applied sciences

Article

Joint Pedestrian and Body Part Detection via Semantic Relationship Learning

Junhua Gu [1,2], Chuanxin Lan [1,3], Wenbai Chen [4] and Hu Han [3,*]

[1] School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China; jhgu@hebut.edu.cn (J.G.); 15122238532@163.com (C.L.)
[2] Hebei Province Key Laboratory of Big Data Computing, Tianjin 300401, China
[3] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[4] School of Automation, Beijing Information Science and Technology University, Beijing 100101, China; chenwb@bistu.edu.cn
* Correspondence: hanhu@ict.ac.cn

Received: 24 December 2018; Accepted: 5 February 2019; Published: 21 February 2019

Abstract: While remarkable progress has been made to pedestrian detection in recent years, robust pedestrian detection in the wild e.g., under surveillance scenarios with occlusions, remains a challenging problem. In this paper, we present a novel approach for joint pedestrian and body part detection via semantic relationship learning under unconstrained scenarios. Specifically, we propose a Body Part Indexed Feature (BPIF) representation to encode the semantic relationship between individual body parts (i.e., head, head-shoulder, upper body, and whole body) and highlight per body part features, providing robustness against partial occlusions to the whole body. We also propose an Adaptive Joint Non-Maximum Suppression (AJ-NMS) to replace the original NMS algorithm widely used in object detection, leading to higher precision and recall for detecting overlapped pedestrians. Experimental results on the public-domain CUHK-SYSU Person Search Dataset show that the proposed approach outperforms the state-of-the-art methods for joint pedestrian and body part detection in the wild.

Keywords: joint pedestrian and body part detection; adaptive joint non-maximum suppression; semantic relationship learning

1. Introduction

Pedestrian and body part (e.g., upper body, head-shoulder and head) detection has wide potential applications, in person re-identification [1], intelligent surveillance, action recognition from body parts [2], etc. Most of the existing methods focus on detecting the whole pedestrians. The early studies on pedestrian detection used hand-crafted features such as Histogram of Oriented Gradients (HOG) [3], Deformable Part Model (DPM) [4] and Integral Channel Features (ICF) [5,6]. HOG has much better performances than other hand-crafted features at the time. DPM uses mixtures of multi-scale deformable part models to solve the deformation problem. ICF method achieves competitive results to previous features result but with low computational complexity due to the oriented gradient and colour feature (HOG+LUV) selection by boosted decision forests. However, the hand-crafted features are intuitively designed based on pedestrian shape characteristics, which limits their ability to distinguish between pedestrian and complicated backgrounds.

In recent years, deep learning methods, e.g., convolutional neural networks (CNNs) are widely used for pedestrian detection, such as Faster R-CNN [7], PVANet [8], YOLO [9], SSD [10], etc. The deep learning based pedestrian detection methods achieve much better results than the traditional hand-crafted feature based methods. However, the performance of the existing deep learning based pedestrian detection methods can drop significantly under unconstrained scenarios. As shown in Figure 1, the main challenge is that a number of pedestrians in one image can be very small or occluded in the surveillance scenarios, which leads to insufficient feature map.

Some efforts have been made to handle occlusions or small objects during pedestrian detection. Repulsion loss was proposed in [11] to improve pedestrian detection performance by utilizing the repulsion of the surrounding pedestrians in the crowd. Ref. [12] employs an attention mechanism across channels to handle various occlusions. Ref. [13] computes an occlusion-aware detection score based on the part detection confidence scores. In order to handle small objects, Ref. [10] utilizes feature pyramids to enhance the representation of various objects across a large range of scales. Further, Feature Pyramid Networks [14] builds high-level semantic feature maps from all scales to solve the problem of detecting small objects that cannot be well captured by the large-scale feature map. Ref. [15] merges detection results from a set of divided views to improve the detection performance for small objects. In addition, Refs. [16,17] utilize the relations between objects, and [18–20] utilize a cascaded network to improve detection performance.

Despite the efforts in improving pedestrian detection robustness against occlusion and small scale, joint pedestrian and body part detection is seldom studied, which can be more challenging than detecting pedestrian alone. For example, some pedestrians' heads can be completely obscured, while the other body parts are still visible as shown in Figure 1. Ref. [21] proposes a HeadNet to jointly detect pedestrian and the head, head-shoulder, and upper body in a multi-task way [22,23] by utilizing the body context information. But the performance of joint pedestrian and body part detection in the case of severe overlap/occlusion, remains unsatisfactory. To this end, we propose a Body Part Indexed Feature (BPIF) to encode the semantic relationship between individual body parts (i.e., head, head-shoulder, upper body, and whole body), leading to improved robustness against partial even full occlusions of the body part. We also propose an Adaptive Joint Non-Maximum Suppression (AJ-NMS) as a replacement of the original NMS algorithms widely used in object detection, leading to higher recall for detection overlapped pedestrians.

In summary, the contributions of this work are as follows:

1. We propose a BPIF representation to encode the semantic relationship between individual body parts (i.e., head, head-shoulder, upper body, and whole body), providing robustness against partial even full occlusions of the body part. While BPIF was used in the image space to perform face landmark detection in [24], our work differs from [24] in that we build BPIF in the feature space in order to allow feature sharing between different modules.
2. We also propose an AJ-NMS to replace the original NMS algorithms widely used in object detection. The traditional NMS is operated on each category of foreground object, i.e., NMS is applied to the head, head-shoulder, upper-body, and body, separately, without considering their correlations. By contrast, the proposed AJ-NMS treat one person's head, head-shoulder, upper-body, and body as a whole unit, leading to higher recall for detecting overlapped pedestrians, and small part such as pedestrian head. In addition, the proposed AJ-NMS possesses an additional advantage of knowing which body parts belong to the same pedestrian. This is useful for succeeding pedestrian analysis applications, such as person re-identification.
3. The proposed approach advances the state-of-the-art in joint pedestrian and body part detection on the widely used CUHK-SYSU Person Search Dataset [25].

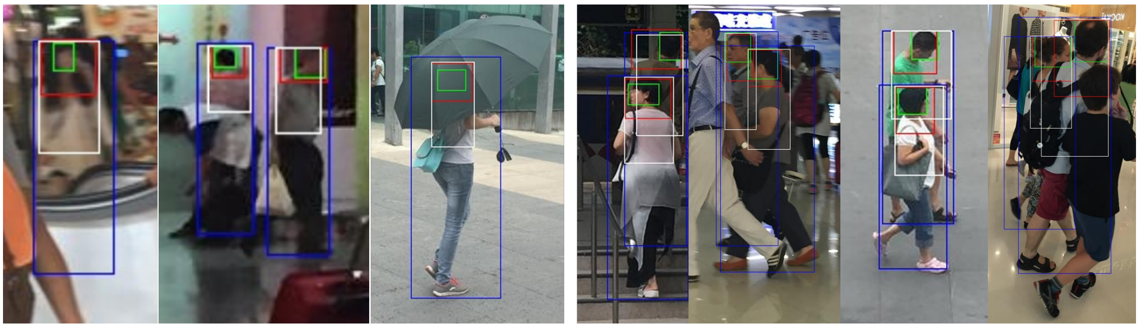


Figure 1. Some example images from the CUHK-SYSU Person Search Dataset [25] illustrate the challenges for joint pedestrian and part detection because of image blurring, occlusion, scale diversity, etc.

2. Related Work

2.1. Pedestrian Detection

Early works [3–6] applied hand-crafted features for pedestrian detection. Dalal et al. [3] proposed the HOG feature which achieves the promising results in pedestrian detection. Felzenszwalb et al. [4] proposed DPM to handle the deformation problem of the pedestrian. The ICF was proposed in [5,6], which shows that oriented gradient and colour feature (HOG+LUV) selection by boosted decision forests are very effective for pedestrian detection. In recent years, there are emerging CNN based methods for pedestrian detection, which achieves great success. Based on the strategy of repulsion-by-surrounding, two types of repulsion losses [11] were proposed to improve network robustness in pedestrian detection in the crowd. Attention [12] was utilized to handle occluded pedestrians by focusing more on non-occluded body parts. Zhang et al. [26] made great efforts on analyzing the failures for a top-performing detector and provided how to engineer better detectors. In addition, HeadNet [21] aimed to utilize the body in context information, i.e., the relationship between individual body parts. Body part detection can be very useful in the succeeding human analysis tasks, such as action recognition. For example, in [2] given the body parts, all possible combination of the body parts are pre-computed to improve the action recognition performance. In addition, HeadNet also introduced beneficial strategies such as Online Hard Example Mining (OHEM) [27] and region of interest align (RoI Align) [28]. These emerging methods perform well under constrained scenarios, but cannot handle well under unconstrained scenarios such as severe overlapped pedestrians detection under surveillance scenarios.

2.2. Non-Maximum Suppression

The success of deep learning based pedestrian detection methods relies on not only the powerful representation ability of CNNs, but also the non-maximum suppression (NMS) [3] to handle the duplicated detections. The principle of NMS is that a bounding box with the local maximum score is selected as the final detection and the other bounding boxes within the pre-defined overlap threshold are suppressed. According to the algorithm, if a foreground object lies within the pre-defined overlap threshold of a local maximum score, it will be suppressed and lead to a miss detection. To solve this problem, Ref. [29] proposed a variant of NMS, named as Soft-NMS, which replaces the suppression strategy in NMS by reducing the scores of the bounding box, which should be suppressed by NMS, according to their overlap degrees. Different from NMS which works as a post-processing step, Ref. [16] implemented a duplicated detection sub-network, making it possible to perform detection and duplication removal in an end-to-end manner. While these traditional NMS methods have been widely used in single-class or multi-class object detection tasks, they are not well suited for the joint detection of pedestrian and its body parts.

2.3. Object Relation Learning

It is believed that object relationship is helpful for object detection. While most of the early works [30,31] utilized object relationships for the post-processing, recent works generally use relationship learning module in the detection network. For example, Ref. [32] uses the early detected objects to assist in detecting the other objects. Further, Refs. [16,17] learn their relationships by parallel. In particular, the relation learning model in [17] considers the relations not only between objects but also between the whole image and individual objects. HeadNet [21] utilized the spatial semantics of the entire body as contextual information. These methods can utilize the relationship between individual objects effectively, but not explicitly model the semantic relationship between individual body parts.

3. Our Approach

3.1. Overview

The overall architecture of the proposed approach for joint pedestrian and body part detection is shown in Figure 2. Give an input image, we first extract the features using a backbone network, such as PVANet [8]. Then the pedestrian region proposals are generated by Region Proposal Network (RPN) [7], and RoI Align [28] is utilized to get the features per proposal that will be used by the Body-in-Context (BiC) module for simultaneously regressing the head, head-shoulder, upper body and pedestrian bounding boxes together with the confidence scores. Based on these initial pedestrian and body part detections, RoI Align is utilized again to perform feature pooling; but this time, we obtain a Body Part Indexed Feature (BPIF) which encodes rich semantic relationship information that is helpful for improving robustness against small object scales or occlusions. Using the BPIF feature, the refinement module again simultaneously regress the head, head-shoulder, upper body and pedestrian bounding boxes. The final bounding boxes are expected to have less false positive detections and false negative detections.

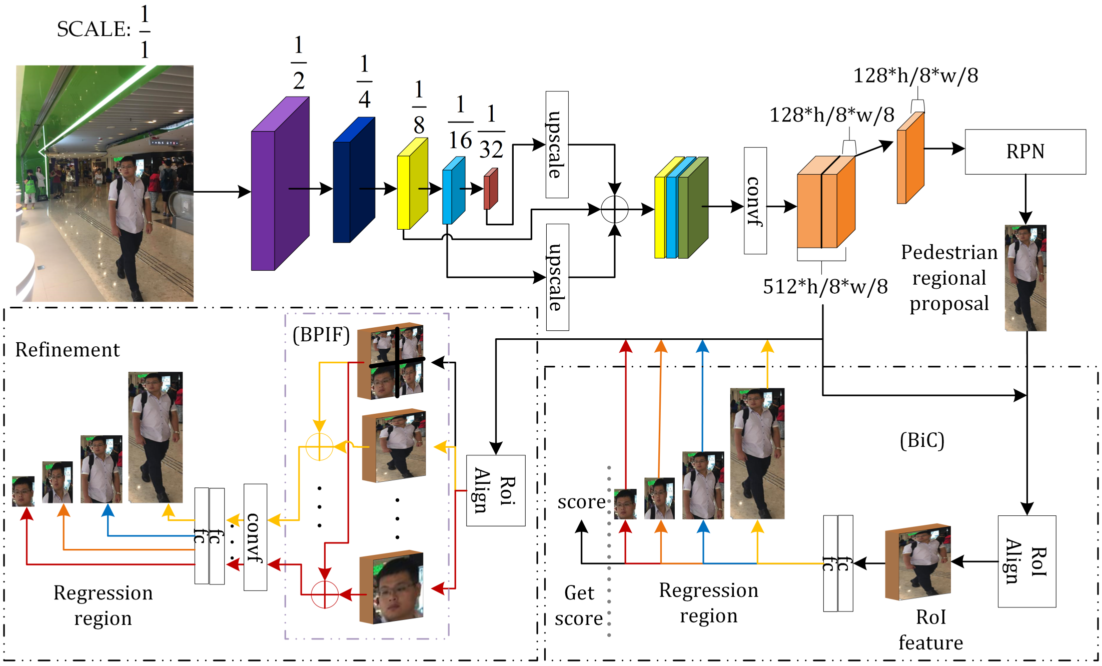


Figure 2. The overall architecture of the proposed approach for joint pedestrian and body part detection in the wild, e.g., with big challenges due to paritial occlusions, diverse scales, etc.

3.2. Body Part Indexed Feature (BPIF)

The aim of BPIF is two-fold: (i) encode the semantic relationship between individual body parts (i.e., head, head-shoulder, upper body, and whole body), and (ii) highlight the features for each body part. As shown in Figure 2, BPIF is composed of two components. One component combines the features from all the body parts and forms a joint feature map, which is inspired by the modified shape indexed features [24] used for face landmark regression. The second component is the per body part features at a more detailed scale. We provide the details of how to compute BPIF in Algorithm 1, and we also give an example of this process in Figure 3. The entire BPIF is followed by two convolutional layers for feature dimensionality reduction.

Algorithm 1 Compute BPIF

```
Input: $F, R_1, R_2, R_3, R_4, N_s$
  $F$ is the image features using a backbone
  $R_1, R_2, R_3$ and $R_4$ are the sets of different body parts' bounding boxes of the BiC results
  $N_s$ is the BPIF scale hyper-parameter
Output: $O_1, O_2, O_3, O_4$
  $O_1, O_2, O_3$ and $O_4$ are the BPIF, which be used to regress the corresponding bounding boxes in the refinement module
  function G_BPIF($F, R_1, R_2, R_3, R_4, N_s$)
      $R'_i = \text{RoI Align}(F, R_i, N_s/2)$ where $i = 1, 2, 3, 4$ and $R_i \in [512, N_s/2, N_s/2]$
      $f_{all} = \begin{bmatrix} R'_1 & R'_2 \\ R'_3 & R'_4 \end{bmatrix}$ where $f_{all} \in [512, N_s, N_s]$
      $R_i = \text{RoI Align}(F, R_i, N_s)$ where $i = 1, 2, 3, 4$ and $R_i \in [512, N_s, N_s]$
      $O_i = \text{concat}(F_{all}, R_i)$ where $i = 1, 2, 3, 4$ and $O_i \in [1024, N_s, N_s]$
      return $O_1, O_2, O_3, O_4$
  end function
```

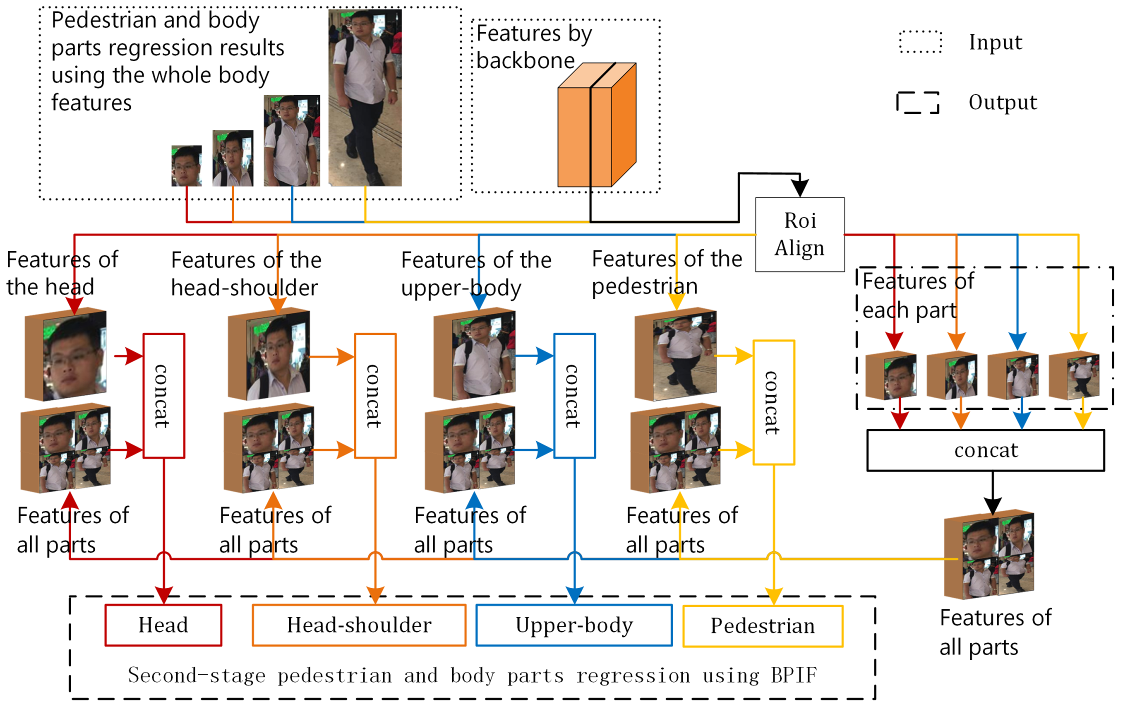


Figure 3. An example of how to compute BPIF features based on the preliminary detection of the pedestrian and body parts.

It is worth noting that unlike HeadNet [21] the Body-in-Context (BiC) and Refinement modules in our model do not share parameters. Our considerations are as follows. The cascaded object detection [33] often shares parameters, but the paper in [16] shows that while iterative bounding box regression is a post-processing procedure used to improve bounding boxes accuracies, cascaded regression is a resampling procedure that changes the distribution of hypotheses to be processed by the different stages. Further, different from normal cascaded object detection, BiC and Refinement modules in our model have different objectives and inputs. While BiC module uses the RoI features per proposal to jointly regress the pedestrian and body part bounding boxes, the refinement module takes the BPIF features to refine the locations of the initial bounding box detections by BiC.

3.3. Adaptive Joint Non-Maximum Suppression

Non-Maximum Suppression (NMS) [3] has been an effective post-processing algorithm in proposal based object detection framework. In our joint pedestrian and body part detection task, we notice the traditional NMS could have the issue that it does not treat the individual parts from one pedestrian as a whole, and thus is not able to leverage the detected body parts with high confidence scores to recover the body parts with low detection scores due to occlusion, etc. As a result, there are false positive and false negative detections that can likely be avoided. In order to reduce such false positive and false negative detections, we propose an Adaptive Joint Non-Maximum Suppression (AJ-NMS) to replace the original NMS. Specifically, AJ-NMS considers individual body parts (i.e., head, head-shoulder, upper body, and whole body) as a whole and use adaptive threshold per body part. We provide the details of AJ-NMS in Algorithm 2. AJ-NMS uses the following joint intersection over union (IOU_{joint}):

$$IOU_{joint} = \lambda_1 IOU_{pedestrian} + \lambda_2 IOU_{head} + \lambda_3 IOU_{head-shoulder} + \lambda_4 IOU_{upper-body}, \tag{1}$$

where: IOU_{joint} is the joint IoU considering both the holistic pedestrian and the associated body parts. $IOU_{pedestrian}$, IOU_{haed}, $IOU_{head-shoulder}$ and $IOU_{upper-body}$ are the traditional IOU defined on pedestrian, head, head-shoulder, and upper-body, respectively. $\lambda_1 - \lambda_4$ are their weights. We empirically determine the weights per part on the training set by greedy search algorithms. Finally, we choose to use λ_1 = 0.6, λ_2 = 0.2, λ_3 = 0.1 and λ_4 = 0.1, respectively.

A brief analysis why the proposed AJ-NMS should work for the joint pedestrian and body part detection task is that as shown in Figure 1, although the two pedestrians' bodies are seriously occluded, their joint IOU considering both the pedestrian and the body parts according to our AJ-NMS remains lower than the threshold to be suppressed because the overlap between the two pedestrians' heads can be very small. As a result, both pedestrian detections can be retained by the proposed AJ-NMS, avoiding the false negative detections and thus improving the recall rate. We also provide some statistical analysis based on the testing set of the CUHK-SYSU Person Search Dataset. Specifically, there are 14,100 ground-truth pedestrian bounding boxes in the whole test set. Among these bounding boxes, 709 pedestrian bounding boxes, for which the IOUs are greater than 0.4 and will be suppressed by the traditional NMS algorithm (leading to missing detections). By contrast, there are only 454 bounding boxes of pedestrian for which the IOUs are above the threshold and to be suppressed by our AJ-NMS. It is worth noting that our AJ-NMS method does not hurt the detection performance while assures higher recall in detecting overlapped pedestrians. Besides improving the pedestrian detection recall under occlusion, the proposed AJ-NMS can be very helpful in improving the precision of small body part detection, such as the head, leveraging the high confident detection for the larger body parts, such as head-shoulder, upper body, etc.

In the proposed approach, since the head, head-shoulder, upper body, and pedestrian bounding boxes are jointly regressed based on one proposal by RPN, we can know which body parts belong to the same pedestrian in the detection results. We maintain this part-pedestrian association information in our detection framework, so that AJ-NMS can work correctly for a pedestrian and its parts. The pedestrian-part association can also be very useful for the succeeding pedestrian analysis applications, such as holistic and local based person re-identification.

3.4. Network Training

Our model can be trained end-to-end by minimizing the following loss function:

$$L_O = \lambda_1 L_{RPN\_cls} + \lambda_2 L_{RPN\_reg} + \lambda_3 L_{BiC\_cls} + \lambda_4 L_{BiC\_reg} + \lambda_5 L_{Refine\_reg}, \tag{2}$$

where L_{RPN_cls} and L_{RPN_reg} are pedestrian region proposals losses in RPN module, which are the same as that those in Faster R-CNN [7]. L_{BiC_cls} is classification loss (cross-entropy loss) in classifying the pedestrian region proposals by the BiC module. L_{BiC_reg} is the regression loss (smooth L1 loss [7]) for pedestrian, head, head-shoulder, and upper body regression in BiC module. Similar to L_{BiC_reg}, L_{Refine_reg} is also a regression loss in regressing four body parts by the Refinement module. We set the hyper-parameters of the loss function λ_1, λ_2, λ_3, λ_4, and λ_5 to 1, 1, 0.25, 0.25 and 0.25, respectively.

During the network training phase, we fine-tune an ImageNet [34] pre-trained PVANet using the training set of the CUHK-SYSU Person Search Dataset [25] with stochastic gradient descent (SGD) optimization. We train the model with an initial learning rate of 0.001 for the first 50,000 iterations and then with a decreased learning rate of 0.0001 for another 30,000 iterations. In order to improve the detection performance for the hard-to-detect objects (occluded pedestrians and small body parts), we generate 2000 pedestrian proposals in RPN, but only 128 hard proposals identified by OHEM will participate in the network training in each iteration. In addition, in order to improve the detector's robustness in handling pedestrians with big scale diversity, we adopt a multi-scale training strategy. Specifically, the shorter edge of an input training image is randomly scaled between 416 pixels and 864 pixels, while the longer edge is retained no more than 1440 pixels.

Algorithm 2 AJ-NMS

```
Input: $B = \{b_1, ..., b_N\}, S = \{s_1, ..., s_N\}, N_t$
  $B$ is the list of initial detection boxes
  $S$ contains corresponding detection scores
  $N_t$ is the AJ-NMS threshold
Output: $B_r = \{b_1, ..., b_m\}, S_r = \{s_1, ..., s_m\}$
  $B_r$ is the list of AJ-NMS results detection boxes
  $S_r$ contains corresponding detection scores
  function AJ-NMS($B$, $S$, $N_t$)
    $B_r = \{\ \}$
    $S_r = \{\ \}$
    while $B \neq empty$ do
      $m = \text{argmax}\ S$
      $B_r = B_r \bigcup b_m$
      $S_r = S_r \bigcup s_m$
      $B = B - b_m$
      $S = S - S_m$
      for $b_i$ in $B$ do
        if $IOU_{joint}(b_m, b_i) > N_t$ then
          $B = B - b_i$
          $S = S - s_i$
        end if
      end for
    end while
    return $B_r, S_r$
  end function
```

4. Experiments

4.1. Datasets and Settings

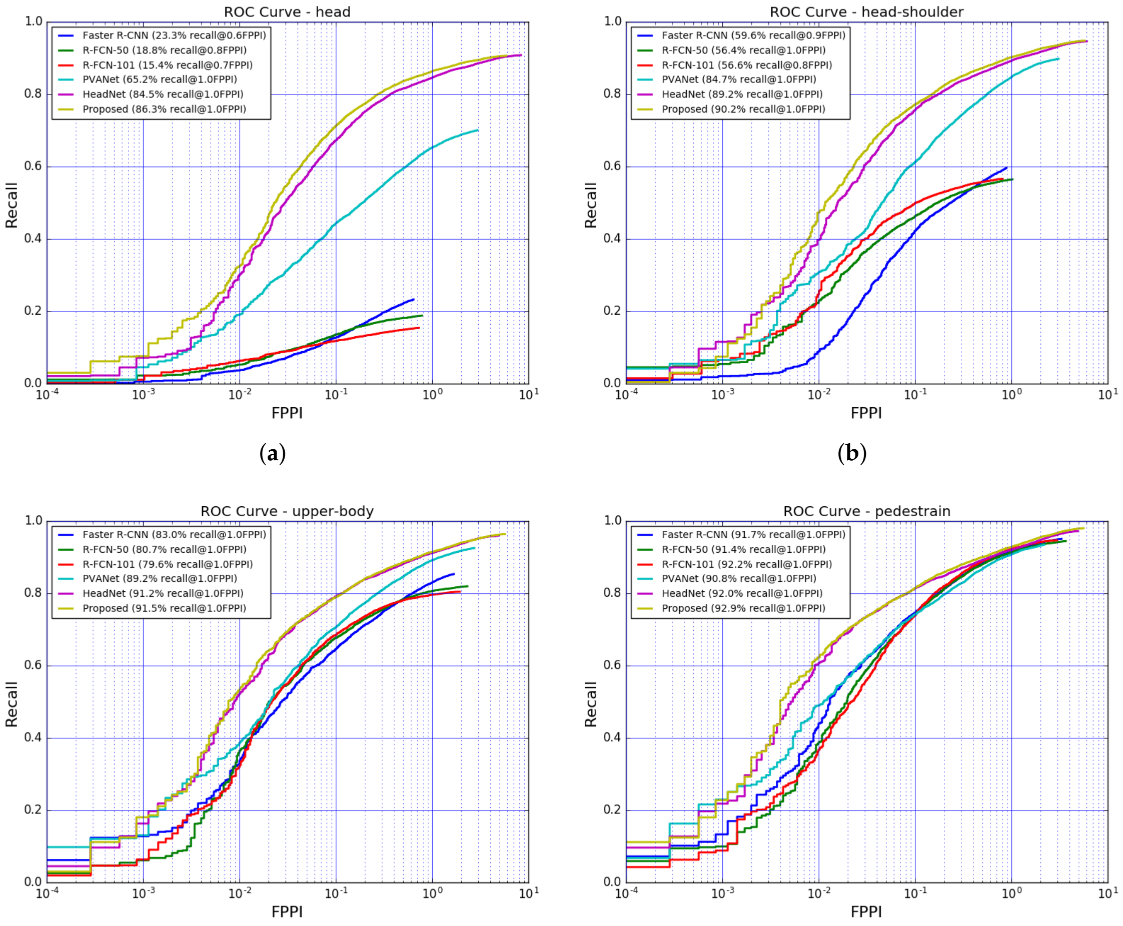
We use the CUHK-SYSU Person Search Dataset [25] for experimental evaluations, which contains 18,184 images and 99,809 annotated pedestrian bounding boxes. The original dataset does not include body parts annotations, but Chen et al. [21] provided head, head-shoulder, and upper body bounding boxes annotations for 16,907 images from CUHK-SYSU Person Search Dataset. Following the protocol in [21], we use 13,399 and 3508 images as the training and testing sets, respectively.

4.2. Comparisons with the State-of-Art

In this section, we provide evaluations of the proposed approach and a number of state-of-the-art algorithms such as Faster R-CNN [7], R-FCN-50 [35], R-FCN-101 [35], PVANet [8], and HeadNet [21] under the same experimental settings, except that Faster R-CNN, R-FCN-50, and R-FCN-101 do not use OHEM and the image shorter side is scaled between 416 pixels and 768 pixels due to memory occupation of R-FCN-101. All evaluation results in terms of recall vs. false positives per image (FPPI) and average precision (AP) are presented in Figure 4 and Table 1. We can observe that although all methods can perform well for the holistic pedestrian detection, the performance of all the methods except for the proposed approach and HeadNet [21] drop significantly in detecting the small body parts. This observation indicates that the spatial semantics of the entire body is useful for a joint pedestrian and body part detection task. Compared with HeadNet, the proposed approach achieves 0.8%, 0.3%, 0.6%, and 1.0% higher AP, and 0.9%, 0.3%, 1.0%, and 1.8% higher recall rate at 1 false positive per image (FPPI) in detecting pedestrian, upper body, head-shoulder, and head detection, respectively. The gain comes from the proposed BPIF which not only encodes the semantic relationship of individual body parts but also highlights the per part features, as well as the proposed AJ-NMS algorithm which jointly considers individual parts instead of doing NMS separately for each part.

Figure 5 gives some detection results by the proposed approach, we can see the proposed joint pedestrian and body part detection approach performs well under the challenging surveillance scenario with partial occlusions, diverse scales, bad illuminations, etc. In addition, the proposed approach can output the body part and pedestrian association information, i.e., which body parts belong to the same pedestrian.

Our approach is implemented using Caffe (https://github.com/BVLC/caffe), with a few layers in the proposal network implemented in python (the python layers run on CPU). On our platform with a GeForce GTX 1080 Ti GPU and Intel Xeon(R) E5-2620 v4 CPU, the state-of-the-art method HeadNet runs in about 10.8 FPS, and our approach without optimization runs in about 4.5 FPS. The increased computational cost mainly comes from the increased feature map size because of the use of BPIF. In the future, we will optimize the BPIF module to reduce its computational cost, i.e., through Depthwise Separable Convolution [36].



(**a**) (**b**)

(**c**) (**d**)

Figure 4. Detection performance (in terms of recall vs. FPPI) by the proposed approach and the state-of-the-art methods on the CUHK-SYSU Person Search Dataset. (**a**) ROC curve for head detection, (**b**) ROC curve for head-shoulder detection, (**c**) ROC curve for upper body detection, and (**d**) ROC curve for the holistic pedestrian detection.

Table 1. Comparisons of detection performance (in terms of AP) for the whole pedestrian, upper-body, head-shoulder, and head by different approaches.

	Pedestrian AP	**Upper-Body AP**	**Head-Shoulder AP**	**Head AP**
Faster R-CNN [7]	92.1	82.3	56.2	20.3
R-FCN-50 [35]	91.4	79.7	54.7	17.2
R-FCN-101 [35]	91.8	78.6	55.3	14.3
PVANet [8]	90.8	89.3	85.1	64.6
HeadNet [21]	94.1	92.8	90.8	86.0
Proposed	**94.9**	**93.1**	**91.4**	**87.0**

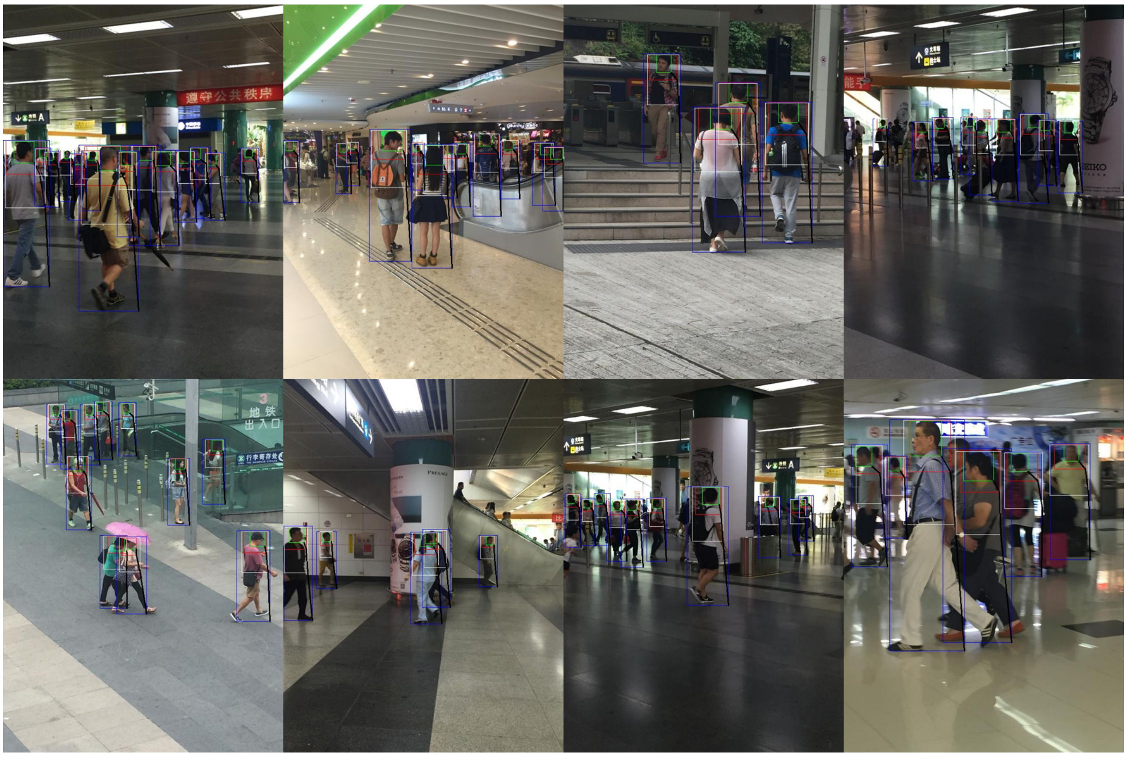


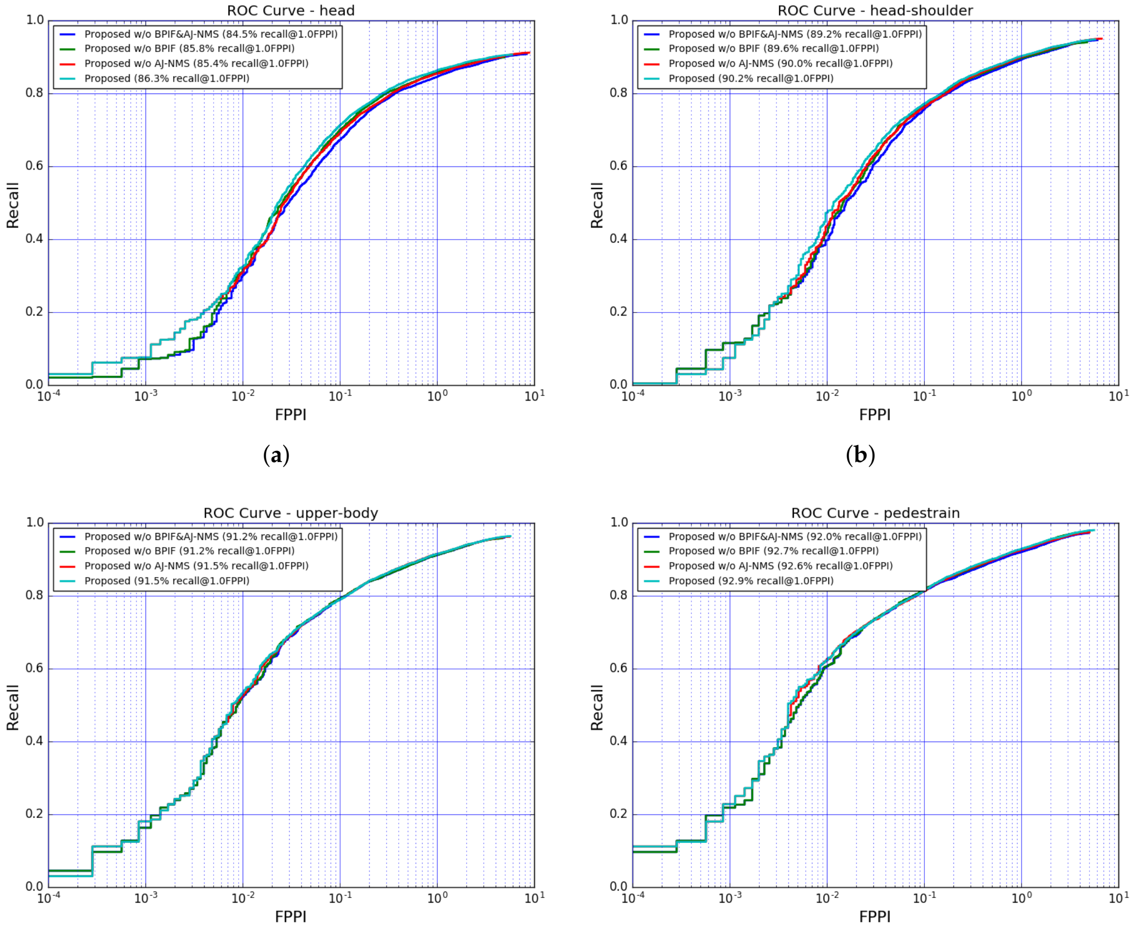
Figure 5. Joint pedestrian and body part detection results by the proposed approach. Blue, white, red and green bounding boxes represent the detection results for pedestrian, upper body, head-shoulder, and head, respectively. The black lines indicate the body parts and pedestrian association information, i.e., referred four body parts belong to the same pedestrian.

4.3. Ablation Study

In this section, we perform ablation experiments to analyze the impact of individual parts in the proposed approach, e.g., BPIF and AJ-NMS. We cover three ablation studies: (i) remove both BPIF and AJ-NMS from the proposed approach (denoted as Proposed w/o BPIF&AJ-NMS), (ii) remove BPIF only (denoted as Proposed w/o BPIF), and (iii) remove AJ-NMS only (denoted as Proposed w/o AJ-NMS). All the experiments follow the same experimental setting as our full method. The results of all methods are shown in Figure 6 and Table 2.

As shown in Figure 6 and Table 2, we can see that under most of the cases, removing either BPIF or AJ-NMS will lead to degraded detection performance. This indicates the usefulness of the proposed BPIF and AJ-NMS for the joint pedestrian and body part detection task. It also indicates that it is reasonable for the BiC and refinement modules to not share parameters.

Compared to 'Proposed w/o BPIF&AJ-NMS' and 'Proposed w/o AJ-NMS', 'Proposed w/o BPIF' and our full method have shown much better detection performance in head and pedestrian. The gain mainly comes from the improved detections for overlapped pedestrians by the proposed approach, and reduced false head detection by the proposed approach. Both are attributed to the designed BPIF and AJ-NMS. Specifically, with traditional NMS a false head detection may not be suppressed because its IOU with a nearby true head detection can lower the threshold to be suppressed. By contrast, the proposed AJ-NMS considers the head IOU together with the other three parts (head-shoulder, upper-body, and whole pedestrian) . As a result, a false head detection can be suppressed with our AJ-NMS because the joint IOU of a false detection can be larger than the suppression threshold, and will be suppressed, leading to reduced false detections (high precision). Figure 7 gives an example in which a false head detection cannot be suppressed by the traditional NMS but can be suppressed by our AJ-NMS.



(**a**) (**b**)

(**c**) (**d**)

Figure 6. Ablation study of the proposed approach w.r.t. BPIF and AJ-NMS on the CUHK-SYSU Person Search Dataset measured in terms of recall vs. FPPI. (**a**) ROC curve for head detection, (**b**) ROC curve for head-shoulder detection, (**c**) ROC curve for upper body detection, and (**d**) ROC curve for the holistic pedestrian detection.

Table 2. Ablation study of the proposed approach w.r.t. BPIF and AJ-NMS measured in terms of AP (in %).

Model	Pedestrian AP	Upper-Body AP	Head-Shoulder AP	Head AP
Proposed w/o BPIF&AJ-NMS	94.1	92.8	90.8	86.0
Proposed w/o BPIF	94.6(↑0.5)	92.7(↓0.1)	90.8(↑0.0)	86.3(↑0.3)
Proposed w/o AJ-NMS	94.5(↑0.4)	93.0(↑0.2)	91.3(↑0.5)	86.6(↑0.6)
Proposed	94.9(↑0.8)	93.1(↑0.3)	91.4(↑0.6)	87.0(↑1.0)

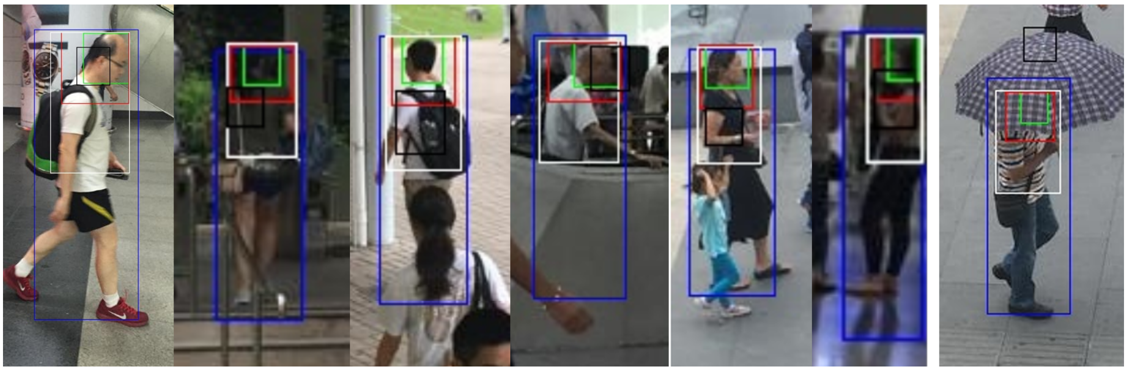


Figure 7. Blue, white, red and green bounding boxes stand for correct pedestrian, upper body, head-shoulder and head detection by our full method, respectively. The black bounding box stands for the incorrect head detection results in ablation study by 'Proposed w/o AJ-NMS'.

5. Conclusions

Joint pedestrian and body parts detection in the wild via a single model is challenging because of partial occlusion, scale diversity, nonuniform illumination, etc. We propose to resolve these challenges via body part indexed feature (BPIF) and adaptive joint Non-Maximum Suppression (AJ-NMS). While BPIF is able to encode both semantic relationship between individual body parts and highlights per part features at detailed scales, AJ-NMS is helpful for handling the pedestrian and its individual parts as a joint unit to reduce false detections and missed detections when each body part is handled separately. The proposed approach achieves promising performance on the public dataset, and outperforms the state-of-the-art methods.

We would like to investigate adaptive relationship learning approach between pedestrian and its body parts to enhance the flexibility in utilizing contextual information in joint pedestrian and body part detection. In addition, we would also like to investigate the effectiveness of the proposed approach for joint pedestrian and retrieval task [37].

Author Contributions: Conceptualization, C.L. and H.H.; methodology, C.L., H.H. and W.C.; validation, C.L.; formal analysis, C.L.; Investigation, C.L.; writing—original draft preparation, C.L., H.H. and J.G.; writing—review and editing, H.H. and W.C.; Funding acquisition, J.G.

Funding: This research received no external funding.

Acknowledgments: This study was supported in part by NSF of Hebei Province through the Key Program under Grant F2016202144.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Y.; Zhao, Q.; Wu, Z. Pooling body parts on feature maps for misalignment robust person re-identification. In Proceedings of the 4th IEEE International Conference on Identity, Security, and Behavior Analysis (ISBA 2018), Singapore, 11–12 January 2018; pp. 1–8.
2. Mousas, C.; Anagnostopoulos, C.N. Performance-Driven Hybrid Full-Body Character Control for Navigation and Interaction in Virtual Environments. *3D Res.* **2017**, *8*, 18. [CrossRef]
3. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, 20–26 June 2005; pp. 886–893.
4. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.A.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
5. Dollár, P.; Appel, R.; Belongie, S.J.; Perona, P. Fast Feature Pyramids for Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1532–1545. [CrossRef] [PubMed]
6. Dollár, P.; Tu, Z.; Perona, P.; Belongie, S.J. Integral Channel Features. In Proceedings of the 20th British Machine Vision Conference (BMVC 2009), London, UK, 7–10 September 2009; pp. 1–11.

7. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
8. Kim, K.; Cheon, Y.; Hong, S.; Roh, B.; Park, M. PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection. *arXiv* **2016**, arXiv:1608.08021.
9. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV 2016), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
11. Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion Loss: Detecting Pedestrians in a Crowd. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7774–7783.
12. Zhang, S.; Yang, J.; Schiele, B. Occluded Pedestrian Detection Through Guided Attention in CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6995–7003.
13. Noh, J.; Lee, S.; Kim, B.; Kim, G. Improving Occlusion and Hard Negative Handling for Single-Stage Pedestrian Detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 966–974.
14. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
15. Tang, C.; Ling, Y.; Yang, X.; Jin, W.; Chao, Z. Multi-View Object Detection Based on Deep Learning. *Appl. Sci.* **2018**, *8*, 1423. [CrossRef]
16. Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. Relation Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3588–3597.
17. Liu, Y.; Wang, R.; Shan, S.; Chen, X. Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6985–6994.
18. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
19. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-Shot Refinement Neural Network for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4203–4212.
20. Ramalingam, B.; Lakshmanan, A.K.; Ilyas, M.; Le, A.V.; Elara, M.R. Cascaded Machine-Learning Technique for Debris Classification in Floor-Cleaning Robot Application. *Appl. Sci.* **2018**, *8*, 2649. [CrossRef]
21. Chen, G.; Cai, X.; Han, H.; Shan, S.; Chen, X. HeadNet: Pedestrian Head Detection Utilizing Body in Context. In Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 556–563.
22. Han, H.; Jain, A.K.; Wang, F.; Shan, S.; Chen, X. Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2597–2609. [CrossRef] [PubMed]
23. Wang, F.; Han, H.; Shan, S.; Chen, X. Deep Multi-Task Learning for Joint Prediction of Heterogeneous Face Attributes. In Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 173–179.
24. Zhang, G.; Han, H.; Shan, S.; Song, X.; Chen, X. Face Alignment across Large Pose via MT-CNN Based 3D Shape Reconstruction. In Proceedings of the 13th International Conference on Automatic Face and Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 210–217.
25. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint Detection and Identification Feature Learning for Person Search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 3376–3385.

26. Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.H.; Schiele, B. Towards Reaching Human Performance in Pedestrian Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 973–986. [CrossRef] [PubMed]
27. Shrivastava, A.; Gupta, A.; Girshick, R.B. Training Region-Based Object Detectors with Online Hard Example Mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
28. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
29. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving Object Detection with One Line of Code. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 5562–5570.
30. Divvala, S.K.; Hoiem, D.; Hays, J.; Efros, A.A.; Hebert, M. An empirical study of context in object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 1271–1278.
31. Galleguillos, C.; Rabinovich, A.; Belongie, S.J. Object categorization using co-occurrence, location and appearance. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, USA, 24–26 June 2008; pp. 1–8.
32. Stewart, R.; Andriluka, M.; Ng, A.Y. End-to-End People Detection in Crowded Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 2325–2333.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
35. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2016), Barcelona, Spain, 5–10 December 2016; pp. 379–387.
36. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2018**, arXiv:1704.04861.
37. Han, H.; Li, J.; Jain, A.K.; Shan, S.; Chen, X. Tattoo Image Search at Scale: Joint Detection and Compact Representation Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1–15. [CrossRef] [PubMed]