Resisting the Distracting-factors in Pedestrian Detection

Zhe Wang¹, Jun Wang¹, Yezhou Yang²

¹{wangzhewz, wangj203}@buaa.edu.cn  Beihang University
²yz.yang@asu.edu  Arizona State University

Abstract. Pedestrian detection has been heavily studied in the last decade due to its wide applications. Despite incremental progress, several distracting-factors in the aspect of geometry and appearance still remain. In this paper, we first analyze these impeding factors and their effect on the general region-based detection framework. We then present a novel model that is resistant to these factors by incorporating methods that are not solely restricted to pedestrian detection domain. Specifically, to address the geometry distraction, we design a novel coulomb loss as a regulator on bounding box regression, in which proposals are attracted by their target instance and repelled by the adjacent non-target instances. For appearance distraction, we propose an efficient semantic-driven strategy for selecting anchor locations, which can sample informative negative examples at training phase for classification refinement. Our detector can be trained in an end-to-end manner, and achieves consistently high performance on both the Caltech-USA and CityPersons benchmark.²

Keywords: Pedestrian Detection, Object Detection

1 Introduction

Pedestrian detection is a canonical sub-problem of object detection that remains a critical research topic in computer vision and has attracted massive research interest in recent years [1,2,3,4,5,6]. It aims to predict accurate bounding boxes enclosing each pedestrian instance and serves as a key component of various real-world applications such as autonomous driving, robotics, and intelligent video surveillance.

Although promising performances have been achieved, several distracting-factors are still challenging state-of-the-art pedestrian detecting models. Since most detection frameworks adopt region-based approach, the accuracy of region localization and classification directly reflects on detection performance. Therefore, we divide distracting-factors into two categories: geometry and appearance, as illustrated in Figure [1]

¹ Code will be publicly available upon publication.
Fig. 1. Pedestrian detection in the wild. Green boxes represent correct predictions. Red boxes indicate missed targets and misplaced prediction caused by geometry distraction. Blue boxes are the false positive predictions caused by appearance distraction.

Geometry distraction is mainly caused by crowd occlusion (also known as *intra-class* occlusion). It is the most significant barrier for accurate pedestrian detection in the wild and also the major occlusion case in most pedestrian datasets. When pedestrians gather together and occlude each other, detector is prone to be disturbed by the instance that is adjacent to the target and generates bounding boxes among their overlaps (as the left red box in Figure 1). Even worse, during non-maximum suppression (NMS) processing, misplaced boxes with higher confidence scores may suppress the accurate ones or bigger boxes may suppress their neighbouring small ones. At the same time, it also makes detector sensitive to the threshold of NMS as a higher threshold brings in more false positives while a lower threshold leads to more missed detections [7]. Appearance distraction represents the false positives that share similar shape with human body (e.g. pillars, light poles). These objects frequently appear in the common scenarios of real world and we define them as human-like structures. Due to the complicated light condition and variant resolutions, detector is unable to correctly recognize these objects and may assign them with higher probabilities to person rather than background (as the blue boxes in Figure 1).

Several efforts have been made to tackle these two challenges. For the former one, previous methods like [7,8] introduce an extra penalty term on the bounding box regressor to constraint each sampled proposals. But their regularization is incomplete and is likely to get conflict with the original regression function. Others like [9,10] try to mitigate negative impact by refining traditional greedy-NMS [11]. However, the effectiveness of these post-processing methods are restricted by the accuracy of the predictions. For the latter one, several hard/soft-sampling methods [12,13] are proposed to either mine hard negative samples or re-weight each proposal to train the classifier. Both methods are loss-driven, which means they may easily neglect the semantic relations with the foregrounds, that can be useful clues for negative example sampling.

In this paper, we put forward a novel D**i**stra**c**ting-factors R**e**sistant model (dubbed as DR-CNN) based on Faster R-CNN framework [14] to tackle the
aforementioned challenges. For geometry distraction, the key point is to generate accurate bounding boxes in occluded scenes. Inspired by the Coulomb Force \cite{15} between two electric charges, we define Attractive Force between proposals and their target ground truth as well as Repulsive Force between proposals and their non-target ground truth. With this insight, we build a physics modeling and use the energy consumption, calculated by work formula, as the measurement of loss value. The new loss function, termed as Coulomb Loss (CouLoss), works as a regulator that constrains each proposal during regression process. As for appearance distraction, the breakthrough could be achieved by avoiding misclassification of human-like structures. To this end, we propose an efficient anchor location selecting strategy functioning as informative negative examples mining. By adding an extra branch on region proposal network (RPN) \cite{14}, a probability map is yielded and we only process anchors whose probabilities are over a dynamic threshold. These informative negative examples not only cause high loss values but also have semantic relations with pedestrian foregrounds.

To validate the effectiveness of these improvements, we conduct extensive experiments on both Caltech-USA \cite{3} and CityPersons \cite{16} benchmarking datasets. The main contributions are as follows:

- For geometry distracting-factor, we design a new CouLoss on the basis of work formula that serves as a regulator for bounding box regression. It enforces proposals to minimize the intra-region distance as well as to maximize the inter-region distance.
- For appearance distracting-factor, we modify RPN \cite{14} with an extra branch for anchor location selecting, and propose a novel sampling method to capture informative negative examples to train the classifier.
- Experimental results show the superiority of the proposed methods on pedestrian detection benchmarks. We also carry out experiments on PASCAL VOC dataset \cite{17} to validate that our approaches are applicable for other general object detection tasks.

2 Related Work

We briefly review recent work on CNN-based pedestrian detector and discuss related researches on the two target distracting-factors: crowd occlusion and human-like structures.

CNN-Based Pedestrian Detectors. Recently, CNN-based methods have dominated the field of pedestrian detection \cite{18,19,20,21,22,23} and achieved state-of-the-art performance \cite{7,24,8} on several benchmarks: INRIA \cite{25}, ETH \cite{26}, Caltech-USA \cite{3}, and CityPersons \cite{16}. Most of these models adopt region-based approaches where detectors are trained to localize and classify sampled regions. Similar to general object detection, there are also two different frameworks in pedestrian detection. Two-stage framework like \cite{7,24,8} first generate a set of candidate proposals and then sample a small batch of proposals for further bounding box regression and classification. One-stage frameworks like \cite{27,20,28}
directly predict bounding box offsets and class scores from all anchors at each coordinate.

The proposal stage in two-stage framework can rapidly narrow down the number of candidate proposals and provide processed proposals for the following stage, which makes it more suitable for handling distracting-factors in pedestrian detection. With this consideration, we choose Faster R-CNN [14] as our baseline model.

**Crowd occlusion resolving.** Attention models are proposed to improve the feature representation of visible parts. [19] generates scale-aware attention masks in semantic segmentation manner. [22] employs a channel-wise attention mechanism from three different attention modules. Anchor-free methods are used to directly predict bounding boxes of each target. In [29], bounding boxes are learned through a single convolutional neural network. [24] predicts the center points of targets and regresses the height and width of them. Other solutions formulate the issue as a regression problem. To better allocate proposals to each pedestrian, [7] proposes Repulsion Loss to keep proposals away from the non-targeted ground truth and their proposals, while [8] comes up with Aggregation Loss that enforces proposals to locate compactly around each other when they belong to the same target.

Our method shares a common spirit with [7,8] where an extra regulation term is used in loss function to guide proposal regression. While the distinctive part is that we simultaneously consider both attraction and repulsion progress in the extra regulation term, which makes our constrains theoretically more complete than [7,8]. What’s more, we propose a physics framework to unify these two progresses and make them compatible with each other.

**Human-like structures handling.** Multi-classifier is a common structure to refine classification results. [21] employs different patterns that can generate a pool of parts for classifier to choose. [30] trains multiple classifiers in parallel phase and fuse the scores to filter candidates. A set of grid score map from multi-stage is generated by [20] to revise final prediction scores. Methods like [31,12,32,13] balance the region of interest (ROI) to train the classifier. [13] proposes a hard example mining method which only samples negative proposals with high loss values. [12] designs Focal Loss which assigns different weights to all proposals based on their probabilities.

We believe the problem is caused by under-sampling of useful negative examples (foreground-background imbalance). Our strategy is mining human-like structures as negative examples to train the classifier. Different from current sampling methods like [31,12,13] which are loss-driven and [33,32] which are IoU-driven, our method samples regions that have high semantic relativity with pedestrians. We term these regions as informative negative examples since they have higher probabilities to contain human-like structures than others.

### 3 Proposed Approach

We transform the aforementioned critical issues into two specific tasks respectively: bounding box localization of pedestrians in crowds and sampling human-
like structures as negative examples. In this section, we systematically analyze each task and then offer our solution. Besides the benefit to the performance, an important advantage of our method is that we do not increase any computational cost during inference phase.

Specifically, we introduce coulomb loss which is especially designed for crowd scenes in Section 3.1. Then, a novel anchor sampling method is proposed in Section 3.2 to mine informative negative examples. Finally, in Section 3.3 we present the network architecture and the loss function for end-to-end training.

3.1 Coulomb Loss

The ideal detection proposals require not only being close to their target ground truth, but also having minimal intra-region distances and limited maximal inter-region distances, which shares the same goal with visual recognition tasks where both intra-class margin minimization and inter-class margin maximization are required. This insight encourages us to leverage regularization term in loss function for improving localization accuracy of pedestrians in crowds.

Inspired by Coulomb Force, we regard each bounding box as a single charge. Then, we define Attractive Force $F_a$ and Repulsive Force $F_r$ as the interaction between a proposal and its target/non-target ground truth respectively. Suppose $\mathcal{P}_+$ is the set of proposals that has high Intersect over Union (IoU) value (e.g., $IoU > 0.5$) with ground truth. We set proposals $P_p, P_n \in \mathcal{P}_+$, and $G_i, G_j$ are the target ground truth of $P_p$ and $P_n$ respectively. For the convenience of analysis, we form a triplet $\langle G_i, P_p, P_n \rangle$, where $G_i$ is the anchor while $P_p$ and $P_n$ are positive and negative sample respectively.

In physics, work\footnote{https://en.wikipedia.org/wiki/Work_(physics)} is used to measure the energy consumption for moving an object from one place to another. Rationally, we can set this value as the cost of pulling $P_p$ toward $G_i$ or pushing $P_n$ away from $G_i$, which is exactly the loss we need. To utilize the work formula $W = F \cos \theta \cdot s$ for calculating, we build a physics modeling at box-level which will be discussed in details in the following.

First and the most important, we have to define the Force between boxes which is related to their distance. Since IoU is a widely used metric for measuring the closeness between two bounding boxes, we refer to the objective function of IoU Loss \cite{34} and formulate the forces as:

$$F_a = -\ln(IoU(G_i, P_p))$$
$$F_r = -\ln(1 - IoU(G_i, P_n)),$$

Note that the forces only exist when there is an overlap between proposal and ground truth (i.e., $IoU > 0$). From Eq. 1 we can see that the lower closeness between a proposal and its target instance, the stronger Attractive Force will be applied to the proposal, whilst the higher closeness between a proposal and its non-target instance, the stronger Repulsive Force will be applied. Numerically, $-\ln(x)$ gets extreme large when $x$ approaches to 0, which will make the training
Fig. 2. (a) The Attractive and Repulsive Forces between proposal and ground truth. The direction of Attractive Force is always toward the target, while the direction of Repulsive Force may deviate from the target. (b) The proposed DR-RPN module with an extra anchor location branch. The new added branch yields a probability map of the existence of human-shaped structures. During training, a dynamic threshold is used to filter out low-probability regions, and the model only select anchors whose centers fall into the remained regions.

process unstable. Here, we propose a re-sampling strategy for proposals that shares the similar spirit with [35], where we only select proposals whose center points fall into the region of their corresponding ground truth boxes.

Originally, \( \cos \theta \) is introduced because the force may not always has the right direction that moving object toward its target location. In this case, only part of the force is effective. It is also reasonable to follow the same definition in box regression, as illustrated in Figure 2(a). The Attractive Force always pulls \( P_p \) at the correct direction, but the Repulsive Force may push \( P_n \) deviated from its original target when its direction is not on the center line of \((G_i, G_j)\). To handle such case, we introduce Effective Force \( (F^e) \) as the component of original force:

\[
F^e_\text{a} = F_\text{a} \cos \theta, \quad \theta = 0 \\
F^e_\text{r} = F_\text{r} \cos \theta, \quad \theta = \angle P_n G_i G_j,
\]

where \( \cos \theta \) can be calculated by the law of cosines since we have the coordinates of each proposal and ground truth. With Eq. 2, \( F^e_\text{a} \) is defined as the force pulling \( P_p \) toward \( G_i \), and \( F^e_\text{r} \) is the force pushing \( P_n \) toward \( G_j \).

At last, we define \( s \) as the distance between proposal and its target ground truth:

\[
s = \sqrt{(1 - \min(l, r) / W_G/2) \cdot (1 - \min(t, b) / H_G/2)}.
\]

In Eq. 5, \( l, r, t, b \) are the distance of the center point of proposal to the left, right, top, bottom border of its ground truth respectively, as shown in Figure 2(a). A merit of Eq. 5 is that in the crowd scenes where there is an intrinsic overlap between \( G_i \) and \( G_j \), it restricts the repulsion from \( G_i \) when \( P_n \) is close to \( G_j \). This is another crucial difference with RepLoss [7]. Under their definition, the
repulsion is always exist as long as there is an overlap between $G_i$ toward $P_n$ which will push the well-regressed $P_n$ away from $G_j$.

The work value and overall CouLoss is calculated as:

$$W_a^{<G_i,P_p,P_n>} = F_a \cos \theta_a \cdot s^{<G_i,P_p>}$$
$$W_r^{<G_i,P_p,P_n>} = F_r \cos \theta_r \cdot s^{<G_i,P_n>}$$

$$L_{cou} = \frac{1}{|G|} \sum_{G \in G} \sum_{P_p,P_n \in P} (W_a^{<G_i,P_p,P_n>} + W_r^{<G_i,P_p,P_n>})$$

(4)

(5)

It is worth noting that we ignore the cases when $W^{triplet} \leq 0$ in Eq. 4 since they do not make any work that move proposals toward their target locations. Last but not least, this new CouLoss can benefit both RPN and Fast-RCNN modules in Faster R-CNN algorithm.

Beside the physics view, we also interpret Eq. 4 in a measurement angle. Since $F$ is related to the IoU value, it can be regard as a scale calibration between $P$ and $G$; And $s$ can be defined as the distance (similar to the center-ness in [35]) which serves as location calibration from $P$ to $G$; At last, $\cos \theta$ is the dynamic weight for the aggregated value.

### 3.2 Anchor Location Selecting

Human-like structures always act as false positives in pedestrian detection due to the foreground-background class imbalance. This problem is caused by the detection framework. For instance, in RPN [14], since the only sampling principle for negative examples is the IoU with ground truth bounding boxes (e.g., $\text{IoU} < 0.3$), there is a high probability for negative proposals to be sampled in easily distinguished area (e.g., sky and road). Classifier trained with these negative examples will soon converge and lose the ability to learn hard ones. To this end, our solution is trying to mine informative negative examples to train the classifier.

To better sample informative negative examples, we put forward a novel scheme that can erase anchors from easily distinguished areas. As shown in Figure 2(b), an anchor localization branch is added in RPN module which can yield a probability map representing the existence of human-shaped structures (including humans and human-like structures) at each coordinate. The root mean square value of the probability map is set as a dynamic threshold ($\epsilon_a$) during the training phase to reserve valuable regions which contain informative negative examples. It is worth noting that, though our modification to RPN is similar to the changes in [37], these two models share completely different designing goals. [37] is trying to generate accurate bounding boxes for foregrounds by learnable shape and location, while we propose to sample negative proposals that have high scores on location confidence map. Our setting is based on the fact that human-like structures usually have similar feature representations with humans. Therefore, there exists strong semantic relations between them, and we make use of these relations as the clues to mine informative negative examples.
To train the anchor localization branch, we employ the ground-truth bounding boxes to generate a binary score map where 1 indicates selected location and 0 indicates the rest. In specific, we categorize three types of regions on each score map as shown in Figure 3.

1. Positive region (PR). We define the areas of visible bounding boxes \((x_v, y_v, w_v, h_v)\) as PR, since these parts provide the most valuable semantic information.

2. Ignored region (IR). The non-visible part is generated by excluding visible part from full-body bounding boxes \(((x_f, y_f, w_f, h_f))\). We mark this area \(((x_f, y_f, w_f, h_f)) - ((x_v, y_v, w_v, h_v))\) as ignored region. These regions are harmful to classifier, because proposals in IR might be labeled as positive but without any human feature representations (see Figure 4(b) in [23] for further details).

3. Negative region (NR). The rest part of the score map only contains background information and is regarded as NR.

The proposed anchor selecting strategy rapidly narrows down the searching space of generated anchor (e.g. \(\sim 100k\)) to a small scale (e.g. \(\sim 10k\)). As shown in Figure 5, it can effectively filter out the low-probability regions and select the anchors that have strong semantic relations with foregrounds (e.g. human-like structures) as negative examples. Please note that we only use this strategy during training phase and we can drop the anchor localization branch for computational cost saving during inference phase.

3.3 Network Architecture

Our DR-CNN detector follows the implementation of Faster R-CNN [38] and uses VGG-16 [39] as the backbone. To better fulfill pedestrian detection task, the detector is modified following the settings in [16].

The final loss function is jointly optimized with the following losses:

\[
L = L_{ori} + \alpha L_{rpn_{cou}} + \beta L_{rcnn_{cou}} + \gamma L_{loc},
\]

where \(L_{ori}\) represents the original classification and regression loss in both RPN and Fast-RCNN modules. \(L_{rpn_{cou}}\) and \(L_{rcnn_{cou}}\) are the extra regularization term for
regression, and $L_{loc}$ is the Focal Loss for training binary classification for anchor location selecting. Coefficients $\alpha$, $\beta$, and $\gamma$ are the hyperparameters used to balance auxiliary losses.

4 Experiments

4.1 Experimental Setting

Datasets. We conduct experiments on two benchmarks: Caltech-USA and CityPersons. Both benchmarks contain annotations for the visible areas. We use Caltech-USA10x which samples 42,782 frames and 4,024 frames as training and testing datasets respectively. The refined annotation provided by [6] is used in related experiments. CityPersons is a more challenging dataset derived from Cityscapes. It includes 5,000 images in total and 2,975, 500, 1,525 images for training, validation and testing respectively.

Implementation details. As a common convention, we horizontally flip training images for pre-processing. The Adam solver with 0.0001 weight decay is adopted to optimize the network on 1 Nvidia TITAN GPU. A mini-batch involves 2 image per GPU for computational resource constraint. We set the base learning rate set to 0.0001 and train the network for 16 epochs and 12 epochs on Caltech-USA and CityPersons respectively. Hyperparameters $\alpha$, $\beta$, and $\gamma$ are empirically set to 1. We only select ground-truth pedestrian examples with height $\geq 50$ pixels and set the rest as ignored examples for training purpose.

Evaluation protocols. The models are evaluated by log-average miss rate ($\text{MR}^{-2}$), which is the average value over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$. The lower value represents better pedestrian detection performance. To further evaluate performances in occluded scenes, pedestrian instances are divided into bare, partial, heavy subset, representing visible ratio $[0.9, 1]$, $[0.65, 0.9)$, $[0.2, 0.65)$ respectively.

4.2 Comparisons with State-of-the-art Methods

Result on CityPersons dataset. We compare our DR-CNN with state-of-the-art pedestrian detection frameworks, including FRCNN, RepLoss, OR-CNN, ATT-part, Bi-Box, MGAN, ALFNet, TLL and CSP on CityPersons validation set. It is noticing that existing pedestrian detection methods employ different detection framework and backbone, and set different input scale, so we also list these components in Table for fair comparison.

The performance results are summarized in Table. It is evident that our model achieves best performance on Reasonable subset, e.g. outperforming the second best results by a margin of 0.6%. Comparing with CSP, which is the current best region-based one-stage detector, our DR-CNN improves the MR$^{-2}$ on Reasonable subset from 11.0% to 10.4%. It is worth mentioning that the extra anchor location branch in DR-RPN is removable during inference, which makes
Table 1. Pedestrian detection results on CityPersons validation set. ALS is the short form of Anchor Location Selecting. All models are trained on the trainset. We use MR$^{-2}$ as the performance comparing each detectors. The best and the second best are highlighted in red and blue.

| Method       | Framework | Scale | Reasonable | Heavy | Partial | Bare |
|--------------|-----------|-------|------------|-------|---------|------|
| ATT-part [22]| VGG-16    | ×1    | 16.0       | 56.7  | -       | -    |
| TLL [43]     | ResNet-50 | ×1    | 15.5       | 53.6  | 17.2    | 10.0 |
| FRCNN [16]   | VGG-16    | ×1    | 12.9       | 50.5  | -       | -    |
| ALFNet [27]  | ResNet-50 | ×1    | 12.0       | 51.9  | 11.4    | 8.4  |
| RepLoss [7]  | ResNet-50 | ×1.3  | 11.6       | 55.3  | 14.8    | 7.0  |
| MGAN [42]    | VGG-16    | ×1    | 11.5       | 51.7  | -       | -    |
| Bi-Box [23]  | VGG-16    | ×1.3  | 11.2       | 44.2  | -       | -    |
| OR-CNN [8]   | VGG-16    | ×1.3  | 11.0       | 51.3  | 13.7    | 5.9  |
| CSP [24]     | ResNet-50 | ×1    | 11.0       | 49.3  | 10.4    | 7.3  |

| Method | CouLoss | ALS | Scale | Reasonable | Heavy | Partial | Bare |
|--------|---------|-----|-------|------------|-------|---------|------|
| Ours   | √       |     | two-stage | ×1.3 | 12.7 | 54.2 | 14.4 | 7.3 |
|        | √       |     |         |       | 10.5 | 51.9 | 11.3 | 5.8 |
|        | √ √     |     |         |       | 11.1 | 51.2 | 12.2 | 5.9 |
|        | √ √     |     |         |       | 10.4 | 46.9 | 10.7 | 5.8 |

We can observe that our DR-CNN surpasses these two models by 2.5% / 1.2% on Reasonable subset and 3.6% / 8.4% on Heavy subset. Models like OR-CNN [8], ATT-part [22], Bi-Box [23], MGAN [42] modify the network architecture in the second stage which lead to better performance under occlusion cases. Our DR-CNN achieves 46.9% on Heavy subset, which is competitive with these models.

Result on Caltech-USA dataset. We conduct extensively comparison with recent methods, including DeepParts [21], RPN+BF [5], MS-CNN [44], SDS-RCNN [18], ATT-part [22], RepLoss [7], Bi-Box [23], and CSP [24]. The results are mainly compared on three occlusion settings: Reasonable, Heavy, and All, where All represents the visible ratio is larger than 0.2.

As shown in Table 2, our DR-CNN achieves superior results comparing with most of the models and performs competitively with state-of-the-art method. Specifically, on Reasonable subset, our model surpasses Bi-Box [23] by a margin of 2.7% but slightly falls behind CSP [24] by 0.4%. Comparing with RepLoss [7], MR$^{-2}$ on Heavy and All subsets reduce from 47.9% to 45.5% and 59.0% to 57.0% respectively.

4.3 Ablation Study

We carry out comprehensive ablation studies on CityPersons dataset to evaluate the contribution of different model components and the training configurations. Coulomb loss. As shown in Table 3, we denote DR-CNN-A as the detector that uses CouLoss as embedded regularization on original regression loss in the
Table 2. Pedestrian detection results on Caltech-USA test set. It is worth mentioning that all models are directly trained on Caltech-USA. We use MR\(^{-2}\) as the performance to compare each detectors. The best and the second best are highlighted in red and blue.

| Method         | Reasonable | Heavy | All  |
|----------------|------------|-------|------|
| DeepParts [21] | 11.9       | 60.4  | 64.8 |
| ATT-part [22]  | 10.3       | 45.2  | 54.5 |
| MS-CNN [44]    | 10.0       | 59.9  | 60.9 |
| RPN+BF [5]     | 9.6        | 74.4  | 64.7 |
| Bi-Box [23]    | 7.6        | 44.4  | -    |
| SDS-RCNN [18]  | 7.4        | 58.6  | 61.5 |
| RepLoss [7]    | 5.0        | 47.9  | 59.0 |
| CSP [24]       | 4.5        | 45.8  | 56.9 |
| Ours           | 4.9        | 45.5  | 57.0 |

Table 3. Comparison between CouLoss with other loss functions on CityPersons. The fourth row represents the baseline model, the fifth and sixth row represent using CouLoss only in RPN stage and only in Fast-RCNN stage respectively.

| Model        | Reasonable | Heavy | Partial | Bare |
|--------------|------------|-------|---------|------|
| IoULoss [34] | 12.4       | 52.0  | 12.6    | 6.9  |
| RepLoss [7]  | 11.6       | 55.3  | 14.8    | 7.0  |
| AggLoss [8]  | 11.4       | 52.6  | 13.8    | 6.2  |

| α | β | Reasonable | Heavy | Partial | Bare |
|---|---|------------|-------|---------|------|
| 0 | 0 | 12.7       | 54.2  | 14.4    | 7.3  |
| 1 | 0 | 10.9       | 53.0  | 11.5    | 5.8  |
| 0 | 1 | 11.0       | 53.8  | 11.7    | 6.0  |
| 0.3| 0.7 | 10.5       | 51.9  | 11.3    | 5.8  |

baseline detector. Different α, β represents different combination of CouLoss on both RPN and Fast-RCNN modules. Comparing the detection result in Table 3 we can observe that when using CouLoss, the MR\(^{-2}\) on four subsets decrease greatly from baseline by a margin of 2.2%, 2.3%, 3.1%, and 1.5% respectively. The results of using CouLoss at different stages show that CouLoss can benefit both stages by better aligning proposals around their ground truth. We train our DR-CNN-A for several rounds and find the best combination of CouLoss when setting α = 0.3, β = 0.7. What’s more, we also study the effectiveness of Attractive Force and Repulsion Force separately, and observe 1.1%, 1.4% improvement on baseline on Reasonable subset respectively.

It is worth mentioning that our CouLoss is superior to the two state-of-the-art methods using AggLoss [8] and RepLoss [7] which also serves as a regularization term on regression function. This proves that the design guideline of CouLoss is more complete and suitable for bounding box regression. We also compare the results with IoULoss [8] since our CouLoss is based on the form of IoULoss. The
result show that the promotions are marginally related with using the form of IoU Loss.

Since CouLoss can pull proposals to their target ground truths and push them away from non-target ones, DR-CNN-A becomes less sensitive to the NMS threshold. To demonstrate this point, we present the miss rate with CouLoss across various NMS threshold at FPPI = 10^{-2}. As mentioned in Section 1, a high NMS threshold may lead to more false positives, while a low NMS threshold may lead to more false negatives. In Figure 4(a), DR-CNN-A always produces lower miss rate than baseline. It is noteworthy that the curve of DR-CNN-A is smoother than that of baseline, indicating that changing NMS threshold has less impact on DR-CNN-A. In addition, we also visualize the predicted bounding boxes before NMS in crowd scenes in Figure 4(b). Compared with baseline, the predictions of DR-CNN-A locate compactly around ground truths and there are fewer proposals lying in the overlaps between adjacent pedestrians.

**Anchor location selecting.** We first demonstrate the effectiveness of DR-RPN architecture by constructing a detector that uses proposed DR-RPN instead of original RPN in baseline. The comparison results are reported in Table 4. In the second row, we set $\epsilon_a = 0$ to bias the anchor location selecting process (labeled as w/o selecting). This model outperforms baseline on MR$^{-2}$ by a margin of 1.0% on Heavy subset, indicating that the extra anchor location branch is helpful in centering anchors around foregrounds. When introducing anchor location selecting in place of sliding window anchoring method to DR-RPN, DR-CNN-B processes further improvements of 2.0% on Reasonable subset. This proves that the proposed anchor sampling strategy can filter out less-informative proposals from negative examples. Since the proposed method is specially designed for
Table 4. Validation of the necessity of anchor location selecting process and ignored region (IR) in DR-CNN-B. The results are reported in the form of $\text{MR}^{-2}/\text{FPPI}$. The FPPI is calculated under $\text{MR} = 0.1$.

| Model               | Reasonable | Heavy | Partial | Bare  |
|---------------------|------------|-------|---------|-------|
| Baseline            | 12.7/0.22  | 54.2/0.65 | 14.4/0.24 | 7.3/0.04 |
| DR-CNN-B w/o selecting | 11.7/0.16  | 53.2/0.54 | 11.7/0.20 | 6.9/0.04 |
| DR-CNN-B w/o IR     | 11.9/0.17  | 54.0/0.53 | 12.7/0.20 | 6.2/0.03 |
| DR-CNN-B w selecting/IR | 11.1/0.14  | 51.2/0.48 | 12.2/0.20 | 5.9/0.02 |

Fig. 5. The visualization of the generated probability maps and the selected anchors sampled by proposed method. We can observe that the probability maps in (b) are highly correlated with human-shape structures in (a), which leads the selected anchors to concentrate more on these objects as shown in (c).

false positive cases which can’t be clearly reflected on $\text{MR}^{-2}$, we use FPPI to evaluate the model and observe consistent lower score on all subsets.

Results mentioned above are also supported by the visualization results shown in Figure 5, where we present input images, the generated probability maps and the selected anchors sampled by the proposed method. It can be seen that the probability maps in Figure 5(b) are highly correlated with the human-shape structures, which leads the selected anchors to concentrate more on these objects as shown in Figures 5(c).

An additional experiment is done to validate the necessity of introducing ignored region (IR) when training the anchor location branch. As shown in the third row in Table 4, model trained with without IR performs consistently worse on all subsets. This is mainly because it miss-labels proposals as positive examples when they are largely occupied by non-visible parts, as discussed in Section 3.2.

5 Extension: Results on PASCAL VOC

In this section, we extend the application of our proposed methods to reveal its universality. The modifications are applied on general object detection application which also suffers from occluded scenes and false positive examples.
Table 5. General object detection results on PASCAL VOC 2007 test set. Note that these results are evaluated under COCO metrics which are different from the original VOC metrics.

| Method                               | \(AP\) | \(AP_{50}\) | \(AP_{75}\) |
|--------------------------------------|--------|-------------|-------------|
| Faster R-CNN                         | 49.2   | 77.2        | 53.8        |
| + CouLoss                            | +0.8   | +0.2        | +1.0        |
| + Anchor location selecting          | +0.7   | +0.2        | +1.1        |
| + All                                | +2.3   | +0.5        | +3.7        |

Our experiments are performed on PASCAL VOC dataset [17], which is a common benchmark for general objection detection. We employ Faster R-CNN with ResNet-101 [45] as the backbone for baseline detector. The model is trained on the training and validation sets of PASCAL VOC 2007 and PASCAL VOC 2012, and is tested on the testing set of PASCAL VOC 2007. To evaluate high quality detection results from our methods, we use the COCO metrics for evaluation. The results in Table 5 show that the proposed methods have significant improvements on general object detection task, especially under high IoU threshold, which demonstrates the universality of the proposed methods.

6 Conclusion

In this paper, we put forward a novel DR-CNN framework to tackle geometry and appearance distracting-factors in pedestrian detection, i.e. crowd occlusion and human-like structures. We transform these factors into two specific tasks respectively: bounding box localization of pedestrians in crowds and sampling human-like structures as negative examples, and devise two general methods to approach them. For geometry distraction, we design a new loss function, termed as CouLoss, to regulate the process of bounding box regression. Specifically, we build a physics framework to unify the proposed Attractive Force and Repulsive Force which can pull proposals towards their target ground truths and push proposals away from non-target ones respectively. For appearance distraction, an efficient semantic-driven strategy for selecting anchor locations is introduced, which can sample human-like structures as informative negative examples at training phase for classification refinement. It is worth mentioning that both methods don’t increase any computational cost during inference.

Our model is trained in an end-to-end fashion and achieves competitive performance on two widely adopted benchmarking datasets, i.e. Caltech-USA and CityPersons. Detailed ablation experiments have demonstrated the effectiveness of each proposed approach respectively. More importantly, the promising preliminary results on PASCAL VOC show that our methods could also be adopted towards other appearance-based object detection tasks.

---

3 The annotations of PASCAL VOC are transformed to COCO format and COCO toolbox is used for evaluation.
References

1. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: European Conference on Computer Vision, Springer (2014) 613–627
2. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. (2009)
3. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 304–311
4. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 6034–6043
5. Zhang, L., Lin, L., Liang, X., He, K.: Is faster r-cnn doing well for pedestrian detection? In: European Conference on Computer Vision, Springer (2016) 443–457
6. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1259–1267
7. 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. (2018) 7774–7783
8. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Occlusion-aware r-cnn: detecting pedestrians in a crowd. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 637–653
9. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms – improving object detection with one line of code. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
10. Liu, S., Huang, D., Wang, Y.: Adaptive nms: Refining pedestrian detection in a crowd. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2019)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
13. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 761–769
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
15. Halliday, D., Resnick, R., Walker, J.: Fundamentals of physics. John Wiley & Sons (2013)
16. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 1. (2017) 3
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision 88(2) (2010) 303–338
18. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. arXiv preprint arXiv:1706.08564 (2017)

19. Lin, C., Lu, J., Wang, G., Zhou, J.: Graininess-aware deep feature learning for pedestrian detection. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 732–747

20. 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. (2018) 966–974

21. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE international conference on computer vision. (2015) 1904–1912

22. 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. (2018) 6995–7003

23. Zhou, C., Yuan, J.: Bi-box regression for pedestrian detection and occlusion estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 135–151

24. Wei Liu, Shengcai Liao, W.R.W.H.Y.Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)

25. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1., IEEE (2005) 886–893

26. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8

27. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: The European Conference on Computer Vision (ECCV). (September 2018)

28. Song, T., Sun, L., Xie, D., Sun, H., Pu, S.: Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: The European Conference on Computer Vision (ECCV). (September 2018)

29. Huang, L., Yang, Y., Deng, Y., Yu, Y.: Densebox: Unifying landmark localization with end to end object detection. arXiv preprint arXiv:1509.04874 (2015)

30. Du, X., El-Khamy, M., Lee, J., Davis, L.: Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. In: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, IEEE (2017) 953–961

31. Li, B., Liu, Y., Wang, X.: Gradient harmonized single-stage detector. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8577–8584

32. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra r-cnn: Towards balanced learning for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 821–830

33. Cao, Y., Chen, K., Loy, C.C., Lin, D.: Prime sample attention in object detection. arXiv preprint arXiv:1904.04821 (2019)

34. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: Proceedings of the 2016 ACM on Multimedia Conference, ACM (2016) 516–520

35. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: The IEEE International Conference on Computer Vision (ICCV). (October 2019)
36. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. (2015) 1440–1448

37. Wang, J., Chen, K., Yang, S., Loy, C.C., Lin, D.: Region proposal by guided anchoring. arXiv preprint arXiv:1901.03278 (2019)

38. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MM Detection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)

39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

40. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. (2017) 2980–2988

41. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 3213–3223

42. Pang, Y., Xie, J., Khan, M.H., Anwer, R.M., Khan, F.S., Shao, L.: Mask-guided attention network for occluded pedestrian detection. In: The IEEE International Conference on Computer Vision (ICCV). (October 2019)

43. Song, T., Sun, L., Xie, D., Sun, H., Pu, S.: Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 536–551

44. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: European conference on computer vision, Springer (2016) 354–370

45. 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. (2016) 770–778