Progressive Latent Models for Self-Learning Scene-Specific Pedestrian Detectors

Qixiang Ye, *Senior Member, IEEE*, Tianliang Zhang, *Student Member, IEEE*, and Wei Ke, *Student Member, IEEE*

***Abstract*— The performance of offline learned pedestrian detectors significantly drops when they are applied to video scenes of various camera views, occlusions, and background structures. Learning a detector for each video scene can avoid the performance drop but it requires repetitive human effort on data annotation. In this paper, a self-learning approach is proposed, toward specifying a pedestrian detector for each video scene without any human annotation involved. Object locations in video frames are treated as latent variables and a progressive latent model (PLM) is proposed to solve such latent variables. The PLM is deployed as components of object discovery, object enforcement, and label propagation, which are used to learn the object locations in a progressive manner. With the difference of convex (DC) objective functions, PLM is optimized by a concave-convex programming algorithm. With specified network branches and loss functions, PLM is integrated with deep feature learning and optimized in an end-to-end manner. From the perspectives of convex regularization and error rate estimation, detailed optimization analysis and learning stability analysis of the proposed PLM are provided. The extensive experiments demonstrate that even without annotation involved the proposed self-learning approach outperforms weakly supervised learning approaches, while achieving comparable performance with transfer learning approaches.**

***Index Terms*— Pedestrian detection, self-learning, progressive latent model, difference of convex.**

I. Introduction

WITH widespread use of surveillance cameras, the need for automatically detecting objects, *e.g.* pedestrians, has significantly increased. Recent methods [1]–[4] have achieved encouraging progress for detecting objects in images, given large-scale training sets and high-capacity deep learning models [5]. However, their performance in video scenes is limited for the following reasons: 1) Supervised learning of detectors for different scenes requires repeated human effort on data annotation; 2) Offline-trained detectors unavoidably degrade with changes in the scene or camera [6], [7]; 3) Scene specific cues including camera viewpoints, object occlusions, and background structures are not incorporated into the detectors [8]–[11].

To robustly detect objects in various of video scenes, transfer learning [6], [7], [12] can be used to adapt the learned detectors to new scenes without using additional data annotation [13]–[15]. Semi-supervised learning uses a small number of instances to initialize detectors and incrementally improves the detectors by mining samples in new domains [7], [16], [17]. However, transfer learning is challenged when the object appearance in the target domains is significantly different with that in the source domains; while semi-supervised models might drift away from the intended aims given noisy or unrelated samples [7]. Most importantly, both methods require partial instance annotations (object bounding boxes), and therefore, do not fully reduce human supervision.

In this paper, we discuss the possibility of self-learning pedestrian detectors in dynamically changing scenes, *e.g.* a city square, to build a pedestrian detection system in a fully unsupervised manner. The inputs of self-learning include video sequences where pedestrians are the dominant moving objects and additional negative images randomly collected from the Web, Fig. 1. The aim is to simultaneously learn pedestrian detectors and pedestrian locations under the hypothesis that positive objects (pedestrians) share an appearance model while the negatives are diverse and do not share any appearance model [18]–[20].

Pedestrian locations in video frames are treated as latent variables which are solved with a progressive latent model (PLM). Accordingly, the self-learning approach is deployed as components of object discovery, object enforcement, and label propagation, which are optimized in a progressive manner. Given the prior that the presence of pedestrian in video frames of significant motion and the absence of pedestrian in negative images, the image-level label is estimated. With estimated image-level labels, object discovery is implemented with a latent SVM [21], which outputs appearance models and coarsely localizes objects by selecting region proposals to minimize image-level classification error. With localized objects, a spatial regularization procedure is explored to reducing the localization ambiguity and discriminate object parts with the objects themselves. A label propagation component is further used to discover harder-positive instances and enables the self-learning approach to find complex sample domains comprising multi-posture and multi-view pedestrians [22].

Manuscript received January 27, 2018; revised August 29, 2018, December 2, 2018, and February 3, 2019; accepted March 31, 2019. Date of publication April 29, 2019; date of current version March 27, 2020. This work was supported in part by the National Nature Science Foundation of China under Grant 61836012 and Grant 61671427, and in part by the Beijing Municipal Science and Technology Commission under Grant Z181100008918014. The Associate Editor for this paper was K. Wang. *(Corresponding author: Wei Ke.)*

The authors are with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 108408, China (e-mail: qxye@ucas.ac.cn; zhangtianliang17@mails.ucas.ac.cn; kewei11@mails.ucas.ac.cn).

Digital Object Identifier 10.1109/TITS.2019.2911315

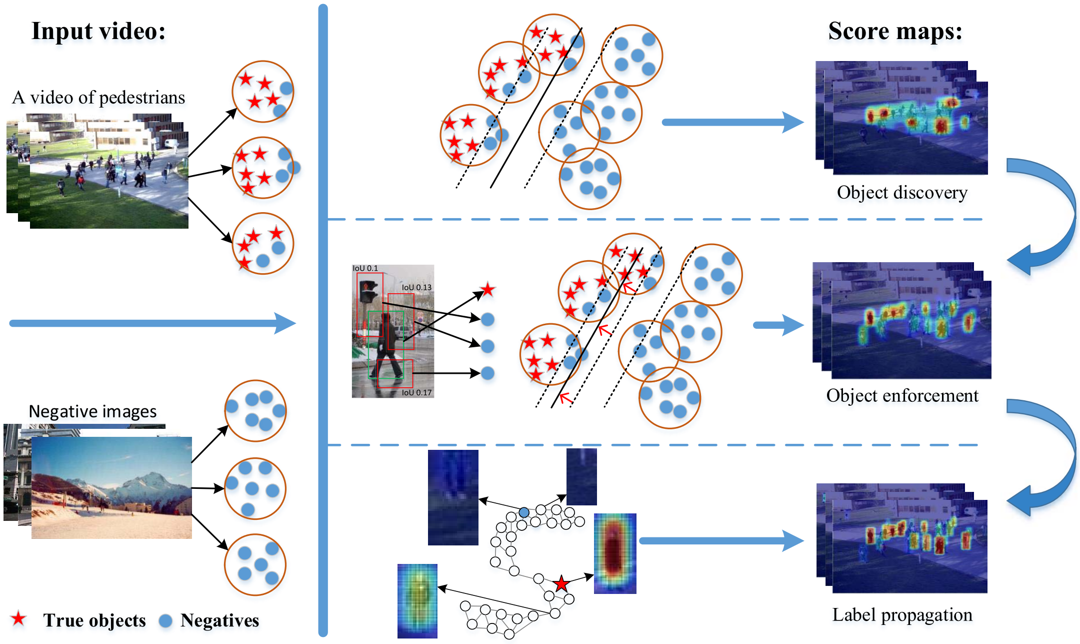


Fig. 1. Proposed self-learning framework. Given a video where pedestrians are dominant moving objects, our proposed approach progressively constructs a scene-specific detector using a self-learning procedure. In the learning procedure, each positive image is decomposed into a "bag" of proposals, and the object discovery, object enforcement, and label propagation procedures are iteratively applied to identify true positives from the proposals.

The self-learning approach was first proposed in our CVPR 2017 paper [23], and is updated to processing deep learning features. With the added optimization analysis of the progressive latent model (PLM) and integration of the PLM with an deep learning framework, the self-learning approach is comprehensively presented. The contributions of this paper include: (1) A self-learning pedestrian detection framework, which is deployed as iterative procedures of object discovery, object enforcement, and label propagation, posing a new direction in the field of (unsupervised) object detection; (2) A progressive latent model (PLM), which uses spatial regularization and label propagation to reduce ambiguity of discovered samples, as well as addressing the stability of self-learning. (3) A deep PLM, which integrates the progressive latent model with a deep learning framework for unsupervised pedestrian modeling. (4) Detailed optimization analysis and learning stability analysis of the proposed PLM.

The remainder of this paper is organized as follows. Related works are presented in Section II. In Section III and Section IV, the PLM and model learning procedure are described, respectively. In Section V, we give two implementations of the self-learning approach. Extensive experiments are conducted in Section VI and we conclude the paper in Section VII.

II. Related Works

Visual representations and classification models are two major research topics in the area of supervised pedestrian detection [22], [24]–[28]. In the long-term research history of pedestrian detection, visual representations including Haar-like features [29], [30], histogram of orientated gradient (HOG) [31], multi-scale orientation features [9], and deep learning features [24], [32] were explored. Classification models including SVMs [31], discriminatively trained part based models (DPMs) [33], Random Forest [34], and deep neural networks [32] were applied.

Our work is based on the most successful supervised methods, *i.e.* DPM [33] and RCNN [3], with a new motivation to learn pedestrian detectors using minimum supervision. It is related to transfer learning, online learning, weakly supervised learning, and unsupervised object discovery methods.

A. Transfer Learning

One conventional strategy of transfer learning was to leverage the object distributions in target domains to improve the performance of pre-trained detectors in source domains. Researchers utilized context cues [13], [15], confidence propagation [15], [35], and virtual-real world adaptation [36] to perform transfer learning. Gaussian process regression [37] and super-pixel region clustering [8] were employed to select "safe" samples in target domains. Large margin embedding [38] and transductive multi-view embedding [39] were explored to expand detector horizons. Generative or discrepant classifiers [40] were used to fine-tune the network parameters so that the feature distributions in the source and target domains become similar.

Transfer learning can obviously reduce human annotations. Nevertheless, it suffers from the domain variation problem, *i.e.* the major differences of object appearance, viewpoint, and illumination between source and target domains. When the gap is significant, the adaptation of pre-trained features and/or models becomes non-smooth. By contrast, the proposed approach initializes and improves models in the same scenes, naturally alleviating the domain variation problem.

B. Online/Semi-Supervised Learning

Online learning and semi-supervised learning improved scene-specific detectors by taking advantage of the continuous incoming data stream from the target domains. Classical detection-by-tracking (DBT) [41], [42] initialized the system using offline trained detectors and leveraged temporal cues to extend sample domains and cancel detection errors. Tracking-Learning-Detection (TLD) [43] initialized the system with a single sample, and used tracking and online learning to boost detectors. Despite the popularity of DBT and TLD approaches, recent studies [7] demonstrated that the simple combination of detection with tracking might introduce poor detectors as the errors from both detection and tracking were amplified in a coupled system.

A P-N expert [43] was used in TLD to control precision and recall rates, guaranteeing the learning stability as a linear dynamic system. The learning stability of our approach can also be guaranteed as the difference of convex (DC) objective functions of PLM converge at each learning iteration.

C. Weakly Supervised Learning

Weakly supervised object detection (WSOD), where only the image-level annotations indicating the presence or absence of a class of objects in images are available, has attracted increased attention [18], [44]. Compared with supervised object detection methods that require annotated bounding-boxes for all samples in all training images, WSOD requires only image-level annotations, and thus can leverage tagged images on Web and significantly reduce human effort about data annotation. To tackle the problem of WSOD, latent variable learning and multi-instance learning (MIL) are two kinds of representative methods. Using redundant object proposals as input, the learning objective of these methods is typically designed to solve latent variables by minimizing the image-level classification error.

Latent variable learning alternates between sample labeling and detector learning in a way similar to Expectation Maximization optimization. Due to the missing annotations, however, this optimization is non-convex and therefore prone to getting stuck in a local minimum and outputting wrong labelings [45]. Cinbis *et al.* [46] used a multi-fold splitting of the training set while Bilen *et al.* [45] used convex clustering to prevent getting stuck to wrong labels. In this paper, we propose alleviating the local optima problem with a more reasonable way by converting the problem into a difference of convex (DC) optimization. We also introduce regularization terms about domain knowledge, *i.e.* intra-frame hard-negative mining and inter-frame similarity propagation.

D. Self-Paced Learning

Inspired by the cognitive principle of humans, Bengio *et al.* [47] proposed self-paced learning (SPL), where a model was learned by gradually including samples from easy to complex. Recently, several works provided more comprehensive understanding of the learning insight underlying CL/SPL, and formulated the learning model as a general optimization problem and SPL was proposed for object detection. Lee and Grauman [48] introduced a self-paced approach to focus on the easiest instances first, and progressively expanded its repertoire to include more complex objects. Sangineto *et al.* [49] presented a self-paced learning protocol for object detection that iteratively selected the most reliable images and boxes according to class-specific confidence levels and inter-classifier competitions. Wang *et al.* [50] used the self-supervision to mine valuable information from unlabeled and partially labeled data.

Existing SPL approaches mainly focus on instance mining in a easy-to-hard manner. Nevertheless, due to the long-tail distribution of samples, hard instances are sparse and it is a sophisticated task to find hard yet informative samples. In this paper, the self-learning approach is deployed as components of object discovery, object enforcement, and label propagation, which are optimized in a progressive manner. It can use spatial regularization and temporal consistence of video objects to mine hard examples.

E. Unsupervised Video Object Discovery

An early approach [51] learned scene-specific object detectors by online boosting, but it required offline learned seed detectors. Recent research [18], [20], [52] formulated unsupervised video object discovery as two complementary steps. The first step established correspondences between prominent regions across video frames, and the second step associated successive similar object regions within the same video. Xiao and Lee [20] proposed a fully unsupervised video object proposal approach which first discovered a set of easy-to-group instances by clustering and then updated the appearance model to gradually detect harder instances by the initial detector and temporal consistency. This unsupervised approach can automatically generate object proposals, but cannot output precise detections. Schulter *et al.* [53] formulated an iterative process that exploits both motion (optical flow) and appearance cues via a joint formulation of conditional random field to extract and segment objects. Researchers also used domain adaptation to construct a self-learning-camera [54]. However, these methods did not incorporate a principle way to model latent objects and often lacked strategies to guarantee the learning stability.

III. Progressive Latent Model

The progressive latent model (PLM) targets at finding accurate object locations given a set of object proposals that have salient object-like appearance and motion, Fig. 2a. To this end, PLM is decomposed into three basic components: object discovery, object enhancement, and label propagation. The object discovery component aims to find region proposals that best discriminates positive video frames from the negative images. The object enhancement component discovers hard negatives that help reducing falsely localized object parts, as well as improving object localization. The label propagation component mines harder instances throughout the video, Fig. 2c and Fig. 2d. The three components iterate until an error rate based stability criteria is met.

Let $x \in \mathcal{X}$ denote a video frame or a negative image. $y \in \mathcal{Y}, \mathcal{Y} = \{0, 1\}$ are labels denoting whether x contains a pedestrian object or not. $y = 1$ indicates that there is at least one pedestrian in the frame while $y = 0$ indicates

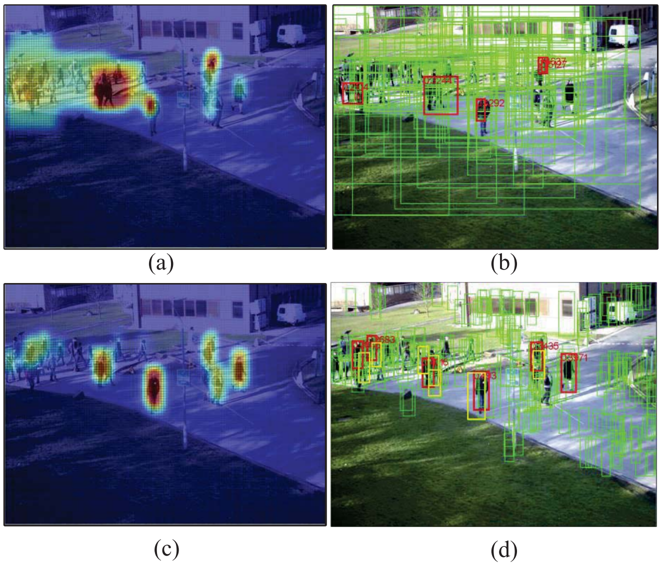


Fig. 2. Object discovery from noisy proposals. (a) The score map in the first learning iteration and (b) candidate objects (red boxes) discovered. (c) The score map and in the fifth learning iteration. (d) Candidate objects (red boxes) and hard negatives (yellow boxes). (Best viewed in color.)

a frame without pedestrian object or a negative image. PLM is formulated with a multi-objective function that targets at jointly determining the latent object $h \in \mathcal{H}$ and a latent model β in a progressively optimization procedure, as

$$\begin{aligned}\left\{h^*, \beta^*\right\} &= \arg\min_{\beta,h} \mathcal{F}_{(\mathcal{X},\mathcal{Y})}(\beta, h)\\ &= \arg\min_{\beta,h} \mathcal{F}_l(\beta, h) - \lambda \mathcal{F}_s(\beta) + \gamma \mathcal{F}_g(\beta, h), \quad (1)\end{aligned}$$

where $\mathcal{F}_l(\beta, h)$, $\mathcal{F}_s(\beta)$, and $\mathcal{F}_g(\beta, h)$,[1] as defined below, are the objective functions for object discovery, spatial regularization, and score propagation, respectively. λ and γ are regularization factors.

A. Object Discovery

The object discovery component is implemented with a latent SVM (LSVM) model to choose object proposals that best discriminate positive frames from negative images, as

$$\left\{y^*, h^*, \beta^*\right\} = \mathop{\arg\max}_{y\in\mathcal{Y}, h\in\mathcal{H}, \beta} \beta^T \cdot \upsilon\left(x, y, h\right), \quad (2)$$

where $\upsilon(x, y, h)$ denotes a normalized feature vector. $\mathcal{H}$ denotes the proposal set, which is made up of proposals $\mathcal{H}_i, i = 1, \ldots, N$ from video frames. Basically, solving Eq. 2 produces a high-scored $\beta^T \cdot \upsilon(x, y, h)$ for each positive frame ($y = 1$) and a low score for each negative image ($y = 0$). Concretely, we learn the model β on a collection of video frames and negative images $X = \{(x_i, y_i), i = 1, \ldots, N\}$ with

$$\min_{\beta,h} \mathcal{F}_l(\beta, h) = \min_{\beta,h} \frac{1}{2}||\beta||^2 + \mathcal{C} \sum_{i=1}^{N} l(\beta, x_i, y_i, h), \quad (3)$$

where $\mathcal{C}$ is a regularization factor and l is a difference-convex loss function defined as

$$\begin{aligned}l(\beta, x_i, y_i, h) = \max_{y,h} \left(\beta^T \cdot \upsilon(x_i, y, h) + \Delta(y_i, y)\right)\\ - \max_{h} \beta^T \cdot \upsilon(x_i, y_i, h). \quad (4)\end{aligned}$$

$\Delta(y_i, y) = 0$ if $y = y_i$, and 1 otherwise. Eqs. 3 and 4 target at choosing the highest scoring proposals h from the other configurations, defining a max-margin formulation to measure the mismatch between the image, the image label, and the object proposals.

B. Object Enforcement

The object discovery component aims at optimizing the image-level classification instead of the object-level classification. Once the image-level classification objective function reaches optimization, whether or not the object-level classification is optimized, the learning procedure stops [21]. Considering that all positive images contain the object parts but none of negative images does, LSVM could falsely select object parts as positive objects. The reason lies in that Eq. 4 is non-convex as it is a difference of two convex functions. It is known that optimizing a non-convex function is easy to get stuck to a local minimum. Such local minimum means that the algorithm might falsely select object parts as positive objects.

Motivated by the success of the application of hard negative mining in visual object detection approaches [55], we propose mining hard negatives those overlap with true positives and using spatial regularization to enforce the localization of objects. Denoting by $\mathcal{H}_i$ object proposals in frame i and h' the hard negatives corresponding to an object h in a video frame, we define a function to maximize the distance between the potential object and its spatial neighbors, as

$$\max_{\beta} \mathcal{F}_s(\beta) = \sum_{i=1}^{N} \sum_{\substack{h\in\mathcal{H}_i\\ h'\in\Omega_{\mathcal{H}_i,h}}} ||\beta^T \cdot \left(\upsilon(x_i, h) - \upsilon(x_i, h')\right)||^2, \quad (5)$$

where $\Omega_{\mathcal{H}_i,h}$ denotes the spatial neighbors of h in $\mathcal{H}_i$. The spatial neighbors are high-scored object parts and surrounding image patches that have IoU (Intersection of Union) with h in the interval (0.0 0.25). Eq. 5 optimizes the model β using fixed h, and thus is a regularization function. Such a function enforces the latent model, yielding a consistent and significant boosts in object localization during progressive learning.

C. Label Propagation

According to Eq. 2, the object discovery component chooses an object proposal that best discriminates the positive frame from negative images. To mine more positives and negatives, we propose using label propagation for incremental learning.

Suppose there are totally l labeled samples from previous learning iterations. We select $u = l \times (r - 1.0)$ high-scored proposals as unlabeled samples, where $r > 1.0$ is the learning rate, related to the expected density of pedestrians. Given labeled samples $\{h_i\}, i = 1, \ldots, l$, and unlabeled proposals $\{h_j\}, j = l, \ldots, l + u$, a kNN graph in the feature space is

[1] In what follows, $(\mathcal{X}, \mathcal{Y})$ is omitted for short.

first constructed. The graph vertex defines the nearest neighbor vertices of samples. h_i and h_j are connected if one of them is among the other's kNN [56]. kNN graph automatically adapts to the density of instances in a feature space: in a dense region, the kNN neighborhood radius will be small; in a sparse region, the radius will be large. The graph-based label propagation procedure is defined as $g(\beta, h_j) = \frac{\sum_{k=l}^{l} w_{jk} g(\beta,h_k)}{\sum_{k=l}^{l} w_{jk}}, j = l + 1, \ldots, l+u$, where w_{ik} denotes the edge weight defined with a Gaussian Function on Euclidean distance between h_i and h_k, $w_{ik} = \exp\left(-\frac{(\|h_i - h_k\|^2)}{2\sigma^2}\right)$, and σ is the bandwidth parameter and controls the speed of weight decrease. This is equivalent to a convex optimization problem [56], as

$$\min_{g(\beta,h)} \mathcal{F}_g(\beta,h) = \min_{g(\beta,h)} \sum_{i=1}^{l} \sum_{j=l}^{l+u} w_{ij} \left(g(\beta,h_i) - g(\beta,h_j) \right)^2$$
$$s.t.\ g(\beta,h_i) = y_i, \quad i = 1, \ldots, l, \qquad (6)$$

where $g(\beta, h_j)$ is the propagated score of proposal h_j and y_i is the label of the frame/image that h_i belongs to.

IV. Model Learning

The procedure of model learning is to solve Eq. 1 with a progressive optimization algorithm. The stability of this algorithm is empirically guaranteed with the monotonically non-increased error rate.

A. Progressive Optimization

In the learning procedure, the optimization of $F_s(\beta)$ (object enforcement) and $F_g(\beta, h)$ (label propagation) depends on the results of $F_l(\beta, h)$ and Eq. 1 is a progressive model, where F_l, F_s and F_g should be alternatively optimized. The objective functions of Eq. 1 could be written as the difference of convex functions. This allows us to optimize it with a two-step Concave-Convex Procedure (CCCP) [21]. The CCCP algorithm applied to latent model gives rise to a very intuitive algorithm that alternates between learning the latent variable h that best explains the training pair (x_i, y_i) and solving the optimization problem while treating the latent variables as completely observed. This is similar to the iterative process of Expectation Maximization (EM), which maximizes the expected log likelihood under the marginal distribution of the latent variables. We minimize the regularized loss against a single latent variable h_i that best explains (x_i, y_i).

The first-step CCCP for $\mathcal{F}_l$ discovers potential pedestrian objects in frames and initializes the latent model, the second-step CCCP for $\gamma\mathcal{F}_g - \lambda\mathcal{F}_s$ performs object enforcement and label propagation. The two-steps of CCCP progressively optimize the PLM until the change of the estimated sample error rate is negligible. Theoretically, the CCCP algorithms guarantee the optimization with difference of convex objective functions converges to a local minimum or a saddle point [21]. Therefore, the iterative usage of the two-steps CCCP algorithm with the constraint of error rate monotonicity (discussed in Section B) can guarantee the stability of self-learning.

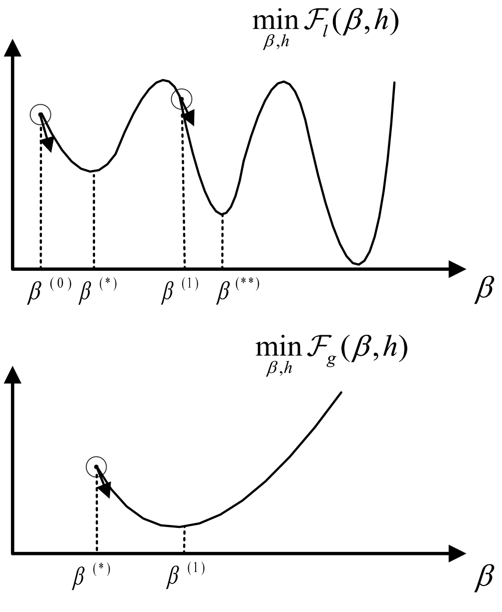


Fig. 3. The objective function (up) of the latent model is non-convex and is easy to get stuck in a local minimum. The proposed progressive latent model (PLM) uses a convex objective function (down) to assist the latent model escaping from local minima and pursuing a stronger global minimum.

As analyzed above, the objective function of a conventional latent model, Eq. 4, is typically non-convex, and thus is easy to fall into a local optimum and falsely select object parts as objects. The proposed PLM incorporates a spatial regularization term and a convex label propagation term, and converts the conventional non-convex optimization to an optimization problem with the difference of convex (DC) objective functions. The essence of this approach is using convex regularization to assist the latent model escaping from the local minimum $\beta^{(*)}$ and pursuing a stronger global minimum $\beta^{(**)}$, Fig. 3. Accordingly, PLM incorporates a spatial regularization term to reduce ambiguities in object proposals and to enforce object localization, and also a graph-based label propagation term to discover harder instances in video frames. This introduces to the objective function the object similarity constraint among video frames, as well as preventing the object parts be falsely localized.

B. Error Rate Analysis

PLM incorporates a label propagation procedure, Eq. 6, which iteratively introduces new samples and updates the model. In this procedure, the primary problems to be solved are avoiding model drift and reducing the error rate. Eq. 1 and Eq. 6 imply that a larger γ value introduces more newly labeled samples, as well as a larger error rate ξ, and vice versa. The number of newly labeled samples u is determined to be an implicit function of γ, $u(\gamma)$. The value of γ needs to essentially guarantee that the error rate of newly labeled samples is smaller than that of existing samples, meaning the error rate of the training set is monotonically non-increased. It is also expected that there is a large γ, which implies that more samples could be labeled in each iteration. To decide the

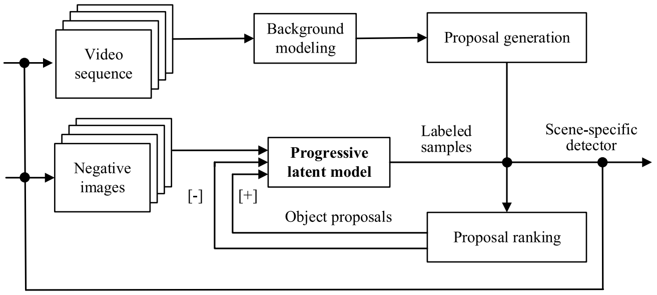


Fig. 4. Block diagram of self-learning. The proposal generation component localizes potential objects using objectness, motion, and appearance cues. The proposal ranking component chooses the high-ranked proposals as positive candidates, and low-ranked ones as negatives. The PLM identifies positives and hard negatives from candidates. PLM iteratively runs until the convergence of the learning procedure.

value of γ, an optimization objective function is defined as

$$\begin{aligned} &\max_{\gamma,\beta,y_j} \gamma \\ &s.t.\ \xi_{u(\gamma)} \leq \xi_l \\ &\qquad \approx \frac{1}{l+u(\gamma)} \sum_{j=1}^{l+u(\lambda)} (f_\beta(h_j) - \widetilde{y}_j) \\ &\qquad \leq \frac{1}{l} \sum_{i=1}^{l} (f_\beta(h_i) - \widetilde{y}_i), \end{aligned} \tag{7}$$

where l and $u(\gamma)$ respectively denote the numbers of labeled samples in previous iterations and unlabeled samples in current iteration.

The optimization of Eq. 7 guarantees that the estimated error rate of newly labeled samples $\xi_{u(\gamma)}$ is empirically smaller than that of labeled samples ξ_l by finding a proper γ in each learning iteration. γ is optimized with a linear searching algorithm [57], which searches in the interval [0.0, 1.0] with step size 0.1 and updates the detector $f_\beta(h_j)$ to $f_{\widetilde{\beta}}(h_j)$ at each step. Meanwhile, $\widetilde{y}_j$ is estimated with $\widetilde{y}_j = f_{\widetilde{\beta}}(\cdot)$, with which the error rate $\xi_{u(\gamma)}$ is calculated.

V. Self-Learning Implementations

Based on the progressive latent model (PLM), two kinds of self-learning implementations are provided in what follows.

A. Self-Learning With HOG Features

With the proposed PLM, a self-learning approach is implemented as shown in Fig. 4. The proposal generation component localizes potential objects using objectness, motion, and appearance cues. The proposals are extended in successive video frames with a Kanade-Lucas-Tomasi (KLT) tracking algorithm. The proposal ranking component chooses the high-ranked proposals as positive candidates and low-ranked proposals as negatives. The PLM that incorporates components of object discovery, object localization, and label-propagation identifies positives and hard negatives from given proposals. With mined positive samples, a DPM detector $f_\beta(h)$ is trained to perform pedestrian detection.

Given a video of static background, a motion score map is calculated for each video frame with a background modeling algorithm. On the motion score map, detection proposals (as shown in Fig. 2b) are extracted using the EdgeBoxes approach [58]. From the second iteration, a pedestrian detector is initialized and a sliding window strategy is used to generated object proposals, as shown in Fig. 2d. To extend the proposals in the temporal domain, a KLT tracking algorithm is employed to track and collect proposals from frame t to frame $t + \tau$, where τ is empirically set to 10. Before feeding these spatial-temporal proposals to the learning algorithm, their aspect ratios are normalized to the average aspect ratio. To prevent falsely choosing static backgrounds in videos of sparse pedestrians, the average background probability of a proposal is required to be larger than a threshold, empirically set to 0.20 in our experiments.

To choose high-ranked proposals and reduce redundancy of object proposals, we propose using a combinatorial score, as

$$f(h) = \alpha^T \cdot (f_\beta(h), f_m(h), f_o(h)), \tag{8}$$

where α^T is a ranking weight vector. $f_\beta(x)$, $f_m(h)$ and $f_o(h)$, respectively, are the detection, motion, and objectness scores. The motion score $f_m(h)$ of a proposal is defined as the averaged motion scores of all pixels in its image region. Objectness score $f_o(h)$ is defined by calculating contours in the proposal regions [58]. A larger score gives higher confidence that the proposal is an object. Detection score $f_\beta(h)$ is calculated from the second learning iteration, by the learned detector. From the second iteration, the proposal region centers are set as root locations, around which we use a sliding window strategy [2] to localize proposals.

In each learning iteration, the ranking weight vector α^T is updated using a zero-space regression method [59], which performs learning without using output values. It basically minimizes the regression error of all samples, as well as maximizing the distance from a hyperplane to the origin. This results in a weight vector which captures regions in the input sample space where the probability density of the data is found, and enables the proposal ranking to be adaptive.

B. Self-Learning With Deep Features

PLM is also implemented with a deep convolutional neural network (CNN), where the network parameter β and object locations h are jointly optimized with the stochastic gradient decent (SGD) algorithm. PLM has two network branches added atop of the FC layers, Fig. 5. The first network branch, designated as the object discovery branch, defines the distribution of object scores and targets at finding high-scored object proposals by optimizing the image classification loss defined by Eq. 4. The second branch, designated as the object localization branch, uses pseudo-objects localized to learn a localization classifier. It targets at finding true objects by optimizing the spatial enforcement terms defined by Eq. 5. The label propagation is independently performed as Eq. 6.

The learning procedure targets at transferring the image-level supervision, *i.e.* the absence or presence of pedestrians in an image/frame, to object locations. In a feed-forward

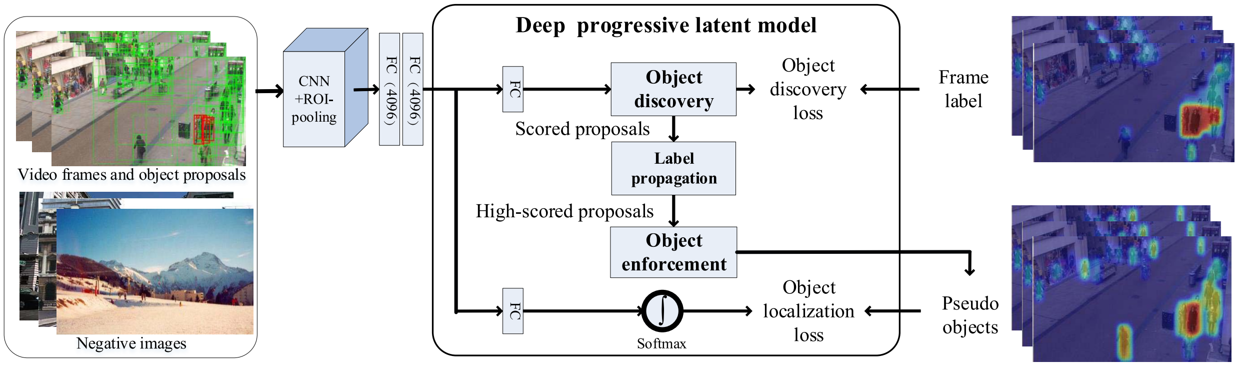


Fig. 5. Block diagram of self-learning with a deep learning framework. Given object proposals and frame-level labels, it targets at learning network parameters and pedestrian locations by minimizing the object discovery and object localization loss. The model is progressively learned by iteratively optimizing the two network branches which using forward propagation to select sparse proposals as object instances, and back-propagation to optimize the network parameters with SGD.

procedure, minimization of image-level classification loss, defined as $l(\beta, x_i, y_i, h)$, discovers high-scored proposals. The object enhancement is then performed by minimizing $-\sum_{\substack{h\in\mathcal{H}_i \\ h'\in\Omega_{\mathcal{H}_i,h}}} \beta^T ||\cdot(v(x_i,h) - v(x_i,h'))||^2$. The label propagation is finally performed based on the scored proposals by the above two steps to mine pseudo objects for detector learning. In the back-propagation procedure, the object discovery and object enhancement 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 β are updated and object detectors are enforced.

In the learning phase, object proposals are firstly generated for each image. An ROI-pooling layer atop the convolutional layer (CONV5) is used for efficient feature extraction for these proposals. The PLM model is progressively learned by optimizing the object discovery branch and object enforcement branch which use forward propagation to select sparse proposals as object instances, and back-propagation to optimize the network parameters. We use the VGG16 [60] trained on ImageNet as a backbone network. For network fine-tuning, the fully connected layers used for soft-max classification from zero-mean Gaussian distributions with standard deviations 0.01. A momentum of 0.9 and parameter decay of 0.0005 (on weights and biases) are used. Biases are initialized to 0. A global learning rate of 0.001 is used. We use 20 epoches for learning and each epoch has 5k iteration. After 12 epoches, the learning rate is reduced to 0.0001. In the detection phase, the learned detector, *i.e.* the parameters of the soft-max and FC layers, are used to classify proposals and localize objects.

VI. Experiments

The proposed approach is evaluated on five video datasets including PETS2009 [61], Towncentre [62], PNN-Parking-Lot2/Pizza [8], CUHK Square [15], and 24Hours [23]. In what follows, the datasets and experimental settings are first described. The evaluation of the model and comparison with relevant approaches are then presented. Finally, we analyze the limitation of the proposed approach.

A. Experimental Settings

For all datasets except the 24Hours, half video frames are used for learning while the other annotated frames are used for testing. These video sequences are captured with surveillance cameras and involve challenges from object occlusions, low resolution, and/or moving distracters. The pedestrians are dominant moving objects, *i.e.*, more than 75% moving objects in a scene are pedestrians. For video scenes of few pedestrians but many moving distracters, we observed that the learned detectors fail to detect pedestrians. The detailed description of these datasets can be found in [23].

The proposed approach is compared with the supervised learning, transfer learning, and weakly supervised learning approaches including:

1) Offline-DPM [2]: A DPM pedestrian detector off-line trained on the PASCAL VOC 2007 dataset.

2) Supervised-DPM: A supervised DPM detector trained with human annotated samples on specific scenes and additional negative samples mined from negative images.

3) Supervised-FasterRCNN [4]: A state-of-the-art detector with region proposal and deep feature networks.

4) Supervised-SLSV [6]: A state-of-the-art scene-specific pedestrian detector learned from virtual pedestrians whose appearance is simulated in the specific scene under consideration. Without public available source code, SLSV is only compared on the Towncentre dataset using the reported results.

5) Transfer-DPM [8]: A scene-specific detection approach based on transfer learning. Detections are originally obtained with a DPM pedestrian detector off-line trained using PASCAL VOC 2007 dataset and then improved using superpixel based clustering and classification.

6) Transfer-SSPD [15]: A state-of-the-art scene-specific pedestrian detector with transfer learning.

7) Weakly-MIL [46]: A widely used weakly supervised approach based on multiple instance learning. A DPM detector is then learned from annotated positive samples.

B. Model Effect

1) Object Enforcement: We first evaluate the module of object enforcement. By using the object enforcement (OE)

TABLE I

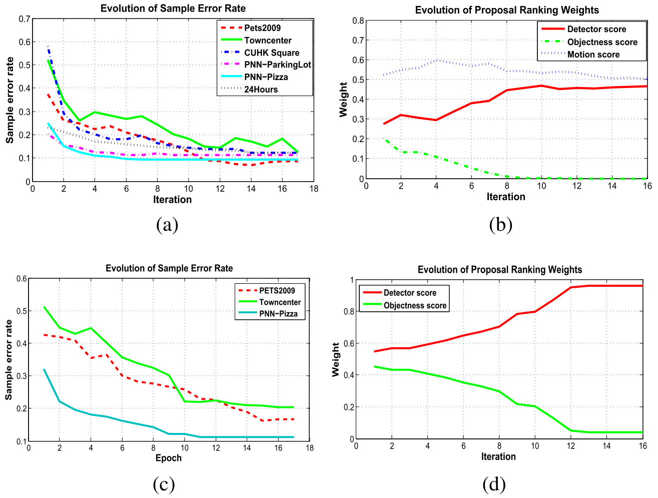
AP (AVERAGE PRECISION) OF PLMS ON THE PETS2009 DATASET. "BASELINE" DENOTES THE DPM DETECTOR TRAINED WITH TOP-RANKED PROPOSALS IN VIDEO FRAMES. "OE" DENOTES THE OBJECT ENFORCEMENT COMPONENT. "ITER1-10" DENOTE THE FIRST TO THE 10-*th* ITERATIONS. "FINAL" DENOTES THE LEARNED DPM DETECTOR BY PLM

Model	Baseline	Iter1	Iter2	Iter5	Iter10	Final
w/o OE	0.482	0.506	0.573	0.614	0.622	0.625
with OE	–	0.530	0.601	0.662	0.691	0.695

TABLE II

AP OF FASTRCNN DETECTOR LEARNED BY DEEP PLMS ON THE PETS2009 DATASET. "EPOCH1-10" DENOTE THE FIRST TO THE 10-*th* EPOCHES

Model	Baseline	Epoch1	Epoch2	Epoch5	Epoch10	Final
AP	0.407	0.524	0.566	0.580	0.584	0.585



(a) (b)

(c) (d)

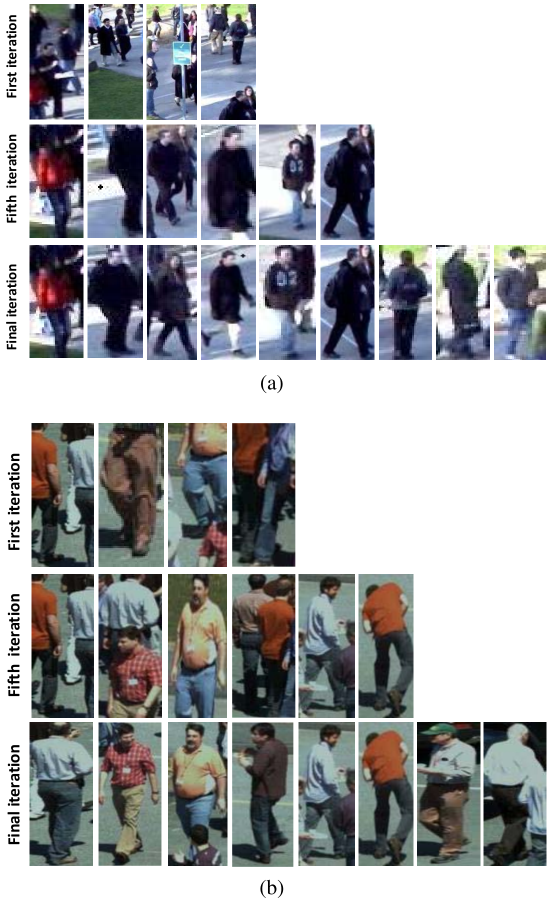
Fig. 6. Learning stability. (a) Monotonic decrease of sample error rates of PLM. (b) Evolution of proposal ranking weights of PLM. (c) Monotonic decrease of sample error rates of deepPLM. (d) Evolution of proposal ranking weights of deepPLM. (The motion cue is not used in deep PLM.)

component, Eq. 5, the performance of the learned detector significantly improves, as shown in Table I. The reason is that pedestrians are more precisely localized and most falsely detected object parts are suppressed. For the final detection models, the average precision (AP) improves about 7%, validating that the object enforcement modules helps learning better detectors.

2) *Label Propagation:* Given ranked object proposals, the label propagation component can incrementally annotate pedestrian samples without supervision. Table I clearly shows that the detection model is iteratively improved, validating the effectiveness of the label prorogation component. The Average Precision (AP) improves 6.7% from the first to the second iteration, and improves 2.9% from the fifth to the tenth iteration. After ten iterations, there are few positive instances can be labeled and the performance remains stable. Table II shows the progressive performance improvement of the end-to-end deep PLM.

3) *Progressive Latent Model:* To show the overall effect of the PLM model, we train baseline DPM [33] and FastRCNN [55] detectors by selecting top-ranked proposals from video frames. The baseline detectors do not use object enforcement or progressive optimization. As shown in Table I and Table II, PLMs outperform the baseline detectors with large margins, which verifies the effectiveness of the proposed model and the self-learning approach.

4) *Error Rate Analysis:* The evolution of sample error rates and proposal ranking weights are used to validate the convergence of the learning procedure. Fig. 6a and Fig. 6c show that the error rates of labeled training samples monotonically decrease, showing the stability of the proposed self-learning approach. From the 10-*th* to the 15-*th* learning iteration (epoch), the sample error rates became small enough, although there is little fluctuation on the Towncentre dataset. Fig. 6b shows the evolution of proposal ranking weights on the PETS2009 dataset. The weight for the objectness score quickly decays to zero, which implies that the objectness score is not as discriminative as the detection and the motion scores.

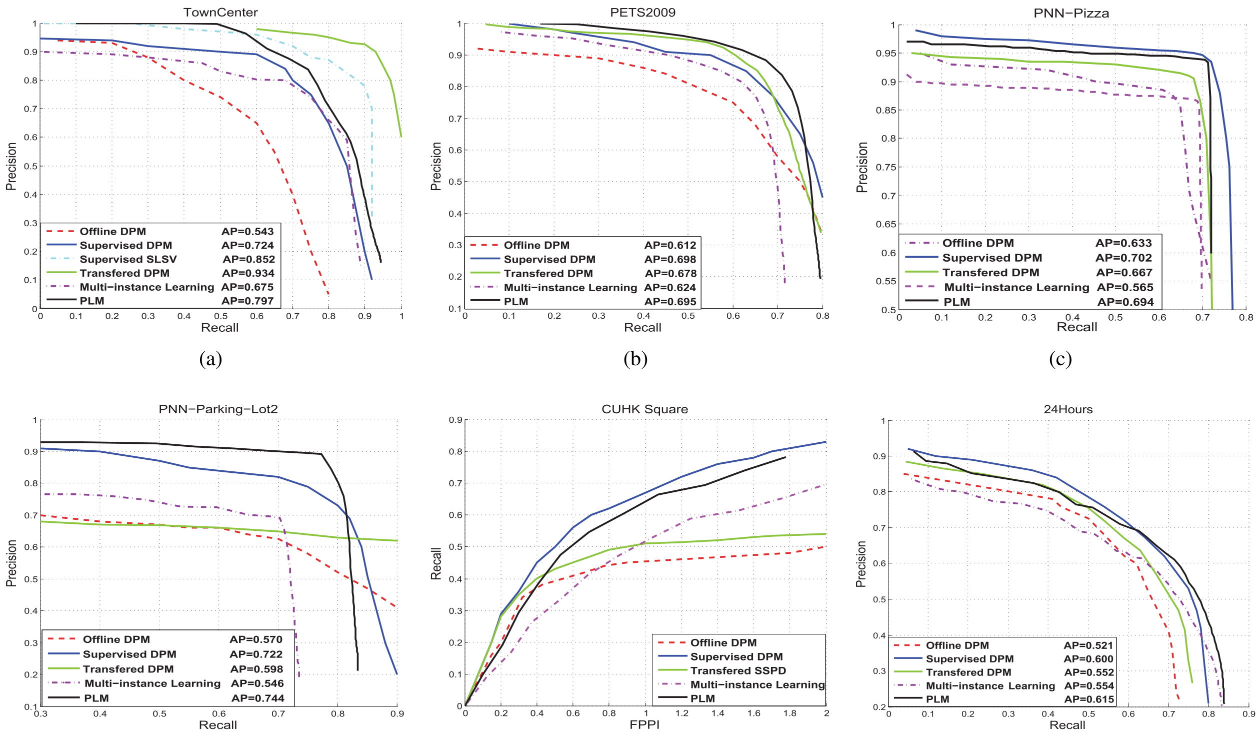


(a)

(b)

Fig. 7. Examples of learned positive instances at different iterations. (a) Pets 2009 dataset. (b) PNN-Pizza dataset.

The weight for the detection score keeps increasing during learning, which indicates that the detector is progressively improved. The weight of motion score decreases to be close to the detector score. In Fig. 6d, the proposal ranking weights of the deep PLM are shown. It can be seen that the weight of



(a) (b) (c)

(d) (e) (f)

Fig. 8. Performance of the PLM (DPM detector) and comparisons with weakly supervised, supervised, and transfer learning approaches. On five datasets the Precision-Recall metric is adopted for evaluaiton. On the CUHK dataset the FPPI-Recall metric used, which is consistent with the state-of-the-art scene-specific detection approach [15].

object score approaches 1.0 and that of objectness decreases to zero.

Fig. 7 shows examples of learned positive instances at different iterations, which also shows that the proposed PLM can remove false instances while mine true positives, progressively.

Experiments show that with linear search defined in Section IV, Part B we can determine a proper propagation parameter γ for a dataset. $\gamma = 0.7$ of the Towncentre dataset is the largest, while $\gamma = 0.3$ of the CUHK dataset is the smallest. A larger γ implies that the object proposals have fewer noises. The Towncentre dataset is a video with little illumination variance and few moving distracters, and therefore a larger γ is proper to it. The CUHK and 24Hours datasets have many moving distracters, so they need a smaller γ [23].

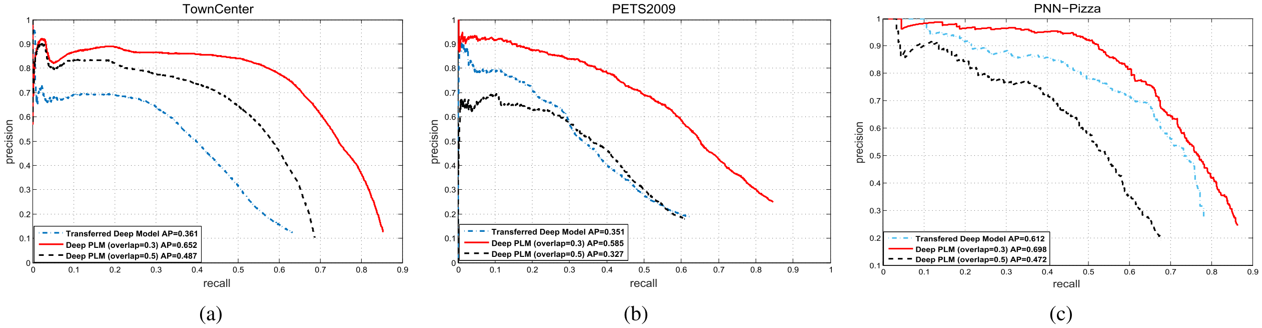
C. Performance and Comparison

The PR and FR curves in Fig. 8a-f show that the PLM significantly outperforms the off-line learned DPM detector. It also significantly outperforms the Weakly supervised (MIL) approach by 14% when FPPI = 1.0. On the PETS2009 and PNN-Parking-Lot2 datasets, PLM outperforms all of the compared approaches except for the fully supervised DPM method. On the CUHK dataset PLM significantly outperforms the scene-specific approach [15] that uses transfer learning to aggregate the detection performance on difference video scenes. It is even comparable to the supervised DPM method.

On the Towncentre dataset, the proposed PLM outperforms the MIL approach as well, *i.e.* AP 0.797 vs. 0.695. However, its performance is lower than that of the fully supervised approach SLSV [6] (AP 0.852) and the transfer learning approach (AP 0.934) [8]. The likely reason is that our approach cannot mine sufficient positive instances when the pedestrians are sparse.

On the 24Hours dataset, the AP of our approach reports the highest performance, Fig. 8e. Its performance is about 6% higher than that of the transfer learning method. The reason lies in that transfer learning suffers from the domain variation problem, *e.g.* adapting a model trained on images with daytime illumination to a video sequence of 24-hour illumination changes. By contrast, the proposed self-learning approach, which applies the detectors learned from the same scenes, can alleviate the domain variation problem. Surprisingly, our approach outperforms the fully supervised approaches in this dataset. The reason lies in that additional motion cues, which are discriminative for video sequences of static backgrounds, are incorporated in the detector.

The PR and FR curves in Fig. 9 show that the self-learning approach with deep PLM is effective. On the TownCentre, PETS2009, and PL-Pizza datasets, the deep PLM respectively achieves 48.7%, 32.7%, and 47.2% mAPs, which are comparable to 36.1%, 35.1%, and 61.2% mAPs of the transferred deep learning models. We also note that the transferred models are trained with precisely annotated instances in other scenes, but deep PLMs require to learn from chaos. When the overlap threshold reduces to 0.3, it can be seen that the performance of our self-learning approach respectively increases to 65.2%, 58.5%, and 69.8% on the three datasets. This shows that the errors of self-learning with deep features mainly arise



(a) (b) (c)

Fig. 9. Performance of the learned FastRCNN detector with deep PLM and comparison with the transferred deep models.

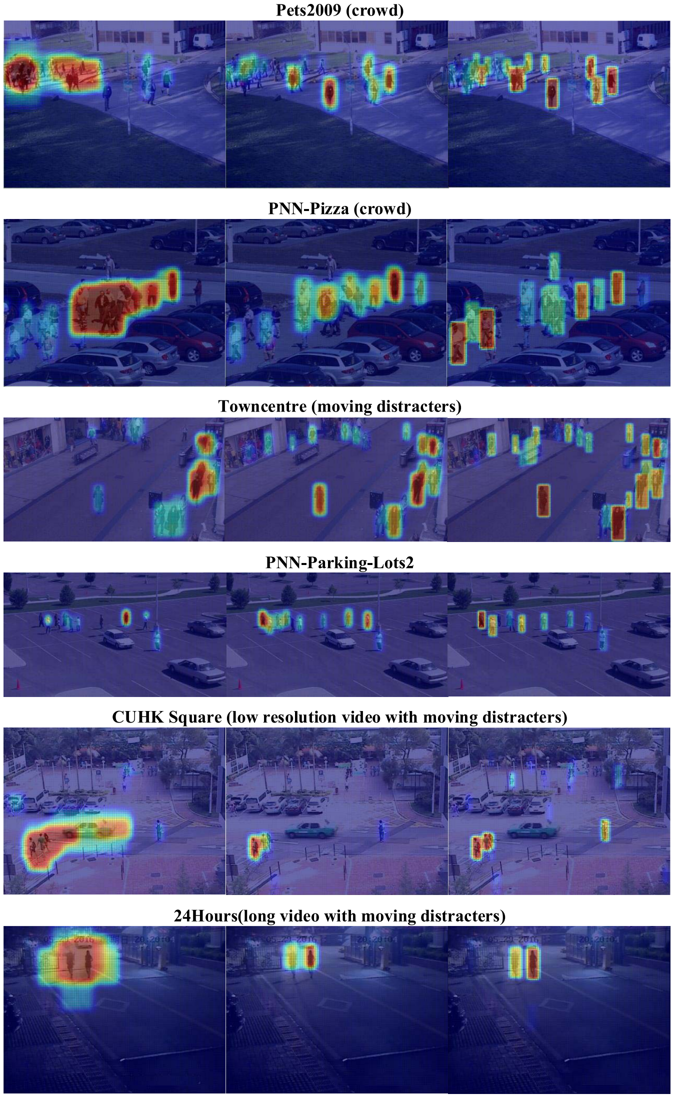

Pets2009 (crowd)
PNN-Pizza (crowd)
Towncentre (moving distracters)
PNN-Parking-Lots2
CUHK Square (low resolution video with moving distracters)
24Hours(long video with moving distracters)

Fig. 10. Illustration of learning procedures. From left to right: score maps in the first, fifth, and tenth learning iterations, respectively. For more learning results, please refer to [23].

from imprecise object localization, *i.e.* the overlap between a detected object with the groundtruth is smaller than 0.5. The deep self-learning approach can well learn pedestrian patterns, which can well discriminate the pedestrians with the clutter backgrounds but experiences difficulty to precisely localize them. In addition, in the deep PLM we do not use motion cues.

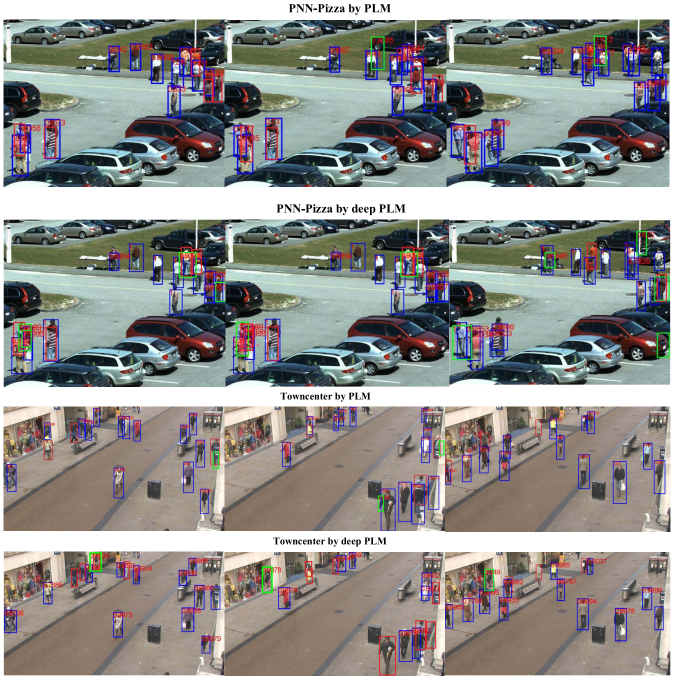


Fig. 11. Pedestrian detection examples using our proposed PLM and deep PLM. Blue boxes indicate correct detections, red boxes missed detections, and green box false detections. The results shows that PLM can effectively discover human objects and most false detections are caused by inaccurate localization, *i.e.* the detectors can localize partial objects or objects with backgrounds. (Best viewed in color.)

In Fig. 10, we illustrate the evolution of proposal scores during the progressive optimization. It can be seen that the positive samples are incrementally labeled and noise samples are reduced. On the crowded PES2009 dataset and the PNN-Pizza dataset which incorporates significant occlusions our approach can accurately localize positive samples, demonstrating effectiveness of the object discovery and label propagation components. On the Towncentre and CUHK datasets, although there exist moving distractors, *e.g.* bicycles and vehicles, our approach correctly localizes the pedestrians, demonstrating its robustness in noisy environments. In the 24Hours dataset, some video frames have dense pedestrians (daytime) but others have sparse pedestrians (at night). Learning from the early morning to the middle of the night, our approach progressively aggregates the performance without model drift. In Fig. 11, the detection results show that the learned scene-specific detectors are discriminative, showing robustness to camera views, occlusions, and background structures.

D. Limitations

In experiments, it is observed that most false detections and missed detections (red boxes in Fig. 11) are caused by false localization. Some detected object boxes are significantly larger or smaller than the objects. This indicates that the deep PLM can discover pedestrians well but fails to localize them precisely. In the learning procedure of deep PLM, object locations could evolve with great randomness, $e.g.$ switching among object parts [63]. Various object parts are capable of optimizing the learning objective by minimizing image classification loss, but experienced difficulty in optimizing object localization. One reason lies in the inconsistency between the frame-level supervision and object-level models. It requires solving non-convex optimization in vast solution spaces, $e.g.$ thousands of images and thousands of object proposals for each frame, which might introduce sub-optimal solutions, as analyzed in Sec. IV, Part A. On the other hand, the deep features are shift invariant and thus tend to learn imprecise localization models given ambiguous object locations.

VII. Conclusion

Supervised learning of detectors for all scenes requires significant human effort on sample annotation. Commonly used transfer learning and semi-supervised learning do not fully reduce human supervision as they require partial object-level annotations. In this work, we show that by leveraging extremely weakly annotated video data it is possible to learn customized pedestrian models for specific video scenes. A progressive latent model (PLM) is proposed by incorporating discriminative and incremental functions. A self-learning approach is implemented by optimizing the model over spatio-temporal object proposals. Experiments demonstrated that the self-learned detectors outperform weakly supervised approaches and transfer learning approaches with significant margins and are comparable to fully supervised ones. The reality behind the superior performance is that the scene-specific object occlusions, camera views, and background structures are well incorporated into the self-learned detectors.

The self-learning approach is also implemented in a deep learning framework by updating the PLM to deep PLM and fine-tuning network parameters with learned objects. Experiments validate that deep PLM is effective to discover pedestrian objects but experiences difficulty to precisely localize them. This indicates future research opportunities for self-learning video detectors using deep learning features.

References

[1] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[3] 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)*, Jun. 2014, pp. 580–587.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[5] S. Zhang, R. Benenson, and B. Schiele. (2017). "CityPersons: A diverse dataset for pedestrian detection." [Online]. Available: https://arxiv.org/abs/1702.05693

[6] H. Hattori, V. N. Boddeti, K. M. Kitani, and T. Kanade, "Learning scene-specific pedestrian detectors without real data," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3819–3827.

[7] I. Misra, A. Shrivastava, and M. Hebert, "Watch and learn: Semi-supervised learning for object detectors from video," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3593–3602.

[8] G. Shu, A. Dehghan, and M. Shah, "Improving an object detector and extracting regions using superpixels," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3721–3727.

[9] Q. Ye, J. Liang, and J. Jiao, "Pedestrian detection in video images via error correcting output code classification of manifold subclasses," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 193–202, Mar. 2012.

[10] A. Prioletti, A. Møgelmose, P. Grisleri, M. M. Trivedi, A. Broggi, and T. B. Moeslund, "Part-based pedestrian detection and feature-based tracking for driver assistance: Real-time, robust algorithms, and evaluation," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1346–1359, Sep. 2013.

[11] W. Ke, J. Chen, J. Jiao, G. Zhao, and Q. Ye, "SRN: Side-output residual network for object symmetry detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1068–1076.

[12] S. Stalder, H. Grabner, and L. Van Gool, "Exploring context to learn scene specific object detectors," in *Proc. PETS*, 2009, pp. 1–8.

[13] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3401–3408.

[14] X. Wang, G. Hua, and T. X. Han, "Detection by detections: Non-parametric detector adaptation for a video," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 350–357.

[15] X. Wang, M. Wang, and W. Li, "Scene-specific pedestrian detection for static video surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 361–374, Feb. 2014.

[16] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell, "Semi-supervised domain adaptation with instance constraints," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 668–675.

[17] Y. Yang, G. Shu, and M. Shah, "Semi-supervised learning of feature hierarchies for object detection in a video," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1650–1657.

[18] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid, "Unsupervised object discovery and tracking in video collections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3173–3181.

[19] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1777–1784.

[20] F. Xiao and Y. J. Lee, "Track and segment: An iterative unsupervised approach for video object proposals," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 933–942.

[21] C.-N. J. Yu and T. Joachims, "Learning structural SVMS with latent variables," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1–21.

[22] Q. Ye, Z. Han, J. Jiao, and J. Liu, "Human detection in images via piecewise linear support vector machines," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 778–789, Feb. 2013.

[23] Q. Ye *et al.*, "Self-learning scene-specific pedestrian detectors using a progressive latent model," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 509–518.

[24] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3361–3369.

[25] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[26] W. Ke, Y. Zhang, P. Wei, Q. Ye, and J. Jiao, "Pedestrian detection via PCA filters based convolutional channel features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., (ICASSP)*, Apr. 2015, pp. 1394–1398.

[27] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1904–1912.

[28] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1259–1267.

[29] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proc. Int. Conf. Image Process. (ICIP)*, 2002, p. 1.

[30] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. 511–518.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[32] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 443–457.

[33] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[34] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 645–659.

[35] X. Zeng, W. Ouyang, M. Wang, and X. Wang, "Deep learning of scene-specific classifier for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 472–487.

[36] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa, and D. Geronimo, "Virtual and real world adaptation for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 797–809, Apr. 2014.

[37] J. Xu, S. Ramos, D. Vázquez, and A. M. López, "Domain adaptation of deformable part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2367–2380, Dec. 2014.

[38] A. Kuznetsova, S. J. Hwang, B. Rosenhahn, and L. Sigal, "Expanding object detector's Horizon: Incremental learning framework for object detection in videos," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 28–36.

[39] Y. Fu, T. M. Hospedales, T. Xiang, Z. Y. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 584–599.

[40] K. Saito, Y. Ushiku, T. Harada, and K. Watanabe, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3723–3732.

[41] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[42] Y. Mao and Z. Yin, "Training a scene-specific pedestrian detector using tracklets," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2015, pp. 170–176.

[43] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[44] H. O. Song, R. B. 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)*, 2014, pp. 1611–1619.

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

[46] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2016.

[47] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 41–48.

[48] Y. J. Lee and K. Grauman, "Learning the easy things first: Self-paced visual category discovery," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1721–1728.

[49] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe. (2016). "Self paced deep learning for weakly supervised object detection." [Online]. Available: https://arxiv.org/abs/1605.07651

[50] K. Wang, X. Yan, D. Zhang, L. Zhang, and L. Lin, "Towards human-machine cooperation: Self-supervised sample mining for object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1605–1613.

[51] B. Wu and R. Nevatia, "Improving part based object detection by unsupervised, online boosting," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[52] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1639–1651, Apr. 2018.

[53] S. Schulter, C. Leistner, P. M. Roth, and H. Bischof, "Unsupervised object discovery and segmentation in videos," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2013, pp. 1–12.

[54] A. Gaidon, G. Zen, and J. A. Rodriguez-Serrano. (2014). "Self-learning camera: Autonomous adaptation of object detectors to unlabeled video streams." [Online]. Available: https://arxiv.org/abs/1406.4296

[55] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[56] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2009.

[57] D. E. Knuth, *The Art of Computer Programming: Sorting and Searching*, vol. 3, 3rd ed. Reading, MA, USA: Addison-Wesley, 1997.

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

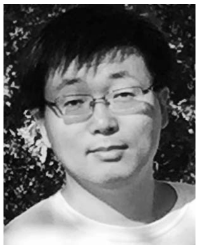
[59] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Sys. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

[60] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

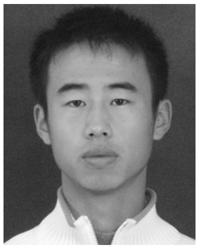
[61] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *Proc. 12th IEEE Int. Workshop Perform. Eval. Tracking Surveill.*, Dec. 2009, pp. 1–6.

[62] B. Benfold and I. D. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3457–3464.

[63] F. Wan, P. Wei, Z. Han, J. Jiao, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1297–1306.



Qixiang Ye (M'10–SM'15) received the B.S. and M.S. degrees in mechanical and electrical engineering from the Harbin Institute of Technology, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2006. He was a Visiting Assistant Professor with the Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD, USA, in 2013, and a Visiting Scholar with IIID of Duke University in 2016. Since 2016, he has been a Professor with the University of Chinese Academy of Sciences. He pioneered the Kernel SVM-based pyrolysis output prediction software which was put into practical application by SINOPEC in 2012. He developed two kinds of piecewise linear SVM methods which were successfully applied into visual object detection. His research interests include image processing, visual object detection, and machine learning. He has published over 100 papers in refereed conferences and journals, including the IEEE CVPR, ICCV, ECCV, the IEEE TRANSACTIONS ITS, TIP, and PAMI. He received the Sony Outstanding Paper Award.



Tianliang Zhang received the B.S. degree in electronic information engineering from the Wuhan University of Technology (WUT) in 2013, and the M.S. degree in industrial engineering from the University of Chinese Academy of Sciences in 2017, where he is currently pursuing the Ph.D. degree with the School of Electronic, Electrical, and Communication Engineering. His research interests include computer vision and deep learning.

Wei Ke (S'15) received the B.S. degree in electrical engineering and automation from Beihang University, Beijing, China, in 2011, and the Ph.D. degree from the University of Chinese Academy of Sciences in 2018. In 2016, he visited the Center for Machine Vision, University of Oulu, as a joint Ph.D. student, supported by the China Scholarship Council (CSC). He is currently a Postdoctoral Researcher with the Human Sensing Laboratory, Carnegie Mellon University (CMU). His research interests include computer vision and deep learning. He has published 10 papers in refereed conferences and journals, including the IEEE CVPR and ECCV. He was a recipient of the President Award of the Chinese Academy of Sciences in 2017.