Body-Face Joint Detection via Embedding and Head Hook

Junfeng Wan*† Jiangfan Deng* Xiaosong Qiu Feng Zhou
Algorithm Research, Aibee Inc.
{jfwan, jfdeng, xsqiu, fzhou}@aibee.com

Abstract

Detecting pedestrians and their associated faces jointly is a challenging task. On one hand, body or face could be absent because of occlusion or non-frontal human pose. On the other hand, the association becomes difficult or even miss-leading in crowded scenes due to the lack of strong correlational evidence. This paper proposes Body-Face Joint (BFJ) detector, a novel framework for detecting bodies and their faces with accurate correspondance. We follow the classical multi-class detector design by detecting body and face in parallel but with two key contributions. First, we propose an Embedding Matching Loss (EML) to learn an associative embedding for matching body and face of the same person. Second, we introduce a novel concept, "head hook", to bridge the gap of matching body and faces spatially. With the new semantical and geometrical sources of information, BFJ greatly reduces the difficulty of detecting body and face in pairs. Since the problem is unexplored yet, we design a new metric named log-average miss matching rate (mMR^{-2}) to evaluate the association performance and extend the CrowdHuman and CityPersons benchmarks by annotating each face box. Experiments show that our BFJ detector can maintain state-of-the-art performance in pedestrian detection on both one-stage and two-stage structures while greatly outperform various body-face association strategies. Code will be available at: https://github.com/AibeeDetect/BFJDet.

1. Introduction

Pedestrian detection has been a long-standing topic in the field of computer vision. Accurate localization of individuals in the scene can effectively facilitate the downstream process such as recognition, re-identification and tracking. By the development of deep learning, methods based on convolutional neural networks [20] have dramatically improved the detecting performance on both general objects like MS-COCO [23] and pedestrians like Crowd-Human [30], leading to the ability of application-level use such as video surveillance and identity authentication. For a pedestrian, the face is the most semantically discriminative part. So finding out body and face jointly becomes meaningful. However, for the joint detection, there exist three main obstacles. First, all persons' bodies and faces are not perfectly visible and not always yielding one-to-one correspondance. As shown in Fig. 1a, the man in green box is side to the camera while the man in blue box is back to the camera, both of their faces can not be seen. This phenomenon makes the simple method of regressing body and face boxes jointly from one proposal [4] impractical. In Fig. 1b, the three people have visually clear faces while their bodies are hardly observed due to the heavy occlusion. This anisotropy limits another intuitive pipeline of first detecting bodies and then finding face from them (we depict it as *cas-*

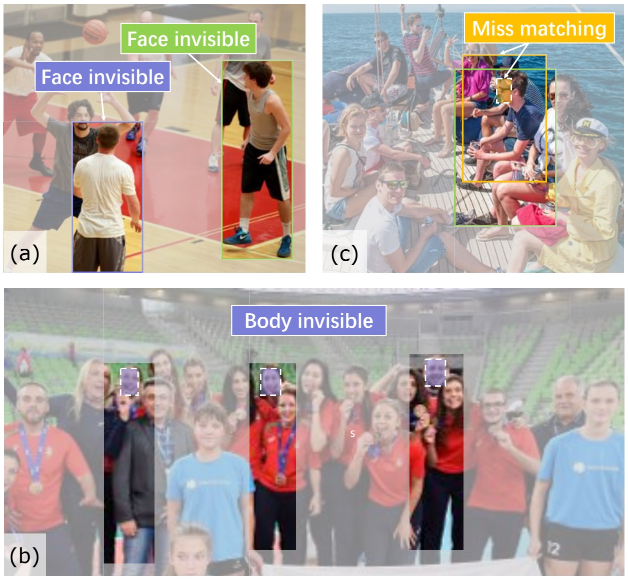

Figure 1. Challenges of detecting pedestrian with associated faces in the wild. (a) Face of pedestrians in green box (side to the camera) and blue box (back to the camera) are invisible. (b) Bodies of the three people are absent despite their faces are clearly visible. (c) Miss-matching between bodies and faces in crowded scenes.

[*] Equal contribution.
[†] Work done during Junfeng's internship at Aibee.

cade mode) since it would seriously affect the face recall. The third solution is detecting bodies and faces respectively and then making associations based on their positional relationships by solving an assignment problem. However, in crowded scenes, this approach (we depict it as *position mode*) would cause severe miss-matching since many faces fall into the body region of other people (Fig. 1c).

To overcome these obstacles, we propose a novel framework named Body-Face Joint (BFJ) detector. In our method, body and face are treated as two independent categories and are detected in parallel. From this design, the incorrespondance issue can be inherently avoided and both categories preserve a good performance since they do not depend on each other (such as sharing the same pre-defined box). Then, we make correlations reasonably from the appearance as well as the geometry level. First, an extra branch is attached to the end of the detector, generating embeddings for all objects detected. We propose an Embedding Matching Loss (EML) to learn the optimal embedding space where the pair of body and face from the same pedestrian are closer to each other. Second, based on the statistical fact that head often appears along with body or face, we introduce a novel conception: "head hook", which means the center-point of the adjunct head of each body and face. According to the information theory [29] that distinct sources of information often provide complementarity. In the association process, we match bodies and faces under the guidance from both the feature level (embeddings) and spatial level (head hooks) above.

Until now, there is neither metric to evaluate the body-face matching quality nor benchmark that has completed annotations for paired bodies and faces. To verify our method, inspired by the log-average miss rate (MR^{-2}) [8] in pedestrian detection, we design a new metric named log-average miss matching rate (mMR^{-2}) to measure errors in body-face association. Moreover, we carefully annotate face box for each pedestrian in two public datasets: CrowdHuman [30] and CityPersons [38]. Experiments show that BFJ outperforms the intuitive methods in both *cascade mode* and *position mode* with a large margin.

In summary, our contributions are two-fold: 1) We propose a joint body-face detection scheme that output body-face pairs for each pedestrian, showing powerful performance in both detection and body-face association. 2) To our best knowledge, we are the first one to systematically investigate the performance of body-face joint detection. Therefore we design a principled metric to evaluate the quality of body-face association. And we annotate faces in CrowdHuman and CityPersons, constructing two new large-scale benchmarks for joint body-face detection.

2. Related Works

Pedestrian Detection. The majority of early approaches for pedestrian detection are part-based [25, 27, 40, 31, 41]. In these methods, the fundamental logic of finding a pedestrian is to detect their semantic parts. Along with the rapid development of object detection driven by deep learning [12, 28, 21, 1, 22], many powerful CNN-based pedestrian detectors have emerged [32, 39, 34, 24, 5, 35], achieving promising results. In recent years, much effort has been spent on the crowdness issue, which also poses great challenges in the body-face association task. For instance, [32] and [39] propose specific loss functions to constrain proposals more closer to the corresponding ground-truth, enhancing discrimination between overlapped individuals. CaSe [34] uses a new branch to count pedestrian number in a region of interest (RoI) and OR-CNN [39] constructs a part occlusion-aware RoI (PORoI) pooling operation to get prior information of body visibility. A group of works focus on alleviating the deficiency of Non-Maximun-Suppressing (NMS) for heavily-overlapped objects. Adaptive-NMS [24] introduces an adaptation mechanism to dynamically adjust the threshold in NMS, leading to better recall in a crowd. In [11] and [16], NMS leverages the less-occluded visible boxes to guide the selection of full boxes, inherently avoiding the crowdness issue. Recently, there are other directions to tackle the crowdness challenges. CrowdDet [5] conducts one proposal to make multiple predictions and use a specially designed Set-NMS to solve heavily-overlapped cases. In [35], the Beta distribution is utilized to modeling the relationship between full box and visible box of a person.

Body-Head Joint Detection. As a key structural part, head plays a vital role in identifying a person. Some recent works [37, 4, 3] concentrate on body-head joint detection to better handle the occlusion because head can provide complementary spatial information to body. JointDet [4] generates head proposals to predict body ones using a statistical ratio, followed by a discriminative module to match them. In Double Anchor R-CNN [37], anchors for body and head are simultaneously produced and cross-optimized with each other. PedHunter [3] takes a different approach to encode the head box in an attention module. Despite the similarity at a glance, body-face joint detection is much more challenging because *body and face are not one-to-one correspondance*. In contrast to previous paradigms, we choose a bottom-up scheme of first detecting them independently and then making reasonable associations.

Embedding Learning. As an effective modeling strategy, embeddings have been widely used in tasks such as image retrieval [10, 33], image captioning [13] and phrase localization [17]. In [26], an associative embedding method is proposed to group joints into different bodies, where the embedding vectors of joints from the same person are encouraged to be close enough. Beyond that, CornerNet [19] introduces the idea of associative embedding to object detection. By pulling corner embeddings from the same object and pushing them across individuals, bounding boxes can be directly generated among peaks in a heatmap instead of the tedious sliding window scheme. Moreover, in the field of segmentation, embeddings are also utilized to group pixels into different objects [36], facilitating instance-level segmentation in one-shot structure. Inspired by [26], we use embedding mechanism to match pairs of body and face. In our approach, each detected instance (either a body or a face) has an embedding vector. These embeddings are grouped into different pedestrians based on their distances.

3. Methodology

As shown in Fig. 2, we extend the head structure of classical one-stage and two-stage detection framework with two sub-modules. First, an extra branch is designed to generate embeddings for all instances detected. Then, a new predictor is added to estimate the centers of the head attached to each body and face (head hooks). After detecting the bodies and faces as two independent categories, a novel association module is employed to estabilish their correspondance using the embedding features and the predicted head hooks.

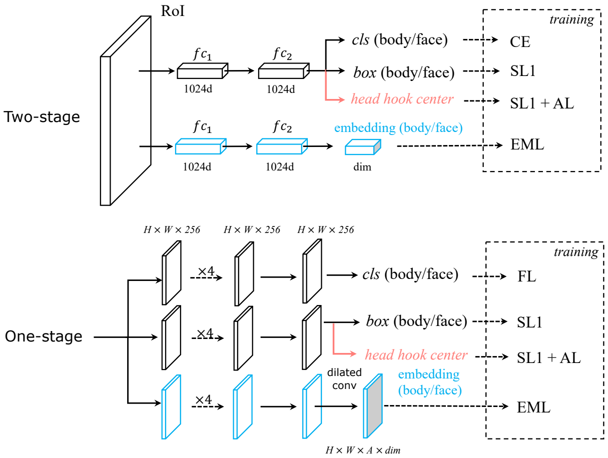

Figure 2. **Detection module of the BFJ detector.** For convenience, we only plot the head structure while the pre-structures of backbone, FPN [21] and RPN [28] are omitted. The branch for generating embeddings for each instance is drawn in blue while the new predictor for estimating the head hooks is in red. Losses used in training are invloved in the dashed box, where CE, FL and SL1 denote the original Cross Entropy loss, Focal loss and Smooth L1 loss while EML and AL denote the newly designed Embedding Matching Loss and Angular Loss.

3.1. Embedding Matching

One key challenge of pairwise matching between face and body boxes is the lack of correlational semantics in the detection output. To alleviate this issue, we extend the idea of learning associative embedding [26], which was originally proposed for grouping joints in multi-person human pose estimation. Instead, we leverage the embeddings to determine if a pair of body and face comes from the same pedestrian. As demonstrated in Fig. 2, the embedding module acts as a parallel branch, producing an embedding vector of dimension dim for each instance. In two-stage detector, the predictor is a fully-connected layer of $dim = 32$ while in one-stage detector, embedding is produced by a dilated convolution layer where the dilation = 2 and $dim = 16$. Below we introduce an Embedding Matching Loss (EML) to pull the embeddings from the same pedestrian and otherwise push them apart across individuals.

Notations. Let $\mathcal{G}$ be the set of all ground truths:

$$\mathcal{G} = \{(g_1^{(b)}, g_1^{(f)}), (g_2^{(b)}, \varnothing), (\varnothing, g_3^{(f)}), ..., (g_n^{(b)}, g_n^{(f)})\},$$

where $(g_i^{(b)}, g_i^{(f)})$ represents the body and face box for the i-th pedestrian. It is worth to notice that some bodies or faces may be absent in real scenarios (denoted as $\varnothing$). In an anchor-based detection framework, there would be a set of proposals $\mathcal{P}(g_k^{(b)})$ corresponding to the ground-truth body (similarly for the face $\mathcal{P}(g_k^{(f)})$) of k-th pedestrian g_k:

$$\mathcal{P}(g_k^{(b)}) = \{p_i^{(b)} \in \mathcal{P} | IoU(p_i^{(b)}, g_k^{(b)}) > \eta\} \tag{1}$$

where $\mathcal{P}$ is the universal set of all proposals and η denotes the overlapping threshold. For each proposal $p_i^{(b)}$, we learn to predict the l_2-normalized embedding $\mathbf{e}_i^{(b)} \in \mathbb{R}^{dim}$.

Pulling Loss. The proposal pairs we need to pull together include three cases: body to body (*bb*), face to face (*ff*), and body to face (*bf*). For the symmetrical *bb* and *ff* cases, since their geometrical locations are naturally aggregated according to Eq. 1, we design the loss weight to be smoothly shifted to the pairs with relatively far distances. Assuming d_{ij} as the distance between the center points of the two proposals (normalized by height of the gt box), we pull them together by minimizing:

$$L_k^{pull_{bb}} = \frac{1}{M_k^2} \sum_{i=1}^{M_k} \sum_{j=1, j \neq i}^{M_k} e^{d_{ij}} \left\| \mathbf{e}_i^{(b)} - \mathbf{e}_j^{(b)} \right\|^2, \tag{2}$$

$$L_k^{pull_{ff}} = \frac{1}{N_k^2} \sum_{i=1}^{N_k} \sum_{j=1, j \neq i}^{N_k} e^{d_{ij}} \left\| \mathbf{e}_i^{(f)} - \mathbf{e}_j^{(f)} \right\|^2, \tag{3}$$

where M_k and N_k are the number of proposals selected from $\mathcal{P}(g_k^{(b)})$ and $\mathcal{P}(g_k^{(f)})$ (we select the top 3 proposals by sorting IoU values in descending order). For the asymmetrical *bf* case, there is no geometrical aggregation effect as in other two cases. Actually, a person's body can probably be closer to the other one's face in crowded scenes. We therefore remove the distance-aware weighting and pull them directly based on the embedding features:

$$L_k^{pull_{bf}} = \frac{1}{M_k N_k} \sum_{i=1}^{M_k} \sum_{j=1}^{N_k} \left\| \mathbf{e}_i^{(b)} - \mathbf{e}_j^{(f)} \right\|^2, \tag{4}$$

Putting them together, the pulling loss can be defined as:

$$L_k^{pull} = \mu L_k^{pull_{bf}} + \beta (L_k^{pull_{bb}} + L_k^{pull_{ff}}), \tag{5}$$

where we set μ to 1.0 and β to 1.5.

Pushing Loss. For different pedestrians, our target is to push their embeddings away. Analogously, there are still three cases as in the pulling loss. However, since the two embeddings in either body-body (*bb*) or face-face (*ff*) cases come from different people, the distance-aware weighting becomes unnecessary. Therefore, we use a unified formulation to represent the three kinds of pushing terms:

$$L_{kl}^{push_{bf}} = \frac{1}{M_k N_l}\sum_{i=1}^{M_k}\sum_{j=1}^{N_l} max(0, \delta - \|\mathbf{e}_i^{(b)} - \mathbf{e}_j^{(f)}\|^2), \quad (6)$$

$$L_{kl}^{push_{bb}} = \frac{1}{M_k M_l}\sum_{i=1}^{M_k}\sum_{j=1}^{M_l} max(0, \delta - \|\mathbf{e}_i^{(b)} - \mathbf{e}_j^{(b)}\|^2), \quad (7)$$

$$L_{kl}^{push_{ff}} = \frac{1}{N_k N_l}\sum_{i=1}^{N_k}\sum_{j=1}^{N_l} max(0, \delta - \|\mathbf{e}_i^{(f)} - \mathbf{e}_j^{(f)}\|^2), \quad (8)$$

where δ is the margin (we set δ to 2 by default) and M_k and N_l follow the similar settings in the pulling loss. The completed pushing loss is then defined as:

$$L_{kl}^{push} = \mu L_{kl}^{push_{bf}} + \beta(L_{kl}^{push_{bb}} + L_{kl}^{push_{ff}}), \quad (9)$$

where the weights μ and β are the same as in Eq. 5.

Whole Loss. Given the above terms of L_k^{pull} and L_{kl}^{push}, we can write the EML function as below:

$$Loss_{em} = \sigma \cdot \frac{1}{|\mathcal{G}|}\sum_{k=1}^{|\mathcal{G}|} L_k^{pull} + \tau \cdot \frac{1}{|\mathcal{G}|^2}\sum_{k=1}^{|\mathcal{G}|}\sum_{\substack{l=1 \\ l\neq k}}^{|\mathcal{G}|} L_{kl}^{push}. \quad (10)$$

In this equation, $|\mathcal{G}|$ is the size of set $\mathcal{G}$, representing the total number of all pedestrians in the ground truth. σ and τ are weighting coefficients.

3.2. Head Hook Prediction

Given a set of predicted body and face boxes, directly establishing their geometrical correspondance is challenging because they yield quite inconsistent geometry properties. After analyzing the statistics of well-known datasets like CrowdHuman and CityPersons, we had an intuitive observation that head almost always virtually exists for each pedestrian. In another word, a head would very likely appear as long as either a body or face is present. Motivated by this observation, we add a predictor to regress the center-point of the adjunct head attached to each body and face and use it as a "hook" to connect body and face. As shown in Fig. 2, this predictor is built with a typical regression structure and the head center denoted as $h \in \mathbb{R}^2$ is derived from the same proposal.

During training, we assume the ground-truth adjunct head center h^* is given along with the ground-truth center of the associated body b^* or face f^*. Let us first consider the constraints of head hook for body b^* as shown in Fig. 3 and the case of face f^* follows the similar conclusion. The straightforward goal is to minimize the Smooth L1 loss between the predict head hook h and the ground-truth h^*. However, this absolute loss would be dominated by the large-scale human bodies, leading to non-robust performance in training. We therefore further constrain the angle spanned by the vector $\mathbf{v}$ from b^* to h and the one $\mathbf{v}^*$ from b^* to h^*. Formally, we introduce the Angular loss (AL) by computing their cross product and minimize the normalized magnitude:

$$Loss_{al} = \sin(\theta) = \frac{\|\mathbf{v} \times \mathbf{v}^*\|}{\|\mathbf{v}\|\|\mathbf{v}^*\|}. \quad (11)$$

Adopting $Loss_{al}$ has two advantages. First, the angular measure is naturally scale invariant, making the training more balanced for small-scale bodies. Second, $Loss_{al}$ is monotonic when $\mathbf{v}$ is approaching to $\mathbf{v}^*$ and it reaches the minimum value 0 when $\mathbf{v}$ and $\mathbf{v}^*$ are parallel. To sum up, the loss function of the head hook predictor is composed as:

$$Loss_{hook} = \alpha Loss_{sl1} + \gamma Loss_{al}, \quad (12)$$

where α and γ are weighting coefficients. We find simply setting $\alpha = 2.0$ and $\gamma = 1.0$ yields reasonably well results. In all, the total loss of training the BFJ detector is a simple addition of $Loss_{emb}$ (Eq. 10), $Loss_{hook}$ (Eq. 12) and the regular detection loss (consists of the cross-entropy loss for classification and the smoothl1 loss for box regression).

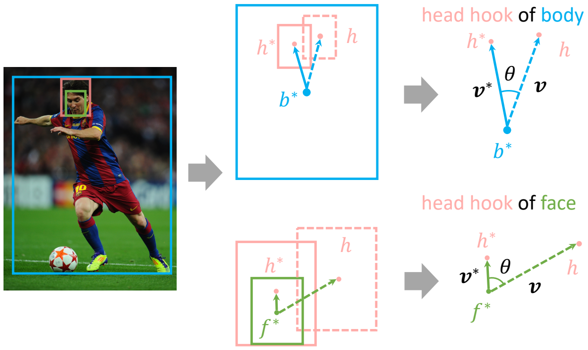

Figure 3. **Angular loss**. We color body, face and head in blue, green and red respectively. Solid boxes represent ground truths while dashed boxes represent detection results (for the head, only the center point is actually predicted instead of the bounding box). The vector $\mathbf{v}$ (or $\mathbf{v}^*$) is constructed from the center point of body (above) or face (below) ground truth to the head hook.

3.3. Association Process

This section explores the utilization of embedding feature and head hook, seeking for the best practice of their collaboration during body-face association.

Before association, we first use a threshold c_{th}=0.3 to filter bodies and faces with low recogonition confidence. Because the geometrical center of heads and the embeddings of face and body lie in different feature space, we take a later fusion approach by first computing the similarity of each cue. To convert the distances into a similarity representative, we feed each normalized distance value through a radial basis function, $s_{ij} = e^{-d_{ij}}$. After computing the similarity value s_{ij} for all pairs of m bodies and n faces, we can construct two similarity matrices $\mathbf{S}_e$ and $\mathbf{S}_h$ for embedding and head hook locations respectively:

$$\mathbf{S}_e = (s^e_{ij}) \in \mathbb{R}^{m \times n}, \quad \mathbf{S}_h = (s^h_{ij}) \in \mathbb{R}^{m \times n},$$

where the element on each location indicates correlation of the specific body-face pairs from its own perspective. Furthermore, in order to utilize the information from two distinct sources as aforementioned, the problem turns to *finding optimal strategy to fuse the two matrices* $\mathbf{S}_e$ *and* $\mathbf{S}_h$.

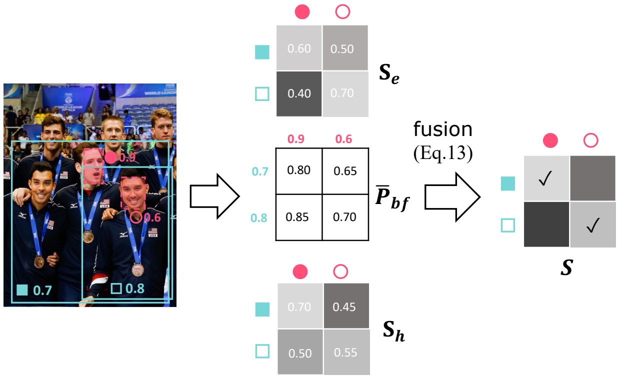

Figure 4. **Association process**. We choose two pedestrians in the image to show the details. Bodies (solid boxes in the image) are denoted by colored squares while faces (dashed boxes in the image) are denoted by colored circles. Two similarity matrices of embedding features $\mathbf{S}_e$ and head hooks $\mathbf{S}_h$ are fused by taking into consideration the averaged confidences (matrix $\bar{\mathbf{P}}_{bf}$).

In general, head hooks make the positional guidance while embedding features provide semantic information. Emperically, positional guidance tends to be accurate when the box prediction is clear, usually including pedestrians near the camera, structurally complete and not crowded. We find this conception of *clear* can largely be indicated by the confidence (classification score) derived from the detection model: the prediction is clear as long as the confidence is high. Following this logic, semantic clues offered by the embedding features are expected to play an important role in more complicated and confused cases, for example, when the pedestrians are in a crowd and their faces cannot be easily allocated into bodies by the head hook distances. Based on the assumptions above, we design a fusing strategy (as shown in Fig. 4) shifting the similarity values to $\mathbf{S}_h$ if the prediction is clear and to $\mathbf{S}_e$ otherwise:

$$\mathbf{S} = \bar{\mathbf{P}}^{\circ\lambda}_{bf} \odot \mathbf{S}_h + (\mathbf{J} - \bar{\mathbf{P}}^{\circ\lambda}_{bf}) \odot \mathbf{S}_e, \quad (13)$$

here the matrix $\bar{\mathbf{P}}_{bf}$ is composed of the average confidence values of body and face along with the index i and j and $\mathbf{J}$ is a unit (all-ones) matrix with the size $m \times n$. The operator $\odot$ represents element-wise multiplication while the operator $[.]^{\circ\lambda}$ is element-wise power by λ (we set λ to 2 by default). With a similarity matrix $\mathbf{S}$ between bodies and faces, the association process is intuitive: for each body, choose the face with maximum similarity. Considering the fact that not every body has an associated face, we set a similarity threshold s_{th}. If the similarity value of the matched face is lower than s_{th}, we think the body does not has a visible face.

4. Experiment

We conduct experiments to systematically evaluate the performance of body-face joint detection from two perspectives: object detection and the association quality.

Datasets. Our benchmark data is built upon two public datasets of pedestrian detection. In CrowdHuman [30], there are 15000 images for training, 4375 images for validation and 5000 images for testing. Annotations for each pedestrian include three boxes: two for visible and full region of the body respectively and one for the head. Following these annotations, we hand label face box for all pedestrians on train and validation sets whose faces can be seen. The CityPersons [38] dataset is a subset of Cityscapes [6], containing person annotations only. There are 2975, 500, and 1525 images for training, validation and testing respectively, in which all pedestrians are annotated by visible and full boxes. Since the test set is withheld, we build face benchmark on train and validation sets. In [4], the authors supply head boxes for this dataset but the annotations are not released yet. We therefore annotate head and face for each pedestrian. In our method, we ignore the visible body boxes and use the full boxes only.

Metrics. For the detection performance, we adopt two commonly used metrics: the *average precision* (AP) [9] (higher is better), and the *log-average miss rate on False Positive Per Image* (FPPI) in the range of $[10^{-2}, 10^{0}]$ shortened as MR^{-2} [8] (lower is better). For body-face association performance, there is no existing metric. Inspired by the principle of MR^{-2}, we propose mMR^{-2} which is abbreviated from *log-average miss matching rate* on FPPI of body-face pairs in $[10^{-2}, 10^{0}]$. The mMR^{-2} exhibits the propotion of body-face pairs that are miss-matched. For a pair of associated body and face, status of *match* is determined on the satisfication of three conditions below:

1) The face box has an IoU higher than 0.5 with a ground-truth face box.

2) The body box has an IoU higher than 0.5 with a ground-truth body box.

3) The two ground-truth boxes belong to the same person.

Otherwise, the status would be regarded as *miss-match*. Based on this explicit definition, we can compute the miss matching rate (mMR) on a specific FPPI point:

$$mMR = 1 - \frac{N_{mp}}{N_p}, \tag{14}$$

where N_{mp} is the number of matched pairs and N_p represents the total number of pairs. Accordingly, the final value of mMR^{-2} can be naturally acquired by log-averaging all mMR values in the FPPI range.

Baseline association method. As mentioned in Sec. 1, we set the straigtforward *cascade mode* (CAS) and *position mode* (POS) as two baseline approaches for body-face association. For the cascade mode, two detectors of the same structure are trained successively, one for detecting body and the other detecting face on each sub-image cropped using body box. For the position mode, we use the same parallel detection scheme for body and face as in BFJ detector. After acquring boxes of the two categories, we pose the body-face association as a Linear Assignment problem. Specifically, first the Intersection-over-Face (IoF) between each body and face is computed as the assignment cost.

$$IoF = \frac{|Box_{body} \bigcap Box_{face}|}{|Box_{face}|}, \tag{15}$$

with these IoF values, we construct a cost matrix $\mathbf{D}$ of size $m \times n$, where m and n are respective body and face number. Then, the efficient Hungarian algorithm [18] (which is first introduced for object detection by [2]) is used to solve this assignment problem and get body-face association. Moreover, since we use head annotations in BFJ detector, we extend the POS baseline by adding heads. In this approach, body, head and face are detected as three independent categories. With the Hungarian algorithm above, we first match body-head pairs then match pairs of head and face, from which body and face are correlated using head as a bridge (simulate the head hook idea in our BFJ detector). We depict this method as *position mode with head* (POSH).

Experimental Settings. We make experiments on both two-stage and one-stage detection frameworks. For two-stage structure, we choose the Faster R-CNN [28] with FPN [21], in which the RoIAlign [14] is used for feature aggregation. For one-stage structure, we adopt RetinaNet [22] as representative. Both of the FPN baseline and RetinaNet use ResNet-50 [15] pre-trained on ImageNet [7] as backbone. Moreover, we implement our method on CrowdDet [5], which is the state-of-the-art detector using full boxes only. We train the networks on 8 Nvidia V100 GPUs with 2 images on each GPU. On CrowdHuman dataset, the short side for each image is resized to 800 and the long side is limited up to 1400. The training process contains 30k iterations, starting from an initial learning rate of 0.02 (FPN) or 0.01 (RetinaNet) and is reduced by 0.1 on 15k and 20k iters respectively. On CityPersons, considering the very small average size of the newly annotated faces, all images are trained using the size of (1536 × 3072) and the input scale of 1.5× is adopted in evaluation. During training, we use an initial learning rate of 0.02 (FPN) or 0.01 (RetinaNet) for the first 5k iterations and reduce it by 0.1 continuously on the next two groups of 2k iterations. Please refer to our code for the detail settings of the similarity threshold s_{th} and σ, τ in Eq. 10.

Method	MR^{-2}	AP@0.5	mMR^{-2}
Two-stage	body/face	body/face	
FPN + CAS	43.0/57.3	85.1/59.3	67.2
FPN + POS	43.5/54.3	87.8/70.3	66.0
FPN + POSH	45.2/57.1	86.9/61.1	65.7
FPN + BFJ	43.4/53.2	88.8/70.0	**52.5**
CrowdDet + CAS	41.7/57.3	90.5/60.3	66.1
CrowdDet + POS	41.9/54.1	90.7/69.6	64.5
CrowdDet + POSH	42.0/57.1	90.0/62.1	64.2
CrowdDet + BFJ	41.9/53.1	90.3/70.5	**52.3**
One-stage	body/face	body/face	
RetinaNet + CAS	52.6/67.1	80.1/53.2	75.0
RetinaNet + POS	52.3/60.1	79.6/58.0	73.7
RetinaNet + POSH	55.6/68.3	75.5/41.1	74.4
RetinaNet + BFJ	52.7/59.7	80.0/58.7	**63.7**

Table 1. Results on the CrowdHuman validation set. All numbers in the table are with the form of percentage (%). CAS: cascade mode, POS: position mode, POSH: position mode with head.

4.1. Results on CrowdHuman

Table. 1 shows main results on CrowdHuman [30]. First, by adding BFJ module, the original detection performance of body and face (detection results in the lines of "+POS") would not be effected. On the FPN baseline, our BFJ detector outperforms the CAS, POS and POSH approaches by **14.7%**, **13.5%** and **13.2%** in mMR^{-2}. In the CAS approach, although the association problem is inherently evaded, there still exist two issues: First, the face detection performance (AP and MR^{-2}) would be damaged due to the missed bodies in the upriver detector. Second, if more than one face emerge in a body region (frequently happens in the crowded scenes), the face detector cannot distinguish them. In the POS approach, since only the IoF information is used for building correlations, it can hardly avoid miss-matchings in a crowd. The third baseline of POSH demonstrates obvious decline of face detection performance. We think it is caused by the confict of anchor assignment between head and face. Results on the state-of-the-art CrowdDet [5] demonstrate similar trends, where performances of the proposed method surpass the baselines with a considerable margin in mMR^{-2}(**13.8%**, **12.2%** and **11.9%** respectively). As shown in the last three lines of Table. 1, our BFJ detector can also work on one-stage RetinaNet [22] detector, in which the mMR^{-2} outperforms the three baselines by **11.3%**, **10.0%** and **10.7%** respectively.

Method	Embed	HHook	mMR^{-2}/%
FPN + BFJ			66.4 (POS)
	✓		55.7
		✓	54.2
	✓	✓	**52.5**
RetinaNet + BFJ			73.8 (POS)
	✓		68.9
		✓	64.5
	✓	✓	**63.7**

Table 2. Ablation study of embedding guidance (Embed) and head hook (HHook) guidance on CrowdHuman validation set.

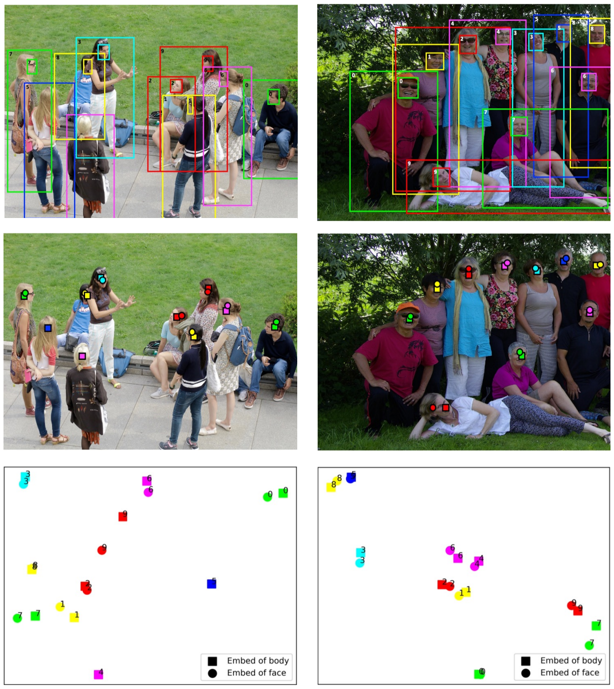

Figure 5. Visualization of the head hooks (the second row) and the embeddings (the third row) of the FPN detector. Body and face are denoted by square and circle respectively.

4.2. Ablation Study

Association guidance. We first make detailed ablations to verify the effectiveness of the two ways for association guidance: the embedding module and the head hook module. Table. 2 demonstrates the comparison results on CrowdHuman [30]. For fairness, result boxes used in the POS baseline (the 1st and 5th row in the table) are produced by the detectors with BFJ module, which are the same with methods in other rows. In summary, the association quality obtains consistant improvement by progressively adding the two modules. On FPN, embedding guidance can bring **10.7%** improvements of mMR^{-2} (66.4% to 55.7%) from the position mode independently. Further analysis indicates that the progress is mainly driven by effective perceptions of semantical body-face correlation, which benefits from the learnable embeddings. We further make PCA-based visualization of embeddings in Fig. 5 (the 3rd line). It is clearly established that embeddings from the same pedestrian are distinctly grouped together. In similar fashion, under head hook guidance alone, the mMR^{-2} can be improved to **54.6%**. The second line in Fig. 5 demonstrates the head hooks predicted by the model. We find that distances between head hooks tend to be more discriminative than the IoFs between face and body, which can provide the association with more accurate reference. Moreover, after similarity fusion in Eq. 13, the mMR^{-2} can be further reduced to **52.5%**. It validates our assumption that these two types of guidance act in a complementary manner, which can be cross optimized through combination. In addition, the relative independence of the two modules offers flexibility of the proposed method. For example, if we do not have the adjunct head annotations, the embedding module can still work on its own. Results on RetinaNet point to the same conclusion, in which the improvement is more significant.

Angular loss. Table. 3 shows ablation studies of the Angular loss. In FPN baseline, the mMR^{-2} obtains **1.2%** improvement by adding Angular loss (53.7% to 52.5%) while in RetinaNet the enhancement is **1.1%** (64.8% to 63.7%). The results confirm our assumption that the angular distances behave well in scale-robustness and promote the head hook prediction.

Method	SL1	AL	mMR^{-2}/%
FPN + BFJ	✓		53.7
	✓	✓	**52.5**
RetinaNet + BFJ	✓		64.8
	✓	✓	**63.7**

Table 3. Ablation study of Angular loss (AL) on CrowdHuman validation set. SL1 represents the original Smooth L1 loss.

Visual Comparison. Fig. 6 makes intuitive visual comparisons of the proposed method with the CAS, POS and POSH baselines respectively. By using our BFJ detector, typical bad cases such miss-matching in the crowded scenes, face miss-recall due to the incorrespondance can be effectively solved or alleviated. Please refer to the appendix for more visual comparisons.

4.3. Results on CityPersons

Table. 4 shows experiments on CityPersons [38]. Following strategies in [32], results are reported on four subsets of *Reasonable* (occlusion $<$ 35%), *Partial* (10% $<$ occlusion $\leqslant$ 35%), *Bare* (occlusion $\leqslant$ 10%) and *Heavy* (occlusion $>$ 35%). In FPN, the BFJ detector surpassess the three baselines of mMR^{-2} by **3.1%**, **0.8%** and **2.5%** re-

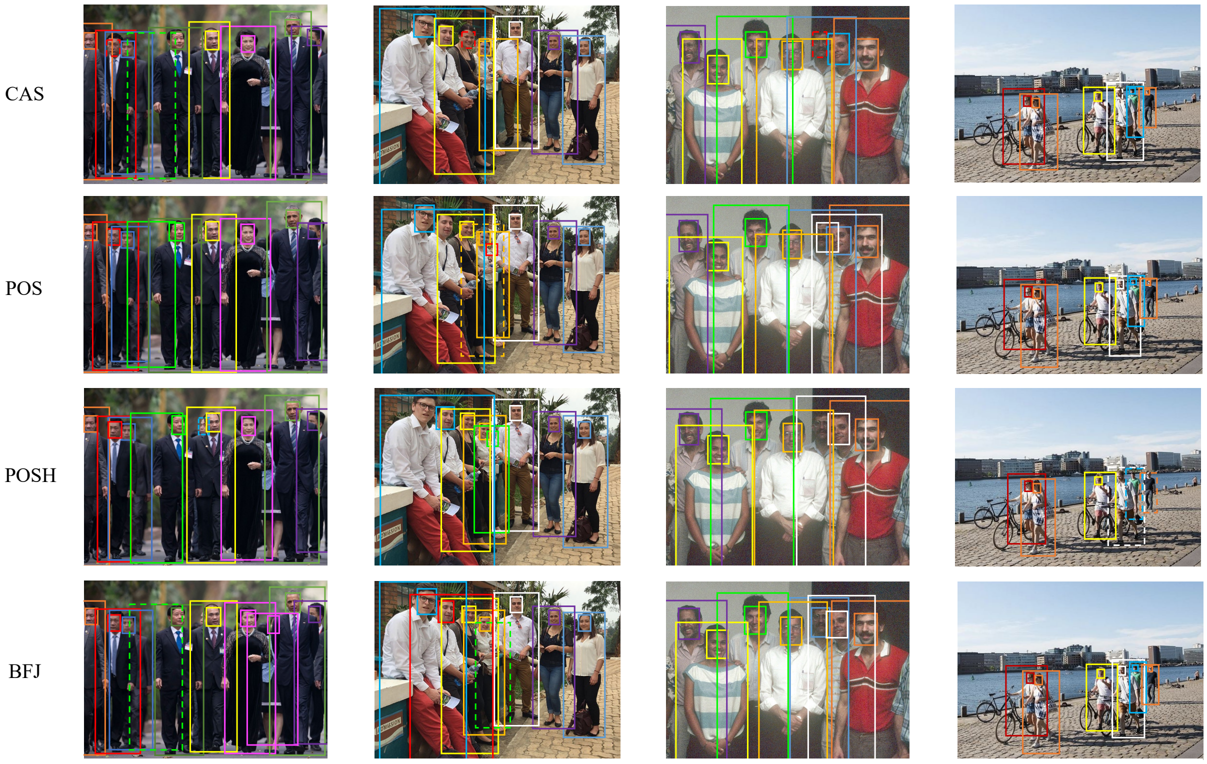

Figure 6. Visual comparisons on the FPN detector of the CAS, POS, POSH baselines and our BFJ respectively. Solid boxes with the same color denote one pair of body and face associated. Dashed box denotes the detected body or face that is not associated successfully.

Method	AP@0.5	*Reasonable*		*Partial*		*Bare*		*Heavy*	
		MR^{-2}	mMR^{-2}	MR^{-2}	mMR^{-2}	MR^{-2}	mMR^{-2}	MR^{-2}	mMR^{-2}
FPN + CAS	81.2/62.8	10.2/23.1	35.8	10.5/19.7	37.3	6.6/22.5	37.0	51.5/39.5	60.5
FPN + POS	80.6/65.5	10.5/20.1	33.5	10.4/18.7	32.7	6.6/20.0	34.1	51.2/38.6	56.6
FPN + POSH	81.5/63.2	10.6/22.5	35.2	10.3/20.5	35.6	6.5/22.8	36.3	50.8/38.5	58.1
FPN + BFJ	84.4/68.0	10.6/17.6	**32.7**	10.8/15.1	**30.6**	6.4/18.7	**33.0**	50.4/26.3	**53.5**
RetinaNet + CAS	78.8/35.7	13.5/33.5	43.7	14.2/28.1	44.5	7.2/30.8	40.4	55.0/38.0	70.0
RetinaNet + POS	78.5/35.3	13.4/25.5	40.0	14.4/23.8	42.8	7.4/25.0	38.7	55.8/36.6	67.0
RetinaNet + POSH	78.1/31.5	13.6/32.8	43.5	14.5/30.1	44.6	7.5/29.9	40.2	55.6/38.3	69.3
RetinaNet + BFJ	79.3/36.2	13.6/23.5	**39.5**	14.3/21.2	**41.5**	7.2/24.4	**38.5**	55.6/35.1	**63.1**

Table 4. Results on the CityPersons validation set. Values on the two sides of the slash / are for body and face respectively. All numbers in the table are with the form of percentage (%).

spectively in the *Reasonable* subset. While in the *Heavy* subset, this superiority can be expanded to **7.0%**, **3.1%** and **4.6%**. These results suppose that our BJF detector has an aptitude of solving miss-matchings in crowded scenes. Results on RetinaNet demonstrate the similar trends, which point to the same conclusion.

5. Conclusion

This paper presents a novel BFJ framework, solving a special yet important body-face joint detection task. Our method not only keeps the great compatibility with the classical one/two-stage detection framework, but also introduces original idea of using embeddings and head hook to guide the association of bodies and faces, yielding promising performances. As the pioneering work, we design a new metric named mMR^{-2} to evaluate the association performance and launch new benchmarks of the body-face joint detection task. In the future, the BFJ framework can be further extended to address other instance-part joint detection tasks (*e.g.*, car body and car plate) and improve structural object detection with part correspondance.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[3] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10639–10646, 2020.

[4] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Relational learning for joint head and human detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10647–10654, 2020.

[5] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12214–12223, 2020.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[10] Andrea Frome, Yoram Singer, Fei Sha, and Jitendra Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[11] Nils Gählert, Niklas Hanselmann, Uwe Franke, and Joachim Denzler. Visibility guided nms: Efficient boosting of amodal object detection in crowded traffic scenes. *arXiv preprint arXiv:2006.08547*, 2020.

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[13] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision*, pages 529–545. Springer, 2014.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Zhida Huang, Kaiyu Yue, Jiangfan Deng, and Feng Zhou. Visible feature guidance for crowd pedestrian detection. *arXiv preprint arXiv:2008.09993*, 2020.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[18] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, March 1955.

[19] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[22] Tsung Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP(99):2999–3007, 2017.

[23] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. 8693:740–755, 2014.

[24] Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6459–6468, 2019.

[25] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc Van Gool. Handling occlusions with franken-classifiers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1505–1512, 2013.

[26] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*, 2016.

[27] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2056–2063, 2013.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems*, pages 91–99, 2015.

[29] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[30] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[31] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1904–1912, 2015.

[32] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018.

[33] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[34] Jin Xie, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Mubarak Shah. Count-and similarity-aware r-cnn for pedestrian detection. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020.

[35] Zixuan Xu, Banghuai Li, Ye Yuan, and Anhong Dang. Beta r-cnn: Looking into pedestrian detection from another perspective. *Advances in Neural Information Processing Systems*, 33, 2020.

[36] Hui Ying, Zhaojin Huang, Shu Liu, Tianjia Shao, and Kun Zhou. Embedmask: Embedding coupling for one-stage instance segmentation. *arXiv preprint arXiv:1912.01954*, 2019.

[37] Kevin Zhang, Feng Xiong, Peize Sun, Li Hu, Boxun Li, and Gang Yu. Double anchor r-cnn for human detection in a crowd. *arXiv preprint arXiv:1909.09998*, 2019.

[38] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017.

[39] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018.

[40] Chunluan Zhou and Junsong Yuan. Non-rectangular part discovery for object detection. In *BMVC*, 2014.

[41] Chunluan Zhou and Junsong Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3486–3495, 2017.