Pedestrian Detection aided by Deep Learning Semantic Tasks

Yonglong Tian[1], Ping Luo[3,1], Xiaogang Wang[2,3], Xiaoou Tang[1,3]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Department of Electronic Engineering, The Chinese University of Hong Kong
[3]Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology,
Chinese Academy of Sciences, Shenzhen, China
{ty014,pluo,xtang}@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

Abstract

*Deep learning methods have achieved great successes in pedestrian detection, owing to its ability to learn discriminative features from raw pixels. However, they treat pedestrian detection as a single binary classification task, which may confuse positive with hard negative samples (Fig.1 (a)). To address this ambiguity, this work jointly optimize pedestrian detection with semantic tasks, including pedestrian attributes (*e.g. *'carrying backpack') and scene attributes (*e.g. *'vehicle', 'tree', and 'horizontal'). Rather than expensively annotating scene attributes, we transfer attributes information from existing scene segmentation datasets to the pedestrian dataset, by proposing a novel deep model to learn high-level features from multiple tasks and multiple data sources. Since distinct tasks have distinct convergence rates and data from different datasets have different distributions, a multi-task deep model is carefully designed to coordinate tasks and reduce discrepancies among datasets. Extensive evaluations show that the proposed approach outperforms the state-of-the-art on the challenging Caltech [9] and ETH [10] datasets where it reduces the miss rates of previous deep models by 17 and 5.5 percent, respectively.*

1. Introduction

Pedestrian detection has attracted wide attentions [5, 31, 28, 7, 8, 9, 17, 6, 36, 13]. This problem is challenging because of large variations and confusions in the human body and background, as shown in Fig.1 (a), where the positive and hard negative patches have large ambiguities.

Current methods for pedestrian detection can be generally grouped into two categories, the models based on hand-crafted features [31, 5, 32, 8, 7, 35, 11] and deep models [21, 23, 28, 22, 16]. In the first category, conventional methods extracted Haar [31], HOG[5], or HOG-LBP [32] from images to train SVM [5] or boosting classifiers [8].

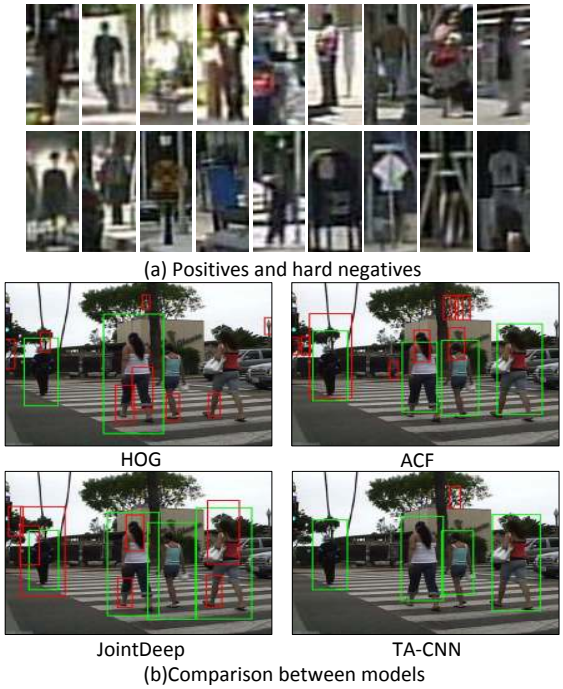


Figure 1: Distinguishing pedestrians from hard negatives is challenging due to their visual similarities. In (a), the first and second row represent pedestrians and equivocal background samples respectively. (b) shows that our TA-CNN rejects more hard negatives than the detectors using hand-crafted features (such as HOG [5] and ACF [7]) and the JointDeep model [22].

The learned weights of the classifier (*e.g.* SVM) can be considered as a global template of the entire human body. To account for more complex poses, the hierarchical deformable part models (DPM) [11, 37, 15] learned a mixture of local templates for each body part. Although they are sufficient to certain pose changes, the feature representations and the classifiers cannot be jointly optimized to improve performance. In the second category, deep neural networks

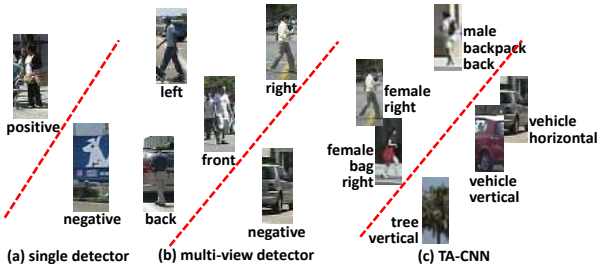


Figure 2: Comparisons between different detectors.

achieved promising results [21, 23, 28, 22, 16], owing to their capacity to learn discriminative features from raw pixels. For example, Ouyang *et al*. [22] learned features by designing specific hidden layers for the Convolutional Neural Network (CNN), such that features, deformable parts, and pedestrian classification can be jointly optimized.

However, previous deep models treated pedestrian detection as a single binary classification task, which are not able to capture rich pedestrian variations, as shown in Fig.1 (a).

This work jointly optimizes pedestrian detection with auxiliary semantic tasks, including pedestrian attributes (*e.g*. 'backpack', 'gender', and 'views') and scene attributes (*e.g*. 'vehicle', 'tree', and 'vertical'). To understand how this work, we provide an example in Fig.2. If only a single detector is used to classify all the positive and negative samples in Fig.2 (a), it is difficult to handle complex pedestrian variations. Therefore, the mixture models of multiple views were developed in Fig.2 (b), *i.e*. pedestrian images in different views are handled by different detectors. If views are treated as one type of semantic tasks, learning pedestrian representation by multiple attributes with deep models actually extends this idea to extreme. As shown in Fig.2 (c), more supervised information enriches the learned features to account for combinatorial more pedestrian variations. The samples with similar configurations of attributes can be grouped and separated in the high-level feature space.

Specifically, given a pedestrian dataset, denoted as $\mathbf{P}$, the positive image patches are manually labeled with several pedestrian attributes, which are suggested to be valuable for surveillance analysis [20]. However, as the number of negatives is significantly larger than the number of positives, we transfer scene attributes information from existing background scene segmentation databases (each one is denoted as $\mathbf{B}$) to the pedestrian dataset, other than annotating them manually. A novel task-assistant CNN (TA-CNN) is proposed to jointly learn multiple tasks using multiple data sources. As different $\mathbf{B}$'s may have different data distributions, to reduce these discrepancies, we transfer two types of scene attributes that are carefully chosen, comprising the shared attributes that appear across all the $\mathbf{B}$'s and the unshared attributes that appear in only one of them. The former one facilitates the learning of shared representation among $\mathbf{B}$'s, whilst the latter one increases diversities of attributes. Furthermore, to reduce the gaps between $\mathbf{P}$ and $\mathbf{B}$'s, we first project each sample in $\mathbf{B}$'s to a structural space of $\mathbf{P}$ and then the projected values are employed as input to train TA-CNN.

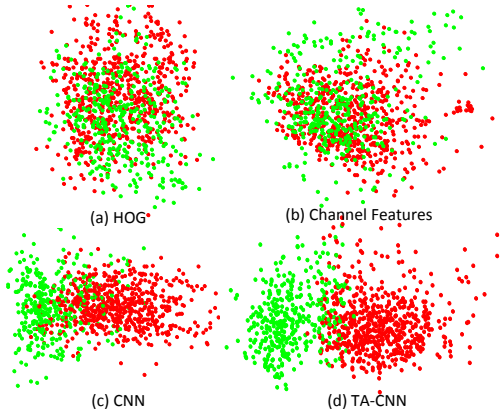


Figure 4: Feature spaces of HOG, channel features, CNN that models pedestrian detection as binary classification, and TA-CNN. Positive and hard negative samples of the Caltech-Test set [9] are represented by red and green, respectively.

This work has the following main **contributions**. (1) To our knowledge, this is the first attempt to learn discriminative representation for pedestrian detection by jointly optimizing it with semantic attributes, including pedestrian attributes and scene attributes. The scene attributes can be transferred from existing scene datasets without annotating manually. (2) These multiple tasks from multiple sources are trained using a single task-assistant CNN (TA-CNN), which is carefully designed to bridge the gaps between different datasets. (3) We systematically investigate the effectiveness of attributes in pedestrian detection. Extensive experiments on both challenging Caltech [9] and ETH [10] datasets demonstrate that TA-CNN outperforms state-of-the-art methods. It reduces miss rates of existing deep models on these datasets by 17 and 5.5 percent, respectively.

1.1. Related Works

We review recent works in two aspects.

Models based on Hand-Crafted Features The hand-crafted features, such as HOG, LBP, and channel features, achieved great success in pedestrian detection. For example, Wang *et al*. [32] utilized the LBP+HOG features to deal with partial occlusion of pedestrian. Chen *et al*. [4] modeled the context information in a multi-order manner. The deformable part models [11] learned mixture of local templates to account for view and pose variations. Moreover, Dollár *et al*. proposed Integral Channel Features

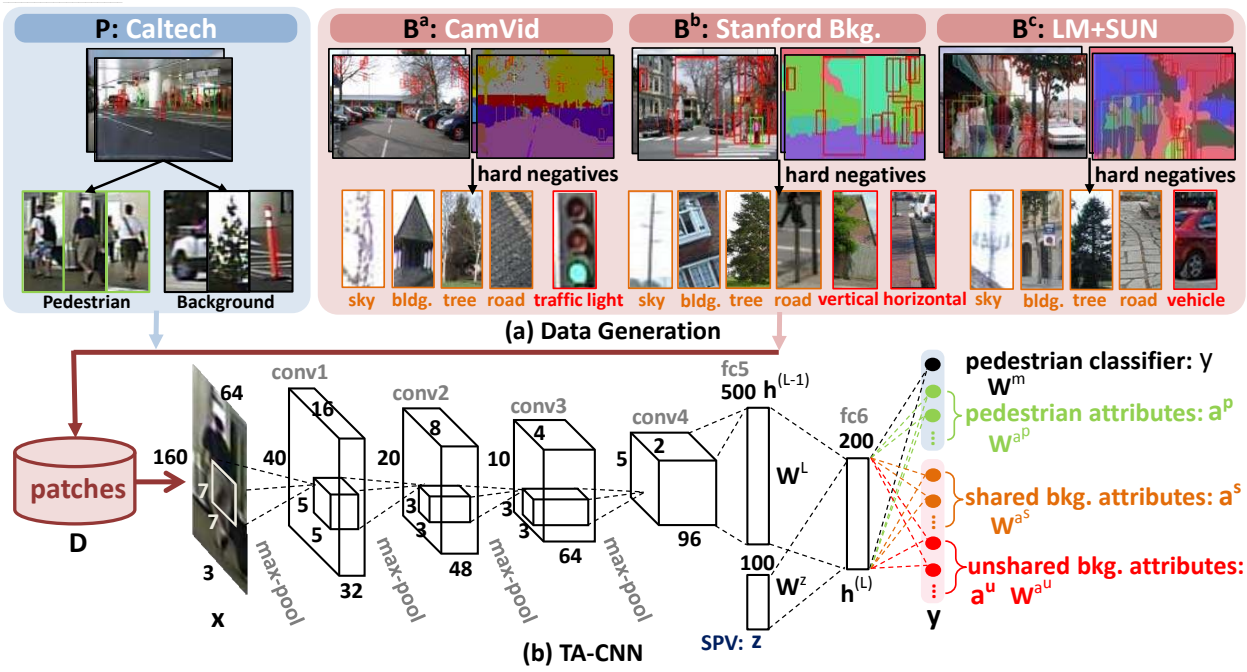


Figure 3: The proposed pipeline for pedestrian detection.

(ICF) [8] and Aggregated Channel Features (ACF) [7], both of which consist of gradient histogram, gradients, and LUV, and can be efficiently extracted. Benenson *et al*. [1] combined channel features and depth information. However, the representation of hand-crafted features cannot be optimized for pedestrian detection. They are not able to capture large variations, as shown in Fig.4 (a) and (b).

Deep Models Deep learning methods can learn features from raw pixels to improve the performance of pedestrian detection. For example, ConvNet [28] employed convolutional sparse coding to unsupervised pre-train CNN for pedestrian detection. Ouyang *et al*. [21] jointly learned features and the visibility of different body parts to handle occlusion. The JointDeep model [22] designed a deformation hidden layer for CNN to model mixture poses information. Unlike the previous deep models that formulated pedestrian detection as a single binary classification task, TA-CNN jointly optimizes pedestrian detection with related semantic tasks. The learned features are more robust to large variations, as shown in Fig.4 (c) and (d). Another contemporaneous deep model [13] seems complementary to our method.

2. Our Approach

Method Overview Fig.3 shows our pipeline of pedestrian detection, where pedestrian classification, pedestrian attributes, and scene attributes are jointly learned by a single TA-CNN. Given a pedestrian dataset $\mathbf{P}$, for example Caltech [9], we manually label the positive patches with nine pedestrian attributes, which are listed in Fig.5. Most of them are suggested by the UK Home Office and UK police and valuable in surveillance analysis [20]. Since the number of negative patches in $\mathbf{P}$ is significantly larger than the number of positives, we transfer scene attribute information from three public scene segmentation datasets to $\mathbf{P}$, as shown in Fig.3 (a), including CamVid ($\mathbf{B}^a$) [3], Stanford Background ($\mathbf{B}^b$) [12], and LM+SUN ($\mathbf{B}^c$) [29], where hard negatives are chosen by applying a simple yet fast pedestrian detector [7] on these datasets. As the data in different $\mathbf{B}$'s are sampled from different distributions, we carefully select two types of attributes, the shared attributes (outlined in orange) that present in all $\mathbf{B}$'s and the unshared attributes (outlined in red) that appear only in one of them. This is done because the former one enables the learning of shared representation across $\mathbf{B}$'s, while the latter one enhances diversities of attributes. All chosen attributes are summarized in Fig.5, where shows that data from different sources have different subset of attribute labels. For example, pedestrian attributes only present in $\mathbf{P}$, shared attributes present in all $\mathbf{B}$'s, and the unshared attributes present in one of them, *e.g*. 'traffic light' of $\mathbf{B}^a$.

We construct a training set $\mathbf{D}$ by combing patches cropped from both $\mathbf{P}$ and $\mathbf{B}$'s. Let $\mathbf{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$ be a set of image patches and their labels, where each $\mathbf{y}_n = (y_n, \mathbf{o}_n^p, \mathbf{o}_n^s, \mathbf{o}_n^u)$ is a four-tuple[1]. Specifically, y_n denotes a binary label, indicating whether an image patch

[1]In this paper, scalar variable is denoted by normal letter, while set, vector, or matrix is denoted as boldface letter.

	Pedestrian Attributes									Scene Attributes							
										Shared				Unshared			
	$p1$	$p2$	$p3$	$p4$	$p5$	$p6$	$p7$	$p8$	$p9$	$s1$	$s2$	$s3$	$s4$	$u1$	$u2$	$u3$	$u4$
	Backpack	Dark-Trousers	Hat	Bag	Gender	Occlusion	Riding	Viewpoint	White-Clothes	Sky	Tree	Building	Road	Traffic-light	Horizontal	Vertical	Vehicle
Caltech (**P**)	√	√	√	√	√	√	√	√	√								
CamVid (**B^a**)										√	√	√	√	√			
Stanford (**B^b**)										√	√	√	√		√	√	
LM+SUN (**B^c**)										√	√	√	√				√

Figure 5: Attribute summarization.

is pedestrian or not. $\mathbf{o}_n^p = \{o_n^{pi}\}_{i=1}^9$, $\mathbf{o}_n^s = \{o_n^{si}\}_{i=1}^4$, and $\mathbf{o}_n^u = \{o_n^{ui}\}_{i=1}^4$ are three sets of binary labels, representing the pedestrian, shared scene, and unshared scene attributes, respectively. As shown in Fig.3 (b), TA-CNN employs image patch $\mathbf{x}_n$ as input and predicts $\mathbf{y}_n$, by stacking four convolutional layers (conv1 to conv4), four max-pooling layers, and two fully-connected layers (fc5 and fc6). This structure is inspired by the AlexNet [14] for large-scale general object categorization. However, as the difficulty of pedestrian detection is different from general object categorization, we remove one convolutional layer of AlexNet and reduce the number of parameters at all remaining layers. The subsequent structure of TA-CNN is specified in Fig.3 (b).

Formulation of TA-CNN Each hidden layer of TA-CNN from conv1 to conv4 is computed recursively by convolution and max-pooling. Each hidden layer in fc5 and fc6 is obtained by a fully-connected transformation. For all these layers, we utilize the rectified linear function [18] as the activation function.

TA-CNN can be formulated as minimizing the log posterior probability with respect to a set of network parameters $\mathcal{W}$

$$\mathcal{W}^* = \arg\min_{\mathcal{W}} -\sum_{n=1}^{N} \log p(y_n, \mathbf{o}_n^p, \mathbf{o}_n^s, \mathbf{o}_n^u | \mathbf{x}_n; \mathcal{W}), \quad (1)$$

where $E = -\sum_{n=1}^{N} \log p(y_n, \mathbf{o}_n^p, \mathbf{o}_n^s, \mathbf{o}_n^u | \mathbf{x}_n)$ is a complete loss function regarding the entire training set. Here, we illustrate that the shared attributes $\mathbf{o}_n^s$ in Eqn.(1) are crucial to learn shared representation across multiple scene datasets $\mathbf{B}$'s.

For clarity, we keep only the unshared attributes, $\mathbf{o}_n^u$, and the loss function becomes $E = -\sum_{n=1}^{N} \log p(\mathbf{o}_n^u | \mathbf{x}_n)$. Let $\mathbf{x}_n^a$ denote the n-th sample of scene dataset $\mathbf{B}^a$. A shared representation can be learned if and only if all the samples share at least one target (attribute). Since the samples are independent, the loss function can be expanded as $E = -\sum_{i=1}^{I} \log p(o_i^{u1} | \mathbf{x}_i^a) - \sum_{j=1}^{J} \log p(o_j^{u2}, o_j^{u3} | \mathbf{x}_j^b) - \sum_{k=1}^{K} \log p(o_k^{u4} | \mathbf{x}_k^c)$, where $I + J + K = N$, implying that each dataset is only used to optimize its corresponding unshared attribute, although all the datasets and attributes are trained in a single TA-CNN. For instance, the classification model of o^{u1} is learned by using $\mathbf{B}^a$ without leveraging the existence of the other datasets. In other words, the probability of $p(o^{u1} | \mathbf{x}^a, \mathbf{x}^b, \mathbf{x}^c) = p(o^{u1} | \mathbf{x}^a)$ because of missing labels. The above formulation is not sufficient to learn shared features among datasets, especially when the data have large differences. To bridge multiple scene datasets $\mathbf{B}$'s, we introduce the shared attributes $\mathbf{o}^s$, the loss function develops into $E = -\sum_{n=1}^{N} \log p(\mathbf{o}_n^s, \mathbf{o}_n^u | \mathbf{x}_n)$, such that TA-CNN can learn a shared representation across $\mathbf{B}$'s because the samples share common targets $\mathbf{o}^s$, *i.e.* $p(o_n^{s1}, o_n^{s2}, o_n^{s3}, o_n^{s4} | \mathbf{x}_n^a, \mathbf{x}_n^b, \mathbf{x}_n^c)$.

Now, we reconsider Eqn.(1), where the loss function can be decomposed similarly as above, $E = -\sum_{i=1}^{I} \log p(\mathbf{o}_i^s, o_i^{u1} | \mathbf{x}_i^a) - \sum_{j=1}^{J} \log p(\mathbf{o}_j^s, o_j^{u2}, o_j^{u3} | \mathbf{x}_j^b) - \sum_{k=1}^{K} \log p(\mathbf{o}_k^s, o_k^{u4} | \mathbf{x}_k^c) - \sum_{\ell=1}^{L} \log p(y_\ell, \mathbf{o}_\ell^p | \mathbf{x}_\ell^p)$. Even though the discrepancies among $\mathbf{B}$'s can be reduced by $\mathbf{o}^s$, this decomposition shows that gap remains between datasets $\mathbf{P}$ and $\mathbf{B}$'s. To resolve this issue, we compute the structure projection vectors $\mathbf{z}_n$ for each sample $\mathbf{x}_n$, and Eqn.(1) turns into

$$\mathcal{W}^* = \arg\min_{\mathcal{W}} -\sum_{n=1}^{N} \log p(y_n, \mathbf{o}_n^p, \mathbf{o}_n^s, \mathbf{o}_n^u | \mathbf{x}_n, \mathbf{z}_n; \mathcal{W}). \quad (2)$$

For example, the first term of the above decomposition can be written as $p(\mathbf{o}_i^s, o_i^{u1} | \mathbf{x}_i^a, \mathbf{z}_i^a)$, where $\mathbf{z}_i^a$ is attained by projecting the corresponding $\mathbf{x}_i^a$ in $\mathbf{B}^a$ on the feature space of $\mathbf{P}$. This procedure is explained below. Here $\mathbf{z}_i^a$ is used to bridge multiple datasets, because samples from different datasets are projected to a common space of $\mathbf{P}$. TA-CNN adopts a pair of data $(\mathbf{x}_i^a, \mathbf{z}_i^a)$ as input (see Fig.3 (b)). All the remaining terms can be derived in a similar way.

Structure Projection Vector As shown in Fig.6, to close the gap between $\mathbf{P}$ and $\mathbf{B}$s, we calculate the structure projection vector (SPV) for each sample by organizing the positive (+) and negative (-) data of $\mathbf{P}$ into two tree structures, respectively. Each tree has depth that equals three and partitions the data top-down, where each child node groups the data of its parent into clusters, for example C_1^1 and C_5^{10}. Then, SPV of each sample is obtained by concatenating the distance between it and the mean of each leaf node. Specifically, at each parent node, we extract HOG feature for each sample and apply k-means to group the data. We partition the data into five clusters (C_1 to C_5) in the first level, and then each of them is further partitioned into ten clusters, *e.g.* C_1^1 to C_1^{10}. As a result, the length of SPV for each sample is $2 \times 5 \times 10 = 100$.

3. Learning Task-Assistant CNN

To learn network parameters $\mathcal{W}$, a natural way is to reformulate Eqn.(2) as the softmax loss functions similar

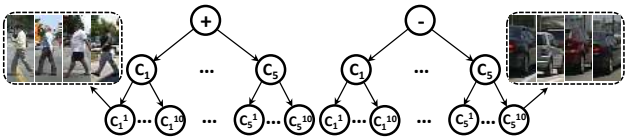


Figure 6: The hierarchical structure of positive and negative samples.

to the previous methods. We have[2]

$$E \triangleq -y \log p(y|\mathbf{x},\mathbf{z}) - \sum_{i=1}^{9} \alpha_i o^{pi} \log p(o^{pi}|\mathbf{x},\mathbf{z}) - \sum_{j=1}^{4} \beta_j o^{sj} \log p(o^{sj}|\mathbf{x},\mathbf{z}) - \sum_{k=1}^{4} \gamma_k o^{uk} \log p(o^{uk}|\mathbf{x},\mathbf{z}), \tag{3}$$

where the *main task* is to predict the pedestrian label y and the attribute estimations, *i.e.* o^{pi}, o^{sj}, and o^{uk}, are auxiliary semantic tasks. α, β, and γ denote the importance coefficients to associate multiple tasks. Here, $p(y|\mathbf{x},\mathbf{z})$, $p(o^{pi}|\mathbf{x},\mathbf{z})$, $p(o^{sj}|\mathbf{x},\mathbf{z})$, and $p(o^{uk}|\mathbf{x},\mathbf{z})$ are modeled by softmax functions, for example, $p(y=0|\mathbf{x},\mathbf{z}) = \frac{\exp(\mathbf{W}_{\cdot 1}^{m\mathsf{T}}\mathbf{h}^{(L)})}{\exp(\mathbf{W}_{\cdot 1}^{m\mathsf{T}}\mathbf{h}^{(L)})+\exp(\mathbf{W}_{\cdot 2}^{m\mathsf{T}}\mathbf{h}^{(L)})}$, where $\mathbf{h}^{(L)}$ and $\mathbf{W}^m$ indicate the top-layer feature vector and the parameter matrix of the main task y respectively, as shown in Fig.3 (b), and $\mathbf{h}^{(L)}$ is obtained by $\mathbf{h}^{(L)} = \text{relu}(\mathbf{W}^{(L)}\mathbf{h}^{(L-1)} + \mathbf{b}^{(L)} + \mathbf{W}^z\mathbf{z} + \mathbf{b}^z)$.

Eqn.(3) optimizes eighteen loss functions together. It has two main drawbacks. First, since different tasks have different convergence rates, training many tasks together suffers from over-fitting. Second, if the dimension of the features $\mathbf{h}^{(L)}$ is high, the number of parameters at the top-layer increases rapidly. For example, if the feature vector $\mathbf{h}^{(L)}$ has H dimensions, the weight matrix of each two-state variable (*e.g.* $\mathbf{W}^m$ of the main task) has $2 \times H$ parameters, whilst the weight matrix of the four-state variable 'viewpoint' has $4 \times H$ parameters[3]. As we have seventeen two-state variables and one four-state variable, the total number of parameters at the top-layer is $17 \times 2 \times H + 4 \times H = 38H$.

To resolve the above issues, we cast learning multiple tasks in Eqn.(3) as optimizing a single multivariate cross-entropy loss,

$$E \triangleq -\mathbf{y}^{\mathsf{T}}\text{diag}(\boldsymbol{\lambda}) \log p(\mathbf{y}|\mathbf{x},\mathbf{z}) - (\mathbf{1}-\mathbf{y})^{\mathsf{T}}\text{diag}(\boldsymbol{\lambda})(\log \mathbf{1} - p(\mathbf{y}|\mathbf{x},\mathbf{z})), \tag{4}$$

where $\boldsymbol{\lambda}$ denotes a vector of tasks' importance coefficients and $\text{diag}(\cdot)$ represents a diagonal matrix. Here, $\mathbf{y} = (y, \mathbf{o}^p, \mathbf{o}^s, \mathbf{o}^u)$ is a vector of binary labels, concatenating the pedestrian label and all attribute labels. Note that each two-state (four-state) variable can be described by one bit (two bits). Since we have seventeen two-state variables and one four-state variable, the weight matrix at the top layer, denoted as $\mathbf{W}^y$ in this case, has $17 \times H + 2 \times H = 19H$ parameters, which reduces the number of parameters by half, *i.e.* $19H$ compared to $38H$ of Eqn.(3). Moreover, $p(\mathbf{y}|\mathbf{x},\mathbf{z})$ is modeled by sigmoid function, *i.e.* $p(\mathbf{y}|\mathbf{x},\mathbf{z}) = \frac{1}{1+\exp(-\mathbf{W}^{y\mathsf{T}}\mathbf{h}^{(L)})}$, where $\mathbf{h}^{(L)}$ is achieved in the same way as in Eqn.(3).

The network parameters are updated by minimizing Eqn.(4) using stochastic gradient descent [14] and back-propagation (BP) [27], where the error of the output layer is propagated top-down to update filters or weights at each layer. The BP procedure is similar to [14]. The main difference is how to compute error at the L-th layer. In the traditional BP algorithm, the error $e^{(L)}$ at the L-th layer is obtained by the gradient of Eqn.(4), indicating the loss, *i.e.* $e^{(L)} = \overline{\mathbf{y}} - \mathbf{y}$, where $\overline{\mathbf{y}}$ denotes the predicted labels. However, unlike the conventional BP where all the labels are observed, each of our dataset only covers a subset of attributes. Let $\widehat{\mathbf{o}}$ signify the unobserved labels. The posterior probability of Eqn.(4) becomes $p(\mathbf{y}_{\setminus\widehat{\mathbf{o}}}, \widehat{\mathbf{o}}|\mathbf{x},\mathbf{z})$, where $\mathbf{y}_{\setminus\widehat{\mathbf{o}}}$ specifies the labels $\mathbf{y}$ excluding $\widehat{\mathbf{o}}$. Here we demonstrate that $\widehat{\mathbf{o}}$ can be simply marginalized out, since the labels are independent. We have $\sum_{\widehat{\mathbf{o}}} p(\mathbf{y}_{\setminus\widehat{\mathbf{o}}}, \widehat{\mathbf{o}}|\mathbf{x},\mathbf{z}) = p(\mathbf{y}_{\setminus\widehat{\mathbf{o}}}|\mathbf{x},\mathbf{z}) \cdot \sum_{\widehat{o_1}} p(\widehat{o_1}|\mathbf{x},\mathbf{z}) \cdot \sum_{\widehat{o_2}} p(\widehat{o_2}|\mathbf{x},\mathbf{z}) \cdot \ldots \cdot \sum_{\widehat{o_j}} p(\widehat{o_j}|\mathbf{x},\mathbf{z}) = p(\mathbf{y}_{\setminus\widehat{\mathbf{o}}}|\mathbf{x},\mathbf{z})$. Therefore, the error $e^{(L)}$ of Eqn.(4) can be computed as

$$e^{(L)} = \begin{cases} \overline{y} - y, & \text{if } y \in \mathbf{y}_{\setminus\widehat{\mathbf{o}}}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

which demonstrates that the errors of the missing labels will not be propagated no matter whether their predictions are correct or not.

We fix the important coefficient $\lambda_1 \in \boldsymbol{\lambda}$ of the main task y, *i.e.* $\lambda_1 = 1$. As the auxiliary tasks are independent, their coefficients can be obtained by greedy search between zero and one. To simplify the learning procedure, we have $\forall \lambda_i \in \boldsymbol{\lambda}, \lambda_i = 0.1, i = 2, 3, ..., 18$ and found that this setting provides stable and reasonable good results.

4. Experiments

The proposed TA-CNN[4] is evaluated on the Caltech-Test [9] and ETH datasets [10]. We strictly follow the evaluation protocol proposed in [9], which measures the log average

[2] We drop the sample index n in the remaining derivation for clarity.

[3] All tasks are binary classification (*i.e.* two states) except the pedestrian attribute 'viewpoint', which has four states, including 'front', 'back', 'left', and 'right'.

[4] http://mmlab.ie.cuhk.edu.hk/projects/TA-CNN/ The corresponding author is Ping Luo (pluo.lhi@gmail.com).

main task	backpack	dark-trousers	hat	bag	gender	occlusion	riding	white-cloth	viewpoint	**All**
31.45	30.44	29.83	28.89	30.77	30.70	29.36	28.83	30.22	28.20	**25.64**

Table 1: Log-average miss rate (%) on Caltech-Test with pedestrian attribute learning tasks.

	main task	sky	tree	building	road	vehicle	traffic-light	vertical	horizontal
Neg.	31.45	31.07	30.92	31.16	31.02	30.75	30.85	30.91	30.96
Attr.		30.79	30.50	30.90	30.54	29.41	28.92	30.03	30.40

Table 2: Log-average miss rate (%) on Caltech-Test with scene attribute learning tasks.

miss rate over nine points ranging from 10^{-2} to 10^{0} False-Positive-Per-Image. We compare TA-CNN with the best-performing methods as suggested by the Caltech and ETH benchmarks[5] on the *reasonable* subsets, where pedestrians are larger than 49 pixels height and have 65 percent visible body parts.

4.1. Effectiveness of TA-CNN

We systematically study the effectiveness of TA-CNN in four aspects as follows. In this section, TA-CNN is trained on Caltech-Train and tested on Caltech-Test.

Effectiveness of Hard Negative Mining To save computational cost, We employ ACF [7] for mining hard negatives at the training stage and pruning candidate windows at the testing stage. Two main adjustments are made in ACF. First, we compute the exact feature pyramids at each scale instead of making an estimated aggregation. Second, we increase the number of weak classifiers to enhance the recognition ability. Afterwards, a higher recall rate is achieved by ACF and it obtains 37.31 percent miss rate on Caltech-Test. Then the TA-CNN with only the main task (pedestrian classification) achieved 31.45 percent miss rate by cascading on ACF, obtaining more than 5 percent improvement.

Effectiveness of Pedestrian Attributes We investigate how different pedestrian attributes can help improve the main task. To this end, we train TA-CNN by combing the main task with each of the pedestrian attributes, and the miss rates are reported in Table 1, where shows that 'viewpoint' is the most effective attribute, which improves the miss rate by 3.25 percent, because 'viewpoint' captures the global information of pedestrian. The attribute capture the pose information also attains significant improvement, *e.g.* 2.62 percent by 'riding'. Interestingly, among those attributes modeling local information, 'hat' performs the best, reducing the miss rate by 2.56 percent. We observe that this result is consistent with previous works, SpatialPooling [25] and InformedHaar [35], which showed that head is the most informative body parts for pedestrian detection. When combining all the pedestrian attributes, TA-CNN achieved 25.64 percent miss rate, improving the main task by 6 percent.

Effectiveness of Scene Attributes Similarly, we study how different scene attributes can improve pedestrian detection. We train TA-CNN combining the main task with each scene attribute. For each attribute, we select $5,000$ hard negative samples from its corresponding dataset. For example, we crop five thousand patches for 'vertical' from the Stanford Background dataset. We test two settings, denoted as "Neg." and "Attr.". In the first setting, we label the five thousand patches as negative samples. In the second setting, these patches are assigned to their original attribute labels. The former one uses more negative samples compared to the TA-CNN (main task), whilst the latter one employs attribute information.

The results are reported in Table 2, where shows that 'traffic-light' improves the main task by 2.53 percent, revealing that the patches of 'traffic-light' are easily confused with positives. This is consistent when we exam the hard negative samples of most of the pedestrian detectors. Besides, the 'vertical' background patches are more effective than the 'horizontal' background patches, corresponding to the fact that hard negative patches are more likely to present vertically.

Attribute Prediction We also consider the accuracy of attribute prediction and find that the averaged accuracy of all the attributes exceeds 75 percent. We select the pedestrian attribute 'viewpoint' as illustration. In Table 3, we report the confusion matrix of 'viewpoint', where the number of detected pedestrians of 'front', ''back', ''left', and 'right' are 283, 276, 220, 156 respectively. We observed that 'front' and 'back' information are relatively easy to capture, rather than the 'left' and 'right', which are more likely to confuse with each other, *e.g.* $21 + 40 = 61$ misclassified samples.

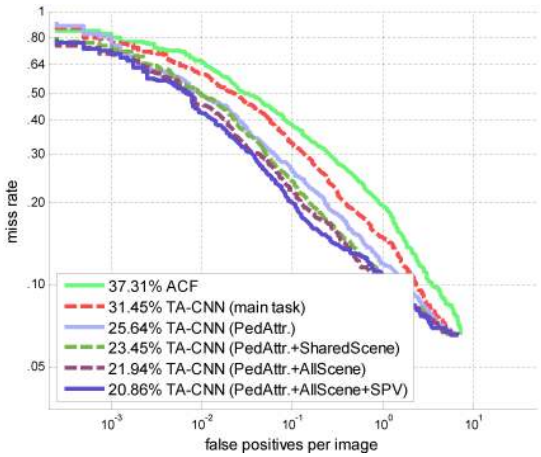
4.2. Overall Performance on Caltech

We report overall results in two parts. All the results of TA-CNN are obtained by training on Caltech-Train and evaluating on Caltech-Test. In the first part, we analyze the performance of different components of TA-CNN. As shown in Fig.7a, the performances show clear increasing patterns when gradually adding more components. For example, TA-CNN (main task) cascades on ACF and re-

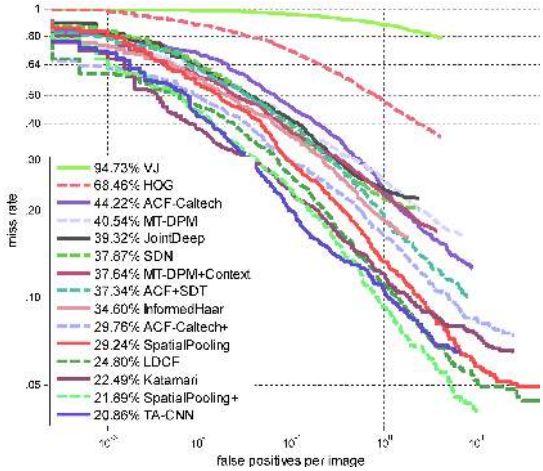
[5] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

		Predict State			
		Frontal	Back	Left	Right
True State	Frontal	226	32	15	10
	Back	24	232	12	8
	Left	22	13	164	21
	Right	5	15	40	96
Accuracy		0.816	0.796	0.701	0.711

Table 3: View-point estimation results on Caltech-Test.



(a) Log-average miss rate reduction procedure



(b) Overall Performance on Caltech-Test

Figure 7: Results under standard evaluation settings

duces the miss rate of it by more than 5 percent. TA-CNN (PedAttr.+SharedScene) reduces the result of TA-CNN (PedAttr.) by 2.2 percent, because it can bridge the gaps among multiple scene datasets. After modeling the unshared attributes, the miss rate is further decreased by 1.5 percent, since more attribute information is incorporated. The final result of 20.86 miss rate is obtained by using the structure projection vector as input to TA-CNN. Its effectiveness has been demonstrated in Fig.7a.

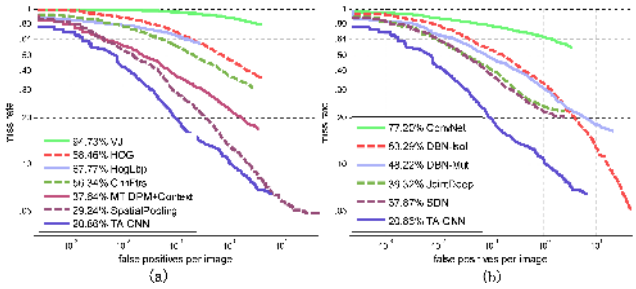


Figure 8: Results on Caltech-Test: (a) comparison with hand-crafted feature based models; (b) comparison with other deep models

In the second part, we compare the result of TA-CNN with all existing best-performing methods, including VJ [30], HOG [5], ACF-Caltech [7], MT-DPM [33], MT-DPM+Context [33], JointDeep [22], SDN [16], ACF+SDT [26], InformedHaar [35], ACF-Caltech+ [19], SpatialPooling [25], LDCF [19], Katamari [2], SpatialPooling+ [24]. These works used various features, classifiers, deep networks, and motion and context information. We summarize them as below. Note that TA-CNN dose not employ motion and context information.

Features: Haar (VJ), HOG (HOG, MT-DPM), ChannelFeature (ACF+Caltech, LDCF); **Classifiers**: latent-SVM (MT-DPM), boosting (VJ, ACF+Caltech, SpatialPooling); **Deep Models**: JointDeep, SDN; **Motion and context**: MT-DPM+Context, ACF+SDT, Katamari, SpatialPooling+.

Fig.7b reports the results. TA-CNN achieved the smallest miss rate compared to all existing methods. Although it only outperforms the second best method (SpatialPooling+ [24]) by 1 percent, it learns 200 dimensions high-level features with attributes, other than combining LBP, covariance features, channel features, and video motion as in [24]. Also, the Katamari [2] method integrates multiple types of features and context information.

Hand-crafted Features The learned high-level representation of TA-CNN outperforms the conventional hand-crafted features by a large margin, including Haar, HOG, HOG+LBP, and channel features, shown in Fig.8 (a). For example, it reduced the miss rate by 16 and 9 percent compared to DPM+Context and Spatial Pooling, respectively. DPM+Context combined HOG feature with pose mixture and context information, while SpatialPooling combined multiple features, such as LBP, covariance, and channel features.

Deep Models Fig.8 (b) shows that TA-CNN surpasses other deep models. For example, TA-CNN outperforms two state-of-the-art deep models, JointDeep and SDN, by 18 and 17 percent, respectively. Both SDN and JointDeep treated pedestrian detection as a single task and thus cannot learn high-level representation to deal with the challenging hard

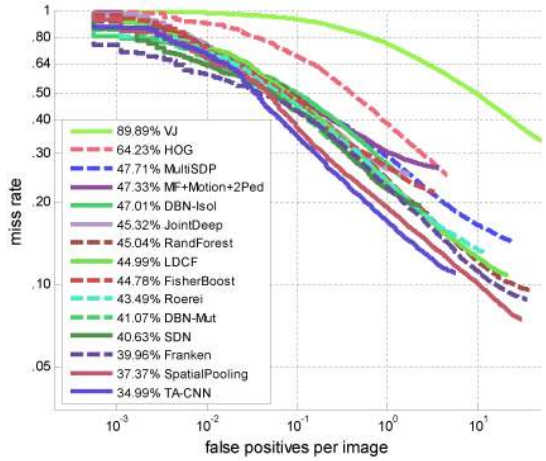


Figure 9: Results on ETH

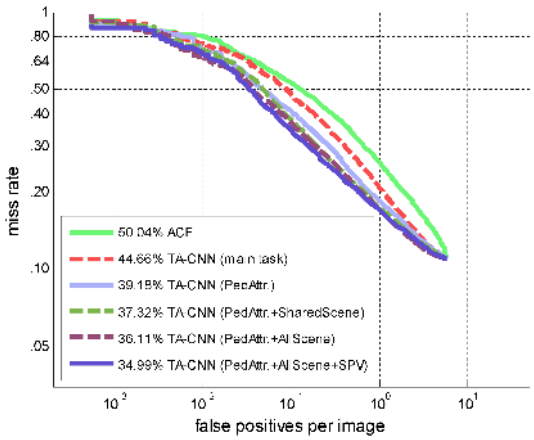


Figure 10: Log-average miss rate reduction procedure on ETH

negative samples.

4.3. Overall Performance on ETH

We compare TA-CNN with the existing best-performing methods (see Sec.4.2) on ETH [10]. TA-CNN is trained on INRIA-Train [5]. This setting aims at evaluating the generalization capacity of the TA-CNN. As shown in Fig.9, TA-CNN achieves the lowest miss rate, which outperforms the second best method by 2.5 percent. It also outperforms the best deep model by 5.5 percent.

Effectiveness of different Components The analysis of the effectiveness of different components of TA-CNN is displayed in Fig.10, where the log-average miss rates show clear decreasing patterns as follows, while gradually accumulating more components. First, TA-CNN (main task) cascades on ACF and reduces the miss rate by 5.4 percent. Second, with pedestrian attributes, TA-CNN (PedAttr.) reduces the result of TA-CNN (main task) by 5.5 percent. Third, when bridging the gaps among multiple scene datasets with shared scene attributes, TA-CNN (PedAttr.+SharedScene) further lower the miss rate by 1.8 percent. Forth, after incorporating unshared attributes, the miss rate is further decreased by another 1.2 percent. TA-CNN finally achieves 34.99 percent log-average miss rate with the structure projection vector.

Comparisons with Deep Models Fig.11 shows that TA-CNN surpasses other deep models on ETH dataset. For example, TA-CNN outperforms other two best-performing deep models, SDN [16] and DBN-Mul [23], by 5.5 and 6 percent, respectively. Besides, TA-CNN even reduces the miss rate by 12.7 compared to MultiSDP [34], which carefully designed multiple classification stages to recognize hard negatives.

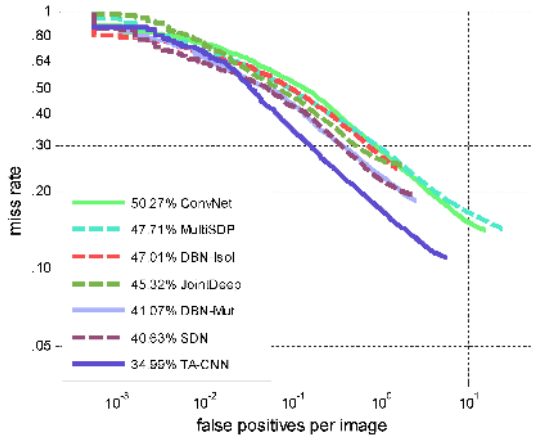


Figure 11: Comparison with other deep models on ETH dataset

5. Conclusions

In this paper, we proposed a novel deep model to learn features from multiple tasks and datasets, showing superiority over hand-crafted features and features learned by other deep models. Extensive experiments demonstrate its effectiveness. Future work tends to explore more attribute configurations. The proposed approach also has potential for attribute prediction and background scene understanding.

Acknowledgement This work is partly supported by Natural Science Foundation of China (91320101, 61472410), Guangdong Innovative Research Team Program (201001D0104648280), Shenzhen Basic Research Program (JCYJ20120903092050890, JCYJ20120617114614438, JCYJ20130402113127496).

References

[1] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *CVPR*, pages 2903–2910, 2012.

[2] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV Workshop*, 2014.

[3] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, pages 44–57, 2008.

[4] G. Chen, Y. Ding, J. Xiao, and T. X. Han. Detection evolution with multi-order contextual co-occurrence. In *CVPR*, pages 1798–1805, 2013.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*. 2005.

[6] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM Multimedia*, 2014.

[7] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *TPAMI*, 2014.

[8] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009.

[9] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34, 2012.

[10] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, pages 1–8, 2007.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[12] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, pages 1–8, 2009.

[13] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *CVPR*, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[15] Z. Lin and L. S. Davis. Shape-based human detection and segmentation via hierarchical part-template matching. *TPAMI*, 32(4):604–618, 2010.

[16] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *CVPR*, pages 899–906, 2014.

[17] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *ICCV*, 2013.

[18] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010.

[19] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved pedestrian detection.

[20] T. Nortcliffe. People analysis cctv investigator handbook. In *Home Office Centre of Applied Science and Technology*, 2011.

[21] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012.

[22] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013.

[23] W. Ouyang, X. Zeng, and X. Wang. Modeling mutual visibility relationship in pedestrian detection. In *CVPR*, 2013.

[24] S. Paisitkriangkrai, C. Shen, and A. v. d. Hengel. Pedestrian detection with spatially pooled features and structured ensemble learning. *arXiv*, 2014.

[25] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *ECCV*, pages 546–561. 2014.

[26] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollár. Exploring weak stabilization for motion feature extraction. In *CVPR*, pages 2882–2889, 2013.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[28] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*. 2013.

[29] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*, pages 352–365. 2010.

[30] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[31] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005.

[32] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, 2009.

[33] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li. Robust multi-resolution pedestrian detection in traffic scenes. In *CVPR*, pages 3033–3040, 2013.

[34] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In *ICCV*, pages 121–128, 2013.

[35] S. Zhang, C. Bauckhage, and A. Cremers. Informed haar-like features improve pedestrian detection. In *CVPR*, pages 947–954, 2013.

[36] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, 2015.

[37] L. Zhu, Y. Chen, and A. Yuille. Learning a hierarchical deformable template for rapid deformable object parsing. *TPAMI*, 32(6):1029–1043, 2010.