Center and Scale Prediction: Anchor-free Approach for Pedestrian and Face Detection

Wei Liu [a,1], Irtiza Hasan [b,1], Shengcai Liao [b,*]

[a] National University of Defense Technology, Changsha, China
[b] Inception Institute of Artificial Intelligence (IIAI), UAE

ARTICLE INFO

ABSTRACT

Object detection traditionally requires sliding-window classifier in modern deep learning based approaches. However, both of these approaches requires tedious configurations in bounding boxes. Generally speaking, single-class object detection is to tell where the object is, and how big it is. Traditional methods combine the "where" and "how" subproblems into a single one through the overall judgement of various scales of bounding boxes. In view of this, we are interesting in whether the "where" and "how" subproblems can be separated into two independent subtasks to ease the problem definition and the difficulty of training. Accordingly, we provide a new perspective where detecting objects is approached as a high-level semantic feature detection task. Like edges, corners, blobs and other feature detectors, the proposed detector scans for feature points all over the image, for which the convolution is naturally suited. However, unlike these traditional low-level features, the proposed detector goes for a higher-level abstraction, that is, we are looking for central points where there are objects, and modern deep models are already capable of such a high-level semantic abstraction. Like blob detection, we also predict the scales of the central points, which is also a straightforward convolution. Therefore, in this paper, pedestrian and face detection is simplified as a straightforward center and scale prediction task through convolutions. This way, the proposed method enjoys an anchor-free setting, considerably reducing the difficulty in training configuration and hyper-parameter optimization. Though structurally simple, it presents competitive accuracy on several challenging benchmarks, including pedestrian detection and face detection. Furthermore, a cross-dataset evaluation is performed, demonstrating a superior generalization ability of the proposed method.

© 2022 Published by Elsevier Ltd.

1. Introduction

Pedestrian detection is a very active research topic in the computer vision. It has applications in many different domains, including autonomous driving [1], video surveillance [2,3], pedestrian detection in underground mines [4], thermal imagery [5] and action recognition [6]. Pedestrian detection also plays a critical role as foundation steps in various other computer vision research areas, such as multi-object tracking [7], human pose estimation [8], person re-identification [9,10], etc. Recently, convolutional neural network (CNNs) based approaches have advanced the field of pedestrian detection a lot, with significantly improved performance.

For a broader domain, pedestrian detection is a special case of general object detection. Starting from the pioneering work of the Viola-Jones detector [11], object detection generally requires sliding-window classifiers in traditional or anchor based predictions in CNN-based methods. These detectors are essentially local classifiers used to judge the pre-defined anchors (windows or anchor-boxes) as being objects or not. However, either of these approaches requires tedious configurations in anchors. Several (but not all) of these configurations, which include the number of scales, the sizes of anchors, the aspect ratios, and the overlap thresholds with ground truth boxes. All of these are task-oriented, and it is difficult to figure out which combination is the optimal one. Generally speaking, single-class object detection is to tell where the object is, and how big it is. Traditional methods combine the "where" and "how" subproblems into a single one through the overall judgement of various scales of anchor boxes. In view of this, we are interesting in whether the "where" and "how"

---

* Corresponding author.
E-mail addresses: liuwei16@nudt.edu.cn (W. Liu), irtiza.hasan@inceptioniai.org (I. Hasan), shengcai.liao@inceptioniai.org (S. Liao).
[1] Equal contribution

subproblems can be separated into two independent subtasks to ease the problem definition and the difficulty of training.

We resort to traditional feature detection. Feature detection is one of the most fundamental problems in computer vision. It is usually viewed as a low-level technique, with typical tasks including edge detection (e.g. Canny [12], Sobel [13]), corner (or interest point) detection (e.g. SUSAN [14], FAST [15]), and blob (or region of interest point) detection (e.g. LoG [16], DoG [17], MSER [18]). Feature detection is of vital importance to a variety of computer vision tasks ranging from image representation, image matching to 3D scene reconstruction. Generally speaking, a feature is defined as an "interesting part" of an image. Therefore, feature detection aims to compute abstractions of image information. Subsequently, feature detection makes local decisions at every image point whether there is an image feature of a given type at that point or not. With the rapid development for computer vision tasks, deep convolutional neural networks (CNN) are believed to be of very good capability to learn high-level image abstractions. CNNs have also been applied for feature detection, and demonstrate attractive successes even in low-level feature detections. For example, there is a recent trend of using CNN to perform edge detection [19–22], which has substantially advanced this field. It shows that clean and continuous edges can be obtained by deep convolutions, which indicates that CNN has a stronger capability to learn higher-level abstraction of natural images than traditional methods. This capability may not be limited to low-level feature detection; it may open up many other possibilities of high-level feature detection.

Therefore, in this paper, we argue in the favor of posing pedestrian detection as high-level semantic feature detection task. However, unlike traditional low-level feature detectors [5,23], the proposed detector goes for a higher-level abstraction, that is, we are looking for central points where there are objects. Besides, similar to the blob detection, we also predict the scales of the central points. However, instead of processing an image pyramid to determine the scale as in traditional blob detection, we also predict object scale with a straightforward convolution in one pass upon a fully convolution network (FCN) [24], considering its strong capability. As a result, pedestrian and face detection is simply formulated as a straightforward center and scale prediction task via CNNs. Therefore, the proposed CSP detector separates the "where" and "how" subproblems into two different convolutions. These design choices enables CSP to enjoy an anchor-free (short for window-free or anchor-anchor-free) setting, considerably reducing the difficulty in training configuration and hyper-parameter optimization. The overall pipeline of the proposed method, denoted as **C**enter and **S**cale **P**rediction (**CSP**) based detector, is illustrated in Fig. 1.

Previously, FCN has already been applied to and made a success in multi-person pose estimation [25,26], where several keypoints are firstly detected merely through responses of full convolutions, and then they are further grouped into complete poses of individual persons. In view of this, recently two inspirational works, CornerNet [27] and TLL [28], successfully go free from windows and anchor-boxes, which perform object detection as convolutional keypoint detections and their associations. Though the keypoint association require additional computations, sometimes complex as in TLL, the keypoint prediction by FCN inspires us to go a step further, achieving center and scale prediction based pedestrian and face detection in full convolutions.

In summary, the main contributions of this work are as follows: (i) We show a new possibility that pedestrian and face detection can be simplified as a straightforward center and scale prediction task through convolutions, which bypasses the limitations of anchor-box based detectors and gets rid of the complex post-processing of recent keypoint pairing based detectors. (ii) The proposed CSP [29] detector achieves the new state-of-the-art performance on two challenging pedestrian detection benchmarks, CityPersons [30] and Caltech [31], it also achieves competitive performance on one of the most popular face detection benchmark– WiderFace [32]. (iii) The proposed CSP detector presents good generalization ability when cross-dataset evaluation is performed.

This work is built upon CSP [29]. The major new content is additional experiments on face detection to demonstrate CSP's capability in detecting objects with various aspect ratios. Besides, we also conduct comparative experiments and analysis to demonstrate the generalization ability of the proposed detector. They are summarized as follows:

- We evaluate the proposed method for face detection on one of the most popular face detection benchmarks, i.e. WiderFace [32]. The model is trained on the official training subset and evaluated on both the validation and test subsets. Comparable performance to other state-of-the-art face detectors on this benchmark are reported and thus demonstrates the proposed method's capability and competitiveness on face detection.
- To further evaluate the generalizability of the proposed CSP detector, we also conduct two cross-dataset evaluation experiments. For pedestrian detection, we compare the proposed CSP detector with the state-of-the-art anchor-box based pedestrian detector (ALFNet [33]). Both of the two detectors are trained on the CityPersons [30] training set and then are directly tested on Caltech [31]. For face detection, we compare the proposed CSP detector with the state-of-the-art anchor-box based face detector (DSFD [34]). Both of the two detectors are trained on
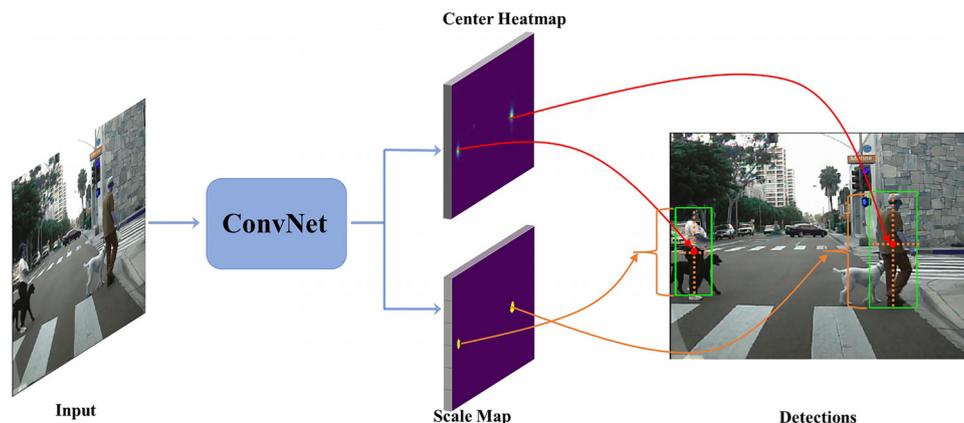

Fig. 1. The overall pipeline of the proposed CSP detector. The final convolutions have two channels, one is a heatmap indicating the locations of the centers (red dots), and the other serves to predict the scales (yellow dotted lines) for each detected center.

the WiderFace [32] training set and then are directly tested on FDDB [35], UCCS [36] and DarkFace [37]. Experimental results show that the proposed CSP detector has a superior generalization ability than the compared methods.

Besides, we considerably reorganize the whole paper, especially the introduction part, for a better and more clear motivation.

2. Related Works

2.1. Anchor-box based object detection

Anchor based object detection is arguably the dominant paradigm in object detection [38,39], popularized by [40]. Architecturally speaking, one of the major component of anchor-box based object detectors are fixed anchor-boxes (pre-defined scales and aspect ratios). Subsequently, these fixed anchor-boxes are used for classification and regression of final object bounding-box. Most two-stage based object detectors, such as Faster R-CNN [40], generates proposals (in region proposal network) and progressively classifies and refines them in its downstream classifier (Fast R-CNN branch). Making it end-to-end trainable, in a single unified framework. In contrast to two stage methods, singe-stage detectors, such as SSD [41], without the proposal generation step and achieved comparable accuracy while are more efficient (computationally) than two-stage detectors.

Pedestrian detection can be broadly categorized into two sub-categories, 1) from the context of autonomous driving [30] and 2) surveillance [42]. In both cases, majority of the methods are direct extensions of Faster R-CNN. For example, [43] was the first work to use RPN [40] along with random forest as a downstream classifier. MS-CNN [44] utilized the Faster R-CNN framework, but additionally exploited the multi-scale feature maps. Moreover, Zhang *et al.* [30] used standard faster R-CNN but with added strategies for training, such as high image resolution, larger receptive field, quantization for anchors etc. [42] suggested a scale-aware Faster R-CNN. Related to [42,45] suggested a complexity-aware method, which combined features of different complexities for pedestrian detection. Moreover, [46] proposed an attribute-aware pedestrian detector to explicitly model people's semantic attributes in a high-level feature detection method. Recently, [47] proposed WSSN, which utilizes Faster R-CNN architecture along with a parallel branch for depth images. WSSN [47] also generates segmentation maps for pedestrians (using only information from depth images), without using accurate pixel-level semantic segmentation masks. Despite achieving high accuracy on non-occluded pedestrians, occluded pedestrian detection stands as one of the major problems in pedestrian detection [48], initial works focused on tailoring of the loss function to address the problems [49,50]. Subsequently, Bi-Box [51] extended Faster R-CNN by proposing a separate branch to predict the visible parts of the pedestrian. In addition to occlusion, small-scale pedestrians are also among the harder cases [52]. Recently, [53] proposed Dual-Region Feature Generation (DRFG) method, the central idea is to generate two sets of proposals for visible body parts and full body. Subsequently visible part is used as a clue to detect full body. Several techniques have been used to address small-scale object detection [54,55], in particular, methods based on reinforcement learning have been successful in detecting small scale objects [56–59], where the general idea is to first propose probable regions containing small-scale objects and subsequently attending only to these regions at a higher resolution for better accuracy and low computational cost. Similar concepts have been utilized for small scale pedestrian detection as well [60,61]. Besides reinforcement learning based approaches, other methods such as [62] utilized pose estimation along with detec-

tion in a joint optimization for achieving better accuracy on small scale pedestrians.

Unlike general object detection, single stage methods [63,64] achieved superior performances than two stage methods in pedestrian detection. For instance, ALFNet [33] and [65] proposes the asymptotic localization fitting strategy to progressively refine anchor-boxes for accurate localization. Finally, [66] ameliorated SSD architecture by focusing on the discriminative feature learning.

For face detection, it is dominated by the single-stage framework in recent years. Most of the advanced face detectors focus on the anchor-box design [67] and matching strategies, because faces in the wild exhibits a large variation in size. Previously, [67] proposed framework the relied on cascade of anchors at various scales to better detect faces of different sizes. Furthermore, FaceBoxes [68] introduces an anchor-box densification strategy to ensure anchor-boxes of different sizes have the same density on an image, and in [69,70], the authors propose different anchor-box matching threshold to ensure a certain number of training examples for tiny faces, further DSFD [34] proposes an improved anchor-box matching strategy to provide better initialization for the regressor. In SSH [71], anchor-boxes with two neighboring sizes share the same detection feature map. PyramidBox [72] designs novel PyramidAnchors to help contextual feature learning. Recently, GSSD [73] was proposed which detects faces of various scales using grouping strategy.

2.2. Anchor-free object detection

Anchor-free detectors bypass the requirement of anchor-boxes and detect objects directly from an image. DeNet [74] proposes to generate proposals by predict the confidence of each location belonging to four corners of objects. Following the two-stage pipeline, DeNet also appends another sub-network to re-score these proposals. Within the single-stage framework, YOLO [75] appends fully-connected layers to parse the final feature maps of a network into class confidence scores and box coordinates. Densebox [76] devises a unified FCN that directly regresses the classification scores and distances to the boundary of a ground truth box on all pixels, and demonstrates improved performance with landmark localization via multi-task learning. Therefore, it is effective pipeline different from Faster R-CNN [40]. However, DenseBox resizes objects to a single scale during training, thus requiring image pyramids to detect objects of various sizes by multiple network passes during inference. Besides, Densebox defines the central area for each object and thus requires four parameters to measure the distances of each pixel in the central area to the object's boundaries. In contrast, the proposed CSP detector defines a single central point for each object, therefore two parameters measuring the object's scale are enough to get a bounding box, and sometimes one scale parameter is enough given uniform aspect ratio in the task of pedestrian detection. Most recently, CornerNet [27] also applies a FCN but to predict objects' top-left and bottom-right corners and then group them via associative embedding [26]. Enhanced by the novel corner pooling layer, CornerNet achieves superior performance on MS COCO object detection benchmark [77]. Similarly, TLL [28] proposes to detect an object by predicting the top and bottom vertexes. To group these paired keypoints into individual instances, it also predicts the link edge between them and employs a post-processing scheme based on Markov Random Field. Applying on pedestrian detection, TLL achieves significant improvement on Caltech [31], especially for small-scale pedestrians.

The proposed methodology is an an anchor-free detection method. However, our work is different from the conventional approaches, as such, that we try to illustrate that a single FCN can be effectively deployed for face and pedestrian detection. In doing so, we also show that a single center point can be used for pre-

cise localization without any post-processing methodologies (except NMS).

2.3. Feature detection

Feature detection is a widely studied problem with extensive literature in computer vision. Broadly, it primarily includes edge detection [12,13], corner detection [15,78], blob detection [18,79] etc. Classical methods [12,13] of feature detection mostly relied on local cues, such as the brightness, colors, gradients and textures. However, after the emergence of CNN-based methods, the task of feature detection was greatly advanced [80–83]. For instance, recently several methods deployed CNN-based methods to perform edge detection, such as, [19–22], which have substantially advanced this field. However, unlike the above mentioned methods that perform low-level feature detection (edge, corners and blobs), the proposed method aims for a higher level of abstraction, that is, our center of attention is localizing central points, where there are pedestrians, for which modern deep models are naturally suited.

3. Proposed Method

3.1. Preliminary

Generally, with an input image I, the detection network may generate several feature maps with different resolutions, which can be defined as follows:

$$\phi_i = f_i(\phi_{i-1}) = f_i(f_{i-1}(...f_2(f_1(I)))), \tag{1}$$

where ϕ_i represents feature maps output by the ith stage. These feature maps are generated by $f_i(.)$, decreasing in size progressively. Given a network with N downsampling stages, all the generated feature maps can be denoted as $\Phi = \{\phi_1, \phi_2, ..., \phi_N\}$, upon which detection heads are further built.

We denote these feature maps that are responsible for detection as Φ_{det}. Specifically, in a object detector, the feature maps responsible for detection can be represented as $\Phi_{det} = \{\phi_L, \phi_{L+1}, ..., \phi_N\}$, where $1 < L < N$. Besides Φ_{det}, in anchor-based detectors, another key component is called anchors (termed as $\mathcal{B}$). Given Φ_{det} and $\mathcal{B}$ in hand, detection can be formulated as:

$$Dets = \mathcal{H}(\Phi_{det}, \mathcal{B}) = \{cls(\Phi_{det}, \mathcal{B}), regr(\Phi_{det}, \mathcal{B})\}, \tag{2}$$

where $\mathcal{B}$ is pre-defined according to the corresponding set of feature maps Φ_{det}, and $\mathcal{H}(.)$ represents the detection head. Generally, $\mathcal{H}(.)$ contains two elements, namely $cls(.)$ which predicts the classification scores, and $regr(.)$ which predicts the scaling and offsets of the anchors.

While in anchor-free detectors, detection is performed merely on the set of feature maps Φ_{det}, that is,

$$Dets = \mathcal{H}(\Phi_{det}) \tag{3}$$

3.2. Overall architecture

The proposed CSP belongs to the group of anchor-free detectors, the overall architecture is illustrated in Fig. 2. The backbone network are truncated from a standard network pretrained on ImageNet [84].

Feature Extraction. Taking ResNet-50 as an example, it can be divided into five stages, where the output feature maps of each stage are downsampled by 2, 4, 8, 16, 32 w.r.t. the input image. As a common practice [28,49], the dilated convolutions are adopted in *stage 5* to maintain the final feature map with a higher resolution. We denote the output of *stage 2, 3, 4 and 5* as ϕ_2, ϕ_3, ϕ_4 and ϕ_5, in which the shallower feature maps can provide more precise localization information, while the coarser ones contain more semantic information with increasing the sizes of receptive fields. We simply fuse these multi-scale feature maps by concatenation, before which a deconvolution layer is adopted to make multi-scale feature maps with the same resolution. Since the feature maps from each stage have different scales, we use L2-normalization to rescale their norms to 10, as adopted in [66]. To investigate the optimal combination from these multi-scale feature maps, we conduct an ablative experiment in Sec. 4.2. Given an input image of size $H \times W$, the size of final concatenated feature maps is $H/r \times W/r$, where r is the downsampling factor. Similarly to [28], $r = 4$ gives the best performance as demonstrated in our experiments, because a larger r means coarser feature maps which struggle on accurate localization. Given its simplicity, more complicated feature fusion strategies in [85–87] can be explored to further improve the performance, but it is not in the scope of this work.

Detection Head. Upon the concatenated feature maps Φ_{det}, a detection head is appended to parse it into detection results. As stated in [88], the detection head plays a significant role. In this

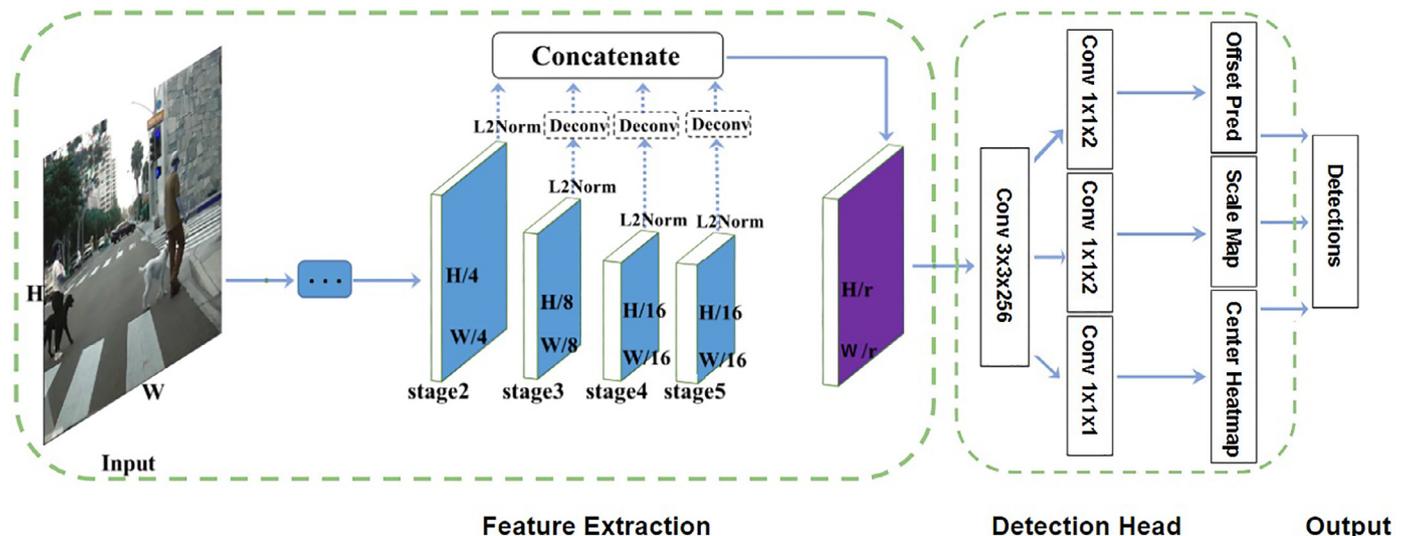**Fig. 2.** Overall architecture of CSP, which mainly comprises two components, i.e. the feature extraction module and the detection head. The feature extraction module concatenates feature maps of different resolutions into a single one. The detection head merely contains a 3x3 convolutional layer, followed by three prediction layers, for the offset prediction, center location and the the corresponding scale.
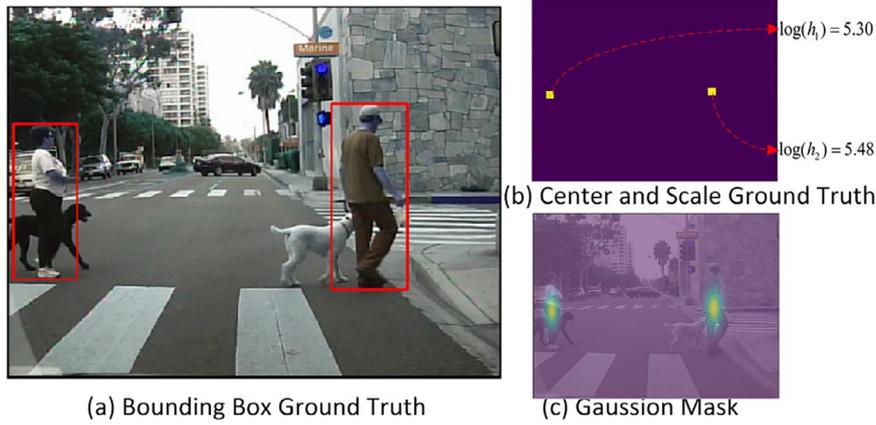
(a) Bounding Box Ground Truth

(b) Center and Scale Ground Truth

(c) Gaussion Mask

Fig. 3. (a) is the bounding box annotations commonly adopted by anchor-box based detectors. (b) is the center and scale ground truth generated automatically from (a). Locations of all objects' center points are assigned as positives, and negatives otherwise. Each pixel is assigned a scale value of the corresponding object if it is a positive point, or 0 otherwise. We only show the height information of the two positives for clarity. (c) is the overall Gaussian mask map M defined in Eq. 4 to reduce the ambiguity of these negatives surrounding the positives.

work, we firstly attach a single 3x3 *Conv* layer on Φ_{det} to reduce its channel dimensions to 256, and then two sibling 1x1*Conv* layers are appended to produce the scale map and center heatmap, respectively. Also, we do this for simplicity and any improvement of the detection head [34,88–90] can be flexibly incorporated into this work to be a better detector.

Optionally, to slightly adjust the center location, an extra offset prediction branch can be appended in parallel with the above two branches. This extra offset branch resolves mismatching of centers that occurs due to downsampled feature maps by learning the shifted offset. We will demonstrate the effectiveness of the extra offset branch in (Sec. 4.2).

3.3. Training

Ground Truth. The predicted heatmaps are with the same size as the concatenated feature maps (i.e. $H/r \times W/r$). An illustration example is depicted in Fig. 3 (b) to illustrate how to generate the center and scale ground truth by the given bounding box annotations. For the center ground truth, the location where an object's center point falls in is assigned as positive while all others are negatives.

Scale can be defined as the height and/or width of objects. In pedestrian detection, line annotation is first proposed in [30,91] to generate high-quality ground truth, where tight bounding boxes are automatically generated with a uniform aspect ratio[2] of 0.41. In accordance to this annotation, we can merely predict the height of each pedestrian instance and generate the bounding box with the predetermined aspect ratio. Therefore, for the scale ground truth, the kth positive location is assigned with the value of $log(h_k)$. To reduce the center ambiguity, $log(h_k)$ is also assigned to the negatives within a radius 2 of the positives, while all other locations are assigned as zeros. Alternatively, we can also predict the width or height+width, which is required for face detection, because face objects in the wild exhibit a large variation in aspect ratios.

When the offset prediction branch is appended, the ground truth for the offsets of those centers can be defined as $(\frac{x_k}{r} - \lfloor \frac{x_k}{r} \rfloor, \frac{y_k}{r} - \lfloor \frac{y_k}{r} \rfloor)$.

Loss Function. The center prediction can be formulated as a classification task. Note that it is difficult to decide an 'exact' center point. In order to reduce the ambiguity of these negatives surrounding the positives, we also apply a 2D Gaussian mask $G(.)$

---

[2] Aspect ratio here is defined as width / height

centered at the location of each positive, which is also adopted in [27,28]. An example of the overall mask map M is depicted in Fig. 3 (c). Formally, it is formulated as:

$$M_{ij} = \max_{k=1,2,\ldots,K} G(i, j; x_k, y_k, \sigma_{w_k}, \sigma_{h_k}),$$
$$G(i, j; x, y, \sigma_w, \sigma_h) = e^{-(\frac{(i-x)^2}{2\sigma_w^2} + \frac{(j-y)^2}{2\sigma_h^2})}, \tag{4}$$

where K is the number of objects in an image, (x_k, y_k, w_k, h_k) is the center coordinates, width and height of the kth object, and the variances (σ_w^k, σ_h^k) of the Gaussian mask are proportional to the height and width of individual objects. If these masks have overlaps, we choose the maximum values for the overlapped locations. To combat the extreme positive-negative imbalance problem, the focal weights [92] on hard examples are also adopted. Thus, the classification loss can be formulated as:

$$L_{center} = -\frac{1}{K} \sum_{i=1}^{W/r} \sum_{j=1}^{H/r} \alpha_{ij}(1 - \hat{p}_{ij})^\gamma log(\hat{p}_{ij}), \tag{5}$$

where

$$\hat{p}_{ij} = \begin{cases} p_{ij} & \text{if } y_{ij} = 1 \\ 1 - p_{ij} & \text{otherwise,} \end{cases}$$
$$\alpha_{ij} = \begin{cases} 1 & \text{if } y_{ij} = 1 \\ (1 - M_{ij})^\beta & \text{otherwise.} \end{cases} \tag{6}$$

In the above, $p_{ij} \in [0, 1]$ is the center prediction in the location (i, j), and $y_{ij} \in \{0, 1\}$ specifies the ground truth label, where $y_{ij} = 1$ represents the positives and $y_{ij} = 0$ the negatives. γ is the focusing hyper-parameter, which is set as 2 as suggested in [92]. To reduce the ambiguity from those negatives surrounding the positives, the α_{ij} according to the Gaussian mask M is applied to reduce their contributions to the total loss, in which the hyper-parameter β controls the penalty. Experimentally, $\beta = 4$ gives the best performance, which is similar in [27].

The scale prediction can be formulated as a regression task via smooth L1 loss [93]:

$$L_{scale} = \frac{1}{K} \sum_{k=1}^{K} SmoothL1(s_k, t_k), \tag{7}$$

where s_k and t_k represents the network's prediction and the ground truth of each positive, respectively.

If the offset prediction branch is appended, the similar smooth L1 loss in Eq. 7 is adopted (denoted as L_{offset}).

To sum up, the full optimization objective is:

$$L = \lambda_c L_{center} + \lambda_s L_{scale} + \lambda_o L_{offset}, \tag{8}$$

where λ_c, λ_s and λ_o are the weights for center classification, scale regression and offset regression losses, which are experimentally set as 0.01, 1 and 0.1, respectively.

3.4. Inference

For inference, CSP simply involves a single forward of the network. Specifically, given the predictions, locations with confidence score above 0.01 in the center heatmap are kept, along with their predicted scale in the scale map. Then bounding boxes are generated and remapped to the original image size, followed by NMS. If the offset prediction branch is adopted, the centers are shifted according to the predicted offsets before being remapped to the original image size.

4. Experiments

4.1. Experiment settings

4.1.1. Datasets and evaluation metrics

We evaluate our approach on two tasks, including pedestrian detection and face detection.

For pedestrian detection, we choose two of the largest benchmark datasets: Caltech [31] and CityPersons [30]. The Caltech dataset contains approximately 2.5 hours of autodriving video with extensively labelled bounding boxes. Following [30,33,49,50,94], we use the training data augmented by 10 folds (42782 frames) and the standard test set with 4024 frames, evaluation results are reported based on the new annotations provided by [95]. The CityPersons dataset is a more challenging benchmark with various occlusion levels. We train the models on the official training set with 2975 images and test on the validation set with 500 images. Evaluation follows the standard Caltech evaluation metric [31], that is log-average Miss Rate (MR) over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^0]$ (denoted as MR^{-2}).

For face detection, we choose one of the most challenging face detection benchmark, i.e. WiderFace [32]. The main reason we choose this dataset is due to its large variability of scale, aspect ratios, illumination and occlusions. The dataset defines three levels of difficulty: Easy, Medium and Hard by the detection rate of Edge-Box [96]. We train the proposed CSP merely on the training subset and test on both validation and testing subsets by the Average Precision (AP).

We first conduct the ablative study on the Caltech dataset, and then compare the proposed CSP detector with the state of the arts on all the above benchmarks.

4.1.2. Training details

The proposed method is implemented in Keras[3]. We use ResNet-50 [97] pretrained on ImageNet [84] as the backbone unless otherwise stated. Adam [98] optimizer is applied and the strategy of moving average weights [99] is used to achieve more stable training. During training, standard data augmentations include random color distortion, horizontal flip, scaling and crop, the input is finally resized to 336x448, 640x1280 and 704x704 pixels for Caltech, CityPersons and WiderFace, respectively.

For Caltech, a mini-batch contains 16 images with one GPU (GTX 1080Ti), the learning rate is set as 10^{-4} and training is stopped after 15K iterations, the model initialized from CityPersons [30] is trained with the learning rate of 2×10^{-5}. For CityPersons, a mini-batch contains 8 images on 4 GPUs with the learning rate of 2×10^{-4} and training is stopped after 37.5K iterations. For Wider-Face, a mini-batch contains 16 images on 8 GPUs with the learning

---

3 https://github.com/fchollet/keras

Table 1
Comparisons of different high-level feature points. Bold number indicates the best result.

Point Prediction	MR^{-2}(%)	
	IoU=0.5	IoU=0.75
Center point	**4.62**	**36.47**
Top vertex	7.75	44.70
Bottom vertex	6.52	40.25

Table 2
Comparisons of different definitions for scale prediction. Bold number indicates the best result.

Scale Prediction	MR^{-2}(%)	
	IoU=0.5	IoU=0.75
Height	**4.62**	**36.47**
Width	5.31	53.06
Height+Width	4.73	41.09

rate of 2×10^{-4} and training is stopped after 99.5K iterations. Similar data augmentations in PyramidBox [72] is applied to increase the proportion of small faces during training.

4.2. Ablation Study

We perform ablation study of the key components of CSP on the Caltech dataset.

Why do we use the Center Point? The center point is capable of locating an individual object. A coming question is how about other high-level feature points. To answer this, we choose two other high-level feature points adopted in [28], i.e. the top and bottom vertexes. Comparisons are reported in Table 1. It is shown that both the two vertexes can succeed in detection but underperform the center point by approximately 2%-3% under IoU=0.5, and the performance gap is even larger under the stricter IoU=0.75. This is probably because the center point is advantageous to perceive the full body information.

How important is the Scale Prediction? Scale prediction is another indespensible component of CSP. In practice, we merely predict the height for each detected center. To demonstrate the generality of CSP, we have also tried to predict Width or Height+Width for comparison. For Height+Width, the only difference lies in that the scale prediction branch has two channels responsible for the height and width respectively. It can be observed in Table 2 that Width and Height+Width prediction can also achieve comparable but suboptimal results to Height prediction. This result may be attributed to the line annotation style [30,91] which provides accurate Height information, while Width is automatically generated by a fixed aspect ratio and thus is not able to provide additional information for training. Anyway, CSP is potentially feasible for other detection tasks requiring both height and width, which will be demonstrated in the following experiments for face detection.

How important is the Feature Resolution? In the proposed method, the final set of feature maps (denoted as Φ_{det}^r) is downsampled by r w.r.t the input image. To explore the influence from r, we train the models with $r = 2, 4, 8, 16$ respectively. For $r = 2$, Φ_{det}^2 are up-sampled from Φ_{det}^4 by deconvolution. For $r = 4, 8, 16$, the offset prediction branch is alternatively appended. Stricter evaluations under IoU=0.75 are included to verify the effectiveness of additional offset prediction. As can be seen from Table 3, without offset prediction, Φ_{det}^4 presents the best result under IoU=0.5, but performs poorly under IoU=0.75 when compared with Φ_{det}^2, which indicates that finer feature maps are beneficial for precise local-

Table 3

Comparisons of different downsampling factors of the feature maps, which are denoted as Φ_{det}^r downsampled by r w.r.t the input image. Test time is evaluated on the image with size of 480x640 pixels. ΔMR^{-2} means the improvement from the utilization of the offset prediction. Bold numbers indicate the best result.

Feature for Detection	+Offset	Test Time (ms/img)	MR^{-2}(%)		ΔMR^{-2}(%)	
			IoU=0.5	IoU=0.75	IoU=0.5	IoU=0.75
Φ_{det}^2		69.8	5.32	30.08	-	-
Φ_{det}^4		58.2	4.62	36.47	+0.08	+7.67
	✓	59.6	**4.54**	**28.80**		
Φ_{det}^8		49.2	7.00	54.25	+0.92	+21.32
	✓	50.4	6.08	32.93		
Φ_{det}^{16}		42.0	20.27	75.17	**+12.86**	**+41.30**
	✓	42.7	7.41	33.87		

Table 4

Comparisons of different combinations of multi-scale feature representations defined in Sec. 3.2. ϕ_2, ϕ_3, ϕ_4 and ϕ_5 represent the output of *stage 2, 3, 4 and 5* of a backbone network, respectively. Bold numbers indicate the best results.

Feature Maps				ResNet-50 [97]			MobileNetV1 [100]		
ϕ_2	ϕ_3	ϕ_4	ϕ_5	# Parameters	Test Time	MR^{-2}(%)	# Parameters	Test Time	MR^{-2}(%)
✓	✓			4.7MB	36.2ms/img	9.96	2.1MB	27.3ms/img	34.96
	✓	✓		16.1MB	44.5ms/img	5.68	6.0MB	32.3ms/img	**8.33**
		✓	✓	37.4MB	54.4ms/img	5.84	10.7MB	34.5ms/img	10.03
✓	✓	✓		16.7MB	46.0ms/img	6.34	6.3MB	33.3ms/img	8.43
	✓	✓	✓	40.0MB	58.2ms/img	**4.62**	12.3MB	38.2ms/img	9.59
✓	✓	✓	✓	40.6MB	61.1ms/img	4.99	12.6MB	40.5ms/img	9.05

ization. Not surprisingly, a larger r witnesses a significant performance drop, which is mainly because coarser feature maps result in poor localization. In this case, additional offset prediction can substantially improve the detector upon Φ_{det}^{16} by 12.86% and 41.30% under the IoU threshold of 0.5 and 0.75, respectively. It also brings an improvement of 7.67% under IoU=0.75 for the detector upon Φ_{det}^4, with negligible extra computation cost, approximately 1ms per image of 480x640 pixels.

How important is the Feature Combination? It is revealed in [28] that multi-scale representation is vital for detection of various scales. In this part, we consider different combinations of multi-scale feature maps from the backbone. In practice we choose the output of stage 2 (ϕ_2) as a start point and the downsampling factor r is fixed as 4. In spite of the ResNet-50 [97] with *stronger* feature representation, a light-weight network like MobileNetV1 [100] is also choosen. Results in Table 4 reveal that the much shallower feature maps like ϕ_2 result in poorer accuracy, while deeper feature maps like ϕ_4 and ϕ_5 are of great importance for superior performance, and the middle-level feature maps ϕ_3 are indispensable to achieve the best results. For ResNet-50, the best performance comes from the combination of $\{\phi_3, \phi_4, \phi_5\}$, while $\{\phi_3, \phi_4\}$ is the optimal one for MobileNetV1.

4.3. Comparison with the State of the Arts

4.3.1. Pedestrian Detection

Caltech. Extensive comparisons are conducted on three settings: Reasonable, All and Heavy Occlusion. As shown in Fig. 4, CSP achieves MR^{-2} of 4.5% on the Reasonable setting, outperforming the best competitor (5.0 of RepLoss [49]) by 0.4%. With the model initialized from CityPersons[30], CSP achieves a new state of the art of 3.8%, compared to 4.0% of RepLoss [49]. A mere pre-training on CityPersons further boosts the performances because Caltech has lower person per images density compared to CityPersons[4]. It presents the superiority on detecting pedestrians of various scales and occlusion levels as demonstrated in Fig. 4(b). Moreover, Fig. 4(c) shows that CSP also performs very well for heav-

ily occluded pedestrians, outperforming RepLoss [49] and OR-CNN [50] which are explicitly designed for occlusion cases.

CityPersons. Table 5 shows the comparison with previous state of the arts on CityPersons. Following [30,49], results on subsets with different occlusion levels and various scale ranges are also reported. It can be observed that CSP beats the competitors and performs fairly well on occlusion cases even without any specific occlusion-handling strategies [49,50]. On the Reasonable subset, CSP performs the best with a gain of 1.0% MR^{-2} upon the closest competitor (ALFNet [33]), while the speed is comparable on the same running environment with 0.33 second per image of 1024x2048 pixels. Additionally, we also compare CSP with state of the arts (RetinaNet [92] and CornerNet [27]) in generic object detection, results presented in Table 5 demonstrate the superiority of CSP in pedestrian detection.

CrowdHuman. We further validated our approach on a recently collected large-scale general purpose (curated from web-crawling) person detection dataset, CrowdHuman [101]. Unlike CityPersons and Caltech, CrowdHuman dataset does not have a fixed aspect ratio. CrowdHuman benchmark is a more diverse and dense dataset than CityPerson and Caltech in terms of person per image and unique pedestrians. We compare with a single stage object detector RetinaNet [92] in Table 6. CSP outperforms RetinaNet [92] on a general person detection benchmark by more than 1 MR^{-2}(%), illustrating the robustness of the proposed approach.

4.3.2. Face Detection

WiderFace. The model trained on the training subset of WiderFace are evaluated on both the validation and test subsets, and the multi-scale testing is also performed in a similar way as in [72]. Comparisons with the state-of-the-art face detectors on WiderFace are shown in Fig. 5. It can be seen that the proposed CSP detector achieves competitive performance among the state-of-the-art face detectors across the three subsets, i.e. 90.7% (Hard), 95.2% (Medium) and 96.1% (Easy) on validation subset, and 89.9% (Hard), 94.4% (Medium) and 94.9% (Easy) on test subset. Note that most of these face detectors are anchor-box based. Therefore, the results indicate a superiority of the proposed CSP detector when complex

---

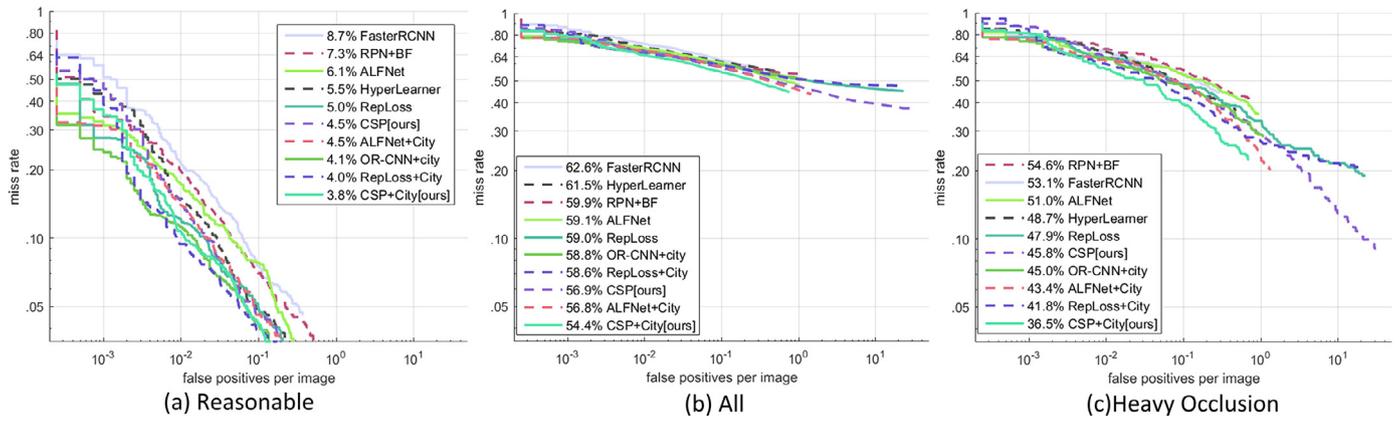[4] CityPersons has roughy 6 persons/image whereas Caltech has 0.3 [101]

Fig. 4. Comparisons with the state of the arts on Caltech using new annotations.

Table 5

Comparison with the state of the arts on CityPersons [30]. Results test on the original image size (1024x2048 pixels) are reported. **Red** and *green* indicate the best and second best performance.

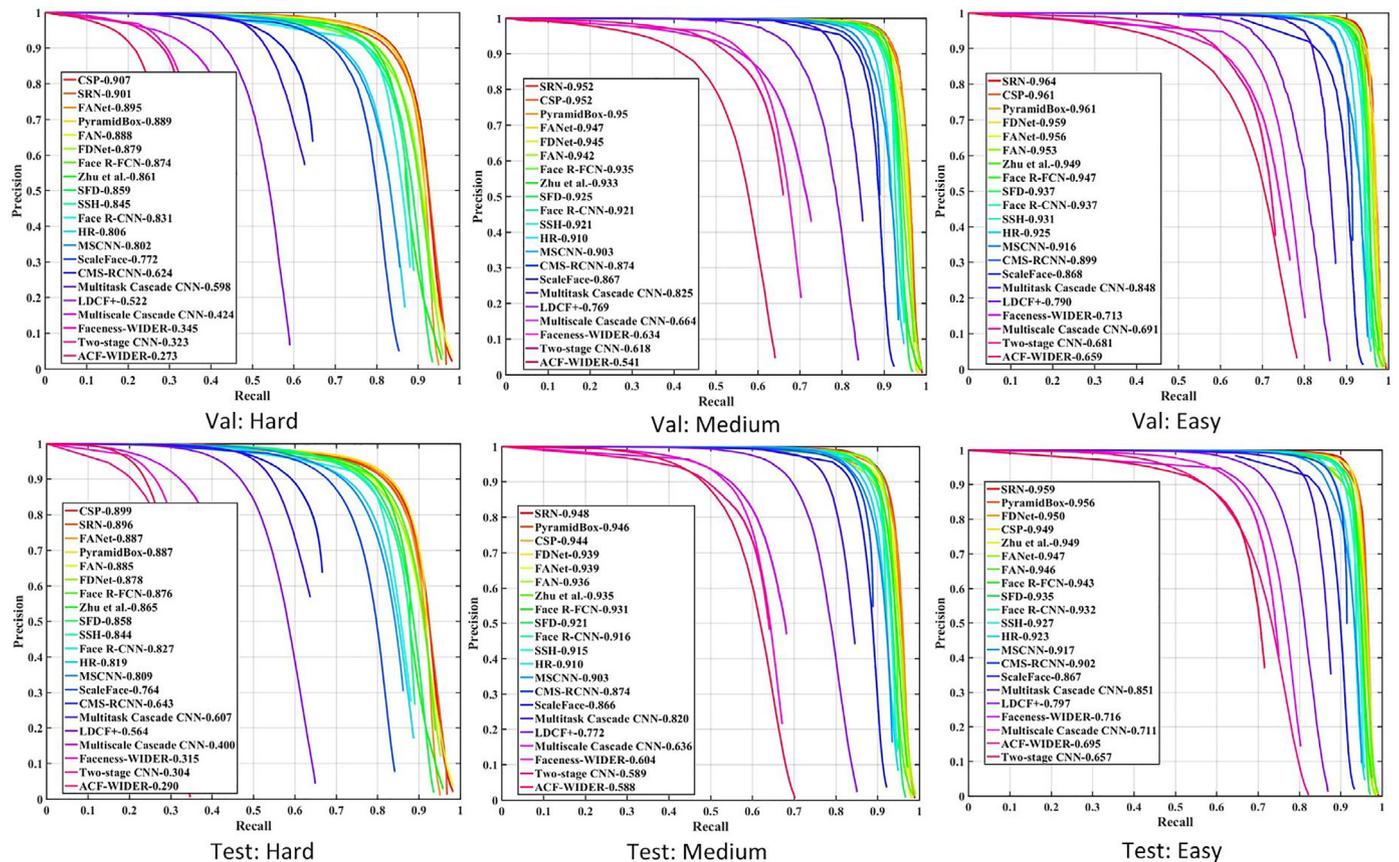| Method	Backbone	Reasonable	Heavy	Partial	Bare	Small	Medium	Large	Test Time
FRCNN [30]	VGG-16	15.4	-	-	-	25.6	7.2	7.9	-
RetinaNet [92]	ResNet-50	15.6	49.98	-	-	-	-	-	-
CornerNet [27]	Hourglass-54	21.0	56.0	-	-	-	-	-	-
FRCNN+Seg [30]	VGG-16	14.8	-	-	-	22.6	6.7	8.0	-
OR-CNN [50]	VGG-16	12.8	55.7	15.3	**6.7**	-	-	-	-
RepLoss [49]	ResNet-50	13.2	56.9	16.8	7.6	-	-	-	-
TLL [28]	ResNet-50	15.5	53.6	17.2	10.0	-	-	-	-
TLL+MRF [28]	ResNet-50	14.4	52.0	15.9	9.2	-	-	-	-
ALFNet [33]	ResNet-50	12.0	51.9	11.4	8.4	19.0	5.7	6.6	0.27s/img
CSP(w/o offset)	ResNet-50	*11.4*	*49.9*	*10.8*	8.1	*18.2*	*3.9*	**6.0**	0.33s/img
CSP(with offset)	ResNet-50	**11.0**	**49.3**	**10.4**	*7.3*	**16.0**	**3.7**	*6.5*	0.33s/img

Fig. 5. Precision-recall curves on WIDER FACE validation and testing subsets.

Table 6

Comparison of CSP with another single stage object detector on general person detection dataset, CrowdHuman).

Methods	MR^{-2}(%)
RetinaNet [92]	63.3
CSP[ours]	**62.1**

Table 7

Comparisons of the generalization ability for pedestrian detection (Evaluation metric: log average Miss Rate; the lower, the better).

Methods	CityPersons → CityPersons	CityPersons → Caltech
ALFNet [33]	12.0	17.8
CSP[ours]	**11.0**	**11.9**

Table 8

Statistics of three face detection datasets for cross-dataset evaluation.

Dataset	FDDB [35]	UCCS [36]	DarkFace [37]
Num. of images	2485	5232	6000
Num. of faces	5171	11109	43849
Mean face size	95x141	225x407	16x17
Mean image size	377x399	5184x3456	1080x720

default anchor-box design and anchor-box matching strategies are abandoned.

4.4. Generalization ability of the proposed method

To further demonstrate the generalization ability of the proposed CSP detector, we perform cross-dataset evaluation on two tasks, i.e. pedestrian detection and face detection. Specifically, models trained on the source dataset are directly tested on the target dataset without further finetuning. Furthermore, unlike humans, faces occur in different scale and aspect ratio, as shown in Table 11. In Table 11, we illustrate that in order to achieve high localization accuracy for faces, CSP needs to predict both height and width jointly, as oppose to simply predicting height (as in pedestrian detection).

4.4.1. Cross-dataset evalutaion for Pedestrian Detection

For pedestrian detection, we compare the proposed CSP detector with the state-of-the-art anchor-box based pedestrian detector (ALFNet [33]). Both of the two detectors are trained on the CityPersons [30] training subset and then are directly tested on the Caltech [31] test subset. For ALFNet [33], we use the source code and models provided by the authors[5]. Results shown in Table 7 are based on the reasonable setting, and the evaluation metric is log average Miss Rate (MR). It can be seen that the gap between the two detectors on the source dataset (CityPersons) is merely 1%, but the gap on the target dataset (Caltech) increases to 5.9%, which gives the evidence that the proposed CSP detector generalizes better to another dataset than the anchor-box based competitor, i.e. ALFNet [33].

4.4.2. Cross-dataset evaluation for Face Detection

For face detection, models trained on the WiderFace [32] training subset are directly tested on three other face detection datasets, i.e. FDDB [35], UCCS [36] and DarkFace [37]. Detailed statistics about these three datasets for testing are listed in Table 8. It can be seen that these three datasets exhibit a large difference in the mean size of face objects.

Table 9

Comparisons on generalization ability of face detectors (Evaluation metric: Average Precision (AP); the higher, the better).

Methods	WiderFace → WiderFace			WiderFace → UCCS	WiderFace → DarkFace
	Hard	Medium	Easy		
DSFD [34]	**90.0**	**95.3**	**96.0**	7.6	25.9
CSP[ours]	89.9	94.4	94.9	**11.3**	**28.0**

FDDB [35] is also a widely adopted face detection benchmark. Comparisons with other advanced face detectors on this benchmark are reported in Fig. 6, results of other face detectors are from FDDB's official website[6]. As shown in Fig. 6, the proposed CSP detector achieves competitive results on both discontinuous and continuous ROC curves, with the true positive rate of 99.2% and 86.0% when the number of false positives equals to 1000, while the results of the most recent anchor-box based face detector (DSFD [34]) are 99.1% and 86.2%, respectively. Since face images of both FDDB and WiderFace are obtained from the Internet, they are basically similar. Therefore, both CSP and DSFD detectors trained on WiderFace perform quite good on FDDB, and there is little performance difference between them.

However, when evaluated on the other two quite different face datasets, UCCS [36] and DarkFace [37], it is interesting to see some difference. UCCS [36], with the full name of UnConstrained College Students (UCCS), is a recently published dataset collected by a long-range high-resolution surveillance camera. The significant difference between UCCS [36] and other face detection datasets is that the data are collected unconstrainedly in surveillance scenes. People walking on the sidewalk did not aware that they were being recorded. As the annotations of the test subset is publicly unavailable, results on the validation subset are reported.

DarkFace [37] is a recently published face detection dataset collected during nighttime, which exhibits an extreme light condition compared to other face detection datasets. The average size of the face objects in this dataset is merely 16x17 pixels. In the official website[7], it totally released 6000 images, on which we test both models and report the results.

For UCCS [36] and DarkFace [37], we compare the proposed CSP with the state-of-the-art anchor-box based face detector (DSFD [34]). For DSFD, we use the source code and models provided by the authors[8]. Results are given in Table 9, and the evaluation metric is the Average Precision (AP). As can be seen from Table 9, though the proposed CSP slightly underperforms DSFD on the WiderFace test subset, it achieves a significant gain over DSFD on these two cross-dataset evaluations, with 3.7% and 2.1% on UCCS [36] and DarkFace [37], respectively. Due to the substantial domain gaps between UCCS, DarkFace and WiderFace, both models trained on WiderFace perform unsurprisingly poor on UCCS and DarkFace, but the proposed CSP detector still outperforms the anchor-box based DSFD, which gives the evidence that CSP generalizes better to unknown domains than the anchor-box based competitor. It is possible that the default configurations of anchor-boxes in anchor-based detectors can not adapt to new scenes especially when the scales and aspect ratios of objects have a large difference as shown in Table 8. In contrast, the proposed detector simply predicts the centers and scales of objects without any considerations of priors of the objects in the dataset, thus shows a better generalization ability.
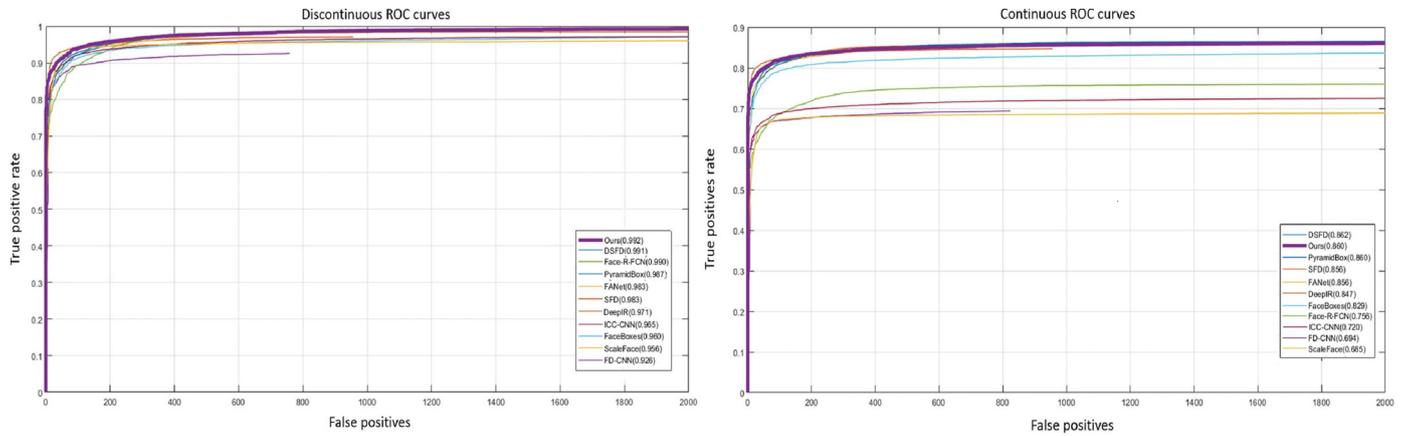
---

[5] https://github.com/liuwei16/ALFNet

[6] http://vis-www.cs.umass.edu/fddb/results.html

[7] https://flyywh.github.io/CVPRW2019LowLight/

[8] https://github.com/TencentYoutuResearch/FaceDetection-DSFD

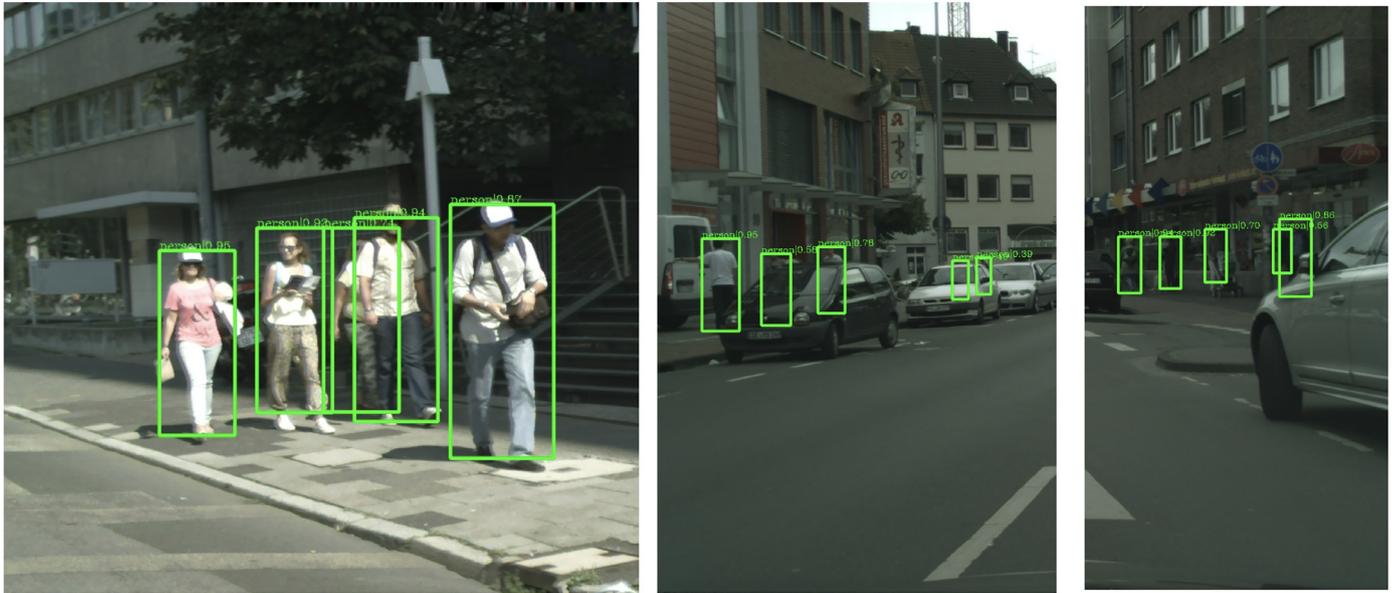**Fig. 6.** Comparisons of ROC results on the FDDB dataset.

Fig. 7. Different type of challenges present in pedestrian detection such as person-over-person occlusion (left), occlusion due to cars and objects (middle) and low illumination (right). CSP is robust enough to handle such challenging scenarios

Table 10
Performance drop with disturbances of the centers.

Disturbance (pixels)	MR^{-2}(%)	ΔMR^{-2}(%)
0	4.62	-
[0, 4]	5.68	↓ 1.06
[0, 8]	8.59	↓ 3.97

Table 11
CSP with H+W vs. CSP with H only.

Method	CSP H + W	CSP H
Easy	0.961	0.903
Medium	0.952	0.898
Hard	0.907	0.840

4.5. Further discussions

Note that CSP only requires object centers and scales for training, though generating them from bounding box or central line annotations is more feasible since centers are not always easy to annotate. Besides, the model may be puzzled on ambiguous centers during training. To demonstrate this, we also conduct an ablative experiment on Caltech, in which object centers are randomly disturbed in the range of [0,4] and [0,8] pixels during training. Results in Table 10 show that performance drops with increasing annotation noise. For Caltech, we also apply the original annotations but with inferior performance to another anchor-free detector, TLL [28]. A possible reason is that TLL includes a series of post-processing strategies in keypoint pairing. As evaluations of TLL on Caltech

with new annotations [95] are not reported in [28], comparison to TLL is given in Table 5 on the CityPersons, which shows the superiority of CSP. Therefore, the proposed method may be limited for annotations with ambiguous centers, e.g. the traditional pedestrian bounding box annotations affected by limbs. In view of this, applying CSP to generic object detection requires further improvement.

When compared with anchor-box based methods, the advantage of CSP lies in two aspects. Firstly, CSP does not require tedious configurations on anchor-boxes specifically for each dataset. Secondly, anchor-box based methods detect objects by overall classifications of each anchor-box where background information and occlusions are also included and will confuse the detector's training. However, CSP overcomes this drawback by scanning for pedes-

trian centers instead of anchor-boxes in an image, thus is more robust to occluded objects as shown in Fig. 7.

5. Conclusion

Deviating from traditional paradigms for feature detection, in this work, we argue in the favor of posing pedestrian detection as a high-level semantic feature detection task through straightforward convolutions for center and scale predictions. This enables a complete anchor.free settings and is also free from complex post-processing strategies as in recent keypoint-pairing based detectors. Consequently, the proposed CSP detector achieves the new state-of-the-art performance on two challenging pedestrian detection benchmarks, namely CityPersons and Caltech. Due to the generic architecture of the CSP detector, we further evaluate it for face detection on the most popular face detection benchmark, i.e. Wider-Face. The comparable performance to other advanced anchor-box based face detectors also shows the proposed CSP detector's competitiveness. Besides, experiments on cross-dataset evaluation for both pedestrian detection and face detection further demonstrate CSP's superior generalization ability over anchor-box based detectors. For future possibilities, it is interesting to further explore CSP's capability in general object detection. Given its superiority on cross-dataset evaluation, it is also interesting to see CSP's potential when domain adaptation techniques are further explored.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] V. Campmany, S. Silva, A. Espinosa, J.C. Moure, D. Vázquez, A.M. López, Gpu-based pedestrian detection for autonomous driving, Procedia Computer Science 80 (2016) 2377–2381.

[2] H. Hattori, V. Naresh Boddeti, K.M. Kitani, T. Kanade, Learning scene-specific pedestrian detectors without real data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3819–3827.

[3] Q. Mou, L. Wei, C. Wang, D. Luo, S. He, J. Zhang, H. Xu, C. Luo, C. Gao, Unsupervised domain-adaptive scene-specific pedestrian detection for static video surveillance, Pattern Recognition 118 (2021) 108038.

[4] X. Wei, H. Zhang, S. Liu, Y. Lu, Pedestrian detection in underground mines via parallel feature transfer network, Pattern Recognition 103 (2020) 107195.

[5] X. Zhao, Z. He, S. Zhang, D. Liang, Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification, Pattern Recognition 48 (6) (2015) 1947–1960.

[6] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1112–1121.

[7] L. Huang, X. Zhao, K. Huang, Bridging the gap between detection and tracking: A unified approach, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 3999–4009.

[8] M. Wang, J. Tighe, D. Modolo, Combining detection and tracking for human pose estimation in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11088–11096.

[9] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1367–1376.

[10] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, L. Shao, Anchor-free person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7690–7699.

[11] P. Viola, M.J. Jones, Robust real-time face detection, International journal of computer vision 57 (2) (2004) 137–154.

[12] J. Canny, A computational approach to edge detection, IEEE Transactions on pattern analysis and machine intelligence (6) (1986) 679–698.

[13] I. Sobel, Camera models and machine perception, Technical Report, Computer Science Department, Technion, 1972.

[14] S.M. Smith, J.M. Brady, Susan: a new approach to low level image processing, International journal of computer vision 23 (1) (1997) 45–78.

[15] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: European conference on computer vision, Springer, 2006, pp. 430–443.

[16] T. Lindeberg, Scale selection properties of generalized scale-space interest point detectors, Journal of Mathematical Imaging and vision 46 (2) (2013) 177–210.

[17] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[18] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image and vision computing 22 (10) (2004) 761–767.

[19] W. Shen, X. Wang, Y. Wang, X. Bai, Z. Zhang, Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3982–3991.

[20] S. Xie, Z. Tu, Holistically-nested edge detection, International Journal of Computer Vision 125 (1-3) (2017) 3–18.

[21] G. Bertasius, J. Shi, L. Torresani, Deepedge: A multi-scale bifurcated deep network for top-down contour detection, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 4380–4389.

[22] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, X. Bai, Richer convolutional features for edge detection, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, 2017, pp. 5872–5881.

[23] S. Wu, R. Laganiere, P. Payeur, Improving pedestrian detection with selective gradient self-similarity feature, Pattern Recognition 48 (8) (2015) 2364–2376.

[24] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

[25] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, arXiv preprint arXiv:1611.08050 (2016).

[26] A. Newell, Z. Huang, J. Deng, Associative embedding: End-to-end learning for joint detection and grouping, in: Advances in Neural Information Processing Systems, 2017, pp. 2277–2287.

[27] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: The European Conference on Computer Vision (ECCV), 2018.

[28] T. Song, L. Sun, D. Xie, H. Sun, S. Pu, Small-scale pedestrian detection based on topological line localization and temporal feature aggregation, in: The European Conference on Computer Vision (ECCV), 2018.

[29] W. Liu, S. Liao, W. Ren, W. Hu, Y. Yu, High-level semantic feature detection: A new perspective for pedestrian detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5187–5196.

[30] S. Zhang, R. Benenson, B. Schiele, Citypersons: A diverse dataset for pedestrian detection, arXiv preprint arXiv:1702.05693 (2017).

[31] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE transactions on pattern analysis and machine intelligence 34 (4) (2012) 743–761.

[32] S. Yang, P. Luo, C.C. Loy, X. Tang, Wider face: A face detection benchmark, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[33] W. Liu, S. Liao, W. Hu, X. Liang, X. Chen, Learning efficient single-stage pedestrian detectors by asymptotic localization fitting, in: The European Conference on Computer Vision (ECCV), 2018.

[34] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, F. Huang, Dsfd: Dual shot face detector, arXiv preprint arXiv:1810.10220 (2018).

[35] V. Jain, E. Learned-Miller, FDDB: A Benchmark for Face Detection in Unconstrained Settings, Technical Report, University of Massachusetts, Amherst, 2010.

[36] (https://vast.uccs.edu/Opensetface/).

[37] C. Wei, W. Wang, W. Yang, J. Liu, Deep retinex decomposition for low-light enhancement, British Machine Vision Conference (2018) 155.

[38] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, X. Bai, Gliding vertex on the horizontal bounding box for multi-oriented object detection, IEEE transactions on pattern analysis and machine intelligence 43 (4) (2020) 1452–1459.

[39] Y. Wang, J. Hou, X. Hou, L.-P. Chau, A self-training approach for point-supervised object detection and counting in crowds, IEEE Transactions on Image Processing 30 (2021) 2876–2887.

[40] S. Ren, K. He, R. Girshick, 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.

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

[42] Y. Qian, M. Yang, X. Zhao, C. Wang, B. Wang, Oriented spatial transformer network for pedestrian detection using fish-eye camera, IEEE Transactions on Multimedia (2019).

[43] L. Zhang, L. Lin, X. Liang, K. He, Is faster r-cnn doing well for pedestrian detection? in: European Conference on Computer Vision, Springer, 2016, pp. 443–457.

[44] Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: European Conference on Computer Vision, Springer, 2016, pp. 354–370.

[45] Z. Cai, M. Saberian, N. Vasconcelos, Learning complexity-aware cascades for pedestrian detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (9) (2019) 2195–2211.

[46] J. Zhang, L. Lin, J. Zhu, Y. Li, Y.-c. Chen, Y. Hu, S.C. Hoi, Attribute-aware pedestrian detection in a crowd, IEEE Transactions on Multimedia 23 (2020) 3085–3097.

[47] Z. Guo, W. Liao, Y. Xiao, P. Veelaert, W. Philips, Weak segmentation supervised deep neural networks for pedestrian detection, Pattern Recognition 119 (2021) 108063.

[48] I. Hasan, S. Liao, J. Li, S.U. Akram, L. Shao, Generalizable pedestrian detection: The elephant in the room, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11328–11337.

[49] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, C. Shen, Repulsion loss: Detecting pedestrians in a crowd, arXiv preprint arXiv:1711.07752 (2017).

[50] S. Zhang, L. Wen, X. Bian, Z. Lei, S.Z. Li, Occlusion-aware r-cnn: Detecting pedestrians in a crowd, in: The European Conference on Computer Vision (ECCV), 2018.

[51] C. Zhou, J. Yuan, Bi-box regression for pedestrian detection and occlusion estimation, in: The European Conference on Computer Vision (ECCV), 2018.

[52] X. Yu, Y. Gong, N. Jiang, Q. Ye, Z. Han, Scale match for tiny person detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1257–1265.

[53] J. Wang, C. Zhao, Z. Huo, Y. Qiao, H. Sima, High quality proposal feature generation for crowded pedestrian detection, Pattern Recognition 128 (2022) 108605.

[54] W. Liu, S. Liao, W. Hu, X. Liang, Y. Zhang, Improving tiny vehicle detection in complex scenes, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2018, pp. 1–6.

[55] W. Liu, S. Liao, W. Hu, Towards accurate tiny vehicle detection in complex scenes, Neurocomputing 347 (2019) 24–33.

[56] S. Liu, J. Tang, Modified deep reinforcement learning with efficient convolution feature for small target detection in vhr remote sensing imagery, ISPRS International Journal of Geo-Information 10 (3) (2021) 170.

[57] S. Mathe, A. Pirinen, C. Sminchisescu, Reinforcement learning for visual object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2894–2902.

[58] M. Najibi, B. Singh, L.S. Davis, Autofocus: Efficient multi-scale inference, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9745–9755.

[59] B. Uzkent, C. Yeh, S. Ermon, Efficient object detection in large images using deep reinforcement learning, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1824–1833.

[60] X. Zhang, L. Cheng, B. Li, H.-M. Hu, Too far to see? not really!pedestrian detection with scale-aware localization policy, IEEE transactions on image processing 27 (8) (2018) 3703–3715.

[61] M. Gao, R. Yu, A. Li, V.I. Morariu, L.S. Davis, Dynamic zoom-in network for fast object detection in large images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6926–6935.

[62] Y. Zhao, Z. Yuan, B. Chen, Accurate pedestrian detection by human pose regression, IEEE transactions on image processing 29 (2019) 1591–1605.

[63] Y. Lee, J. Park, Centermask: Real-time anchor-free instance segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13906–13915.

[64] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9627–9636.

[65] W. Liu, S. Liao, W. Hu, Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding, IEEE transactions on image processing 29 (2019) 1413–1425.

[66] C. Lin, J. Lu, G. Wang, J. Zhou, Graininess-aware deep feature learning for pedestrian detection, in: The European Conference on Computer Vision (ECCV), 2018.

[67] B. Yu, D. Tao, Anchor cascade for efficient face detection, IEEE Transactions on Image Processing 28 (5) (2018) 2490–2501.

[68] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S.Z. Li, Faceboxes: A cpu real-time face detector with high accuracy, International Journal of Central Banking (2017) 1–9.

[69] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S.Z. Li, Sfd: Single shot scale-invariant face detector, International Conference on Computer Vision (2017) 192–201.

[70] C. Zhu, R. Tao, K. Luu, M. Savvides, Seeing small faces from robust anchor's perspective, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 5127–5136.

[71] M. Najibi, P. Samangouei, R. Chellappa, L.S. Davis, Ssh: Single stage headless face detector, International Conference on Computer Vision (2017) 4885–4894.

[72] X. Tang, D.K. Du, Z. He, J. Liu, Pyramidbox: A context-assisted single shot face detector, in: The European Conference on Computer Vision (ECCV), 2018.

[73] X. Ming, F. Wei, T. Zhang, D. Chen, N. Zheng, F. Wen, Group sampling for scale invariant face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[74] L. Tychsen-Smith, L. Petersson, Denet: Scalable real-time object detection with directed sparse sampling, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 428–436.

[75] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[76] L. Huang, Y. Yang, Y. Deng, Y. Yu, Densebox: Unifying landmark localization with end to end object detection, arXiv preprint arXiv:1509.04874 (2015).

[77] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[78] E. Rosten, R. Porter, T. Drummond, Faster and better: A machine learning approach to corner detection, IEEE transactions on pattern analysis and machine intelligence 32 (1) (2010) 105–119.

[79] H. Deng, W. Zhang, E. Mortensen, T. Dietterich, L. Shapiro, Principal curvature-based region detector for object recognition, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–8.

[80] A. Kumawat, S. Panda, Feature detection and description in remote sensing images using a hybrid feature detector, Procedia computer science 132 (2018) 277–287.

[81] F. Qingyun, W. Zhaokui, Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery, Pattern Recognition (2022) 108786.

[82] F. He, Q. Li, X. Zhao, K. Huang, Temporal-adaptive sparse feature aggregation for video object detection, Pattern Recognition 127 (2022) 108587.

[83] Z. Fan, Q. Liu, Adaptive region-aware feature enhancement for object detection, Pattern Recognition 124 (2022) 108437.

[84] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.

[85] Y.-T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, arXiv preprint arXiv:1612.03144 (2016).

[86] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, S.-J. Ko, Parallel feature pyramid network for object detection, in: The European Conference on Computer Vision (ECCV), 2018.

[87] T. Kong, F. Sun, C. Tan, H. Liu, W. Huang, Deep feature pyramid reconfiguration for object detection, in: The European Conference on Computer Vision (ECCV), 2018.

[88] S. Liu, D. Huang, a. Wang, Receptive field block net for accurate and fast object detection, in: The European Conference on Computer Vision (ECCV), 2018.

[89] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A.C. Berg, Dssd: Deconvolutional single shot detector, arXiv preprint arXiv:1701.06659 (2017).

[90] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, J. Sun, Light-head r-cnn: In defense of two-stage object detector, arXiv preprint arXiv:1711.07264 (2017).

[91] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, Towards reaching human performance in pedestrian detection, IEEE transactions on pattern analysis and machine intelligence 40 (4) (2018) 973–986.

[92] Y.-T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, arXiv preprint arXiv:1708.02002 (2017).

[93] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[94] J. Mao, T. Xiao, Y. Jiang, Z. Cao, What can help pedestrian detection? in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, 2017, p. 3.

[95] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, How far are we from solving pedestrian detection? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1259–1267.

[96] C.L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: European Conference on Computer Vision, Springer, 2014, pp. 391–405.

[97] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[98] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[99] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in neural information processing systems, 2017, pp. 1195–1204.

[100] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[101] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, J. Sun, Crowdhuman: A benchmark for detecting human in a crowd, arXiv preprint arXiv:1805.00123 (2018).

Dr Wei Liu finished his PhD from National University of Defense Technology, Changsha, China. Currently working as a postDoc. His interest includes face and pedestrian detection.

Dr Irtiza Hasan finished his PhD from university of Verona, Italy. Currently he is a research associate at Inception Institute of Artificial Intelligence.

Dr Shengcai Liao is a lead scientist at Inception Institute of Artificial Intelligence.