S^3D: Scalable Pedestrian Detection via Score Scale Surface Discrimination

Xiao Wang, Chao Liang *Member, IEEE,*, Chen Chen *Member, IEEE,*, Jun Chen* *Member, IEEE,*, Zheng Wang, Zhen Han, Chunxia Xiao *Member, IEEE,*

***Abstract*—Pedestrian detection has remained an important research topic in both the computer vision and multimedia communities because of its importance in practical applications, such as driving assistance and video surveillance. Existing methods compare the response score with a fixed threshold to determine whether a candidate region contains pedestrians and produce dissatisfactory results that contain either missed detections or false detections, which are difficult to balance. This situation has a serious impact under the condition of variable scale. This paper investigates the functional relationship between the scores and scales of pedestrians. By designing experiments with multiple scales, we have found a discriminant surface in the score scale space. Pedestrians can be distinguished at various scale levels according to their locations on the discriminant surface. The proposed approach is evaluated using four challenging pedestrian detection datasets, including Caltech, INRIA, ETH and KITTI, and superior experimental results are achieved when compared with baseline methods.**

***Index Terms*—Pedestrian detection, Multiple scales, Score scale curve, Discriminant surface**

I. Introduction

Pedestrian detection, which aims to find and locate all pedestrians in an image, has aroused increasing interest in the computer vision and multimedia analysis communities [1]–[3]. This topic is also the basis for many advanced multimedia applications, such as pedestrian retrieval [4], pedestrian tracking [5] and behavior analysis [6]. A direct application of pedestrian detection is that we can automatically locate pedestrians with cameras, which is particularly important in criminal investigations and for driving assistance. Although considerable progress has been achieved in recent years [7]–[22], this task remains challenging due to complex conditions, such as occlusions, deformations and illumination changes. Moreover, the various scales of pedestrians cause additional difficulties.

Representative methods include traditional methods, such as histogram of oriented gradient (HOG) [14], aggregated channel features (ACF) [23], locally decorrelated channel features (LDCF) [24] and Checkerboards [25], and convolutional neural network (CNN) methods, such as region-based CNN (R-CNN) [22], Fast R-CNN [26], Faster R-CNN [27], you only look once (YOLO) [28]–[30] and single shot detector (SSD) [31]. These methods typically contain three stages: candidate region generation, feature extraction and classification decision. The candidate region generation stage aims to generate all possible pedestrian candidates. The feature extraction stage aims to construct discriminative and robust feature descriptions. Subsequently, the classification decision stage focuses on seeking an optimal decision to determine whether the candidate region contains pedestrians according to features in the candidate region. The final goal of these stages is to obtain one response score for each candidate region. The response score is the basis for determining whether this region contains one pedestrian. A fixed threshold is often used to assist this decision. If one score is greater than the fixed threshold, then the corresponding candidate region will be considered to contain one pedestrian; otherwise, it will be considered to contain no pedestrians.

The score changes with the various scales of the pedestrian region. As is known, the score decreases as the pedestrian scale decreases because the discriminative features have degraded.

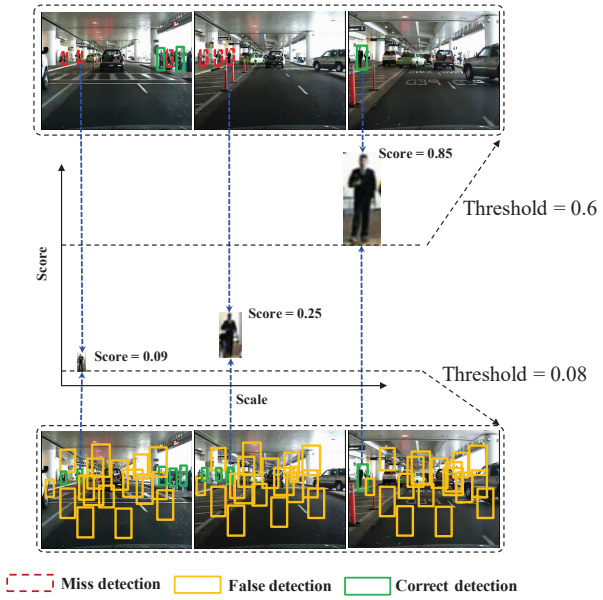

Fig. 1. A higher threshold causes missed detections, whereas a lower threshold leads to a large number of false positives.

X. Wang, C. Liang, J. Chen (Corresponding author), Z. Han and C. Xiao are with National Engineering Research Center for Multimedia Software, School of Computer Science, and Hubei Key Laboratory of Multimedia and Network Communication Engineering, and Collaborative Innovation Center of Geospatial Technology (e-mail: hebeiwangxiao@whu.edu.cn; cliang@whu.edu.cn; chenj.whu@gmail.com; hanzhen_1980@163.com; cxxiao@whu.edu.cn). C. Chen is with Department of Electrical and Computer Engineering, University of North Carolina at Charlotte (e-mail: chen.chen@uncc.edu). Z. Wang is with National Institute of Informatics, Japan (e-mail: wangz@nii.ac.jp)

This work is supported by National Nature Science Foundation of China (No. U1611461, 61876135, 61801335, 61672390, U1736206), National Key R&D Program of China(No. 2017YFC0803700), Hubei Province Technological Innovation Major Project (2018AAA062, 2018CFA024, 2017AAA123), and Nature Science Foundation of Jiangsu Province (No. BK20160386).

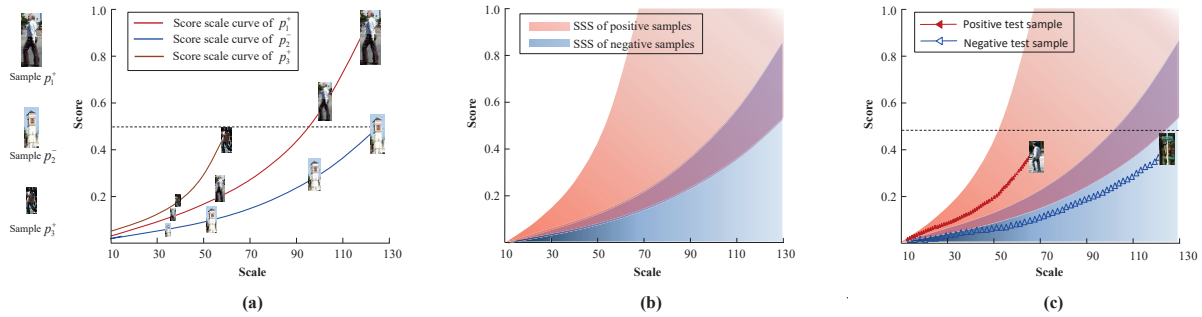

Fig. 2. The discriminant process of the score scale space is shown. In the above figures, the horizontal axis represents the sample height, which is the relevant amount of sample scale, and the vertical axis denotes the response score. (a) **Three score scale curves**. The three curves are generated by comparing the response scores with the different scales of pedestrian samples p_1^+ , p_3^+ and non-pedestrian sample p_2^-. A large threshold can correctly identify p_1^+ and p_2^-, but it cannot identify p_3^+ (missed pedestrian). A small threshold can correctly identify p_1^+ and p_3^+, but it cannot identify p_2^- (false pedestrian). (b) **The score scale surface**. We choose more positive/negative samples to generate score scale curves and obtain expansion areas of these score scale curves in score scale space. *The interface between positive and negative score scale curves is displayed.* (c) Two test samples can be distinguished in the score scale surface according to whether it lies in the positive area or the negative area.

Their scores are usually lower than the threshold and the corresponding regions will be determined as non-pedestrian. In this case, these pedestrians are missed by even the best detectors. Intuitively, a lower threshold is adopted that can recall missed pedestrians. However, it will result in false detections. These methods are inappropriate for dealing with pedestrians with this variable scale and thus result in a certain degree of missed detections and false detections, as shown in Figure 1. There is an urgent need for a discriminant model that can accommodate various scales. The discriminant model must recall missed pedestrians without result in false positives.

A. Motivation

Since the scale affects the response score, we attempt to determine the score variation rule as the scale changes. To investigate this issue, we conducted a preliminary experiment. One pedestrian sample p was randomly selected from the Caltech pedestrian dataset [32]. A subset of samples $\mathbf{p} = \{p^{10}, p^{11}, \cdots, p^{o}\}$ could be obtained after we changed the scale of p step by step, where the superscript represents the height of the new sample.

The height can reflect the scale of pedestrians because the aspect ratio of pedestrians is fixed. The score of each new sample can be obtained by matching the new sample with the learned pedestrian model [33]. This model can discriminate samples of any scale. A series of response scores $\mathbf{s} = \{s^{10}, s^{11}, \cdots, s^{o}\}$ corresponding to the set of samples $\mathbf{p}$ can be obtained. To visualize the changes of scores with scales, we fit a score scale curve based on the distribution of scores with various scales. The curve fitting details are presented in *Section III C*.

Figure 2 (a) shows three score scale curves that are respectively generated by samples p_1^+, p_2^- and p_3^+. The positive and negative samples appear to be separated from each other in Figure 2 (a). The same method is applied to all positive and negative samples of the dataset. Then, score scale curves for the positive and negative samples are obtained. The coverage area of positive/negative score scale curves can be obtained through the distribution of positive/negative score scale curves. In Figure 2 (b), the upper shaded area represents the corresponding spread of positive score scale curves, which is called the score scale surface (SSS) for positive samples. Meanwhile, the bottom shaded area represents the SSS for the negative samples. *The interface between positive and negative SSSs is displayed.* Fortunately, we found that positive and negative samples can be authentically separated in Figure 2 (b). To test the validity of the discriminant surface, we selected two test samples for verification. These two samples can easily be separated, as shown in Figure 2 (c). The scores of both test samples on the original scale are lower than the selected threshold (as shown by the dotted line in Figure 2(a) and (c)), and the scores of pedestrian samples are even lower than those of non-pedestrian samples. It is impossible for existing methods to distinguish them. However, SSS can correctly identify these samples. Although it is difficult to separate some samples that have landed in the overlapping area, they are the most valuable samples for reducing the overlapping area and optimizing the discriminant surface. The smaller the overlap area, the better the discrimination performance. The proposed SSS discrimination (S^3D) method accommodates various scales and explores the discriminative power for pedestrian detection.

The remainder of this paper is organized as follows: Section II provides a brief review of related work for pedestrian detection under various scales. Section III presents the technical details of the proposed S^3D method. The representation of a score scale curve is introduced in Section III C, and SSS discrimination is described in Section III D. The experimental results are shown in Section IV, and conclusions are discussed in Section V.

II. Related Work

In this section, some related studies are briefly reviewed and discussed to illustrate the novelties and contributions of this paper. Pedestrian detection can be regarded as a special image classification problem that needs to locate the position of all pedestrians in image. Therefore, the discriminant region in detection is not the whole image but rather a certain region in the

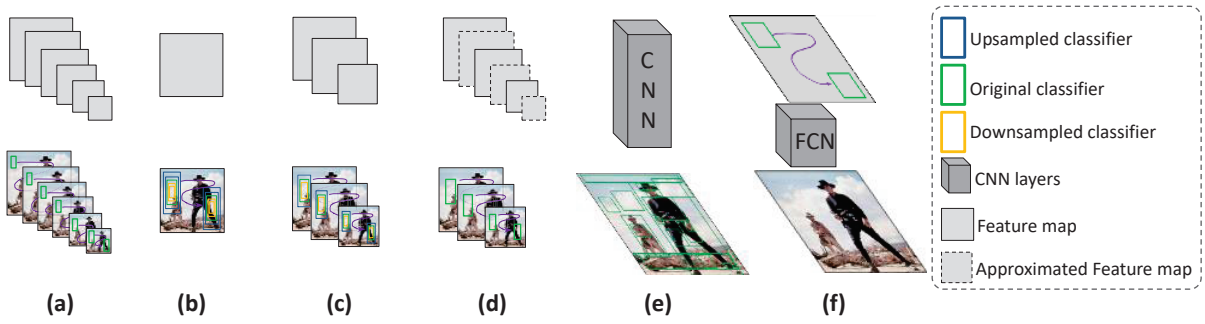

Fig. 3. Different strategies for multiscale detection. (a) Densely sampled image pyramid. (b) Densely sampled classifier pyramid. (c) Densely sampled image and classifier pyramid. (d) Approximating features at multiple scales. (e) Region proposal. (f) Region proposal network.

image. An important process is feature extraction. Well-known features, such as Haar-like features [7], scale-invariant feature transform (SIFT) [34], HOG [14], local binary patterns (LBP) [11] and integral channel features (ICF) [20], are designed to be robust to intraclass variations while remaining sensitive to interclass variations. Recently, CNNs have been successfully applied in generic object recognition [22], [26], [27], [33] because of their power in learning features. Therefore, most researchers tend to focus on improving the performance of pedestrian detection [22], [35]–[38] by using deep learning models. The final stage is to make a proper discrimination based on these features.

Pedestrian detection has recently made a breakthrough. Zhang et al. [39] introduced CityPersons annotations, which enable generalization over multiple benchmarks and are conducive to properly adapting the model for pedestrian detection and pre-training. Zhou et al. [40] learned a deep convolutional neural network (CNN) that consists of two branches: a branch for full body estimation and a branch for visible part estimation. Tian et al. [41] proposed the utilization of the candidate head-top locating stage to efficiently identify the plausible head-top locations and a DMH representation that encodes three channels of information for each candidate region. Zhang et al. [42] proposed a two-staged approach for human detection, which has a physical blob (P-Blob) to identify plausible human heads and used a combination of human upper-body features to filter false positives. This paper focuses on pedestrian detection, primarily pedestrians under various types of scales. Brazil et al. [43] provided an in-depth analysis to demonstrate how shared layers are shaped by the segmentation supervision and to show that the resulting feature maps become more semantically meaningful and robust to shape and occlusion. The relevant studies of pedestrian detection for this topic can generally be categorized into two types: handcrafted models [7]–[11], [13]–[18], [20], [21], and deep learning models [44]–[48]. These two types of models use handcrafted sliding windows and a region proposal network, respectively, to extract candidate regions with different scales.

Handcrafted sliding windows. In traditional methods, candidate regions are primarily obtained using the sliding window scanning method, which mainly consists of three strategies. The first strategy is to learn a single classifier that can match possible pedestrian positions by rescaling the image multiple times [14] (as shown in Figure 3 (a)). This strategy requires repeated feature computations at multiple scales during the testing process. The second strategy is to apply multiple classifiers with different scales to a single input image [16] (as shown in Figure 3 (b)), which avoids the repeated computation of feature maps. However, training detectors with different scales introduces a complicated computational cost. Several approaches have been proposed to balance the computational costs of the testing and training processes. The third strategy is to rescale the input image a few times and learn a number of different scale detectors (as shown in Figure 3 (c)). The representative works are FPDW [21], ACF [23], LDCF [24] and Checkerboards [25]. To reduce the computational complexity of feature extraction during the testing process, a feature approximation strategy (as shown in Figure 3 (d)) was proposed in [23]. This strategy interpolates the missing feature maps and achieves considerable speed-ups with almost no loss in terms of detection accuracy. This method is also applied to this paper. In the aforementioned strategies, windows slide from left to right and from top to bottom in each level of the pyramid to obtain candidate regions, which requires classifying approximately 10^4 to 10^5 candidate regions [49] per image (640×480) if the step of the sliding window (64×128) is 1 for each layer of the pyramid.

Therefore, there are multiple candidate windows of different scales in a pedestrian region. However, the response score of each candidate with different scales is extracted separately, and the maximum response is used for comparison with the threshold value. These strategies have difficulty balancing missed detection and false detection and ignore the relationships among candidate regions at different scales.

Region proposal network. The solutions of the deep learning pedestrian detection model for candidate regions are region proposal and region proposal network (RPN). Deep learning has made major breakthroughs in computer vision. CNNs are particularly prominent in the recognition field [38], [50], [51]. Extracting CNN features requires very complicated calculations because there are millions of parameters [50]. It is not practical in applications to extract CNN features 10^4 times [49] if we obtain candidate regions via handcrafted sliding windows.

A better solution uses the internal structure of the im-

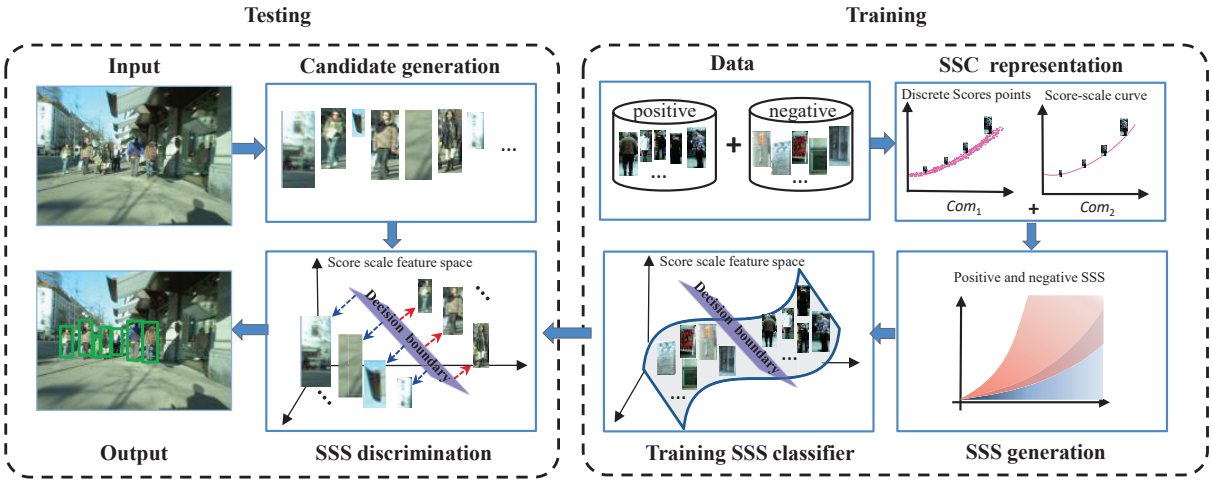

Fig. 4. The architecture of S^{3}D for pedestrian detection, which consists of training and testing phases. In the training phase, positive and negative samples have been expressed as score scale curves in the score scale feature space. The purposes of score scale curve representation and generation are to train the SSS classifier on the score scale feature space. During the testing phase, candidate regions are represented as score scale curves and distinguished by SSS classifiers in the score scale space.

age to extract candidate regions and offers fewer candidate windows and almost no reduction in recall rate compared to handcrafted methods. Selective search [52], EdgeBoxes [53] and Objectness [54]. R-CNN [22] and Fast R-CNN [26] are classic methods that have used this strategy (Figure 3 (e)) and have achieved remarkable breakthroughs. RPN (Figure 3 (f)) was first proposed in Faster R-CNN [27]. RPN is designed to predict candidate regions with different scales and aspect ratios. With almost no effect on the recall rate, the number of candidate areas has been reduced by two orders of magnitude. This approach introduces anchor boxes that serve as candidates at multiple scales and aspect ratios. This scheme avoids enumerating images/filters and performs well. The anchor mechanism has also been applied to multiple feature layers for detecting pedestrians under various scales, such as scale-aware fast (SAF) R-CNN [55] and multiscale CNN (MS-CNN) [56]. These approaches divided pedestrian scales into two different scales: large scale and small scale. YOLOv3 [30] combines features of three different scales for prediction. SSD [31] combines features of six different scales for prediction. There are far more than six scales in the actual scene data. Song et al. [57] devised a fully convolutional network (FCN) with somatic topological line localization (TLL), which takes multi-scale feature representations and regresses the confidence of topological elements. Zhang et al. [58] proposed an active pedestrian detector (TFTS) that explicitly operates over multiple-layer neuronal representations of the input still image. Lin et al. [59] introduced scale-aware pedestrian attention masks and a zoom-in-zoom-out module to improve the capability of the feature maps to identify small pedestrians. Moreover, these methods deal with predictions at different scales independently.

III. Score Scale Surface Discrimination (S^3D)

A. Overview of the Proposed Model

In this section, we describe our S^3D method. First, the candidate generation approach is introduced. It contains the original RPN structure and the modified RPN structure. The purpose of the modification is to restrict the candidate area to the shape of pedestrians and exclude the interference of other forms to provide more suitable candidate regions for pedestrians. Second, the score scale curve representation is described with an exponential function that combines the score and the scale. The coverage areas of the sample's score scale curves constitute the SSS. Last, the SSS discrimination method is provided to distinguish the score scale curve as a pedestrian candidate region or a non-pedestrian candidate region. The whole paradigm is shown in Figure 4.

B. Candidate Generator

For the implementation described in this section, we extract candidate regions that cover possible pedestrian locations. Inspired by [39], the aspect ratio (width to height) of the anchor is changed to 0.41, which is a reasonable aspect ratio for pedestrians [32], [48]. The candidate generator can generates a large pool of candidates with the goal of containing all possible pedestrians, The width-to-height ratio of pedestrians is fixed; thus, the height directly reflects the pedestrian scale. The distribution is as follows:

$$h_n = h_1 q^{(n-1)} \tag{1}$$

where h_n represents the n-th anchor's height, $n \in \{1, 2, 3, \cdots, N\}$, N is the number of pedestrian candidate boxes in the anchor, and q is the ratio among adjacent scales. According to the statistics of the training data [32], in our settings, the values of N and q are 9 and 0.83, respectively. h_1 is the maximum height of the ground truth in the datasets. The same as the structure of [24] and [25], the end of the candidate generator is fed into two fully connected layers, i.e., a box-regression layer and a box-classification layer, followed by the score scale curve representation.

The framework in this paper is a multitask learning framework with three main branches: the extraction of candidate

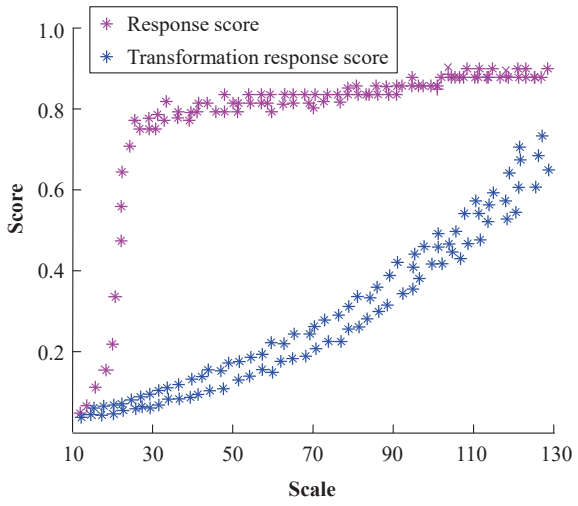

Fig. 5. The comparison between scale and response score of pedestrian samples with different scales is with purple dots. In score scale space, the response score has been transformed by $\bar{s} = \exp(s * \sigma - 1)$, which is more discriminative.

regions, the acquisition of pedestrian scores, and the discrimination of SSS. During training, the correspondence between the ground truth and the predicted pedestrian position needs to be established. This branch is constrained by positional regression loss. Pedestrian response scores are used in score scale curve representation multiple times. This branch is constrained by classification loss. The two branch structures rely on previous work [27]. The loss function of this part is as follows:

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (2)$$

where i is the index of predicted pedestrians in a minibatch and p_i is the predicted probability. The ground-truth label p_i^* is 1 if the candidate region from the anchor is positive and 0 if the candidate region is negative. t_i is a vector representing the 4 parameterized coordinates of the predicted pedestrian region, and t_i^* is that of the ground truth associated with a positive anchor. The classification loss L_{cls} is log loss over two classes (pedestrian vs. non-pedestrian). The regression loss is $L_{reg}(t_i, t_i^*) = L_1(t_i - t_i^*)$, where L_1 is the robust loss function defined in [26].

C. Score Scale Curve Representation

Since the scale affects the response score, the response score alone is not sufficient for making accurate pedestrian detection decisions under varying scale conditions. To this end, we introduce the score scale curve representation to characterize and analyze the trend of the score scale changes of pedestrian and non-pedestrian samples to provide more accurate guidance for pedestrian detection with various scales. In the following, we elaborate the score scale curve representation details.

A pedestrian classifier [27] is initialized with the Caltech dataset [32]. One pedestrian sample p is selected randomly from the training dataset. To obtain samples of different scales, p is resized into various scales to form a set of samples, $\mathbf{p} = \{p^{10}, p^{11}, \cdots, p^o\}$, where the superscript denotes the height of the new resized sample, and o denotes the original height of the pedestrian sample p. With the above setting, the pedestrian classifier is adopted to generate a score vector $\mathbf{s} = \{s^{10}, s^{11}, \cdots s^o\}$. To accurately describe score changes under various scales, we can obtain response score vectors for all pedestrians in the training dataset. The mean score vector $\bar{\mathbf{s}}$ can be obtained by averaging the corresponding bits, and an example is shown as the pink stars in Figure 5 (a). To make a more discriminative and easier discriminative, the response score was transformed by $\bar{s} = \exp(s * \sigma - 1)$, where σ is the scale ratio of sampled sample to original sample. That is, if the height and the width of the downsampled pedestrian are half those of the pedestrian, then the scale ratio $\sigma = 0.5$. The comparison in the score scale space is shown in Figure 5 (b).

To decrease noise and enhance robustness, this paper adopts several smooth curves to describe the dynamic relationship between response scores and scales. Specifically, several common curves are evaluated as shown in Table I. R-square is an objective indicator of the effectiveness of curve fitting. The larger the R-square value, the better the fit. Through the above comparison, the exponential-2 function, $f(h) = ae^{bh} + ce^{dh} + k$, is adopted to describe the score scale relationship of various response scores under different pedestrian sample scales. The fitted curve is named the proposed score scale curve.

TABLE I
DIFFERENT CURVE FITTING METHODS AND EVALUATION PERFORMANCES. *R-square* INDICATES THE FITTING PERFORMANCE OF THE CURVE AND SCORE POINTS. WHEN *R-square* IS CLOSE TO 1, THE FIT IS GOOD.

Fitting method	Function	R-square
Linear fitting	$f(h) = ah + k$	0.6349
Sine fitting	$f(h) = a\sin(bh + c) + k$	0.7288
Gaussian fitting	$f(h) = ae^{-(h-b)^2/2\delta^2} + k$	0.8937
Exponential-1 fitting	$f(h) = ae^{bh} + k$	0.9253
Exponential-2 fitting	$f(h) = ae^{bh} + ce^{dh} + k$	0.9682

D. SSS discrimination

SSS generation. During the training phase, score scale curves for positive and negative samples corresponding to pedestrian and non-pedestrian samples are obtained, as shown in Figure 6 (b). With the derived score scale curves for both positive and negative samples, the positive and negative SSSs can be outlined by the variance spread of their standard deviations with their mean score scale curves. The mean score scale curves of positive and negative samples are shown as black and yellow curves, respectively, in Figure 6 (b).

We are even more concerned with the interface between the positive and negative score scale curves. For this purpose, the standard deviation of the positive score scale curves has been used to obtain the lower bound, and the standard deviation of the negative score scale curve has been used to obtain the upper bound. The interface between the positive and negative samples is shown in Figure 6 (c).

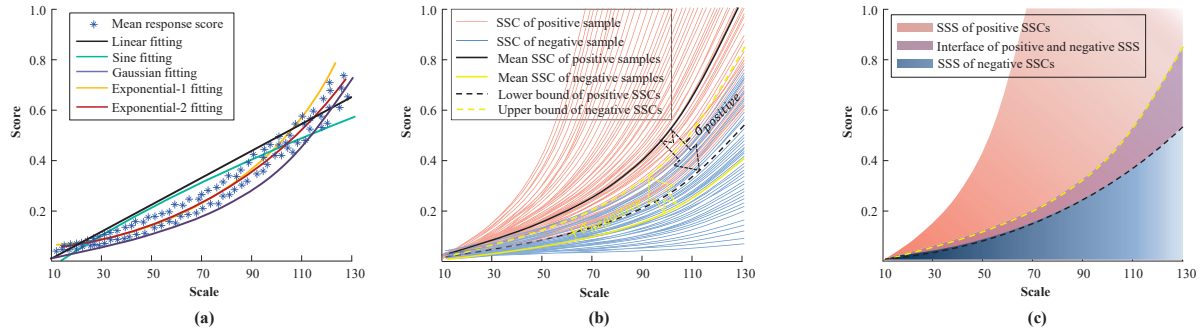

Fig. 6. Score scale curve fitting and SSS generation. (a) The distribution is shown from one average distribution of pedestrian samples from the Caltech dataset [32]. (b) The red score scale curve is generated by positive samples. The blue score scale curve is generated by negative samples. The black score scale curve is the mean of the positive score scale curves. The black dotted score scale curve is the lower bound of positive SSS generated as the spread from the variances of the positive score scale curves (σ_{positive}). The upper bound of negative SSS is generated using the same method. (c) Positive/negative SSSs and their interface.

Algorithm 1 SSS classifier learning

Input: $\{(p_1, l_1), (p_2, l_2), \cdots, (p_i, l_i), \cdots (p_n, l_n)\}$,$l_i \in (0, 1)$
Output: SSS classifier.

for Each sample (p_i, l_i) in training data **do**
  Construct a subset **p** by sampling p_i;
  Obtain scores **s** for subset **p**;
  Form score scale curve according to the distribution of h-s pairs;
  Express score scale curve as a feature vector;
  Optimize SSS classifier according to loss function.
end for
return SSS classifier

Algorithm 2 S^{3}D pedestrian detection framework

Input: image I
Output: Candidate regions that contain pedestrians

Extract candidate regions using the modified RPN;
for Each pedestrian candidate region p **do**
  Construct a subset **p** by sampling the candidate region;
  Obtain corresponding scores **s** for the subset **p**;
  Form the score scale curve according to the distribution of h-s pairs;
  Express the score scale curve as a feature vector;
  Determine whether the candidate region contains pedestrians with the discrimination of the SSS classifier.
end for
return candidate region that contains one pedestrian.

Feature descriptor. To express all properties of the score scale curve in score scale feature space, this paper exploits a feature expression that can accommodate various scales. The feature expression of this part has two components: response scores and curve fitting parameters. The first component reflects the scores of the samples of different scales. The second component shows the relationships among samples of different scales. Specifically, the two components are Com_1 and Com_2. The first component is used to represent the direct relationship corresponding to scores with different scales and expressed as $Com_1 = \mathbf{s}$. The second component is used to represent the indirect relationship corresponding to scores with different scales and is expressed as $Com_2 = [a, b, c, d, k]$. This feature descriptor can accommodate changing scales and be described as $[Com_1, Com_2]$.

Training SSS classifier. With the above feature representation, this paper uses random forest (RF) [60] as the classifier for S^{3}D. RF is a combination of some binary decision trees built based on bootstrap samples. A subset of the parameter vector is randomly chosen for each node of the decision tree, and the best split is calculated with this subset. According to the distribution of parameters in separating positive and negative score scale curves in SSS, RF makes effective decisions; thus, the classifier has been named the SSS classifier.

The SSS interface between positive and negative score scale curve areas reflects the discrimination ability of S^{3}D. The loss of this part is expressed as follows:

$$L_{SSS}(\hat{p}_i, \hat{p}_i^*) = -\frac{1}{N_{cls}} \left(\hat{p}_i \ln \hat{p}_i^* + (1 - \hat{p}_i) \ln (1 - \hat{p}_i^*)\right) \quad (3)$$

where $L_{SSS}(\hat{p}_i, \hat{p}_i^*)$ is the loss function in SSS classifiers learning, $\hat{p}_i$ denotes the distribution of the i-th pedestrian sample in score scale space, and $\hat{p}_i^*$ denotes the prediction probability of the distribution in the score scale space. The optimal loss function directly reflects the size of the overlapping area in Figure 6 (c). The smaller this overlapping region is, the more discriminative the SSS classifier will be. The SSS classifier's learning process is illustrated in Algorithm 1.

SSS for pedestrian detection. The pedestrian detection process in this paper is summarized in three steps: candidate generation, score scale curve representation and SSS discrimination. The detection process is illustrated in Algorithm 2. To avoid unnecessary operations, the scale of the feature map is varied (as shown in Figure 3) (d)) rather than directly changing the candidate regions. Therefore, the proposed method has almost the same computational time as previous pedestrian detection methods.

IV. Experiments

In this section, the proposed approach is validated through comparison with several classical pedestrian detection methods

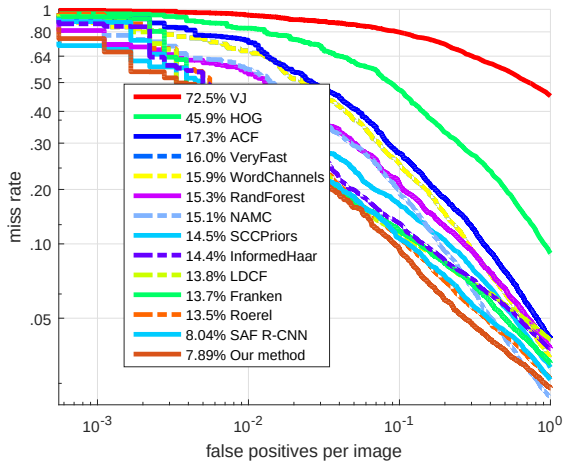

Fig. 7. Quantitative evaluation results (miss rate versus false positives per image) on the INRIA dataset. Performance in the legend is evaluated in terms of average miss rate.

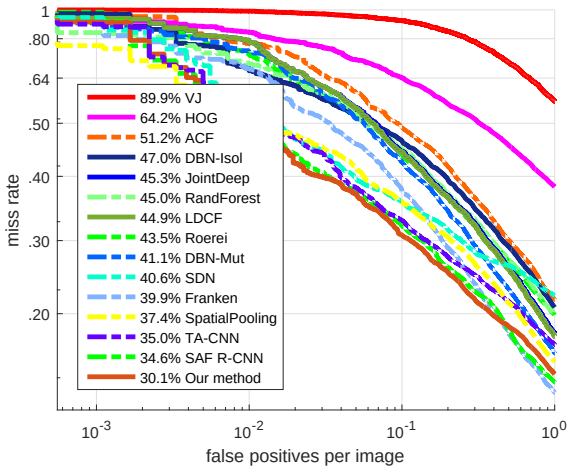

Fig. 8. Quantitative evaluation results (miss rate versus false positives per image) on the ETH dataset. Performance in the legend is evaluated in terms of average miss rate.

on four pedestrian datasets: Caltech [32], INRIA [14], ETH [61] and KITTI [62]. We chose these datasets because they provide many challenges faced in practical surveillance, such as viewpoint, pose, illumination changes, background variation and occlusions. More detailed experimental analyses of the effectiveness of each component in S^3D are further given on the challenging Caltech dataset [32].

A. Implementation details

The scalable ideas are added to the candidate regions rather than a single scale as in [26], [27], [33]. To obtain a fair comparison with most of the existing algorithms [22], [26], [27], the pretrained VGG16 model [37] was used to initialize the modified RPN for candidate regions. The convolutional layers and max-pooling layers are used as the shared convolutional layers before S^3D to produce feature maps from the input image. To obtain richer features, the fourth max-pooling layer is removed to produce larger feature maps.

For RPN training, an anchor is considered to be a positive example if it has an intersection-over-union (IoU) ratio greater than 0.5 with one ground-truth box, and others are considered as negative examples. As in [26], [27], each minibatch consists of 1 image and 120 randomly sampled anchors for computing loss. The ratio of positive to negative samples is 1:5 in a minibatch. To extract the candidate regions for pedestrians, we modified the scale and aspect ratio of the anchor in RPN as the candidate generator in this paper. Because there is only one pedestrian category in our work, the original anchors are changed to the aspect ratio of pedestrians. The candidates from the modified RPN are selected as the pedestrian candidate regions in this paper rather than selecting the proposals from the original RPN. The main concern in our work is multiple scales, which is different from [48]. Thus, there are more scales of the anchors in this paper relative to [48].

To compare the scores of candidate regions with different scales, the features of each pedestrian sample are represented with the same dimension. Region of interest (RoI) pooling has been adopted to extract fixed-length features, and the score of each candidate region was calculated by the classification layer. In more detail, the data are augmented by horizontal flipping and color altering. Dropout [32] is used on the two fully connected layers. A momentum of 0.9 and a weight decay of 0.0005 were used in this paper. The learning rate starts from 0.01 and is divided by 10 when the error plateaus. Our experiments are run on a GeForce GTX Titan X GPU. The layers update parameters with an initial learning rate of 0.001, which is lowered to 1/10 of the current rate after every 4 epochs.

B. Comparisons with State-of-the-art Methods

ETH and INRIA datasets. The INRIA dataset [14] includes a training set and a test set. The training set consists of 614 positive images and 1,218 negative images. The test set consists of 288 images. Our model is evaluated on the test set. The ETH [61] dataset is also used to verify the discrimination capability of the S^3D models. The data were recorded using a pair of AVT Marlins mounted on a chariot. Approximately 12,298 annotated pedestrians have been labeled. The size of these images is 640×480, and the frame rate of the data is 13–14 FPS. The ETH dataset consists of 3 testing video sequences. Our models are evaluated on the 1,804 images in the ETH dataset. Many studies [50], [63] have found that using more training data is beneficial for training models. To evaluate the discrimination capacity of the S^3D model, the INRIA dataset is added to the training set following the approach commonly adopted by the superior approaches [16], [44], [64]. Although these training data have been added, the quantity is not sufficient to train our model. We implement Gaussian blurring and motion blurring on the training set for data augmentation. The evaluation metric of the INRIA [14] and ETH [61] datasets for the following experiments is the average miss rate on false positives per image (FPPI) [32]. An IoU of 0.5 is used to determine true positives.

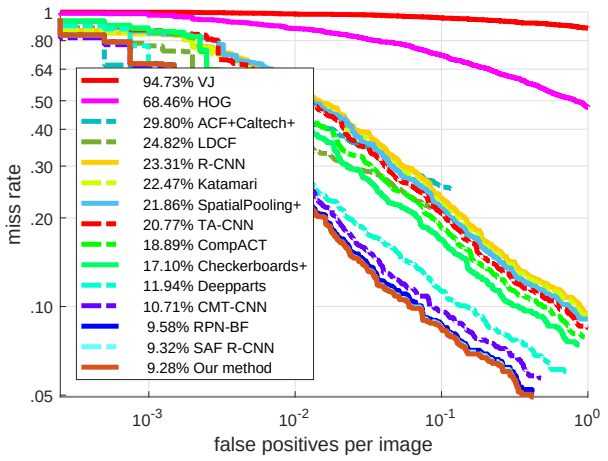

Fig. 9. Quantitative evaluation results (miss rate versus false positives per image) on the Caltech dataset. Performance is evaluated in terms of average miss rate.

Both the INRIA and ETH test sets are used to evaluate our S^{3}D model. Figure 7 and Figure 8 provide the results of the comparison of the S^{3}D model with several methods with state-of-the-art performance. On the INRIA set, our method obtains an average miss rate of 7.89%, which is considerably better than that of the SAF R-CNN [55] method. For the ETH dataset, the S^{3}D model in this paper obtains an average miss rate of 30.12%, which outperforms the second-best method (SAF R-CNN) [55] by 4.52%.

It can be observed that the S^{3}D model achieves a better average miss rate on both datasets. In general, the method proposed in this paper outperforms other popular methods and achieves satisfactory performance on both datasets.

Caltech dataset. The Caltech pedestrian detection dataset [32] and its associated benchmark are among the most popular. This dataset consists of approximately 10 hours of 30-Hz video that has been taken from a vehicle driving through regular traffic in an urban environment. Every frame has been annotated with bounding boxes indicating whether the frame is a pedestrian or not a pedestrian. Approximately 250,000 frames (640×480) with a total of 350,000 bounding boxes and 2,300 unique pedestrians were annotated. The dataset contains a training set and test set. The training set consists of six training subsets (set00–set05). We sample the training data every 4th frame for training S^3D in this paper. The test set consists of five subsets (set06–set10). According to the evaluation method, the test data with the 30th frame are sampled to verify the effectiveness of S^3D.

The evaluation metric of the Caltech [32] dataset for the following experiments is the average miss rate on FPPI [32]. An IoU of 0.5 is used to determine true positives.

The Caltech training set was used to train our method, and the test set was used to evaluate the S^{3}D method. The overall experimental results are presented in Figure 9. Existing methods that achieved superior performance on the Caltech test set are compared with the S^{3}D method proposed in this paper. The comparison algorithms include VJ [7], HOG [14], ACF [23], LDCF [24], R-CNN [22], Katamari [65], SpatialPooling+ [66], task-assistant CNN (TA-CNN) [67], CompACT [47], Checkerboards+ [25], DeepParts [68], cross-modality transfer CNN (CMT-CNN) [69], RPN-BF [48] and SAF R-CNN [55]. We find that S^3D outperforms the other methods and achieves the lowest average miss rate of 9.28%, which is significantly better than the current state-of-the-art approaches.

TABLE II
AVERAGE PRECISION (%) ON THE KITTI DATASET.

Methods	Easy	Moderate	Hard
R-CNN	61.61	50.13	44.79
pAUCEnsT	65.26	54.49	48.6
FilteredICF	67.65	56.75	51.12
DeepParts	70.49	58.67	52.78
CompACT-Deep	70.69	58.74	52.71
Regionlets	73.14	61.15	55.21
SAF R-CNN	77.93	65.01	60.42
Our method	**77.94**	**65.6**	**60.45**

KITTI dataset. The KITTI dataset [62] contains images with stereo data available. There are 7,481 training images and 7,518 test images in the KITTI dataset [62], which are captured from an autonomous driving platform. The evaluations have three levels of difficulty: easy, moderate and hard. The evaluation setting is used to rank the competing methods in the benchmark. The KITTI training set was split into training and validation subsets, as used in [55], [70].

The evaluation metric of the KITTI [62] dataset for the following experiments is the mean average precision (mAP) [62]. The detection results and performance comparisons of the proposed method with several state-of-the-art methods, such as R-CNN [22], pAUCEnsT [66], FilteredICF [25], DeepParts [68], CompACT-Deep [47], Regionlets [71] and SAF R-CNN [55], are presented in Table II. As shown, the S^{3}D model achieves promising results, i.e., 77.08%, 61.12% and 55.09% in terms of average precision (AP) on the easy, moderate and hard subsets, respectively, which outperforms most of the existing methods tested on this benchmark.

C. Ablation study

We conduct ablation experiments on the Caltech dataset [32] in this subsection. These experiments investigate the effectiveness of different components of S^{3}D. The performances achieved by different variants of the S^{3}D and structure settings are reported as follows.

Candidate region extraction. The first phase of pedestrian detection is the selection of candidate regions. As mentioned above, there are two types of methods for extracting candidate regions: handcrafted sliding window methods, whose representative work is LDCF [24], and related deep learning methods, whose representative works include selective search [52] and RPN [27]. We have investigated in terms of candidate region quality and evaluated recall rates with different IoU thresholds. These three algorithms are currently the most representative works. As shown in Figure 10, the modified RPN performs better than the three leading methods.

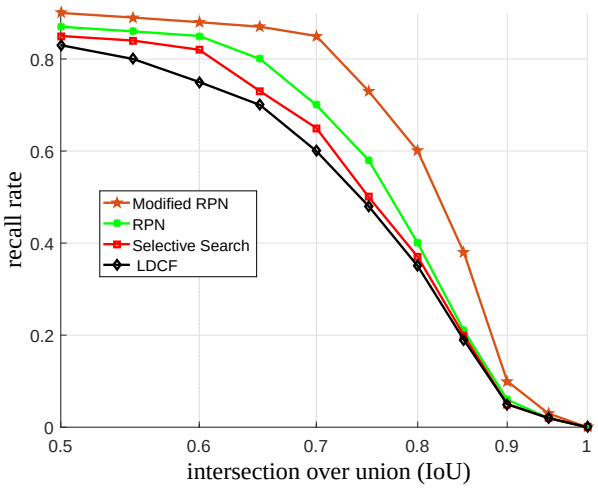

Fig. 10. Comparison of modified RPN and three existing methods on the Caltech pedestrian dataset.

RPN makes full use of the robust feature expression capabilities of deep learning to extract candidate regions at the convolutional layer. Candidate generator effectively correlates the scales of the anchor and the pedestrian. The above experiment has verified the effectiveness of our modification.

Feature selection. SSS can take advantage of the features of various scales from the convolutional network and is flexible. We analyzed the impact of different feature maps. Table III shows the results of using different feature maps in S^3D. In the first experimental combination, Conv3_3 alone yields a good average miss rate of 12.3%, showing the effects of higher resolution features. In the second experiment, the combination of Conv3_3 and Conv4_3 achieves the best average miss rate of 11.5%. In the following experiment, Conv3_3, Conv4_3 and Conv5_3 achieve satisfactory results of 9.28%. In a further combination, the performance is unsatisfied. Our analysis suggests that high-resolution features should provide good performance. The combinations of Conv3_3, Conv4_3 and Conv5_3 validate this analysis. Although Conv2_2 has high-resolution features, it shows degraded performance because of the weaker representation of shallower layers, which is the main reason why the performance is less than satisfactory when we add Conv 2_2 to the experiment.

The score scale curve descriptor consists of two parts: direct score scale descriptor Com_1 and indirect score descriptor Com_2. Both have auxiliary effects on the pedestrian detection in this paper. The detailed effects are shown in Table IV.

Robust classifier. The next phase of pedestrian detection is to classify the extracted candidate regions from the above phase. More importantly, modified RPN as a detector achieves an average miss rate of 14.9%. We set up different combinations by testing the current popular classifier with the modified RPN. These results are compared in Table V. Modified RPN+SSS+RF is better than all combinations. For fair comparisons, we use the same set of modified RPN for all methods in this section. R-CNN [22] was reported earlier. All pedestrian candidate regions are extracted from the modified RPN discussed above. It has an average miss rate of 13.1% and is better than modified RPN alone. The Fast R-CNN classifier with the same set of the modified RPN provided worse results. The small-scale features are caused by the pooling in the network. Therefore, the last pooling is discarded in the settings below. This paper takes full advantage of the relationships of the different scales. The modified RPN+SSS+RF combination achieves an average miss rate of 9.28%.

TABLE III
COMPARISONS OF DIFFERENT FEATURES IN THE S^3D MODEL ON THE CALTECH DATASET. THESE METHODS ARE BASED ON VGG-16, AND THE OTHER STRUCTURE SETTINGS ARE THE SAME. MISS RATE (%) IS USED TO EVALUATE THE PERFORMANCES OF DIFFERENT COMBINATIONS.

Conv2_2	Conv3_3	Conv4_3	Conv5_3	Miss Rate
√				15.7
	√			**12.3**
		√		12.6
			√	18.2
√	√			12.9
	√	√		**11.5**
	√		√	13.6
√	√	√		9.34
	√	√	√	**9.28**
√	√	√	√	9.56

TABLE IV
DIFFERENT SCORE SCALE CURVE DESCRIPTORS

Score scale curve descriptor	Average Miss Rate (%)
Com1	9.57
Com2	9.43
Com1+Com2	**9.28**

TABLE V
MISS RATES OF DIFFERENT COMBINATIONS. ALL METHODS ARE BASED ON VGG-16 AND THE SAME SET OF RPN.

Methods	**Average Miss Rate (%)**
Modified RPN stand alone	14.9
Modified RPN+R-CNN	13.1
Modified RPN+Fast R-CNN	16.2
Modified RPN+RF	9.60
Modified RPN+SSS+RF	**9.28**

Time efficiency. Table VI compares the running times on the Caltech dataset. Our method is faster than current popular methods such as LDCF [24], CCF [35], CompACT-Deep [47] and RPN+BF [48]. To ensure fairness, in this subsection, we report the time using data published in the public literature. The times of LDCF and CCF were reported in [35], and that of CompactACT-Deep was reported in [47]. The time of RPN+BF was reported in [48]. S^3D achieves a satisfactory speed.

Comparison on Caltech-new. Zhang *et al* [72] manually sanitized the Caltech training annotations and improved the training set alignment quality. Aiming for a more complete evaluation, they extended the evaluation FPPI range from

TABLE VI
COMPARISONS OF RUNNING TIME ON THE CALTECH SET.

Methods	**hardware**	time /img (s)
LDCF	CPU	0.6
CCF	Titan Z GPU	13
CompACT-deep	Tesla K40 GPU	0.5
RPN+BF	Tesla K40 GPU	0.5
Our method	Titan X GPU	**0.42**

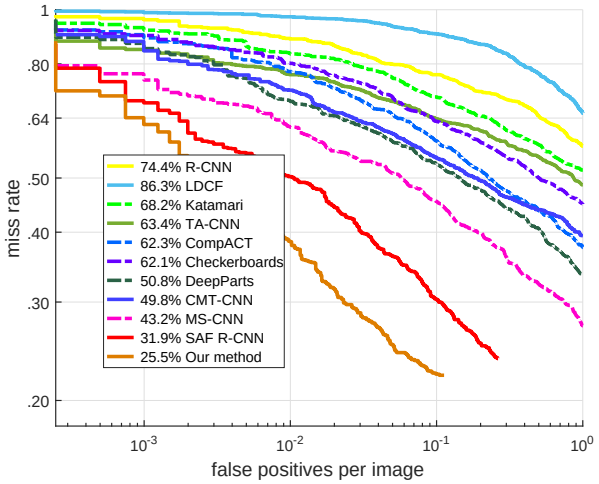

Fig. 11. Quantitative evaluation results (miss rate versus false positives per image) on the Caltech-small dataset. Performance in the legend is evaluated in terms of the average miss rate.

traditional $[10^{-2}; 10^{0}]$ to $[10^{-4}; 10^{0}]$, and they denoted MR_{-2} and MR_{-4} as the corresponding average miss rates. In this section, the new annotations are used and evaluated on the S^{3}D method proposed in this paper.

Table VII shows the results on the Caltech-new pedestrian dataset [72]. In the case of the original annotations, S^{3}D has an average miss rate of 9.28%. In the case of the corrected annotations, S^{3}D has an MR_{-2} of 7.2% and MR_{-4} of 16.7%, which are better than those of the previous methods.

Comparison on pedestrians (height smaller than 80 pixels). Although pedestrian scales vary, negative impact on performance mainly results from small-scale pedestrians.

TABLE VII
COMPARISONS ON THE CALTECH-NEW SET.

Methods	MR_{-2} (%)	MR_{-4} (%)
SCF+AlexNet	29.7	47.4
ACF+Caltech+	27.6	41.9
DeepCascade+	26.2	44.6
LDCF	23.7	38.3
TA-CNN	18.8	34.3
DeepParts	12.9	25.2
MS-CNN	9.5	23.5
CompACT-Deep	9.2	18.6
Our method	**7.2**	**16.7**

TABLE VIII
COMPARISON EXPERIMENT ON PEDESTRIANS (HEIGHT SMALLER THAN 80 PIXELS)

Methods	MR (%) (Without SSS)	MR (%) (With SSS)
MOCO	85.17	80.56
MultiFtr+Motion	84.22	79.73
ChnFtrs	81.63	72.68
ACF+SOt	76.35	70.47
MT-DPM	74.05	69.42
MT-DPM-Context	71.93	68.34
TA-CNN	71.68	66.78
RPN+BF	64.12	62.69
CompACT-Deep	63.63	61.39
SAF R-CNN	62.57	58.83
TLL	60.79	57.42
MS-CNN	60.51	56.34
TFTS	41.85	40.93

Compared with previous research, this paper has three advantages. (1) We sampled a variety of pedestrian samples, which is more consistent with real circumstances. (2) The relationship among different scales is simultaneously considered by curve fitting. The score scale curve can be separated into positive and negative curves by mapping it onto a scale score space rather than using the response score as the final evaluation, which achieves a balance between missed detections and false detections. (3) This method can be applied to any detector. We supplemented the comparison experiments on small-scale partition (height smaller than 80 pixels) according to the division in the Caltech pedestrian dataset. The comparisons are shown in Table VIII, where these similar methods has been improved on our approach (S^{3}D). These methods make predictions on several discrete scale features map. The relationship among different scales has modeled in S^{3}D. The experimental results verify the effectiveness of the S^{3}D model in this paper.

Comparison on Caltech-small. Most pedestrians are not observed at a small scale in the Caltech pedestrian dataset. However, small-scale pedestrians are very common in surveillance video of practical applications, such as criminal investigation and automatic driving. To simulate the small-scale scenarios, we resized the images to a quarter of the original scale, with heights and widths of half the original resolution. The results are shown in Figure 11. We can see that most comparison algorithms suffer declines in the situation of small scales pedestrian, and the performances of these methods dropped considerably. Several studies of pedestrian detection have sought optimization with regard to scale, such as MS-CNN, SAF R-CNN and S^{3}D, as proposed in this paper. S^{3}D involved more scale changes and identified optimized scales per the changes. The S^{3}D model in this paper (25.53%) can be seen to outperform the top two best methods, MS-CNN [56] (43.19%) and SAF R-CNN [55] (31.86%), by a margin of over 6.33%. The maximum score is selected and matched with a fixed threshold. This discriminant method leads to more missed detection, especially in the case of a large number

of small-scale pedestrians. Pedestrians can be distinguished at various scale levels according to their locations on the discriminant surface rather than relying on a separate score. Our method achieved promising results and outperformed most of the existing methods evaluated under this condition.

Therefore, we believe our proposed method provides an effective solution to the limitations of the existing approaches and provides useful insights for researchers and practitioners in this field. Pedestrian detection scores are affected by different factors, such as occlusion, deformation and illumination changes. However, the proposed SSS in this paper does not affected by the above various factors. More precisely, the change of the score comes solely from the varying scale since we only change the scale factor in the score calculation. These factors will not affect the change rule of SSS. Therefore, the proposed SSS learning method of this paper is still feasible given the existence of other interference factors.

Comparing the results shown in Figure 12 on the ETH, INRIA, Caltech and KITTI datasets, it shows that the S^3D method is suitable for pedestrian detection at various scales. The experimental results illustrate that the traditional model has a significant loss in terms of performance when the scales of the images are decreased and demonstrate the effectiveness of the proposed method.

V. Conclusion

This paper raises a new issue which is pedestrian detection at various scales and that has not been investigated previously to the best of our knowledge. A discrimination method based on a single threshold is inappropriate. A discrimination method based on variable thresholds is only minimally effective, and it includes missed detections and false positives. S^3D has learned a discriminant surface to address this problem. We used the change rule of scores and scales to learn a score scale curve. The score scale curve can be separated into positive and negative curves by mapping it onto SSS rather than using the response score as the final evaluation. In the future, we will try to convert it to a multi-task learning framework, which is more suitable for this variable scale.

References

[1] M. Bilal, A. Khan, M. U. K. Khan, and C. M. Kyung, "A low complexity pedestrian detection framework for smart video surveillance systems," *Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 10, pp. 2260 – 2273, 2017.

[2] W. Si, H. S. Wong, and S. Wang, "Variant semiboost for improving human detection in application scenes," *Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 10, pp. 1595 – 1608, 2018.

[3] S. Motiian, F. Siyahjani, R. Almohsen, and G. Doretto, "Online human interaction detection and recognition with multiple cameras," *Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 649–663, 2017.

[4] A. J. Bency, S. Karthikeyan, C. D. Leo, S. Sunderrajan, and B. S. Manjunath, "Search tracker: Human-derived object tracking in the wild through large-scale search and retrieval," *Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1803–1814, 2017.

[5] T. Billah, S. M. M. Rahman, M. O. Ahmad, and M. N. S. Swamy, "Recognizing distractions for assistive driving by tracking body parts," *Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–11, 2018.

[6] H. Fradi, B. Luvison, and Q. C. Pham, "Crowd behavior analysis using local mid-level visual descriptors," *Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 589–602, 2017.

[7] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[8] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[9] P. Felzenszwalb, D. Mcallester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," pp. 1–8, 2008.

[10] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *European Conference on Computer Vision*, 2008, pp. 423–436.

[11] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *International Conference on Computer Vision*, 2010, pp. 32–39.

[12] P. Dollar, Z. Tu, H. Tao, and S. Belongie, "Feature mining for image classification," in *Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[13] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[15] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," in *Joint Pattern Recognition Symposium*, 2008, pp. 82–91.

[16] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models." *Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, p. 1627, 2010.

[17] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *International Conference on Computer Vision*, 2010, pp. 24–31.

[18] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Computer Vision and Pattern Recognition*, 2010, pp. 1030–1037.

[19] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, "Part-based feature synthesis for human detection," in *European Conference on Computer Vision*, 2010, pp. 127–142.

[20] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conference*, 2009, pp. 1–11.

[21] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *British Machine Vision Conference*, 2010, pp. 1–11.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Computer Science*, pp. 580–587, 2014.

[23] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[24] W. Nam, P. Dollr, and J. H. Han, "Local decorrelation for improved pedestrian detection," *Advances in Neural Information Processing Systems*, vol. 1, pp. 424–432, 2014.

[25] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," pp. 1751–1760, 2015.

[26] R. Girshick, "Fast r-cnn," in *International Conference on Computer Vision*, 2015, pp. 1440–1448.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[29] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.

[30] J. Redmon, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, 2016, pp. 21–37.

[32] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[34] D. G. Lowe and D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

Fig. 12. Visual comparison of our S^{3}D detection results vs. those of the of LDCF [24] and SAF R-CNN [55] on the INRIA [14], ETH [61] and Caltech [32] datasets. Because the pedestrian candidate box is constrained in this paper, the pedestrian's position will not be too aligned when the pedestrian part appears in the image; the box covers most pedestrians and does not affect the evaluation. Since we are concerned with the scale issue, S^{3}D is not effective at handling occlusions and cyclists.

[35] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *International Conference on Computer Vision*, 2015, pp. 82–90.

[36] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, and C. C. Loy, "Deepid-net: Deformable deep convolutional neural networks for object detection," in *Computer Vision and Pattern Recognition*, 2015, pp. 2403–2412.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2015.

[39] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3213–3221.

[40] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *European Conference on Computer Vision*, 2018, pp. 135–151.

[41] L. Tian, M. Li, Y. Hao, J. Liu, G. Zhang, and Y. Q. Chen, "Robust 3-d human detection in complex environments with a depth camera," *Transactions on Multimedia*, vol. 20, no. 9, pp. 2249–2261, 2018.

[42] G. Zhang, J. Liu, L. Tian, and Y. Q. Chen, "Reliably detecting humans with rgb-d camera with physical blob detector followed by learning-based filtering," pp. 1–8, 2016.

[43] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *International Conference on Computer Vision*, 2017, pp. 4950–4959.

[44] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Computer Vision and Pattern Recognition*, 2012, pp. 3258–3265.

[45] W. Ouyang, "Joint deep learning for pedestrian detection," in *International Conference on Computer Vision*, 2014, pp. 2056–2063.

[46] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Computer Vision and Pattern Recognition*, 2015, pp. 5079–5087.

[47] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *International Conference on Computer Vision*, 2015, pp. 3361–3369.

[48] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" pp. 443–457, 2016.

[49] X. Wang, J. Chen, W. Fang, C. Liang, C. Zhang, and R. Hu, "Pedestrian detection from salient regions," in *International Conference on Image Processing*, 2015, pp. 2423–2426.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[52] J. R. R. Uijlings, K. E. A. V. D. Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[53] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*, 2014, pp. 391–405.

[54] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, p. 2189, 2012.

[55] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2018.

[56] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European conference on computer vision*, 2016, pp. 354–370.

[57] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation," *arXiv preprint arXiv:1807.01438*, 2018.

[58] X. Zhang, L. Cheng, B. Li, and H. M. Hu, "Too far to see? not really!-pedestrian detection with scale-aware localization policy," *Transactions on Image Processing*, vol. 1, no. 99, pp. 1–11, 2018.

[59] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *European Conference on Computer Vision*, 2018, pp. 732–747.

[60] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[61] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[62] A. Geiger, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[63] M. Ranzato, F. J. Huang, Y. L. Boureau, and Y. Lecun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[64] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Computer Vision and Pattern Recognition*, 2013, pp. 3626–3633.

[65] R. Benenson, M. Omran, J. Hosang, and B. Schiele, *Ten Years of Pedestrian Detection, What Have We Learned?* Springer International Publishing, 2014.
[66] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel, *Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features*. Springer International Publishing, 2014.
[67] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5079–5087.
[68] Y. Tian, P. Luo, and X. Wang, "Deep learning strong parts for pedestrian detection," in *International Conference on Computer Vision*, 2016, pp. 1904–1912.
[69] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5363–5371.
[70] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *International Conference on Neural Information Processing Systems*, 2015, pp. 424–432.
[71] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2071–84, 2015.
[72] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Computer Vision and Pattern Recognition*, 2016, pp. 1259–1267.

Xiao Wang received the B.E. degree from Information Technology School of Hebei University of Economics and Business in 2011. She is currently pursuing his Ph.D degree in School of Computer Science, Wuhan University. Her research interests focus on computer vision, multimedia analysis and machine learning, where she has published several conference papers including ICME, ICIP, and PCM.

Chao Liang received the Ph.D degree from National Lab of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2012. He is currently working as an associate professor at National Engineering Research Center for Multimedia Software (NERCMS), Computer School of Wuhan University, Wuhan, China. His research interests focus on multimedia content analysis and retrieval, computer vision and pattern recognition, where he has published over 60 papers, including premier conferences such as CVPR, ACM MM, AAAI, IJCAI and honorable journals like TNNLS, TMM and TCSVT, and won the best paper award of PCM 2014.

Chen Chen received the B.E. degree in automation from the Beijing Forestry University, Beijing, China, in 2009, and the M.S. degree in electrical engineering from the Mississippi State University, Starkville, MS, USA, in 2012, and the Ph.D. degree in the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, USA, in 2016. He held a Post-Doc position in the Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA. He is currently an assistant professor in the Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, NC, USA. He has published more than 50 papers in refereed journals and conferences in these areas. His research interests include compressed sensing, signal and image processing, pattern recognition, and computer vision.

Jun Chen received the M.S. degree in Instrumentation from Huazhong University of Science and Technology, Wuhan, China, in 1997, and Ph.D degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008. Dr. Chen is the deputy director of National Engineering Research Center for Multimedia Software, and a professor in school of computer science, Wuhan University. His research interests include multimedia analysis, computer vision and security emergency information processing, where he has published over 50 papers.

Zheng Wang received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and is currently working toward the Ph.D. degree at the National Engineering Research Center for Multimedia Software (NERCMS), School of Computer, Wuhan University. His research interests include multimedia content analysis and retrieval, computer vision, and pattern recognition. Mr. Wang was the recipient of the Best Paper Award at the 15th Pacific-Rim Conference on Multimedia (2015).

Zhen Han received the B.S. degree in computer science and technology and Ph.D. degree in computer application technology from Wuhan University, Wuhan, China, in 2002 and in 2009 respectively. Now he is an associate professor in school of computer, Wuhan University. His research interests include image/video compressing and processing, computer vision and artificial intelligence.

Chunxia Xiao received his B.S. and M.S. degrees from the Mathematics Department of Hunan Normal University in 1999 and 2002, respectively, and his PhD degree from the State Key Lab of CAD and CG of Zhejiang University in 2006. Currently, he is a professor in the School of Computer, Wuhan University, China. From October 2006 to April 2007, he worked as a postdoc at the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, and during February 2012 to February 2013, he visited University of California-Davis for one year. His main interests include computer graphics, computer vision and machine learning.