Relational Learning for Joint Head and Human Detection

Cheng Chi,[1,3*] **Shifeng Zhang,**[2,3*†] **Junliang Xing,**[2,3] **Zhen Lei,**[2,3] **Stan Z. Li,**[2,3,4] **Xudong Zou**[1,3]
[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China
[2]CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Macau University of Science and Technology, Macao, China
chicheng15@mails.ucas.ac.cn, {shifeng.zhang, jlxing, zlei, szli}@nlpr.ia.ac.cn, xdzou@mail.ie.ac.cn

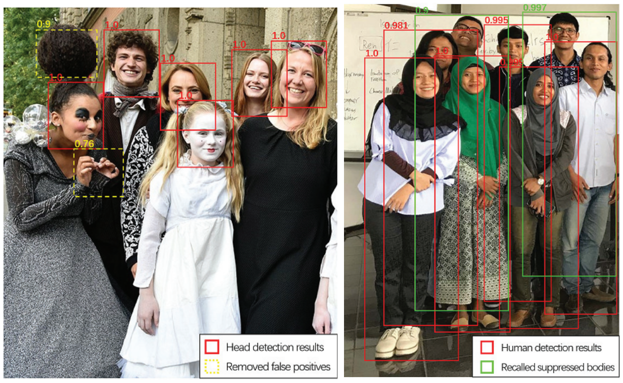
Abstract

Head and human detection have been rapidly improved with the development of deep convolutional neural networks. However, these two tasks are often studied separately without considering their inherent correlation, leading to that 1) head detection is often trapped in more false positives, and 2) the performance of human detector frequently drops dramatically in crowd scenes. To handle these two issues, we present a novel joint head and human detection network, namely JointDet, which effectively detects head and human body simultaneously. Moreover, we design a head-body relationship discriminating module to perform relational learning between heads and human bodies, and leverage this learned relationship to regain the suppressed human detections and reduce head false positives. To verify the effectiveness of the proposed method, we annotate head bounding boxes of the CityPersons and Caltech-USA datasets, and conduct extensive experiments on the CrowdHuman, CityPersons and Caltech-USA datasets. As a consequence, the proposed JointDet detector achieves state-of-the-art performance on these three benchmarks. To facilitate further studies on the head and human detection problem, all new annotations, source codes and trained models are available at https://github.com/ChiCheng123/JointDet.

Introduction

Head and human detection are two important research topics in computer vision field with various applications, such as human behavior analysis, intelligent video surveillance and automatic driving. Although great progress has been made by deep convolutional neural networks (CNNs) on general object detection (Ren et al. 2017; Dai et al. 2016; Lin et al. 2017a; Liu et al. 2016; Lin et al. 2017b; Zhang et al. 2018b), research in the realm of these two subtasks remains challenging due to their characteristics.

Head detection has experienced tremendous development in recent years. The context-aware CNN model (Vu, Osokin, and Laptev 2015) employs a pairwise CNN to model pairwise relations among heads. The HeadNet (Chen et al. 2018) utilizes spatial semantic relations between pedestrian head and other body parts. However, how to reduce the false positives, such as hair, hands and elbows shown in Figure 1(a), still remains an active research direction. Tracing the main cause, the lack of adequate features and contextual informations is the main difficulty.



(a) Remove head false positives (b) Recall suppressed bodies

Figure 1: Effectiveness of JointDet. (a) Remove head false positives: red bounding boxes are the head detection results, and yellow dotted bounding boxes are the removed false positives. (b) Recall missing bodies: red bounding boxes are the human detection results after NMS, and green bounding boxes are the recalled results from suppressed detections.

As for human detection, occlusion is one of the main challenges, especially in the crowded scenes (Zhang et al. 2019b). Some efforts have been made to handle occlusion. The repulsion loss proposed in (Wang et al. 2018) pushes each proposal not only to approach its designated target, but also to keep it away from the other ground truth objects and their corresponding designated proposals. The attention model (Zhang, Yang, and Schiele 2018) employs an attention network employing an attention mechanism across channels with guidances. The Bi-box model (Zhou and Yuan 2018) proposes a network to simultaneously detect pedestrian and estimate occlusion by regressing two bounding

[*]These authors contributed equally to this work.

[†]Corresponding author

boxes for full body and visible part estimation respectively. These methods can alleviate the occlusion issue to some extent. However, while it comes to extremely crowded scenes where the overlaps between humans become large, the Non-Maximum Suppression (NMS) post-process method will result in missing a very large portion of targets, as shown in Figure 1(b).

In this paper, we propose a novel joint head and human detection network, namely JointDet, which detects head and human body simultaneously and performs relational learning between them to improve the performance of both two tasks. As shown in Figure 2, we tile a small quantity of anchors with only one scale and one aspect ratio in each pyramid level to generate head proposals in RPN and then a statistical ratio is applied on head proposals to obtain human body proposals, which significantly accelerates both training and inference. These two classes of proposals are sent into two parallel R-CNN to perform second-stage detection. Moreover, we design a head-body Relationship Discriminating Module (RDM) to predict the relationship between heads and bodies. Since even in extremely crowded scenes, the occlusion between heads is not very serious, we utilize the head location to regain the suppressed body detections. On the other hand, due to the lack of adequate features, head detection usually has false positives on elbows, hands and knees. The proposed post-process strategy also reduces these head false positives via the learned relationship.

As mentioned above, both head and human body annotations are necessary for the proposed method, and only the CrowdHuman dataset (Shao et al. 2018) is publicly available to conduct experiments. To further verify the effectiveness of the proposed model, we annotate head bounding boxes of CityPersons (Zhang, Benenson, and Schiele 2017) and Caltech-USA (Dollár et al. 2009). However, the commonly used 10× training annotations (Zhang et al. 2016b) of Caltech-USA are refined automatically with relatively poor quality, it is hard to annotate head bounding boxes based on the original annotations. Therefore, we also re-annotate Caltech-USA with the full-body bounding-box and the visible-region bounding-box, which serves as a satisfied version of Caltech-USA.

To summarize, this work has five main contributions: 1) proposing an effective framework for joint detection of head and human; 2) designing a RDM to perform relational learning between head and human; 3) introducing a post-process strategy to recall suppressed human detections and reduce head false positives simultaneously; 4) providing the re-annotated body annotations of Caltech-USA, and the head annotations of CityPersons and Caltech-USA; 5) achieving state-of-the-art performance on CrowdHuman, CityPersons and Caltech-USA.

Related Work

Generic Object Detection. Early generic object detectors rely on the sliding window paradigm based on hand-crafted features and classifiers to find objects of interest. In recent years, a new generation of more effective object detectors based on deep convolutional neural network (CNN) significantly improve the state-of-the-art performances, which can be roughly divided into two categories, *i.e.*, the one-stage approach and the two-stage approach. The one-stage approach (Liu et al. 2016; Lin et al. 2017b) directly predicts object class label and regresses object bounding box based on the pre-tiled anchor boxes using deep CNNs. The main advantage of the one-stage approach is its high computational efficiency. In contrast to the one-stage approach, the two-stage approach (Ren et al. 2017; Dai et al. 2016; Lin et al. 2017a) always achieves top accuracy on several benchmarks, which first generates a pool of object proposals by a separated proposal generator, and then predicts the class label, accurate location and size of each proposal.

Head Detection. Early head detectors are used for crowd counting. Merad *et al.* (Merad, Aziz, and Thome 2010) combine positive points of all previous techniques in the head detector. Venkatesh *et al.* (Venkatesh, Descamps, and Carincotte 2012) train a head detector using a cascade of boosted integral features. However, their performance is severely affected under high scene and scale variations because of the usage of handcrafted features. With the arrival of deep learning, some CNN-based methods are proposed. Stewart *et al.* (Stewart, Andriluka, and Ng 2016) introduce a proposal-free head detector that is produced from CNN encoders, where the regression is generally composed of LSTM so that the variable length output prediction is possible. Le *et al.* (Le et al. 2018) introduce a pairwise head detector based on key parts context of the human head and shoulder, and assisted by priority of scene geometry structure. Vu *et al.* (Vu, Osokin, and Laptev 2015) predict the scales and the positions of the head directly from the image, then model the pairwise relationships among the objects. Recently introduction of context information is attractive to improve performance. Some methods (Chen et al. 2016) exploit depth information for head detection with depth images. Nghiem *et al.* (Nghiem, Auvinet, and Meunier 2012) conduct head detection on 3D data as first step for a fall detection system.

Human Detection. One of the key challenges in human detection is occlusion. Several methods (Tian et al. 2015) use part-based model to describe the pedestrian in occlusion handling, which learn a series of part detectors and design some mechanisms to fuse the part detection results to localize partially occluded pedestrians. Besides the part-based model, Zhou *et al.* (Zhou and Yuan 2017) propose to jointly learn part detectors to exploit part correlations and reduce the computational cost. Wang *et al.* (Wang et al. 2018) introduce a novel bounding box regression loss to detect pedestrians in the crowd scenes. Zhang *et al.* (Zhang, Yang, and Schiele 2018) propose to utilize channel-wise attention in convnets allowing the network to learn more representative features for different occlusion patterns in one coherent model. Zhang *et al.* (Zhang et al. 2018a) design an aggregation loss to enforce proposals to be close and locate compactly to the corresponding objects. Zhou *et al.* (Zhou and Yuan 2018) design a method to detect full body and visible part estimation simultaneously to further estimate occlusion. Although numerous pedestrian detection methods (Zhang et al. 2019a) are presented in literature, how to robustly detect each individual human in extremely crowded scenarios is still one of the most critical issues for human detectors.

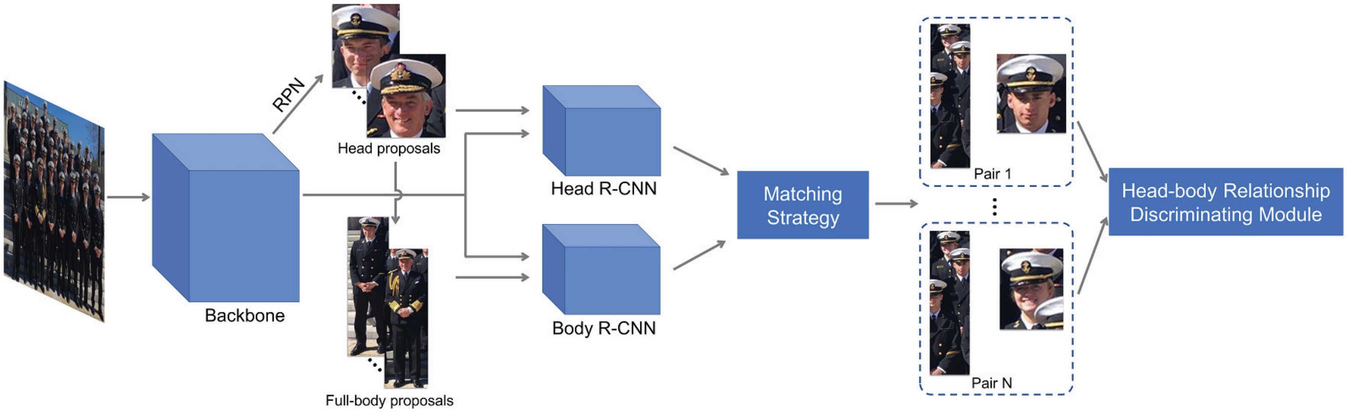


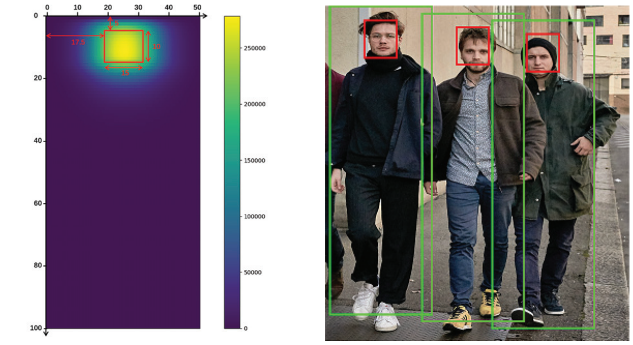
Figure 2: Network structure of JointDet. It consists of RPN, Head R-CNN, Body R-CNN and RDM. RPN only generates head proposals, then a statistical head-body ratio is applied to obtain full-body proposals. After that, head and full-body proposals are sent into two parallel R-CNN branches to obtain temporary results. These temporary results are further processed to get final results as follows: (1) matching them using the proposed strategy to output the matched body-head pairs as Pair 1 to Pair N; (2) extracting corresponding features of each pair for RDM to discriminate their relation (*i.e.*, whether they belong to the same person); (3) according to the learned relationship to reduce head false positives and recall suppressed human detections.

JointDet

Framework Overview

The overall framework is shown in Figure 2. We first utilize RPN to generate head proposals, then apply a statistical head-body ratio on these head proposals to obtain full-body proposals. The specific head-body ratio is shown in Figure 3(a) that is statistically obtained based on all human head-body pairs in the CrowdHuman dataset. After that, the head and full-body proposals are sent into two parallel R-CNN branches, respectively. Since the body proposals are obtained coarsely according to the head proposals, we adopt the cascade training strategy proposed by (Cai and Vasconcelos 2018) for the full-body R-CNN branch to regress more accurate results, where the full-body branch is passed through twice in the training and inference phases. The advantages of this framework can be summarized as follows:

- *A more efficient way to get head and human proposals.* The aspect ratio of head is almost fixed and we just need to tile anchors with one aspect ratio to obtain head proposals. In contrast, the human body has a wide range of aspect ratios because of its deformability and various postures. Tiling anchors to generate human proposals needs to preset a couple of aspect ratios that greatly reduce efficiency. To solve this issue, we use a statistical head-body ratio on head proposals to directly obtain human proposals for free.
- *Decoupling the classification task.* The two parallel R-CNNs only concentrate on detection of one class of object, *i.e.*, head or human body. This design decouples two tasks into separate branches, which is beneficial to make targeted optimization respectively, *e.g.*, using cascade strategy to improve the accuracy of calculated human proposals.



(a) Statistical head-body ratio

(b) Examples

Figure 3: (a) The statistical head-body ratio to calculate the human body proposal based on the head proposal. (b) Red: head proposals; Green: inferred body proposals from head proposals via the statistical head-body ratio.

Relationship Discriminating Module

RDM is designed to learn to discriminate the relationships between the head-body pairs with larger Intersection over Head-box (IoH). The detail expression of IoH is as below:

$$\text{IoH} = \frac{\text{Area of Overlap}}{\text{Area of Head-box}}. \tag{1}$$

The structure of RDM is three stacked fully-connected layers, whose channel setting is same as the classification branch in R-CNN. The process of head-body pair matching and relationship prediction is described in Line 1 to 9 of Algorithm 1. During training, when the matched pair belongs to one person, its ground-truth is 1, otherwise 0. We use the binary cross-entropy loss to optimize RDM. In addition, we set the batch size to be 512, where the proportion of positive and negative examples is set as 1:3.

In the inference phase, we gather the mismatched head detections that have low relationship score or are not matched through IoH, which is demonstrated in Line 10 to 12 of Algorithm 1. There are two situations with the mismatched head detection: 1) The human body corresponding to this head is suppressed by NMS; 2) This head detection is a false positive without a corresponding body. Therefore, we use these mismatched heads and the body detections before NMS to perform matching and relationship discrimination again, as described in Line 14 to 22 of Algorithm 1. If one mismatched head gets strong relationship response in the second time, then we recall the corresponding body detection (described in Line 23 to 25). In contrast, if one mismatched head fails again, then we treat it as a false positive, removing it from the results (described in Line 26 to 28).

Implementation Detail

Anchor Design. At each location of the detection layer, we only associate one specific scale of anchors (*i.e.*, $2S$, where S represents the downsampling factor of the detection layer) and one aspect ratio (*i.e.*, 1.25). In total, there are $A = 1$ anchors per level and they cover the scale range $8 - 128$ pixels across different levels with respect to the input image.

Sample Matching. During training, anchors and proposals need to be divided into positive and negative samples. Specifically, samples are assigned to ground-truth boxes using an IoU threshold of θ_p, and to background if their IoU is in $[0, \theta_n)$. If an anchor is unassigned, which may happen with overlap in $[\theta_n, \theta_p)$, it is ignored during the training phase. We set $\theta_n = 0.3$ and $\theta_p = 0.7$ for the RPN stage, and $\theta_n = 0.5$ and $\theta_p = 0.5$ for the R-CNN stage.

Loss Function. The whole network is optimized by $\mathcal{L} = \mathcal{L}_{\text{RPN}} + \lambda_1 \mathcal{L}_{\text{Head}} + \lambda_2 \mathcal{L}_{\text{Body}} + \lambda_3 \mathcal{L}_{\text{RDM}}$, where $\mathcal{L}_{\text{RPN}}$, $\mathcal{L}_{\text{Head}}$ represent the classification and regression loss of RPN and the head R-CNN branch, which are the same as those proposed in (Ren et al. 2017). $\mathcal{L}_{\text{Body}}$ contains two-stage classification and regression loss as Cascade R-CNN (Cai and Vasconcelos 2018). $\mathcal{L}_{\text{RDM}}$ is the log softmax loss over two classes, which indicates whether the head-body pair belong to one person. The loss weight coefficients λ_1, λ_2 and λ_3 are used to balance different loss terms and we empirically set them as 1 in all the experiments.

Initialization. The backbone network is initialized by the ImageNet (Russakovsky et al. 2015) pretrained ResNet-50. The parameters of newly added layers in RPN are initialized by the normal distribution method, and the parameters in R-CNN are initialized by the MSRA normal distribution method (He et al. 2015).

Optimization. We fine-tune the model using SGD with 0.9 momentum, 0.0001 weight decay. The proposed JointDet is trained on 16 GTX 1080Ti GPUs with a mini-batch 2 per GPU for CrowdHuman and Caltech-USA, and the mini-batch size for Citypersons is 1 per GPU. Each mini-batch involves 512 RoIs per image. Multi-scale training and testing are not applied to ensure fair comparisons with previous methods. We implement JointDet using the PyTorch library. The specific settings of training process for different datasets are described in next sections.

Algorithm 1 Relationship Discriminating Module

```
Require: $\mathcal{H}, \mathcal{B}_1, \mathcal{B}_2, \mathcal{F}, \lambda, \beta_1, \beta_2$
    $\mathcal{H}$ is a set of head detections after NMS
    $\mathcal{B}_1, \mathcal{B}_2$ are a set of body detections before and after NMS
    $\mathcal{F}$ is the feature map of the P2 level in the FPN structure
    $\lambda$ is the IoH threshold while matching
    $\beta_1, \beta_2$ are the low and high relationship score thresholds
Ensure: $\mathcal{D}_h, \mathcal{D}_b$
    $\mathcal{D}_h, \mathcal{D}_b$ are final head and body detections after post process
    /*- - - - - - -Find Mismatched Head for Post-process- - - - -*/
 1: $\mathcal{H}_m \leftarrow \varnothing$ ($\mathcal{H}_m$ is the set of mismatched head detections)
 2: for $h_i \in \mathcal{H}$ do
 3:     $Score \leftarrow \varnothing$
 4:     for $b_j \in \mathcal{B}_2$ do
 5:         if $\text{IoH}(h_i, b_j) > \lambda$ then
 6:             $Feat \leftarrow \text{Concat}(\text{RoIPool}(\mathcal{F}, h_i), \text{RoIPool}(\mathcal{F}, b_j))$
 7:             $Score \leftarrow Score \cup \{\text{RDM}(Feat)\}$
 8:         end if
 9:     end for
10:     if $\max(Score) < \beta_1$ or $Score = \varnothing$ then
11:         $\mathcal{H}_m \leftarrow \mathcal{H}_m \cup h_i$
12:     end if
13: end for
    /*- - - - - - - - - - - - -Post-process Method- - - - - - - - - - - - -*/
14: $\mathcal{D}_h \leftarrow \mathcal{H}, \mathcal{D}_b \leftarrow \mathcal{B}_2$
15: for $h_i \in \mathcal{H}_m$ do
16:     $Score \leftarrow \varnothing$
17:     for $b_j \in \mathcal{B}_1$ do
18:         if $\text{IoH}(h_i, b_j) > \lambda$ then
19:             $Feat \leftarrow \text{Concat}(\text{RoIPool}(\mathcal{F}, h_i), \text{RoIPool}(\mathcal{F}, b_j))$
20:             $Score \leftarrow Score \cup \{\text{RDM}(Feat)\}$
21:         end if
22:     end for
    /*- - - - - - -Recall Suppressed Body Detections- - - - - -*/
23:     if $\max(Score) > \beta_2$ then
24:         $\mathcal{D}_b \leftarrow \mathcal{D}_b \cup$ {$b$ is the body detection with $\max(Score)$}
25:     end if
    /*- - - - - - - -Remove Head False Positives- - - - - - - - -*/
26:     if $\max(Score) < \beta_1$ or $Score = \varnothing$ then
27:         $\mathcal{D}_h \leftarrow \mathcal{D}_h \setminus \{h_i\}$
28:     end if
29: end for
30: return $\mathcal{D}_h, \mathcal{D}_b$
```

Evaluation Metric. Following (Dollár et al. 2009), the log-average miss rate over 9 points ranging from 10^{-2} to 10^{0} FPPI (*i.e.*, MR^{-2}) is used to evaluate the performance of the detectors. We report the detection performance for instances in head and full-body (*i.e.*, human) categories.

Table 1: MR^{-2} performance of different methods on CrowdHuman. Lower MR^{-2} mean better performance.

Source	Method	Head	Human
CrowdHuman	FPN-Head	52.1	-
	FPN-Human	-	50.4
Ours	FPN-Head	48.9	-
	FPN-Human	-	49.7
	FPN-Human-Cascade	-	49.2
	JointDet w/o RDM	48.7	47.0
	JointDet	**48.3**	**46.5**

Experiments

In this section, we perform extensive experiments on the CrowdHuman, CityPersons and Caltech-USA datasets to verify the effectiveness of the proposed framework.

CrowdHuman Dataset

CrowdHuman is a benchmark dataset to better evaluate detectors in crowd scenarios. It is large, rich-annotated, high-diversity and contains $15,000$, $4,370$ and $5,000$ images for training, validation and testing subsets, respectively. There are totally $470k$ human instances from the training and validation subsets, and 22.6 persons per image, with various kinds of occlusions in the dataset. Each human instance is annotated with a head bounding-box, human visible-region bounding-box and human full-body bounding-box. The images and annotations of the training and validation subsets are made freely available to academic for scientific use, while only the images of the testing subset are released and the corresponding annotations are held-out. Since the online evaluation server is not available until now, all our models are trained on the CrowdHuman training subset and tested on the validation subset. During the training phase, the input images are resized so that their short edges are at 800 pixels while the long edges should be no more than 1333 pixels at the same time. We train JointDet with the initial learning rate 0.04 for the first 16 epochs, and decay it by 10 and 100 times for another 6 and 3 epochs.

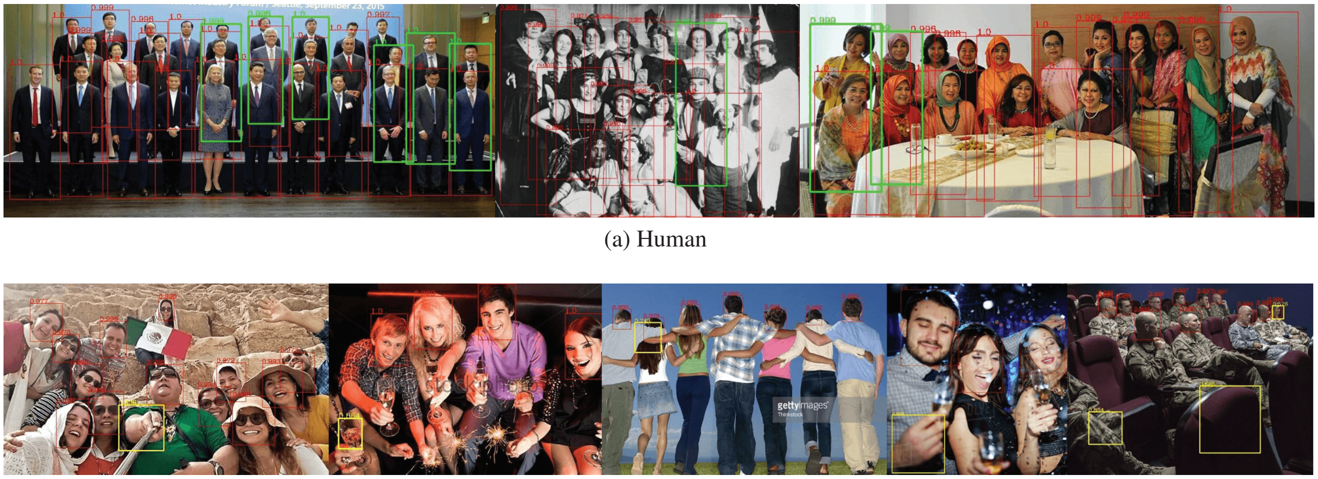
Baseline. Before delving into our proposed framework of joint head and human detection, we first build two strong baselines based on FPN (Lin et al. 2017a) for these two tasks, respectively. We set anchor scale to $2S$ in the head baseline and $8S$ in the full body baseline, where S represents the stride size of each pyramid level. After considering the human body shape, we modify the height *vs.* width ratios of anchors as $\{0.5{:}1, 1{:}1, 2{:}1\}$ for all the experiments related to human detection. While for head detection, the ratios are set to 1.25:1. As shown in Table 1, the baseline of head detection, denoted as FPN-Head, achieves 48.9% MR^{-2} that is 3.2% better than the head detection baseline in CrowdHuman (*i.e.*, 52.1%). And the baseline of human detection, denoted as FPN-Human, obtains 49.7% MR^{-2} that is 0.7% better than the full-body detection baseline in CrowdHuman (*i.e.*, 50.4%). Thus, the detectors trained for the head and human respectively are two strong baselines to verify the effectiveness of our proposed framework.

Ablation Study on Joint Detection. As illustrated in Table 1, after jointing head and human detection in a single detection framework, we achieve 48.7% MR^{-2} for head detection and 47.0% MR^{-2} for human detection. Comparing to the baselines that each task is executed with a separate network, the proposed joint framework not only merges these two tasks into a single network so as to greatly improve the detection efficiency, but also has better MR^{-2} performance, *i.e.*, from 48.9% to 48.7% for head detection and from 49.7% to 47.0% for human detection. The 2.7% improvement on human detection demonstrates the effectiveness of proposed proposal generation method. Notably, we use the cascade training strategy on the full-body R-CNN branch in our joint framework. To have a fair comparison, we train another human detection baseline, denoted as FPN-Human-Cascade, where the R-CNN branch is also passed through twice in the training and inference phases. FPN-Human-Cascade obtains 49.2% MR^{-2}, which still has a large gap with the joint result of human detection. These results demonstrate the effectiveness of the joint framework of head and human detection.

Ablation Study on RDM. The final model of our proposed method is formed by adding the RDM on the joint framework of head and human detection. All the training and testing settings are consistent with previous experiments. The three hyperparameter is set as below: matching IoH threshold λ is set to 0.7, relationship score thresholds β_1 and β_2 are set to 0.1 and 0.9, respectively. As demonstrated in Table 1, after utilizing the head location information to recall the suppressed human bounding boxes, the MR^{-2} of human detection is improved from 47.0% to 46.5%. The advancement indicates that the proposed RDM does recall some human detections suppressed by NMS as shown in Figure 4(a), making our JointDet robust to heavy occlusion in human detection. On the other hand, using the learned head-body relationship can also reduce some head false positives as shown in Figure 4(b), boosting the MR^{-2} of head detection from 48.7% to 48.3% and allowing our head detector to perform well in complex scenarios.

CityPersons Dataset

CityPersons serves as a widely used benchmark dataset for pedestrian detection, which is built upon the semantic segmentation dataset Cityscapes. It is recorded across 18 different cities in Germany with 3 different seasons and various weather conditions. The dataset includes $5,000$ images ($2,975$ for training, 500 for validation, and $1,525$ for testing) with $\sim 35,000$ manually annotated persons plus $\sim 13,000$ ignore region annotations. Both the bounding boxes and visible parts of pedestrians are provided and there are approximately 7 pedestrians in average per image. **For each annotated pedestrian instance, we additionally label the corresponding head bounding box**. The newly annotated head bounding box is within the scope of the original body bounding box. If the head is partly occluded, the annotators are asked to complete the invisible part. Some illustrations of additional head annotations are shown in Figure 5. The proposed JointDet detector is trained on the train-



(a) Human

(b) Head

Figure 4: Qualitative results of JointDet on CrowdHuman. Red bounding boxes represent original results. Green bounding boxes represent recalled human results via RDM. Yellow bounding boxes represent removed head results via RDM.

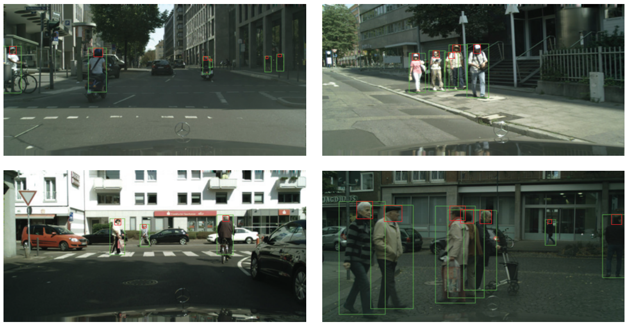


Figure 5: Illustrations of additional head annotations of CityPersons dataset. Green: original pedestrian annotations. Red: additional head annotations.

ing set and evaluated on the validation set. Following the experiment settings in previous works (Wang et al. 2018; Zhang et al. 2018a), we enlarge input images by 1.3 times. The initial learning rate is set to 0.02 for the first 26 epochs, and is decreased to 0.002 and 0.0002 for another 9 and 5 epochs, respectively.

We compare JointDet with TLL(MRF) (Song et al. 2018), Adapted FasterRCNN (Zhang, Benenson, and Schiele 2017), ALFNet (Liu et al. 2018), Repulsion Loss (Wang et al. 2018), PODE+RPN (Zhou and Yuan 2018), OR-CNN (Zhang et al. 2018a) on the CityPersons validation subset in Table 2. Similar with previous works, we evaluate the final model on the Reasonable subset of the CityPersons dataset. The proposed method surpasses all published methods and reduces the MR^{-2} of state-of-the-art results from 11.0% to 10.23% with 0.77% improvement compared with the second best method (Zhang et al. 2018a), demonstrating the superiority of the proposed method in pedestrian detection.

Table 2: MR^{-2} performance on the CityPersons validation set. The scale indicates the enlarge number of original images in training and testing.

Method	Backbone	Scale	*Reasonable*
TLL(MRF)	ResNet-50	-	14.40
Adapted FasterRCNN	VGG-16	×1.3	12.97
ALFNet	VGG-16	×1	12.00
Repulsion Loss	ResNet-50	×1.3	11.60
PODE+RPN	VGG-16	-	11.24
OR-CNN	VGG-16	×1.3	11.00
JointDet (Ours)	ResNet-50	**×1.3**	**10.23**

Caltech-USA Dataset

Caltech-USA is one of the most popular and challenging datasets for pedestrian detection, which comes from approximately 10 hours 30Hz VGA video recorded by a car traversing the streets in the Los Angeles. The training and testing sets contain $42,782$ and $4,024$ frames, respectively. The commonly used $10\times$ training annotations (Zhang et al. 2016b) of Caltech-USA are refined automatically with only $16,376$ poor-quality instances in the training set. **We re-annotate the dataset manually, with a total of** $32,273$ **instances in the training set and** $1,123$ **instances in the testing set**. The labeling rule and method are consistent with the original ones (Dollár et al. 2009). **With the help of new pedestrian annotations, we also label their corresponding head bounding boxes**. Figure 6 shows the comparison of our new annotations and the annotations provided by (Zhang, Benenson, and Schiele 2015). It is obvious that the quality of our new annotations is higher. The detailed analysis of the impact of our new annotations is described below and we use the new annotations to analyze the proposed method in next section. Following the experiment settings in (Wang et al. 2018; Zhang et al. 2018a), we train the proposed method using $2\times$ scale of the image size. The initial

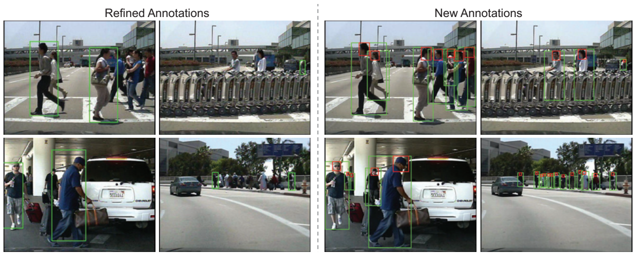


Figure 6: Left: refined annotations provided by Zhang *et al.* Right: our new pedestrian body and head annotations. For simplicity, we do not draw the annotations of visible regions.

learning rate is 0.04 for the first 4 epochs, and is reduced by 10 and 100 times for another 2 and 1 epochs.

Table 3: Effects of different training annotations on different validation annotations of the Caltech-USA dataset. We use the FPN baseline detector for clarification. 'Refined Annotation' indicates the refined annotations by (Zhang et al. 2016b). 'New Annotation' indicates our new annotations.

MR^{-2} Testing / Training	Refined Annotation	New Annotation
Refined annotation	4.31	16.52
New annotation	3.26	14.26

Here we first analyze the effect of our sanitised version of the annotations. As shown in Table 3, using the refined annotations provided by Zhang *et al.* (Zhang et al. 2016b) for training, the FPN detector achieves 4.31% MR^{-2} on the refined testing set. It is reduced to 3.26% with our re-annotated annotations as training set, indicating our new annotations possesses higher quality. When evaluating on our new testing annotations, performances of both detectors drop significantly, *i.e.*, from 4.31% to 16.52% and from 3.26% to 14.26%, which also verify the higher quality of our testing annotations. By statistics, our new annotations have a total of $32,273$ and $1,123$ ground truths in training and testing sets respectively, while the refined version in (Zhang et al. 2016b) only has $16,376$ and 912 instances. Since the benchmark of the original annotations are reaching saturation, our new annotations can serve as a new evaluation metric.

Figure 7 shows the comparison of the JointDet method with other state-of-the-art methods (Cai et al. 2016; Cai, Saberian, and Vasconcelos 2015; Costea and Nedevschi 2016; Du et al. 2017; Li et al. 2018; Mao et al. 2017; Ohn-Bar and Trivedi 2016; Tian et al. 2015; Wang et al. 2018; Zhang et al. 2016a; Zhang, Benenson, and Schiele 2015) on the Caltech-USA refined testing set. All the reported results are evaluated on the widely-used Reasonable subset, which only contains pedestrians with at least 50 pixels tall and occlusion ratio less than 35%. The proposed method outperforms all other methods by producing 2.95% MR^{-2}.

Discussion

Head and Human Annotation. The proposed method requires both head and human annotations, which is feasible in practical applications and academic research with the consideration of the following two aspects: 1) If the human or the heads are annotated, another kind of annotations is easy to obtain via the automatic labelling method (*e.g.*, using a trained head or human detector) or the semi-automatic labelling method (*e.g.*, manual correcting after pre-labelling); 2) We release all new annotations of head and human to facilitate further studies of head and human detection.

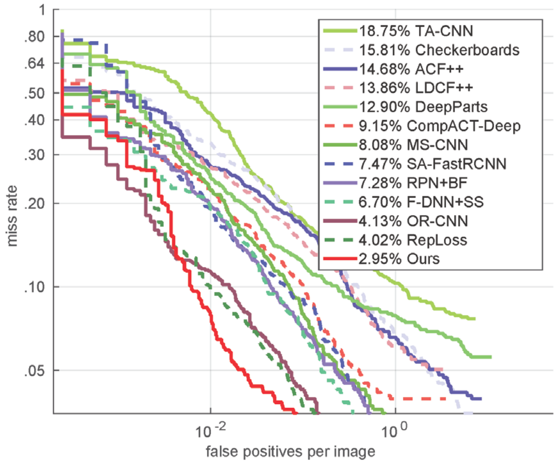


Figure 7: Comparisons with the state-of-the-art methods on the Caltech-USA dataset. The scores in the legend are the MR^{-2} scores of the corresponding methods.

Occluded Head. The proposed method generates the human proposals based on the corresponding head proposals. The results of FPN-Human-Cascade (49.2%) and JointDet w/o RDM (47.0%) in Table 1 have verify that this way of generating human proposals is better than using the RPN proposals. If the head is occluded, it maybe cause some human miss detection but has a slight impact due to: 1) Human with only head occluded is small number case, while occluded body is more common; 2) With help of human body context, RPN can generate proposals for some occluded heads. Thus, the occluded head has ignorable impact and our state-of-the-art human detection performance also confirms the above statement.

Conclusion

In this paper, we have presented a novel joint detection network to detect head and human simultaneously, which utilizes the learned relationship between heads and human bodies to recall the suppressed human detections and reduce head false positives. To sufficiently verify the effectiveness of these proposed components, we have made some efforts in the dataset: 1) providing a better version of Caltech-USA annotations with full body and visible region; 2) annotating the head bounding boxes of CityPersons and Caltech-USA. Consequently, the proposed JointDet detector achieves state-of-the-art performance on CrowdHuman, CityPersons and Caltech-USA. All new annotations, source codes and trained models are public to facilitate further studies.

Acknowledgement

The work is supported by the Natural Science Foundation of China (Grant No.61876178, No.61971399, No.61672519, No.61806196, No.61872367, No.61572501, No.61872367, No.61806203), and Science and Technology Development Fund of Macau (No.0008/2018/A1, No.0025/2019/A1, No.0019/2018/ASC, No.0010/2019/AFJ, No.0025/2019/AKP).

References

Cai, Z., and Vasconcelos, N. 2018. Cascade R-CNN: delving into high quality object detection. In *CVPR*.

Cai, Z.; Fan, Q.; Feris, R. S.; and Vasconcelos, N. 2016. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*.

Cai, Z.; Saberian, M. J.; and Vasconcelos, N. 2015. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*.

Chen, S.; Brémond, F.; Nguyen, H.; and Thomas, H. 2016. Exploring depth information for head detection with depth images. In *AVSS*.

Chen, G.; Cai, X.; Han, H.; Shan, S.; and Chen, X. 2018. Headnet: Pedestrian head detection utilizing body in context. In *FG*.

Costea, A. D., and Nedevschi, S. 2016. Semantic channels for fast pedestrian detection. In *CVPR*.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-FCN: object detection via region-based fully convolutional networks. In *NIPS*.

Dollár, P.; Wojek, C.; Schiele, B.; and Perona, P. 2009. Pedestrian detection: A benchmark. In *CVPR*.

Du, X.; El-Khamy, M.; Lee, J.; and Davis, L. S. 2017. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In *WACV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*.

Le, C.; Ma, H.; Wang, X.; and Li, X. 2018. Key parts context and scene geometry in human head detection. In *ICIP*.

Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; and Yan, S. 2018. Scale-aware fast R-CNN for pedestrian detection. *TMM*.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017a. Feature pyramid networks for object detection. In *CVPR*.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *ICCV*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. E.; Fu, C.; and Berg, A. C. 2016. SSD: single shot multibox detector. In *ECCV*.

Liu, W.; Liao, S.; Hu, W.; Liang, X.; and Chen, X. 2018. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *ECCV*.

Mao, J.; Xiao, T.; Jiang, Y.; and Cao, Z. 2017. What can help pedestrian detection? In *CVPR*.

Merad, D.; Aziz, K.; and Thome, N. 2010. Fast people counting using head detection from skeleton graph. In *AVSS*.

Nghiem, A.; Auvinet, E.; and Meunier, J. 2012. Head detection using kinect camera and its application to fall detection. In *ISSPA*.

Ohn-Bar, E., and Trivedi, M. M. 2016. To boost or not to boost? on the limits of boosted trees for object detection. In *ICPR*.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *TPAMI*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. Imagenet large scale visual recognition challenge. *IJCV*.

Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; and Sun, J. 2018. Crowdhuman: A benchmark for detecting human in a crowd. *CoRR*.

Song, T.; Sun, L.; Xie, D.; Sun, H.; and Pu, S. 2018. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *ECCV*.

Stewart, R.; Andriluka, M.; and Ng, A. Y. 2016. End-to-end people detection in crowded scenes. In *CVPR*.

Tian, Y.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning strong parts for pedestrian detection. In *ICCV*.

Venkatesh, B. S.; Descamps, A.; and Carincotte, C. 2012. Counting people in the crowd using a generic head detector. In *AVSS*.

Vu, T.; Osokin, A.; and Laptev, I. 2015. Context-aware cnns for person head detection. In *ICCV*.

Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; and Shen, C. 2018. Repulsion loss: Detecting pedestrians in a crowd. In *CVPR*.

Zhang, S.; Benenson, R.; and Schiele, B. 2015. Filtered channel features for pedestrian detection. In *CVPR*.

Zhang, S.; Benenson, R.; and Schiele, B. 2017. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*.

Zhang, L.; Lin, L.; Liang, X.; and He, K. 2016a. Is faster R-CNN doing well for pedestrian detection? In *ECCV*.

Zhang, S.; Benenson, R.; Omran, M.; Hosang, J. H.; and Schiele, B. 2016b. How far are we from solving pedestrian detection? In *CVPR*.

Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; and Li, S. Z. 2018a. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In *ECCV*.

Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; and Li, S. Z. 2018b. Single-shot refinement neural network for object detection. In *CVPR*.

Zhang, L.; Liu, Z.; Zhang, S.; Yang, X.; Qiao, H.; Huang, K.; and Hussain, A. 2019a. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*.

Zhang, S.; Xie, Y.; Wan, J.; Xia, H.; Li, S. Z.; and Guo, G. 2019b. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *TMM*.

Zhang, S.; Yang, J.; and Schiele, B. 2018. Occluded pedestrian detection through guided attention in cnns. In *CVPR*.

Zhou, C., and Yuan, J. 2017. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *ICCV*.

Zhou, C., and Yuan, J. 2018. Bi-box regression for pedestrian detection and occlusion estimation. In *ECCV*.