Crowded Human Detection via an Anchor-pair Network

Jinguo Zhu[1], Zejian Yuan[1], Chong Zhang[2], Wanchao Chi[2], Yonggen Ling[2], Shenghao Zhang[2]
[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
lechatelia@stu.xjtu.edu.cn, yuan.ze.jian@xjtu.edu.cn
[2]Tencent, Robotics X, China
aerentzhang@gmail.com, wanchaochi@tencent.com, ylingaa@connect.ust.hk, popshzhang@pku.edu.cn

Abstract

This paper presents an anchor-pair network for crowded human detection, which can overcome and solve the difficulties caused by occlusion in crowded scenes. Specifically, we use a function-aware network structure to extract more distinctive and discriminative features for head and full-body respectively, and then a CNN module is also exploited to fuse the features by learning the correlations between head and full-body to reduce crowd errors. Meanwhile, a novel paired form for anchors, denoted as anchor-pair, is proposed to estimate the head regions and full-body regions simultaneously. Furthermore, a new ingenious Joint-NMS is introduced to perform on the detected head and full-body box pairs, which produces significant performance improvement in heavily occluded scenarios at tiny computational cost. Our anchor-pair network achieves a state-of-the-art result on the CrowdHuman dataset which reduces the $\boldsymbol{MR}^{-2}$ *to 55.43%, achieving 11.59% relative improvement over our dataset baseline.*

1. Introduction

As a key component of wide real-world applications such as automatic driving, robotics, and video surveillance, human detection has attracted increasing attention in recent years [15, 25, 2]. Although great progress has been made in recent years, crowd occlusion remains one of the most difficult challenges in human detection [2, 22, 26, 5, 23].

In real life crowded scenes, human, especially pedestrians and party crowd, often gather together and occlude each other [31, 22]. The main impact of crowd occlusion is that it severely harms the performance of human detector and significantly increases the difficulty in locating each individual accurately [8, 26, 20]. Just as shown in Figure 1, the predicted boxes in crowded scenes with occlusion by using generic detectors often shift dramatically to a neighboring individual, or cover several mutually occluded individuals. A large amount of occlusion condition in the lower parts of body regions undoubtedly increases the difficulty of separating individuals in crowd, due to their similar but not discriminating appearance shared by these people [9, 4]. To make matters worse, the occluded regions of ground truths will have a large number of invalid pixels that actually do not belong to the target individual, but the background or other individuals, which will confuse the detector [31].

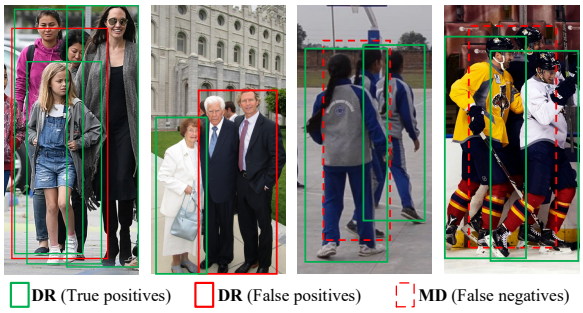

Figure 1. Some typical detection results with crowd errors in crowded scenes detected by Retinanet [11]. Detection results (DR) predicted by the detector and missing detections (MD) that are missed in the results are visualized on the input images.

The choice of a suitable threshold of NMS is still a big deal: the results output by the anchor-based CNN detector have many false positives [17, 18], and it is difficult for the vanilla greedy-NMS method based IOU of full bodies to select a suitable threshold [22]. Just as shown in Figure 1, when two persons are close enough, one of detections is thought to be a repeated detection and would be filtered out. This reveals a limitation of the current CNN-based detector using NMS to get the final detections: a higher threshold brings in more false positives while a lower threshold leads to more missed highly overlapped human [1, 12]. So how to reduce the sensitivity of the NMS threshold to adapt different population densities is also critical to improve the performance of human detector [22].

We do some statistics analysis on the *CrowdHuman* [19] dataset and the results are shown in Figure 2. The consis-

tency of the luminance region in the Figure 2(a) and Figure 2(b) proves that the head region is visible easily and thus its location regression task faces few difficulties caused by crowd occlusion. Moreover, the head regions seem to be a powerful guidance information to direct the detector to estimate the full-body regions. The discriminating features extracted from head region even will assist the features of full-body that are not characteristic enough to estimate the full-body region of each individual in a crowd, thereby reducing crowd errors.

We propose an anchor-pair network to alleviate the impact of crowd occlusion. Specifically, we use function-aware network branches to extract discriminating features from head and full-body region of human and then fuse these features in an appropriate way. Besides, the anchor-pair, shown in Figure 2(c), is designed to simultaneously predict the head regions and body regions. The network can implicitly learn the correlation between the head and body part by sharing features with this dedicated pairing design. Through these designs, we want our detector to judge whether there is a human at this location based on the head region, and then will speculate on the region of the human full-body.

We also propose Joint-NMS to solve inherent defects of Greedy-NMS in crowded scenes. To be specific, we use not the full-body part alone, but both the head and full-body predicted box pairs, to do the NMS procedure. Because head regions face less serious occlusion and can be detected successfully without too many difficulties, there is a lower overlap between the head boxes of different individuals compared with the full-body regions, while the overlap between all repeated head boxes detected from one individual can maintain a high overlap. So the Joint-NMS will make it easier to distinguish between repeated predictions and crowded true positives caused by crowding.

Our main contributions are as follows:

- An anchor-pair network is proposed to detect human in crowded scenes. The network extract more distinctive features through a function-aware network structure and then fuses features via feature fusion module in which the correlation between different regions are learned to improve the performance for detecting human. Moreover, anchor-pair is also proposed to predict the head and full-body regions simultaneously.
- A novel Joint-NMS procedure is proposed which suppresses the candidate boxes based on the IOU of paired head and body predicted boxes, which is still robust to NMS thresholds even with heavy occlusion.
- Several experiments are carried out on *CrowdHuman* dataset to demonstrate the effectiveness of our proposed methods.

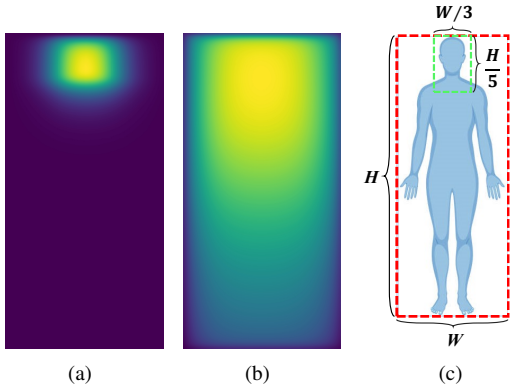

Figure 2. **The distribution of head region and full body region, and the design of our Anchor-pair.** (a) Probability maps of the possibility that the head regions appear (the brighter the color, the more likely the area is the head region); (b) Probability maps of the possibility that the regions is visible (the brighter the color, the more likely the location is to be visible); (c) The proposed anchor-pair design (the red box and the green box), which is designed based on the prior statistics in (a) and (b).

2. Related work

2.1. Part-Based Human Detectors

It's proved to be effective to handle occlusion by learning a set of part detectors which can be integrated properly to detect partially occluded human [23, 5]. The detectors of the parts which are still visible may give a high detection confidence when a human is partially occluded [29]. However, part detectors are usually learned independently so that the correlations between parts are ignored, which can reduce the reliability of the learned part detector [21, 16]. Additionally, the computational cost of applying a set of part detectors increases linearly with the number of parts. Our method firstly extracts the features of different regions through the function-aware structure, followed by he feature fusion module which fuses the features together, thus achieving the purpose of explicitly learning the correlations between different parts with minimum computational cost. In addition, this special anchor-pair can also implicitly learn this correlations to some extent.

2.2. CNN-Based Human Detectors

Recently, the CNN-based detectors [24, 9, 27, 28] show great potential in dominating the field of human detection. Zhou and Yuan [30] proposes a multi-label learning approach to jointly learn part detectors to capture partial occlusion patterns. Wang *et al*. [22] and Zhang *et al*. [26] propose a learning method to improve the robustness of NMS. Among these approaches, most of the effort mainly focuses on end-to-end mapping with deep but plain network architectures [3, 14, 7]. However, learning with pipeline CNN structures is inefficient to train, blindness to tune [4], espe-

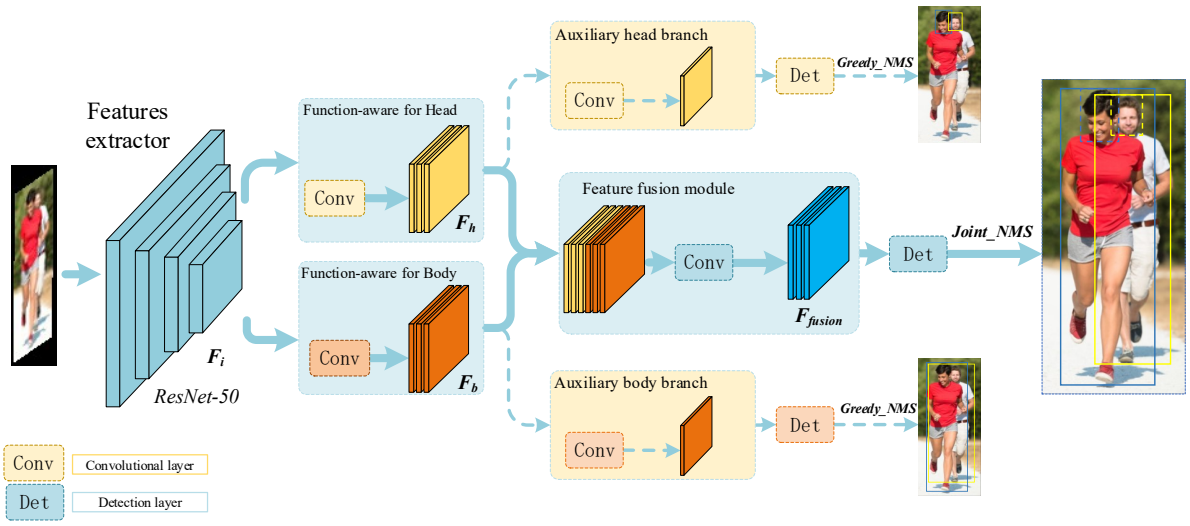

Figure 3. **An overview of our proposed approach for human detection.** The one-stage detector will first localize the head region and full-body simultaneously. And then the redundant detections will be filtered out by our Joint-NMS based on the IOU of head boxes. Auxiliary branches with dotted line only work during training.

cially for detecting human in crowded scenes .

Instead of designing a pipeline network to extract task-unclear features, we propose a function-aware network structure to extract features from input images. By adding auxiliary branches to the function-aware features extractors, the supervision information can be introduced during the network training and the function-aware network will be capable of extracting features purposefully.

3. Approach

In this section, we first introduce our architecture of anchor-pair network (APN). Then we introduce our anchor pairs and how to regress the ground truth based on them. Next, loss function and train strategy of our network are also discussed. Finally, a new joint-NMS approach is proposed for improving the robustness of the NMS threshold to adapt to different crowd densities.

3.1. Network architecture

The network architecture of our proposed CNN-based human detector is shown in Figure 3. The APN is based on the Feature Pyramid Network with ResNet-50 backbone network [11]. Unlike other general detectors, our approach predicts two bounding boxes for each human in the input image which specify his or her full-body and head region respectively.

Specifically, the *ResNet-50* network will extract the initial features F_i from the input image as a shared feature extractor. The initial feature F_i will then be fed into two parallel function-aware network branches to extract features F_h and F_b, which are specifically extracted to characterize the head-region and full-body of a human. These two targeted features will be fused later in the feature fusion module. Finally, based on the fused features F_{fusion}, the APN will regress the anchor-pair into two bounding boxes which specify the regions of head and full-body respectively.

Function-aware Branches: When one person judges how many people there are and predicts their spatial location in the crowd, he will naturally guess the number according to the human heads which are the most distinguishing part and can be easily captured. And then he can proceed from the head region to predict the full-body region depending on his common sense. In order for our network to have the same intuition as human beings, the network cannot treat all parts of the body indiscriminately.

To make the features extracted by the *Resnet-50* more focused on attention-grabbing regions, we build two function-aware branches which are responsible for estimating head-region and full-body region. In order to explicitly model these two tasks of extracting different features from different targeted parts, two auxiliary branches are used during network training to guide the learning of these two feature-extracting branches.

It is worth noting that these auxiliary branches will only work during training procedure and will be pruned after training to reduce the inference time.

Fusion of Features: Just as the result of the auxiliary body branch in Figure 3 shows, the crowd errors usually occur when a predicted box shifts dramatically to neighboring non-target ground-truth human, or bounds the union of

several overlapping human. Clearly it is not enough to only rely on the extracted features for full-body F_b to predict the full-body regions in crowded scenes. We speculate that this is because when extracting features for full-body, the network may pay more attention to whether someone exists or not, and does not care about the boundaries between individuals.

It indicates that to improve the detection performance, more discriminative features are needed to represent the density of the crowd. So we design a feature fusion module to fuse the features of head-regions and full-body regions. With the help of features of head-regions F_h, the location performance of human based on the features of fusion F_{fusion} is greatly refined. By this way, the prior, that the head region can be captured easily at a glance and can guide the detector to determine whether a person exists or not and speculate the location of full bodies, is explicitly integrated into the APN.

3.2. Anchor-pair

Unlike other anchor-based one-stage detectors, our detector will regress two types of bounding boxes, one for the head region and the other for the full-body region. It seems straightforward that using one anchor to regress one ground-truth with two bounding boxes which is similar in SSD [13] and YOLO [17]. However, the large difference between position and size between head ground truth and the anchor which is suitable for the full-body ground truth will make the regression result for locating head regions be less satisfying.

Inspired by part detectors [23, 5], it is advisable to use two anchors with different size and location to regress head region and the full-body region. But how to associate the detection results of these two types of anchors to form a human individual is still a problem. Instead of learning the association between the detections of head region and full-body region, We simply embed this association directly into the anchor-pair we will discuss later. By this way, the APN can learn the correlations between parts implicitly and uses the correlations to refine the detection results.

Anchor-pair Parameterization: Based on the prior knowledge of the position and the size of the human head relative to the body, we propose anchor-pair as shown in Figure 2(c) to detect human. One of the paired anchors is responsible for the regression of head region denoted as P_H (the green box in 2(c)), and the other is responsible for the body regression denoted as P_B (the red box in 2(c)). So all anchor-pairs placed on the feature maps can be denoted as *N* pairs $\left\{(P_H^i, P_B^i)\right\}_{i=1,\dots,N}$, where $P_H^i = \left(P_{H_x}^i, P_{H_y}^i, P_{H_w}^i, P_{H_h}^i\right)$ specifies the pixel coordinates of the center of *i*-th P_H together with the *i*-th P_H's width and height, and $P_B^i = \left(P_{B_x}^i, P_{B_y}^i, P_{B_w}^i, P_{B_h}^i\right)$ specifies the i*th* P_B in the same way.

To avoid the designing complexity, internal restriction for the anchor-pair are set as shown below, which is based on the position and size of the head region relative to the full-body region:

$$\left(P_{B_x}, P_{B_y}, P_{B_w}, P_{B_h}\right) = \left(P_{H_x}, P_{H_y} + 2P_{H_h}, 3P_{H_w}, 5P_{H_h}\right).$$

This restriction has some limitations inevitably due to the different poses and occlusion conditions. Based on the plain yet ingenious design, however, the regression procedure of network can learn the positional offset and size scaling of every anchor in one pair relative to the ground truth to adapt to different poses and occlusion conditions.

Output Definition: Consistent with the denotion of anchor-pair, the ground truth of our task should also be a set of *M* bounding-box pairs $\left\{(G_H^i, G_B^i)\right\}_{i=1,\dots,M}$. More precisely, the two bounding boxes specify the location information of the head region and the full-body region respectively. Our goal is to learn a transformation that maps a human anchor-pair (P_H, P_B) to a ground-truth (G_H, G_B).

Inspired by the idea of parameterized transformation [6] that to learn the scale-invariant translation of the center and the log-space translation of size of the anchors, we parameterize the regression targets for the head region and full-body as $\overline{h} = \left(\overline{h^x}, \overline{h^y}, \overline{h^w}, \overline{h^h}\right)$ and $\overline{b} = \left(\overline{b^x}, \overline{b^y}, \overline{b^w}, \overline{b^h}\right)$. In detail, we define $\overline{h}$ as

$$\begin{aligned} \overline{h^x} &= \frac{G_{H_x} - P_{H_x}}{P_{H_w}}, \quad \overline{h^y} = \frac{G_{H_y} - P_{H_y}}{P_{H_h}} \\ \overline{h^w} &= \log(\frac{G_{H_w}}{P_{H_w}}), \quad \overline{h^h} = \log(\frac{G_{H_h}}{P_{H_h}}) \end{aligned} \tag{1}$$

Similarly, the targets for full-body $\overline{b}$ is also defined as:

$$\begin{aligned} \overline{b^x} &= \frac{G_{B_x} - P_{B_x}}{P_{B_w}}, \quad \overline{b^y} = \frac{G_{B_y} - P_{B_y}}{P_{B_h}} \\ \overline{b^w} &= \log(\frac{G_{B_w}}{P_{B_w}}), \quad \overline{b^h} = \log(\frac{G_{B_h}}{P_{B_h}}) \end{aligned} \tag{2}$$

3.3. Training

Loss function. Based on the fusion features F_{fusion}, our APN will regress the classification predictions c_i, full-body predicted boxes p_b and head predicted boxes p_h simultaneously. Let $x_{ij} = \{0, 1\}$ be an indicator for matching the *i*-th anchor-pair to the *j*-th ground truth of a human instance. So we get a multi-task loss from this branch:

$$\begin{aligned} L_f =& L_{class}\left(x_{ij}, c_i, \overline{c_j}\right) + \\ & \lambda_b L_{loc1}\left(x_{ij}, p_b, \overline{b}\right) + \lambda_h L_{loc2}\left(x_{ij}, p_h, \overline{h}\right) \end{aligned} \tag{3}$$

where L_{class} is the classification loss, and L_{loc1} and L_{loc2} are the bounding box regression loss for the full-body estimation and head region estimation respectively. L_{class} is

focal loss which can address foreground-background class imbalance [11]. For L_{loc1} and L_{loc2}, we use the smooth $L1$ loss proposed in Fast R-CNN [6]. These two hyper parameters used to balance the three loss are set as $\lambda_b = \lambda_h = 1$ empirically.

Note that two auxiliary branches also have losses L_b and L_h for their own task during training, which are the same as loss in [11]. Their losses are added to the total loss L_{total} with discount weights α_1 and α_2 respectively (the losses of the auxiliary branches are weights by coefficients). We learn this model optimal parameters by minimizing the following total loss during the training procedure:

$$L_{total} = \alpha_0 L_f + \alpha_1 L_b + \alpha_2 L_h$$

Multi-step training strategy. Training the network with a fixed L_{total} from scratch may make the network not adjusted according to the original intention we envisioned. So we develop a multi-step training strategy to optimize our network. In the first step, We set $\alpha_0 = \alpha_2 = 0$ while $\alpha_1 = 1$. So the network can be fine-tuned end-to-end for the estimation of full-body regions. In the second step, the network are guided to extract more features from head regions by setting $\alpha_0 = \alpha_1 = 0$ while $\alpha_2 = 1$. Finally, We train the network jointly by reducing these two weights α_1 and α_2 gradually during the training process. Specifically, $\alpha_1 = \alpha_2 = 0.5$ during the first 10 epochs. Then $\alpha_1 = \alpha_2 = 0.1$ during the next 20 epochs. After 30 epochs, $\alpha_1 = \alpha_2 = 0.0$ and these two auxiliary branches will no longer work.

3.4. Joint-NMS

Traditional Greedy-NMS starts with a list of detection boxes $\mathcal{B}$ for full-body with scores $\mathcal{S}$. After selecting the detection with the maximum score $\mathcal{M}$, append it to the set of final detections $\mathcal{F}$. Additionally, it removes $\mathcal{M}$ and any box which has IOU with $\mathcal{M}$ greater than a threshold N_t from the set $\mathcal{B}$. This process is repeated for the remaining detections until the $\mathcal{B}$ is empty. However, Greedy-NMS method is sensitive to the threshold N_t: a higher threshold brings in more false positives while a lower threshold leads to more missing highly overlapped objects.

As shown in Figure 2, the occlusion of head regions is far less serious than that of full bodies in the crowded scenes. But if the performance of detection is good enough, the head regions of the repeated detections of one human instance will have a high overlap. That is to say, the overlap between head boxes can be used as a cue to distinguish between occlusions and repeated predictions. Using this intuition, we propose Joint-NMS to reduce the sensitivity to the threshold. This Joint-NMS algorithm is shown in Algorithm 1, whose main improvement is how to suppress the remaining detections. The Joint-NMS suppresses the candidate detections whose head boxes have IOU with the currently most credible head region $\mathcal{D}$ greater than a threshold N_{t1} or whose full-body boxes have IOU with the currently most credible full-body region $\mathcal{M}$ greater than another threshold N_{t2}. The main difference between Joint-NMS and previous greedy-NMS is that the former utilizes the complementary information from the head boxes, while the greedy-NMS does not use the head boxes at all.

Note that Joint-NMS is also greedy-NMS essentially, and their iterative process is similar, in which a little additional computational cost is introduced. However, Joint-NMS is a *generalized* version of NMS and can reduce the sensitivity to the IOU threshold, thus maintaining the excellent performance of the human detector.

Algorithm 1: Joint-NMS

Input: $\mathcal{B} = \{b_1, ..., b_N\}, \mathcal{H} = \{h_1, ..., h_N\}$,
$\mathcal{S} = \{s_1, ..., s_N\}, N_{t1}, N_{t2}$
$\mathcal{B}$ is the list of initial boxes for full-body regions
$\mathcal{H}$ is the list of initial boxes for head regions
$\mathcal{S}$ contains corresponds detection scores
N_{t1}, N_{t2} are the *IOU* thresholds
begin
  $\mathcal{F} \leftarrow \{\}$
  while $\mathcal{B} \neq$ *empty* **do**
    $m \leftarrow argmax\,(\mathcal{S})$
    $\mathcal{M} \leftarrow b_m; \mathcal{D} \leftarrow h_m$
    $\mathcal{F} \leftarrow \mathcal{F} \cup \mathcal{M}; \mathcal{B} \leftarrow \mathcal{B} - \mathcal{M}; \mathcal{H} \leftarrow \mathcal{H} - \mathcal{D}$
    for h_i *in* $\mathcal{H}$ **do**
      if $iou(\mathcal{D}, h_i) \geqslant N_{t1}$ *or*
       $iou(\mathcal{M}, b_i) \geqslant N_{t2}$ **then**
        $\mathcal{B} \leftarrow \mathcal{B} - b_i; \mathcal{H} \leftarrow \mathcal{H} - h_i; \mathcal{S} \leftarrow \mathcal{S} - s_i$
      end
    end
  end
  return $\mathcal{F}, \mathcal{S}$
end

4. Experiments

4.1. Dataset and Evaluation metric

Recently, *CrowdHuman* [19] has been released specifically for human detection in a crowd. It collects 15000, 4370 and 5000 images from the Internet for training, validation and testing respectively. There are totally 470*k* individual persons in the train and validation subsets, and this dataset contains about 22.6 persons on average per image as well as 2.4 pairwise crowd instances (density higher than 0.5). More importantly, CrowdHuman provides head region bounding-box annotation along with full-body bounding-box annotation for each person, which can be used to train our model to detect the head region and full-body simultaneously.

Table 1. **Selection thresholds of Joint-NMS by adjusting N_{t1} and N_{t2}.** The results are obtained on *CrowdHuman* validation set.

N_{t1}	0.1	0.2	0.3	0.4	0.5
N_{t2}	1	1	1	1	1
MR^{-2}	57.23	56.95	57.02	57.54	58.27
N_{t1}	1	1	1	1	1
N_{t2}	0.4	0.5	0.6	0.7	0.8
MR^{-2}	62.35	61.37	63.52	64.15	65.72
N_{t1}	0.1	0.2	0.2	0.2	0.3
N_{t2}	0.7	0.6	0.7	0.8	0.7
MR^{-2}	56.21	56.03	55.43	55.95	56.09

The log miss rate (MR) average over false positive per-image (FPPI) range of $\left[10^{-2}, 10^{0}\right]$, denoted as MR^{-2}, is used to evaluate the detection performance (lower is better). Additionally, average precision (AP) and recall of the models we train are also shown in the results.

4.2. Implementation Details

We use the same setting of anchor scales as [11]. Considering the human body shape, the anchors' ratios are modified as $\{1:1, 2:1, 3:1\}$. For the input images, we resize them so that their short edge is at 800 pixels while the long edge should be no more than 1400 pixels at the same time.

To make more use of the guidance information of the head region, we choose a more strict matching strategy between anchor-pairs and ground truth. An anchor-pair P is matched to human individual G if it aligns well with G both in head region and body region. Specifically, their difference contributes to the loss if they satisfy

$$IOU\left(P_H, G_H\right) \geq 0.2 \quad and \quad IOU\left(P_B, G_B\right) \geq 0.5,$$

where IOU is the intersection over union of two regions.

We optimize our networks using Stochastic Gradient Descent (SGD) with 0.9 momentum and 0.0005 weight decay. For fair comparison, we set the batch size to 8 on 4 RTX 2080Ti GPUs for all the experiments and train models for 60 epochs, with the base learning rate set to 0.02 and decreased by a factor of 10 after the first 15, 30 and 45 epochs. Multi-scale training/testing are not applied to ensure fair comparisons.

4.3. Ablation experiments

In short, to solve the problem of detection difficulties caused by occlusion in crowded scenarios, we propose three methods: using two function-aware branches to extract features; using anchor-pair to regress human directly from the fusion features; and using Joint-NMS to suppress the false positives. First, we construct our baseline detector (*model_v0* in Table 2), which is based on the RetinaNet [11] to predict full-bodies directly without the assistance of estimation of head regions (This model can be considered to be output from auxiliary body branch). Then, we run several ablation experiments to analyze these methods and discuss their contributions in detail.

4.3.1 The effectiveness of function-aware branches

On the basis of the baseline model, we add an extra branch for regressing the head regions in parallel with the existing branch for regressing the full-body regions, while the feature fusion module and Joint-NMS are still not adopted. In this model, the detection results can be thought as the outputs from two auxiliary branches. This model with multiple regression tasks by using two parallel branches, denoted as *model_v1* in Table 2, outperforms the baseline with an improvement of 2.94 MR^{-2}. Although there is no feature fusion module to help the head regions to refine the features of full-body, the guidance information from the head regions can still exploit the shared feature extractor *ResNet-50* to learn the correlation between different parts.

This demonstrates that head parts, as the most recognizable regions of human, can be well recognized even in the crowd and can help to estimate the full-body. Through the network design of two regression tasks, our model implicitly encodes this capability by improving the shared *ResNet* extractor's capability of features extraction.

4.3.2 The effectiveness of feature fusion

Similarly, we validate the effectiveness of fusing the features F_h and F_b through the feature fusion module. And the output from the fusion features F_{fusion} with greedy-NMS is thought to be the detection results. As a result, this model *model_v2* improves the detection performance with a reduction of 1.33 MR^{-2}, which is shown in Table 2.

This result suggests that the features of head region can refine the full-body location based on guidance information from head features as well as body posture of a human. In other words, the fusion module guides the network to estimate the full-body regions from the characteristic head regions with powerful features, rather than the body torso where the features are not discriminating or even missing.

4.3.3 Joint-NMS

The Balance of Joint-NMS' Thresholds: Our Joint-NMS introduces an extra threshold by using both the IOU of head boxes and full-body boxes. But how to find the most appropriate set of values remains to be discussed. To better investigate the role of the Joint-NMS in post-processing, we experiment with different settings of N_{t1} and N_{t2}, reported in the Table 1.

When set N_{t2} to 1.0, the Joint-NMS screens out the redundant detections only based on the IOU of head-region

(a) GT (b) w/o NMS (c) H-NMS$_{T=0.3\sim0.7}$

(d) B-NMS$_{T=0.3}$ (e) B-NMS$_{T=0.7}$ (f) J-NMS$_{T=(0.2,0.7)}$

Figure 4. **Illustration of different NMS results.** (a) the detection ground truth of the sample image; (b) the raw detections without NMS; (c) results using head boxes to do NMS with different IOU threshold from 0.3 to 0.7; (d) results using body boxes to do NMS with threshold 0.3; (e) results using body boxes to do NMS with threshold 0.7; (f) results using Joint-NMS with threshold $N_{t1} = 0.2$ and $N_{t2} = 0.7$. The yellow box shows the missing human, and the red one highlights the false positive.

boxes. Under such a condition, APN yields the best performance of 56.95 MR^{-2}. When set N_{t1} to 1.0, the Joint-NMS is equivalent to Greedy-NMS. APN with Greedy-NMS yields the best performance of 61.37 MR^{-2} When $N_{t2} = 0.5$.

By comparing the performance of above experiments with control variable, we can find that using the human head boxes is more effective when performing NMS than using the body boxes. This confirms our original intention, the head regions faces less occlusion caused by crowding.

By adjusting there two thresholds, the APN achieves the improved results with 56.88 MR^{-2}. It is undeniable that most of the effectiveness of the Joint-NMS is due to the predicted boxes of head regions. In the latter experiments, we set $N_{t1} = 0.2$ and $N_{t1} = 0.7$, if not specifically pointed out.

The Advantage of Joint-NMS: To evaluate our proposed Joint-NMS, we first use the Joint-NMS to replace the Greedy-NMS in *model_v1* to form *model_v3*. As expected, the model with Joint-NMS the performance of the detector and reduces the MR^{-2} by 2.78.

In the final version model (*model_v4*) with all three methods, we use this special Joint-NMS to suppress repetitive detections based on the *model_v2*. The *model_v2* and *model_v4* use the same trained network and final learning weights, except the different NMS are used in the post-processing procedure when testing. In the end, the *model_v4* achieves 55.43 MR^{-2}, which is an absolute 7.27 point improvement over our baseline.

In order to demonstrate the superiority of the head boxes for NMS, some visualization results using NMS with different region boxes are shown in Figure 4. The large numbers of false positives produced by this model near the ground truth are expected to be filtered out by the traditional greedy-NMS. As the Figure 4 (c) shown, NMS only with head boxes is less sensitive to the threshold than NMS only with body boxes, and is easier to get satisfying results even the threshold changes. That is because that the head regions are not easily occluded and there is a lower overlap between the head boxes of different individuals compared with the full bodies, while the overlap between all head boxes detected from one person can maintain a high overlap. However, there may be some false positives of head detections just as shown in Figure 4 (c). This is due to the head as a separate body part, especially when its size is small, lacks features sufficient to distinguish it from other items with similar shape. When the body part and the head region are considered simultaneously, as shown in 4 (f), these two problems are solved, which can guarantee missing fewer highly overlapped ground truths while ensure less false positive predictions.

4.4. Comparisons with state-of-the-art results

For fair comparison, we use the two evaluation results on the CrowdHuman validation set by using one-stage detectors as reference models, which are both based on the Feature Pyramid Network (FPN) [10] with a ResNet-50 backbone network. It is worth noting that $\{1:1, 1.5:1, 2:1, 2.5:1, 3:1\}$ 5 anchors ratios are used in these two reference networks to accommodate the more complex body shape [19, 12], instead of $\{1:1, 2:1, 3:1\}$ only three ratios we use in our network to make the network more concise.

In Table 2, our baseline $model\_v0$ achieves comparable results as shao *et al*. [19] does. The slight difference in accuracy may be due to the difference in the deep learning framework, the number of anchor ratios and the training iterations.

The APN $model\_v4$ in Table 2 is the final version of our proposed network that integrates all the three method we discussed before. This model significantly reduces the MR^{-2} to 55.43, achieving a $\sim 11.59\%$ relative improvement over the baseline model. The $model\_v4$ outperforms the baseline network and these two reference models with great margin, which verifies the effectiveness of the head region assistance to address the crowd occlusion problem.

In addition, we also show some visual results of *model_v4* and baseline *model_v0* for comparison. As shown in Figure 5, the APN can achieve better detection results. Although blurring or even missing of the features of the full-body region due to crowd occlusion may cause location error of individual easily, the location regression of full-body

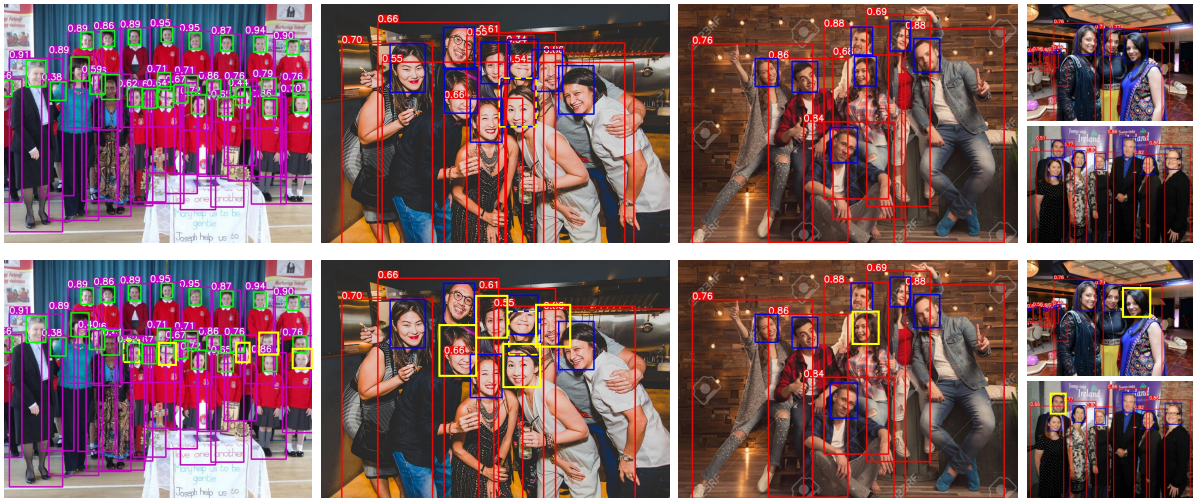

Figure 5. **Visual comparisons of the human detection results.** Top row: our APN *model_v4*; Bottom row: baseline network. Yellow solid boxes are missing objects and yellow dotted boxes are false positives.

Table 2. **Human detection results on the *CrowdHuman* validation set.** The performance of detectors is compared mainly by MR^{-2} (lower is better). Several ablation experiments are run to validate the effectiveness of our proposed methods: adopting a function-aware branches to extract features (FWB), fusing features by feature fusion module(FFM) and using Joint-NMS for suppression (J-NMS).

methods	FWB	FFM	J-NMS	MR^{-2}	Recall	AP
Shao *et al*. [19]				63.33	93.80	80.83
Liu *et al*. [12]				63.03	**94.77**	79.67
APN *model_v0*				62.70	93.33	79.83
APN *model_v1*	√			59.76	93.80	81.84
APN *model_v2*	√	√		61.37	93.55	81.11
APN *model_v3*	√		√	56.98	94.12	81.93
APN *model_v4*	√	√	√	**55.43**	94.66	**82.47**

can be greatly improved and the false positives can even be corrected because of the fusion of discriminating head-region features. In addition, the heads are rarely occluded in crowded scenes with heavy occlusion. So the overlap rate of heads of different individuals is very low. At the same time, the overlap rate of head regions from the same individual is very high when detected repeatedly. As a result, Joint-NMS keeps more crowded true positives and still screens out false positives at the same time. This important effect is in accordance with our intention that APN is specifically designed to address the occlusion problem.

4.5. Inference Latency

We also evaluate the inference time of our proposed APN on a single GTX-2080 Ti GPU. Our APN *model_v4* runs in 157 ms per image, while the baseline model runs 142 ms. Our detecter only increases **15ms** inference time, while improving the detection performance with great margin. The extra latency is mainly spent in the fusion module and relatively complex NMS procedure, which is worthwhile due to their huge gains.

5. Conclusion

In this paper, we propose a novel anchor-pair network to detect human with crowd occlusion. To utilize the correlations between different parts, we use function-aware network structure to extract more powerful features from head and body regions, which then are fused by feature fusion module. Moreover, anchor-pair is designed to localize the full-body and head regions of human simultaneously at tiny computational cost. In addition, we present a new Joint-NMS method to better suppress candidate detections, which is more robust to NMS thresholds. Extensive experiments on the *CrowdHuman* dataset demonstrate the effectiveness of our methods.

Acknowledgments: This work was supported by the National Key R&D Program of China (No.2016YFB1001001), the National Natural Science Foundation of China (No.91648121, No.61976170, No.61573280), and Tencent Robotics X Lab Rhino-Bird Joint Research Program (No.201902, No.201903).

References

[1] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, pages 5561–5569, 2017.

[2] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300:17 – 33, 2018.

[3] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, pages 354–370. Springer, 2016.

[4] X. Du, M. El-Khamy, J. Lee, and L. Davis. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. In *WACV*, pages 953–961. IEEE, 2017.

[5] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR*, pages 990–997. IEEE, 2010.

[6] R. Girshick. Fast r-cnn. In *CVPR*, pages 1440–1448, 2015.

[7] P. S. R. Kishore, S. Das, P. S. Mukherjee, and U. Bhattacharya. Cluenet: A deep framework for occluded pedestrian pose estimation. 2019.

[8] C. Li, D. Song, R. Tong, and M. Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

[9] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20(4):985–996, 2017.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[12] S. Liu, D. Huang, and Y. Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *CVPR*, pages 6459–6468, 2019.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[14] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *CVPR*, pages 5187–5196, 2019.

[15] D. T. Nguyen, W. Li, and P. O. Ogunbona. Human detection from images and videos: A survey. *Pattern Recognition*, 51:148 – 175, 2016.

[16] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, pages 2056–2063, 2013.

[17] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[18] 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*, pages 91–99, 2015.

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

[20] 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.

[21] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *ICCV*, pages 1904–1912, 2015.

[22] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen. Repulsion loss: Detecting pedestrians in a crowd. In *CVPR*, pages 7774–7783, 2018.

[23] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, volume 1, pages 90–97. IEEE, 2005.

[24] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, pages 443–457. Springer, 2016.

[25] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *CVPR*, pages 1259–1267, 2016.

[26] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *ECCV*, pages 637–653, 2018.

[27] Y. Zhao, Z. Yuan, and B. Chen. Training cascade compact cnn with region-iou for accurate pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[28] Y. Zhao, Z. Yuan, H. Zhang, and S. F. Innovation. Joint holistic and partial cnn for pedestrian detection. In *BMVC*, page 81, 2018.

[29] C. Zhou and J. Yuan. Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In *Asian Conference on Computer Vision*, pages 305–320. Springer, 2016.

[30] C. Zhou and J. Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *ICCV*, pages 3486–3495, 2017.

[31] C. Zhou and J. Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *ECCV*, pages 135–151, 2018.