016.

[24] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Image colocalization by mimicking a good detectors confidence score distribution. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 19–34. Springer, 2016.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 936–944, 2017.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 21–37. Springer, 2016.

[27] K. Miller, M. P. Kumar, B. Packer, D. Goodman, and D. Koller. Max-margin min-entropy models. In *Artificial Intelligence and Statistics*, pages 779–787, 2012.

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? weakly supervised learning with convolutional neural networks. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 685–694, 2015.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 779–788, 2016.

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. in Neural Inf. Process. Syst. (NIPS)*, pages 91–99, 2015.

[31] W. Ren, K. Huang, D. Tao, and T. Tan. Weakly supervised large scale object localization with multiple instance learning and bag splitting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):405–416, 2016.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[33] M. Shi, H. Caesar, and V. Ferrari. Weakly supervised object localization using things and stuff transfer. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[34] M. Shi and V. Ferrari. Weakly supervised object localization using size estimates. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 105–121. Springer, 2016.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[36] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, pages 1611–1619, 2014.

[37] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly supervised discovery of visual pattern configurations. In *Adv. in Neural Inf. Process. Syst. (NIPS)*, pages 1637–1645, 2014.

[38] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3059–3067, 2017.

[39] J. R. Uijlings, K. E. Van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *Int. J. Comput. Vis*, 104(2):154–171, 2013.

[40] C. Wang, K. Huang, W. Ren, J. Zhang, and S. Maybank. Large-scale weakly supervised object localization via latent category learning. *IEEE Trans. Image Process.*, 24(4):1371–1385, 2015.

[41] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 431–445. Springer, 2014.

[42] X. Wang, Z. Zhu, C. Yao, and X. Bai. Relaxed multiple-instance svm with application to object discovery. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1224–1232, 2015.

[43] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3460–3469, 2015.

[44] Q. Ye, T. Zhang, Q. Qiu, B. Zhang, J. Chen, and G. Sapiro. Self-learning scene-specific pedestrian detectors using a progressive latent model. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2057–2066, 2017.

[45] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proc. 26st Int. Conf. Mach. Learn. (ICML)*, pages 1169–1176, 2009.

[46] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Soft proposal networks for weakly supervised object localization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1841–1850, 2017.