華東師範大學
EAST CHINA NORMAL UNIVERSITY

Pedestrian Detection

—— 张致恺

Outline

- Challenges in Pedestrian Detection
- What is anchor-free?

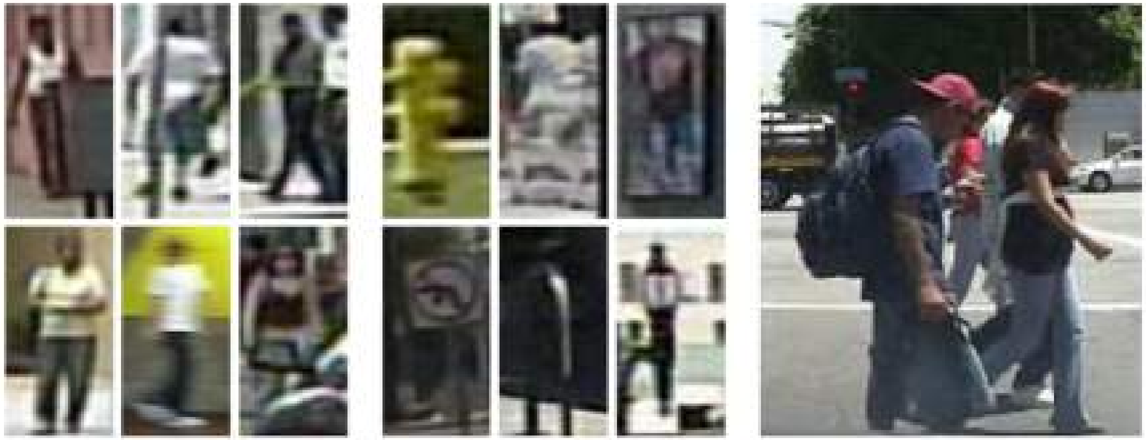
Challenges in Pedestrian Detection

Challenges in Pedestrian Detection

(a)Small pedestrian

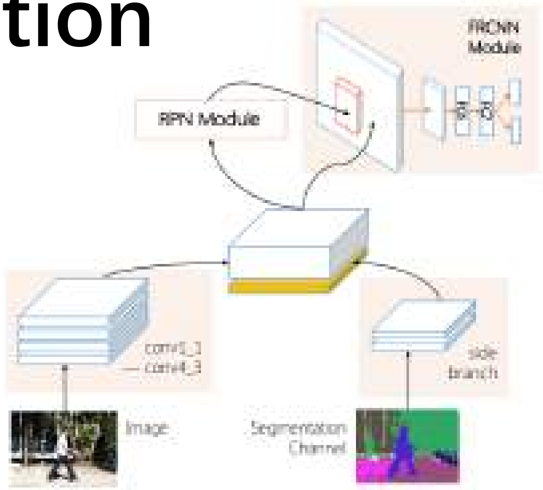
(b)Hard negatives

(c)Dense and occluded pedestrian



(d)Real-time detection

Challenges in Pedestrian Detection



(a)Small pedestrian

1. feature fusion

2. introducing extra high-resolution handcrafted features（E.g. HOG, LUV, heatmap ...）

3. ensembling detection results on multiple resolutions

Challenges in Pedestrian Detection

(b)Hard negatives

1. integration of boosted decision tree

2. semantics segmentation

3. cross-modal learning

Challenges in Pedestrian Detection

(c)Dense and occluded pedestrian

1. design new loss function by considering the repulsion of other surrounding obj

2. part detectors

3. attention mechanism

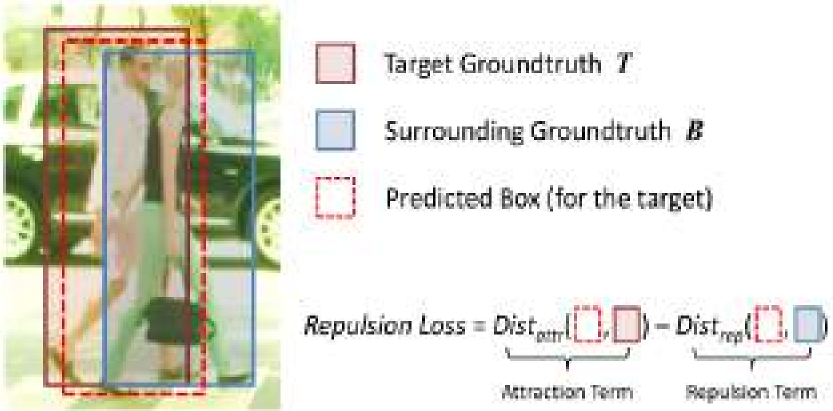

Target Groundtruth T
Surrounding Groundtruth B
Predicted Box (for the target)
Repulsion Loss = Dist$_{attr}$(,) − Dist$_{rep}$(,)
Attraction Term
Repulsion Term

Figure 1. Illustration of our proposed repulsion loss. The repulsion loss consists of two parts: the attraction term to narrow the gap between a proposal and its designated target, as well as the repulsion term to distance it from the surrounding non-target objects.

Challenges in Pedestrian Detection

(d)Real-time detection

1. simplify network structure

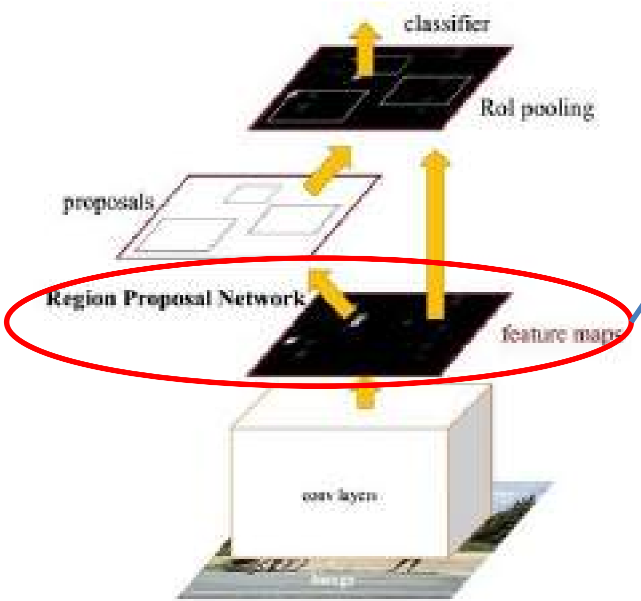
2. one-stage

3. anchor-free

What is anchor-free?

- One-stage :例如yolo, SSD, Retina-Net。
  单阶段算法直接同时实现目标分类和坐标回归。

- Two-stage : 例如Faster-RCNN。
  先提取proposals， 再对proposals做分类和坐标回归。



Anchor-free算法：不使用滑动窗口提取proposals，而改用关键点+尺度、多个关键点等方式提取出proposals

CornerNet: Detecting Objects as Paired Keypoints (ECCV2018)

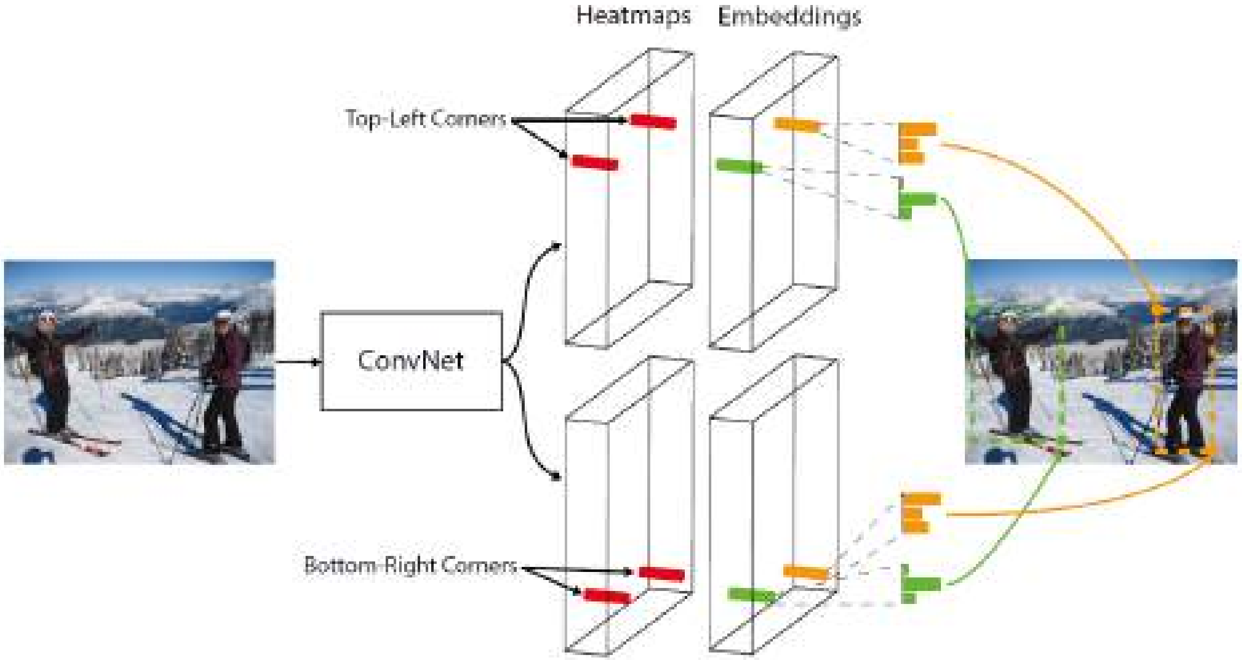


Fig. 1 We detect an object as a pair of bounding box corners grouped together. A convolutional network outputs a heatmap for all top-left corners, a heatmap for all bottom-right corners, and an embedding vector for each detected corner. The network is trained to predict similar embeddings for corners that belong to the same object.

CornerNet: Detecting Objects as Paired Keypoints (ECCV2018)

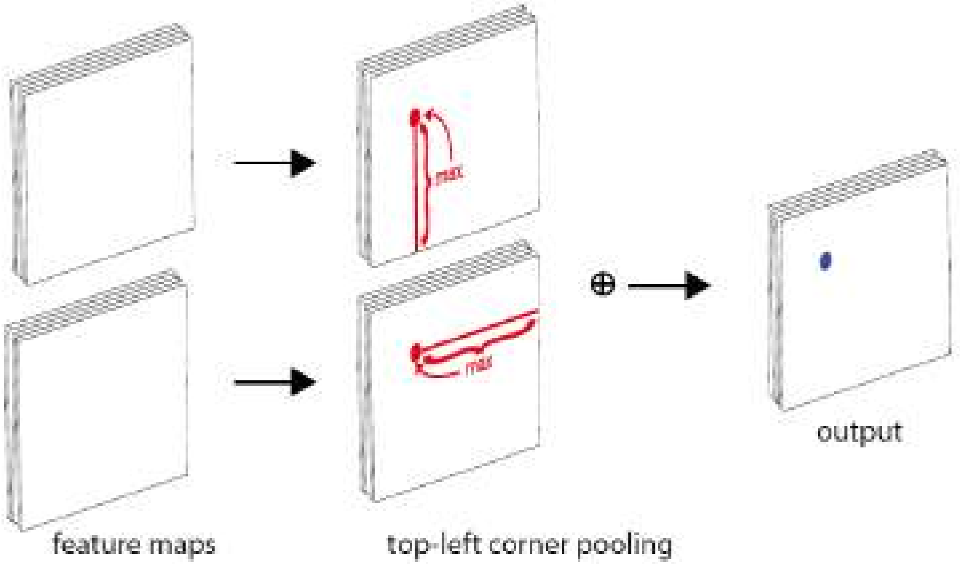
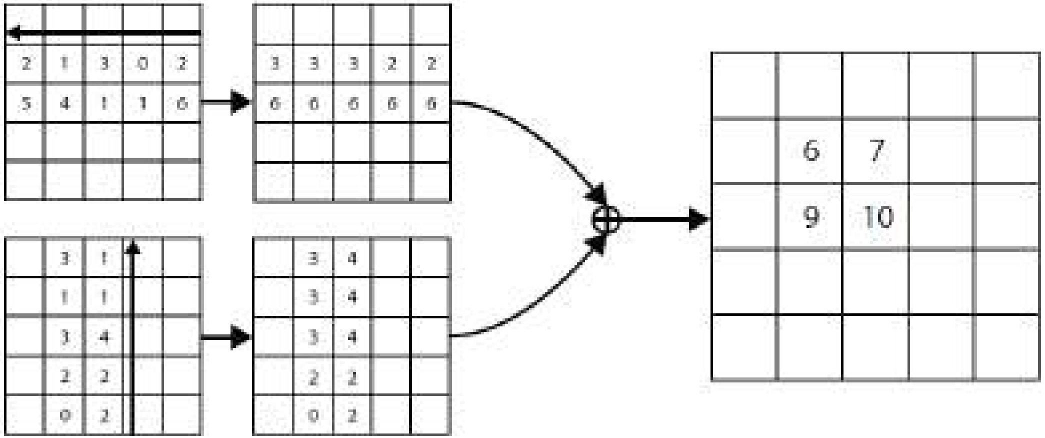


Fig. 3 Corner pooling: for each channel, we take the maximum values *(red dots)* in two directions *(red lines)*, each from a separate feature map, and add the two maximums together *(blue dot)*.

CornerNet: Detecting Objects as Paired Keypoints (ECCV2018)



2	1	3	0	2
5	4	1	1	6

3	3	3	2	2
6	6	6	6	6

	3	1		
	1	1		
	3	4		
	2	2		
	0	2		

	3	4		
	3	4		
	3	4		
	2	2		
	0	2		

	6	7		
	9	10		

Fig. 6 The top-left corner pooling layer can be implemented very efficiently. We scan from right to left for the horizontal max-pooling and from bottom to top for the vertical max-pooling. We then add two max-pooled feature maps.

$$t_{ij} = \begin{cases} \max\left(f_{t_{ij}}, t_{(i+1)j}\right) & \text{if } i < H \\ f_{t_{Hj}} & \text{otherwise} \end{cases} \qquad (6)$$

$$l_{ij} = \begin{cases} \max\left(f_{l_{ij}}, l_{i(j+1)}\right) & \text{if } j < W \\ f_{l_{iW}} & \text{otherwise} \end{cases} \qquad (7)$$

CornerNet: Detecting Objects as Paired Keypoints (ECCV2018)

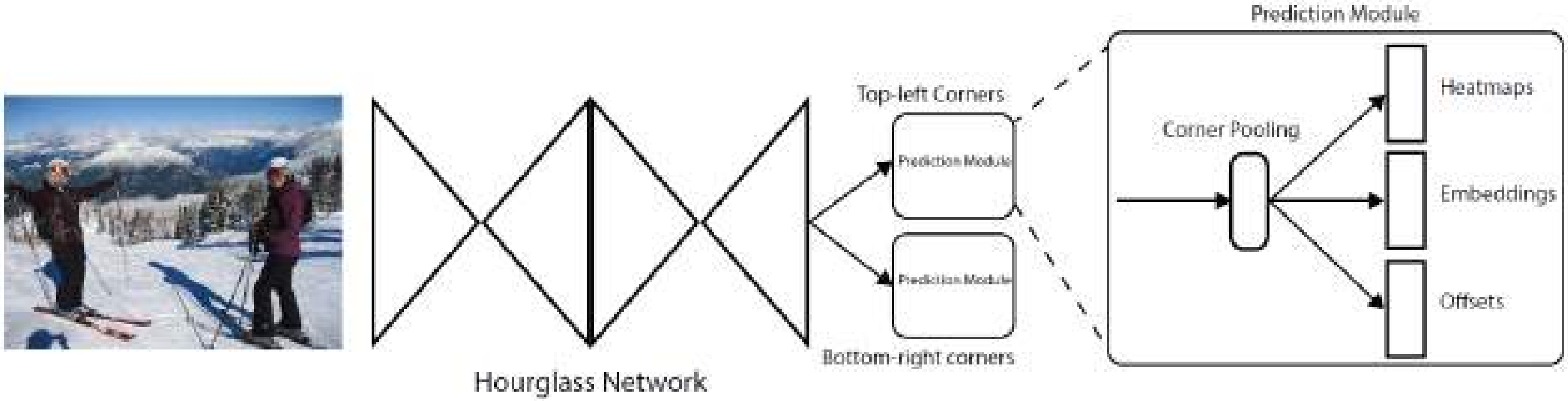


Fig. 4 Overview of CornerNet. The backbone network is followed by two prediction modules, one for the top-left corners and the other for the bottom-right corners. Using the predictions from both modules, we locate and group the corners.

CenterNet: Objects as Points (CVPR2019)

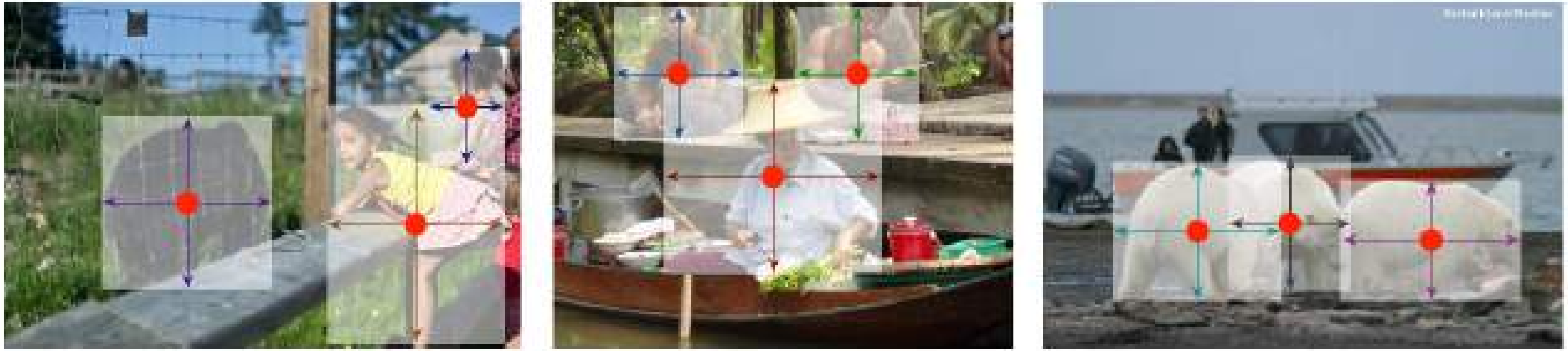


Figure 2: We model an object as the center point of its bounding box. The bounding box size and other object properties are inferred from the keypoint feature at the center. Best viewed in color.

CenterNet: Objects as Points (CVPR2019)

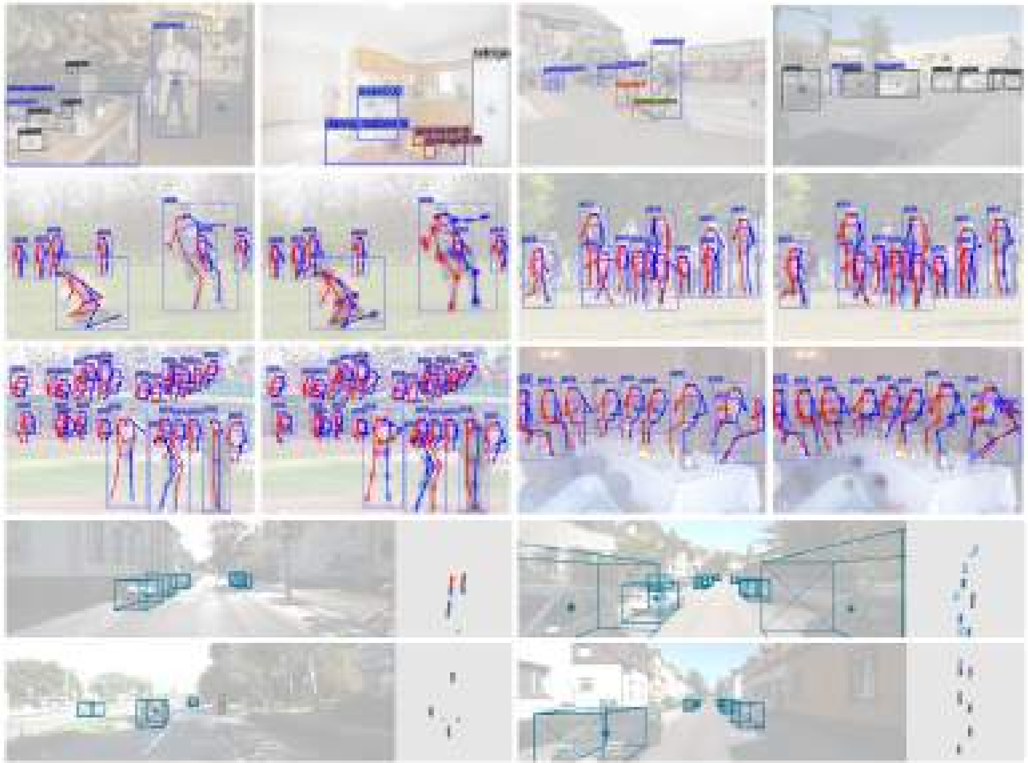


Figure 5: Qualitative results. All images were picked thematically without considering our algorithms performance. *First row:* object detection on COCO validation. *Second and third row:* Human pose estimation on COCO validation. For each pair, we show the results of center offset regression (left) and heatmap matching (right). *fourth and fifth row:* 3D bounding box estimation on KITTI validation. We show projected bounding box (left) and bird eye view map (right). The ground truth detections are shown in solid red solid box. The center heatmap and 3D boxes are shown overlaid on the original image.

CenterNet: Objects as Points (CVPR2019)

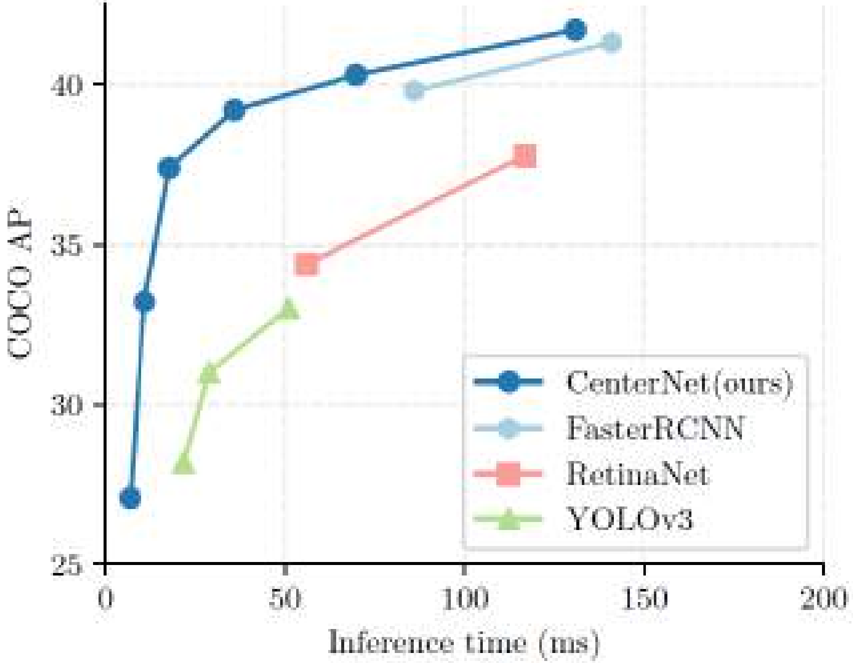


Figure 1: Speed-accuracy trade-off on COCO validation for real-time detectors. The proposed CenterNet outperforms a range of state-of-the-art algorithms.

ExtremeNet: Bottom-up Object Detection by Grouping Extreme and Center Points(CVPR2019)

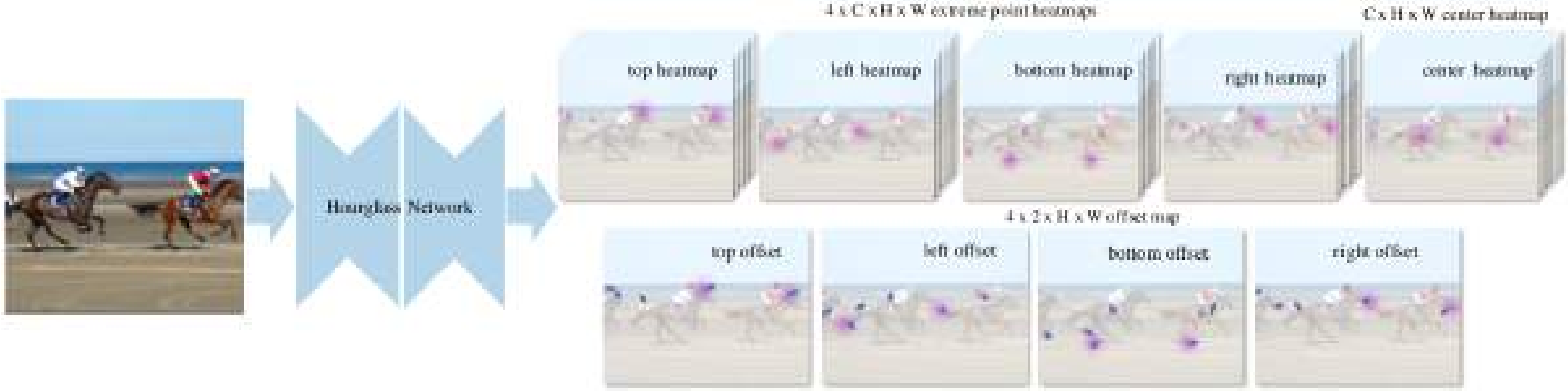


Figure 3: Illustration of our framework. Our network takes an image as input and produces four C-channel heatmaps, one C-channel heatmap, and four 2-channel category-agnostic offset map. The heatmaps are trained by weighted pixel-wise logistic regression, where the weight is used to reduce false-positive penalty near the ground truth location. And the offset map is trained with Smooth L1 loss applied at ground truth peak locations.

High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection(CVPR2019)

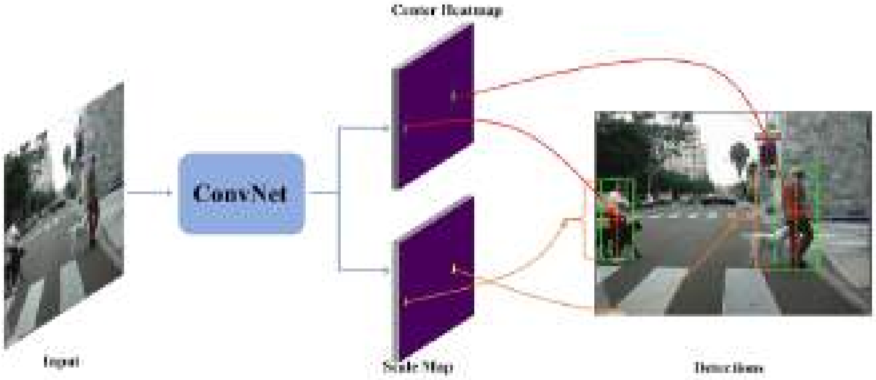


Figure 1. The overall pipeline of the proposed CSP detector. The final convolutions have two channels, one is a heatmap indicating the locations of the centers (red dots), and the other serves to predict the scales (yellow dotted lines) for each detected center.

High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection(CVPR2019)

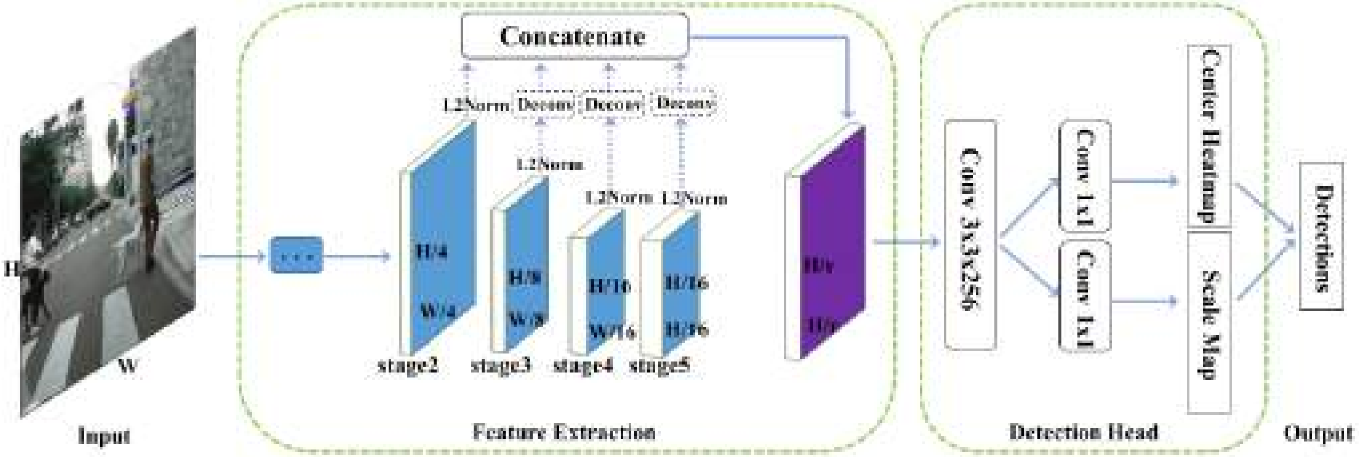
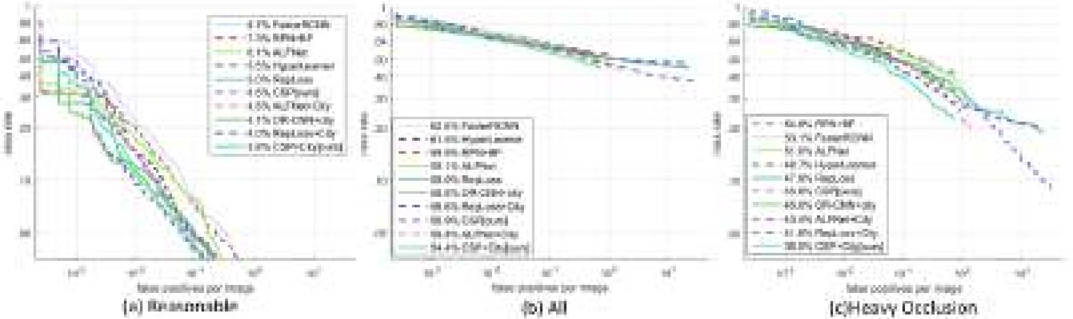


Figure 2. Overall architecture of CSP, which mainly comprises two components, i.e. the feature extraction module and the detection head. The feature extraction module concatenates feature maps of different resolutions into a single one. The detection head merely contains a 3x3 convolutional layer, followed by two prediction layers, one for the center location and the other for the corresponding scale.

High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection(CVPR2019)

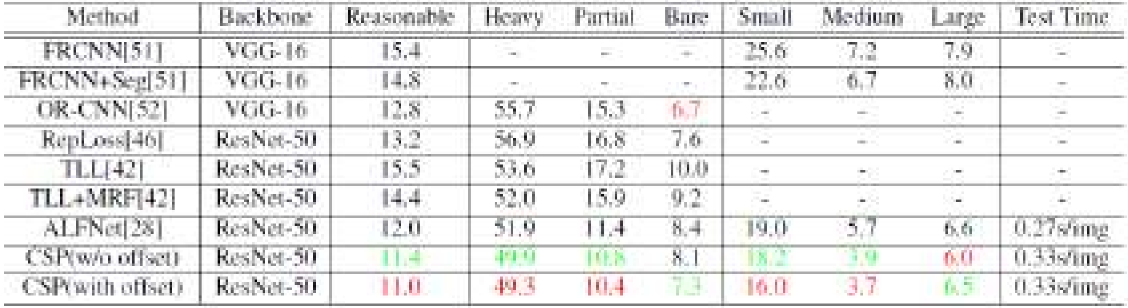


(a) Reasonable

(b) All

(c)Heavy Occlusion

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



Method	Backbone	Reasonable	Heavy	Partial	Bare	Small	Medium	Large	Test Time
FRCNN[51]	VGG-16	15.4	-	-	-	25.6	7.2	7.9	-
FRCNN+Seg[51]	VGG-16	14.8	-	-	-	22.6	6.7	8.0	-
OR-CNN[52]	VGG-16	12.8	55.7	15.3	6.7	-	-	-	-
RepLoss[46]	ResNet-50	13.2	56.9	16.8	7.6	-	-	-	-
TLL[42]	ResNet-50	15.5	53.6	17.2	10.0	-	-	-	-
TLL+MRF[42]	ResNet-50	14.4	52.0	15.9	9.2	-	-	-	-
ALFNet[28]	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

Table 5. Comparison with the state of the arts on CityPersons[51]. Results test on the original image size (1024x2048 pixels) are reported. Red and green indicate the best and second best performance.

Thank You!