Proceedings Paper:
Htike, KK and Hogg, DC orcid.org/0000-0002-6125-9564 (2015) Weakly supervised pedestrian detector training by unsupervised prior learning and cue fusion in videos. In: 2014 IEEE International Conference on Image Processing (ICIP). 2014 IEEE International Conference on Image Processing (ICIP), 27-30 Oct 2014, Paris. IEEE , pp. 2338-2342.

https://doi.org/10.1109/ICIP.2014.7025474

WEAKLY SUPERVISED PEDESTRIAN DETECTOR TRAINING BY UNSUPERVISED PRIOR LEARNING AND CUE FUSION IN VIDEOS

Kyaw Kyaw Htike and David Hogg

University of Leeds
Leeds, UK

ABSTRACT

The growth in the amount of collected video data in the past decade necessitates automated video analysis for which pedestrian detection plays a key role. Training a pedestrian detector using supervised machine learning requires tedious manual annotation of pedestrians in the form of precise bounding boxes. In this paper, we propose a novel weakly supervised algorithm to train a pedestrian detector that only requires annotations of estimated centers of pedestrians instead of bounding boxes. Our algorithm makes use of a *pedestrian prior* learnt in an unsupervised way from the video and this prior is fused with the given weak supervision information in a principled manner. We show on publicly available datasets that our weakly supervised algorithm reduces the cost of manual annotation by over 4 times while achieving similar performance to a pedestrian detector trained with bounding box annotations.

Index Terms— Pedestrian detection, weak supervision, unsupervised prior, cue fusion.

1. INTRODUCTION

Pedestrian detection is often posed as a binary classification problem in which one class is pedestrians and the other is non-pedestrians. To detect pedestrians in an image, the trained classifier is used to score each image patch corresponding to the multi-scale sliding windows and the local modes of the score space give the locations and spatial extent of pedestrians in the image [1, 2, 3]. The most popular way to train a pedestrian detector is using supervised machine learning techniques which require groundtruth annotations of pedestrians. For most state-of-the-art research, this groundtruth annotation is typically given in the form of bounding boxes tightly fitting the pedestrians [4, 5, 6, 7, 8, 9, 10, 11]. However, manually annotating with bounding boxes can be time-consuming.

In this paper, we propose an algorithm for training pedestrian detectors for videos that requires a *weaker* form of supervision than bounding box annotations, namely, approximate center locations of pedestrians (as shown in Fig. 1). This allows for a much easier and faster annotation compared to bounding box annotation. Despite the weak supervision, our algorithm performs comparably with the bounding box supervision (termed in this paper as *strong supervision*) despite having a much lower cost (measured in terms of the time it takes to complete the annotation).

Fig. 1. Strong versus weak annotation (best viewed in color). On the left is the standard way of annotating pedestrians for training a pedestrian detector. On the right is the weak supervision (only approximate centers of pedestrians) required by our proposed algorithm. Note that pedestrians are of different sizes in the video due to projective distortion and hence our algorithm has to cope with *both* noisy locations and unknown scales. Weak supervision on the right is much faster and easier for a human annotator than the strong supervision on the left.

2. RELATED WORK

Compared to training object detectors using strong supervision, the literature concerning weakly supervised training is fairly limited. Furthermore, most of the literature on weakly supervised learning in images use a different setting than our proposed approach. In the existing approaches, supervision is given in the form of image-level labels where the exact locations and spatial extents of objects of interest are considered unknown and treated as latent variables to be inferred from data during training. One of the ways of solving this is by formulating it as a Multiple Instance Learning (MIL) problem [12, 13] in which supervision labels are given at the *bag* level rather than at the instance level. Each positive bag is assumed to contain at least one positive instance and each negative bag is assumed to contain all negative instances. In

order to generate positive bags and because the space of all possible object locations and sizes is too large to be tractable during training, many existing approaches use an ensemble of low-level segmentations to generate numerous candidate regions with the assumption that at least one of them contain the desired object [14, 15]. The performance of such a system, however, depends heavily on the results of segmentation.

Furthermore, in most existing approaches, datasets are assumed in which an object occupies a large central portion of each image in most of the training images [16, 17, 15, 14]. This is in contrast to our approach which is dealing with far-field videos where there are often multiple objects of varying sizes in each frame and each object occupies only a very small portion of the frame. Deselaers *et al*. [18] propose an iterative algorithm to learn object classes from weakly labelled images using a conditional random field that progressively adapts to the new classes. Chum and Zisserman [16] give an algorithm that locates image regions corresponding to object classes of a set of training images by optimizing an objective function that computes similarity between pairs of images. Considering classifier parameters and subwindows of objects jointly as latent variables in a SVM classification objective function, Nguyen *et al*. [17] optimizes the function to infer the variables. Weakly supervised learning is tackled as a structured output learning framework in [19]. All of the aforementioned approaches deal only with images and do not make use of information that can be exploited in surveillance-type videos. We summarize our key ***contributions*** as follows:

1. A weakly supervised training algorithm that makes use of (potentially noisy) center location annotation for training pedestrian detectors for videos.
2. Unsupervised learning of a pedestrian prior for a given video.
3. Combining cues from the unsupervised learnt prior and weak supervision in an optimization framework.
4. Our algorithm can work with low resolution videos that do not allow sophisticated part-based modelling and discovery and, that have multiple objects of varying sizes in each frame.
5. The algorithm is not sensitive to low-level segmentation unlike many state-of-the-art weak-supervision approaches using MIL.
6. Our approach is efficient since it does not require jointly solving all the weak supervisions.

3. OUR APPROACH

The overview of our algorithm is illustrated in Fig. 2. Let $\mathbf{V} = [I_1, I_2, \ldots, I_N]$ be a given video of N frames and let $\mathbf{C} \in \mathbb{R}^{M\times 3}$ be M given center annotations (on sampled frames of $\mathbf{V}$) stored in the form of a matrix. Each row of $\mathbf{C}$ is a vector $[n, x, y]$ containing the frame number n in $\mathbf{V}$ and the x and y coordinates of the weak supervision corresponding to the approximate center of a pedestrian. The goal is to obtain a

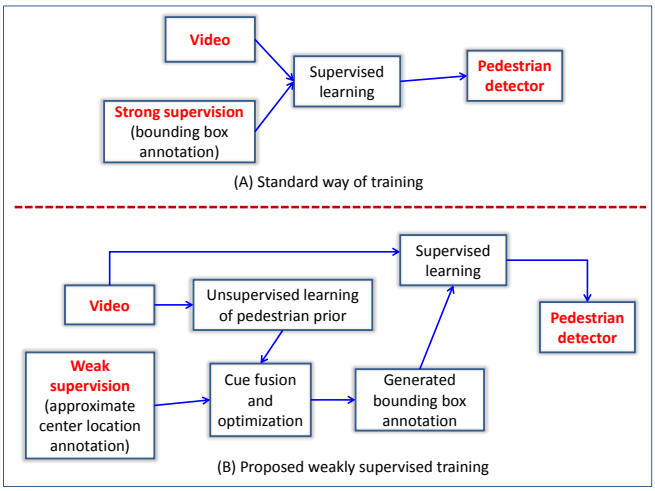

Fig. 2. (A) shows the standard way of training pedestrian detectors. In comparison, (B) illustrates the overview of our proposed algorithm.

pedestrian detector given only $\mathbf{C}$ without being provided any bounding box annotations (which the traditional supervised training requires). Our algorithm is made up of 3 stages.

In the 1st stage, we learn a *pedestrian prior* in an unsupervised way using knowledge that can be automatically extracted from $\mathbf{V}$. This knowledge comes in the following form: for any video captured with a static uncalibrated camera, the dynamic background of the scene can be effectively modelled. Although this model is usually noisy, by considering $\mathbf{V}$ as a whole, we can get some idea about foreground objects in the video and thus effectively build a distribution over objects in $\mathbf{V}$. Furthermore, we can easily slightly bias this model towards the pedestrian class by introducing a few simple constraints (detailed in Section 3.1). After obtaining this model, the pedestrian prior can be represented as $P(\text{pedestrian}|\text{patch})$, *i.e*. given any patch in $\mathbf{V}$, the pedestrian prior gives the *prior* probability[1] that the patch depicts a pedestrian. Due to noises and inaccuracies in the background modelling process, the pedestrian prior is error prone. However, we do not make any hard decisions at this stage and any errors and uncertainties in the pedestrian prior are resolved in the next stage.

The 2nd stage involves an optimization framework with an the objective function that is a mixture of two terms: (1) the score of the pedestrian prior obtained in the 1st stage and (2) the agreement with the centers $\mathbf{C}$. We perform the optimization independently for each center in $\mathbf{C}$, *i.e*. we process each row in $\mathbf{C}$ independently. Formulating in this way is very efficient compared to having to solve them jointly. After optimizing each weak supervision annotation (described in Section 3.2), we automatically obtain the bounding box annotations. Therefore, the 2nd stage is in essence automat-

[1]It is the probability or belief *before* seeing any (weak) supervision.

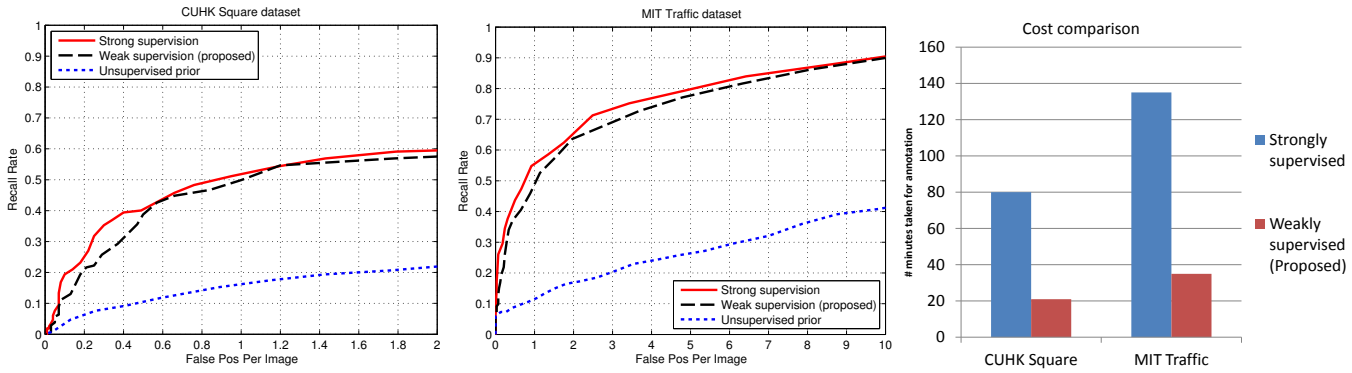

Fig. 3. Detection performance curves (left and middle) and cost comparison (right)

Algorithm 1 Overview of the weakly supervised training

Input: Video $\mathbf{V} = [I_1, I_2, \ldots, I_N]$ and weak supervision $\mathbf{C} \in \mathbb{R}^{M\times 3}$

Output: Pedestrian detector

1: $\mathbf{F} \leftarrow$ LearnPrior($\mathbf{V}$), where LearnPrior is the function to learn unsupervised pedestrian prior. Described in Algorithm 2.
2: $\mathbf{B} \leftarrow$ FuseOptimize($\mathbf{V}, \mathbf{C}, \mathbf{F}$), where FuseOptimize is the function to convert weak center supervision $\mathbf{C}$ to bounding box annotations $\mathbf{B}$. Detailed in Algorithm 3.
3: Train a pedestrian detector using $\mathbf{V}$ and $\mathbf{B}$ using any supervised learning algorithm.
4: **return** Pedestrian detector

Algorithm 2 Unsupervised pedestrian prior learning

Input: Video $\mathbf{V} = [I_1, I_2, \ldots, I_N]$

Output: Unsupervised pedestrian prior $\mathbf{F}$

$\mathcal{D}_p \leftarrow \varnothing$
$\mathcal{D}_n \leftarrow \varnothing$
Initialize background model G.
for $I_i \in \mathbf{V}$ **do**
  $BW \leftarrow$ background subtraction using I_i and G, where BW is a binary image.
  $R \leftarrow$ Connected Component Analysis on BW, where $R = \{r_1, r_2, \ldots, r_T\}$ is a set of T bounding boxes.
  Update G.
  for $j = 1$ **to** T **do**
    if height$(r_j) >$ width(r_j) **then**
      $\vec{d} \leftarrow$ compute feature vector on patch corresponding to the bounding box r_j.
      $\mathcal{D}_p \leftarrow \mathcal{D}_p \cup \{\vec{d}\}$
    end if
  end for
  $\mathcal{D}_q \leftarrow$ compute feature vectors from patches randomly sampled from regions not intersecting with R.
  $\mathcal{D}_n \leftarrow \mathcal{D}_n \cup \{\mathcal{D}_q\}$
end for
Train a binary classifier $\mathbf{F}$ on $\mathcal{D}_p$ and $\mathcal{D}_n$.
Calibrate $\mathbf{F}$ to produce valid probabilities.
return $\mathbf{F}$

ically converting the weak center annotations $\mathbf{C}$ to bounding box annotations $\mathbf{B} \in \mathbb{R}^{M\times 5}$ where each row of $\mathbf{B}$ is a vector $[\text{n}, x_1, y_1, x_2, y_2]$ denoting a bounding box with x_1 and y_1 representing the top left corner of the bounding box and x_2 and y_2 the bottom right corner.

After obtaining $\mathbf{B}$, we can now use any supervised learning algorithm to train a pedestrian detector. This is the 3rd stage. We formalize our approach in Algorithms 1-3 and give further descriptions in the coming sections.

3.1. Unsupervised pedestrian prior learning

For a given video $\mathbf{V}$, background subtraction is performed for each frame I_i and followed by Connected Component Analysis (CCA). Although any background subtraction technique could be used, since the unsupervised prior learning stage is offline and does not need real-time processing, a highly accurate and robust yet reasonably fast background subtraction algorithm (such as [20]) is recommended. The CCA gives a set of bounding boxes and for image patches corresponding to each of them, we compute features after appropriate resizing of each patch. The feature extraction is general and any suitable mechanism can be used. In this paper, Histograms of Oriented Gradients (HOGs) features [3] are used. In order to slightly *bias* the (unknown) multi-modal distribution of foreground classes (and any "noise" classes) towards pedestrian class, we perform a simple filtering by aspect ratio, discarding any bounding box whose height is less than its width. Our goal in this stage is *not* to cluster or discover foreground object classes. Instead, it is simply to capture some information about the pedestrian class (which does *not* require any object class discovery). We achieve this by training a 2-class classifier $\mathbf{F}$ in which the positive class is the set of features of the filtered bounding boxes and the negative training data comes from background regions. This implicitly captures the multi-modal distribution about objects in the scene and from an-

Algorithm 3 Cue Fusion and Optimization

Input: Video $\mathbf{V} = [I_1, I_2, \ldots, I_N]$, weak supervision $\mathbf{C} \in \mathbb{R}^{M\times 3}$ and unsupervised pedestrian prior $\mathbf{F}$

Output: Bounding box annotations $\mathbf{B} \in \mathbb{R}^{M\times 5}$

$\mathbf{B} \leftarrow [\,]$

Let $\{w_{\text{min}}, w_{\text{max}}, h_{\text{min}}, h_{\text{max}}\}$ be estimates of min and max possible widths w and heights h of pedestrians in $\mathbf{V}$.

for $i = 1$ **to** M **do**

  n $\leftarrow \mathbf{C}_{i1}$ % get frame num of i^{th} weak supervision %

  $x \leftarrow \mathbf{C}_{i2}$ % get x position of the center %

  $y \leftarrow \mathbf{C}_{i3}$ % get y position of the center %

  $\vec{e} \leftarrow [x - w_{\text{max}}/2, y - h_{\text{max}}/2, x + w_{\text{max}}/2, y + h_{\text{max}}/2]$

  $\mathcal{W} \leftarrow$ get multiscale sliding windows (larger than w_{min} and h_{min}) in the area surrounded by rectangle $\vec{e}$.

  $\mathcal{W} = \{\vec{w}_1, \ldots, \vec{w}_K\}$ is the set of K bounding boxes and $\vec{w} = [x_1, y_1, x_2, y_2]$ is a vector denoting coordinates of top left and bottom right corners.

  Let Y be the function to compute a feature vector given a patch in frame I_n corresponding to $\vec{w}$.

  Let $\vec{w}^{[j]}$ be a scalar denoting the j^{th} element of $\vec{w}$.

  Let $G(\bullet) = \mathcal{N}(\bullet; \vec{\mu}, \Sigma) = \mathcal{N}(\bullet; [x, y], \left[\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right])$

  $p_1 \leftarrow \sum_{\vec{w}\in\mathcal{W}} \mathbf{F}(Y(\vec{w}))$

  $p_2 \leftarrow \sum_{\vec{w}\in\mathcal{W}} G([\frac{\vec{w}^{[1]}+\vec{w}^{[3]}}{2}, \frac{\vec{w}^{[2]}+\vec{w}^{[4]}}{2}])$

  $\vec{w}_{\text{best}} = \arg\max_{\vec{w}\in\mathcal{W}} \frac{\mathbf{F}(Y(\vec{w}))}{p_1} + \frac{G([\frac{\vec{w}^{[1]}+\vec{w}^{[3]}}{2}, \frac{\vec{w}^{[2]}+\vec{w}^{[4]}}{2}])}{p_2}$

  Add to matrix $\mathbf{B}$ a new row given by $[n, \vec{w}_{\text{best}}]$

end for

return $\mathbf{B}$

other perspective, the classifier $\mathbf{F}$ gives some measure about *objectness* in the scene. This can also be considered as a *pedestrian prior* after the aforementioned biasing. Another potential benefit of the biasing is that it may allow us to use a simple linear classifier to obtain $\mathbf{F}$. If $\mathbf{F}$ does not output valid probabilities (such as when using a SVM), we calibrate it to produce probabilities by simple Platt scaling.

3.2. Cue Fusion and Optimization

For each weak supervision center, we first compute a large rectangle $\vec{e}$ surrounding it. This can be computed by setting for the whole video, an estimate of the width and height of the largest possible pedestrian in the scene. This does not need to be accurate and it can be easily determined by a human. Then multi-scale sliding windows $\mathcal{W}$ are generated within $\vec{e}$. We seek the best sliding window $\vec{w}_{\text{best}} \in \mathcal{W}$ such that $\vec{w}_{\text{best}}$ is scored highest by two terms in the objective function: (1) score of the pedestrian prior given by $\frac{\mathbf{F}(Y(\vec{w}))}{p_1}$ and (2) the closeness (in distance) to the given center supervision given by $\frac{G([\frac{\vec{w}^{[1]}+\vec{w}^{[3]}}{2}, \frac{\vec{w}^{[2]}+\vec{w}^{[4]}}{2}])}{p_2}$ where p_1 and p_2 are the normalization terms. The relative weighing of the two terms in the objective function is set equal (see Algorithm 3 for details). Informally, the optimization objective prefers the sliding windows which are scored highly by $\mathbf{F}$ but they are penalized more, the further the centers of the sliding windows are from the given center supervision. This penalization is achieved by a gaussian weighing function $G([x, y])$ which is given by a bivariate normal distribution with mean $[x, y]$ and covariance $\left[\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right]$.

4. EXPERIMENTAL RESULTS

We have used the challenging CUHK Square [21] and MIT Traffic [22] video datasets which contain a variety of object classes including low resolution pedestrians in the scene. The CUHK and MIT videos are 60 and 90 minutes long respectively and for each video, we split it to two equal halves. During training (including unsupervised prior learning), we only use the first half. The second half is kept purely for evaluating the resulting pedestrian detectors which is summarized in terms of recall-FPPI (False Positives Per Image) curves. To score the bounding boxes, we use the PASCAL 50% overlap criteria [23]. We perform 3 different types of experiments on each dataset: (1) the pedestrian detector obtained by our weakly supervised algorithm (2) the detector obtained by strong supervision (manual bounding box annotation) and (3) the detector corresponding to the unsupervised prior (as described in Algorithm 2). These experiments are respectively named *Weak supervision*, *Strong supervision* and *Unsupervised prior* in the curves shown in Fig. 3. In addition, the cost comparison between weak and strong supervisions is also shown in Fig. 3.

As illustrated, the detection performance of the proposed algorithm closely matches that of the strong supervision. Yet, the time it took to manually annotate training data for the proposed algorithm is less than one quarter of the time taken for the strong supervision. This means that our algorithm reduces the manual human annotation effort by over 4 times to get the same performance as the standard strongly supervised training in literature. We also evaluated unsupervised prior in order to show the effectiveness of our fusion and optimization framework. The unsupervised prior alone performs poorly; however, when fused with the weak supervision, the resulting detector has a much higher performance than the unsupervised prior.

5. CONCLUSION

We have proposed a novel weakly supervised learning algorithm for training pedestrian detectors for videos. The algorithm consists of learning an unsupervised prior using unlabelled data in the video and then fusing the prior with the weak supervision in an optimization framework to generate bounding box annotations. We showed that the weakly supervised algorithm reduces the amount of human annotation effort by over 4 times without sacrificing the accuracy of the resulting detector.

6. REFERENCES

[1] Constantine Papageorgiou and Tomaso Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[2] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. IEEE, 2001, vol. 1, pp. I–511.

[3] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR (1)*, 2005, pp. 886–893.

[4] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2012.

[5] Markus Enzweiler and Dariu M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, 2009.

[6] Ross B Girshick, Pedro F Felzenszwalb, and David A Mcallester, "Object detection with grammar models," in *Advances in Neural Information Processing Systems*, 2011, pp. 442–450.

[7] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona, "Fast feature pyramids for object detection," *PAMI*, 2014.

[8] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[9] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *ECCV*, 2012.

[10] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool, "Pedestrian detection at 100 frames per second," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2903–2910.

[11] Marco Pedersoli, Andrea Vedaldi, and Jordi Gonzalez, "A coarse-to-fine approach for fast deformable object detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1353–1360.

[12] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[13] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2002, pp. 561–568.

[14] Carolina Galleguillos, Boris Babenko, Andrew Rabinovich, and Serge Belongie, "Weakly supervised object localization with stable segmentations," in *Computer Vision–ECCV 2008*, pp. 193–207. Springer, 2008.

[15] Yixin Chen and James Z Wang, "Image categorization by learning and reasoning with regions," *The Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[16] Ondrej Chum and Andrew Zisserman, "An exemplar model for learning object classes," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[17] Minh Hoai Nguyen, Lorenzo Torresani, Fernando de la Torre, and Carsten Rother, "Weakly supervised discriminative localization and classification: a joint learning process," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1925–1932.

[18] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari, "Localizing objects while learning their appearance," in *Computer Vision–ECCV 2010*, pp. 452–466. Springer, 2010.

[19] Matthew Blaschko, Andrea Vedaldi, and Andrew Zisserman, "Simultaneous object detection and ranking with weak supervision," in *Advances in neural information processing systems*, 2010, pp. 235–243.

[20] Jian Yao and J-M Odobez, "Multi-layer background subtraction based on color and texture," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[21] Meng Wang, Wei Li, and Xiaogang Wang, "Transferring a generic pedestrian detector towards specific scenes," in *CVPR*, 2012, pp. 3274–3281.

[22] Meng Wang and Xiaogang Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *CVPR*, 2011, pp. 3401–3408.

[23] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.