Repulsion Loss: Detecting Pedestrians in a Crowd

Xinlong Wang[1*]   Tete Xiao[2*]   Yuning Jiang[3]   Shuai Shao[3]   Jian Sun[3]   Chunhua Shen[4]

[1]Tongji University
1452405wxl@tongji.edu.cn

[2]Peking University
jasonhsiao97@pku.edu.cn

[3]Megvii, Inc.
jyn, shaoshuai, sunjian@megvii.com

[4]The University of Adelaide
chunhua.shen@adelaide.edu.au

Abstract

Detecting individual pedestrians in a crowd remains a challenging problem since the pedestrians often gather together and occlude each other in real-world scenarios. In this paper, we first explore how a state-of-the-art pedestrian detector is harmed by crowd occlusion via experimentation, providing insights into the crowd occlusion problem. Then, we propose a novel bounding box regression loss specifically designed for crowd scenes, termed repulsion loss. This loss is driven by two motivations: the attraction by target, and the repulsion by other surrounding objects. The repulsion term prevents the proposal from shifting to surrounding objects thus leading to more crowd-robust localization. Our detector trained by repulsion loss outperforms the state-of-the-art methods with a significant improvement in occlusion cases.

1. Introduction

Occlusion remains one of the most significant challenges in object detection although great progress has been made in recent years [10, 9, 24, 19, 1, 20, 11, 3]. In general, occlusion can be divided into two groups: *inter-class occlusion* and *intra-class occlusion*. The former one occurs when an object is occluded by stuff or objects of other categories, while the latter one, also referred to as *crowd occlusion*, occurs when an object is occluded by objects of the same category.

In pedestrian detection [31, 14, 6, 5, 7, 21], crowd occlusion constitutes the majority of occlusion cases. The reason is that in application scenarios of pedestrian detection, *e.g.*, video surveillance and autonomous driving, pedestrians often gather together and occlude each other. For instance, in the CityPersons dataset [33], there are a

Figure 1. Illustration of our proposed repulsion loss. The repulsion loss consists of two parts: the attraction term to narrow the gap between a proposal and its designated target, as well as the repulsion term to distance it from the surrounding non-target objects.

total of $3,157$ pedestrian annotations in the validation subset, among which 48.8% of them overlap with another annotated pedestrian whose Intersection over Union (IoU) is above 0.1. Moreover, 26.4% of all pedestrians have considerable overlaps with another annotated pedestrian whose IoU is above 0.3. The highly frequent crowd occlusion severely harms the performance of pedestrian detectors.

The main impact of crowd occlusion is that it significantly increases the difficulty in pedestrian localization. For example, when a target pedestrian T is overlapped by another pedestrian B, the detector is apt to get confused since these two pedestrians have similar appearance features. As a result, the predicted boxes which should have bounded T will probably shift to B, leading to inaccurate localization. Even worse, as the primary detection results are required to be further processed by non-maximum suppression (NMS), shifted bounding boxes originally from T may be suppressed by the predicted boxes of B, in which T turns into a missed detection. That is, crowd occlusion makes the detector sensitive to the threshold of NMS: a higher threshold brings in more false positives while a

---

lower threshold leads to more missed detections. Such undesirable behaviors also harms most instance segmentation frameworks [11, 18], since they also require accurate detection results. Therefore, how to robustly localize each individual person in crowd scenes is one of the most critical issues for pedestrian detectors.

In state-of-the-art detection frameworks [9, 24, 3, 19], the bounding box regression technique is employed for object localization, in which a regressor is trained to narrow the gap between proposals and ground-truth boxes measured by some kind of distance metrics (*e.g.*, Smooth_{L1} or IoU). Nevertheless, existing methods only require the proposal to get close to its designated target, without taking the surrounding objects into consideration. As shown in Figure 1, in the standard bounding box regression loss, there is no additional penalty for the predicted box when it shifts to the surrounding objects. This observation makes one wonder *whether the locations of its surrounding objects could be taken into account if we want to detect a target in a crowd?*

Inspired by the characteristics of a magnet, *i.e.*, *magnets attract and repel*, in this paper we propose a novel localization technique, referred to as repulsion loss (RepLoss). With RepLoss, each proposal is required not only to approach its designated target T, but also to keep away from the other ground-truth objects as well as the other proposals whose designated targets are not T. In other words, the bounding box regressor with RepLoss is driven by two motivations: attraction by the target and repulsion by other surrounding objects and proposals. For example, as demonstrated in Figure 1, the red bounding box shifting to B will be given an additional penalty since it overlaps with a surrounding non-target object. Thus, RepLoss can prevent the predicted bounding box from shifting to adjacent overlapped objects effectively, which makes the detector more robust to crowd scenes. Our main contributions are as follows:

- We first experimentally study the impact of crowd occlusion on pedestrian detection. Specifically, on the CityPersons benchmark [33] we analyze both false positives and missed detections caused by crowd occlusion quantitatively, which provides important insights into the crowd occlusion problem.
- Two types of repulsion losses are proposed to address the crowd occlusion problem, namely RepGT Loss and RepBox Loss. RepGT Loss directly penalizes the predicted box for shifting to the other ground-truth objects, while RepBox Loss requires each predicted box to keep away from the other predicted boxes with different designated targets, making the detection results less sensitive to NMS.
- With the proposed repulsion losses, a crowd-robust pedestrian detector is trained end-to-end, which out-performs all the state-of-the-art methods on both CityPerson and Caltech-USA benchmarks [7]. It should also be noted that the detector with repulsion loss significantly improves the detection accuracy for occlusion cases, highlighting the effectiveness of repulsion loss. Furthermore, our experiments on the PASCAL VOC [8] detection dataset show that the RepLoss is also beneficial for general object detection, besides pedestrians.

2. Related Work

Object Localization. With the recent development of convolutional neural networks (CNNs) [16, 26, 12], great progress has been made in object detection, in which object localization is generally framed as a regression problem that relocates an initial proposal to its designated target. In R-CNN [10], a linear regression model is trained with respect to the Euclidean distance of coordinates of a proposal and its target. In [9], the Smooth_{L1} Loss is proposed to replace the Euclidean distance used in R-CNN for bounding box regression. [24] proposes the region proposal network (RPN), in which bounding box regression is performed twice to transform predefined anchors into final detection boxes. Densebox [15] proposes an anchor-free, fully convolutional detection framework. IoU Loss is proposed in [29] to maximize the IoU between a ground-truth box and a predicted box. We note that a method proposed by Desai *et al.* [4] also exploits the attraction and repulsion between objects to capture the spatial arrangements of various object classes, still, it is to address the problem of object classification via a global model. In this work, we will demonstrate the effectiveness of the Repulsion Loss for object localization in crowd scenes.

Pedestrian Detection. Pedestrian detection is the first and an critical step for many real-world applications. Traditional pedestrian detectors, such as `ACF` [5], `LDCF` [22] and `Checkerboard` [32], exploit various filters on Integral Channel Features (`IDF`) [6] with sliding window strategy to localize each target. Recently, the CNN-based detectors [17, 30, 21, 14, 28] show great potential in dominating the field of pedestrian detection. In [28, 30], features from a Deep Neural Network rather than hand-crafted features are fed into a boosted decision forest. [21] proposes a multi-task trained network to further improve detection performance. Also in [23, 27, 34], a part-based model is utilized to handle occluded pedestrians. [13] works on improving the robustness of NMS, but it ends up relying on an additional network for post-processing. In fact, few of previous works focus on studying and overcoming the impact of crowd occlusion.
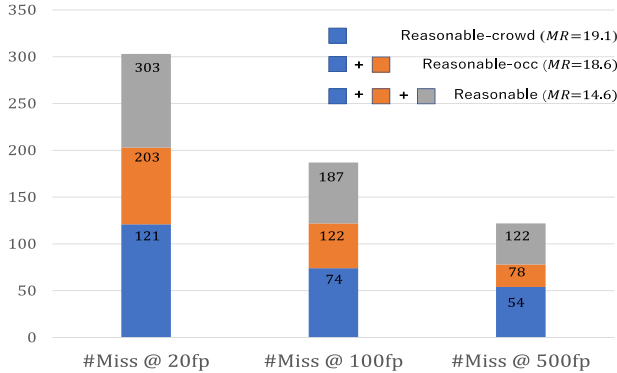
Figure 2. Missed detection numbers and MR^{-2} scores of our baseline on the reasonable, reasonable-occ, reasonable-crowd subsets. Of all missed detection in reasonable-occ subset, crowd occlusion accounts for ~60%, making it a main obstacle for addressing occlusion issues.

3. What is the Impact of Crowd Occlusion?

To provide insights into the crowd occlusion problem, in this section, we experimentally study how much crowd occlusion influences pedestrian detection results. Before delving into our analysis, first we introduce the dataset and the baseline detector that we use.

3.1. Preliminaries

Dataset and Evaluation Metrics. CityPersons [33] is a new pedestrian detection dataset on top of the semantic segmentation dataset CityScapes [2], of which $5,000$ images are captured in several cities in Germany. A total of $\sim35,000$ persons with an additional $\sim13,000$ ignored regions, both bounding box annotation of all persons and annotation of visible parts are provided. All of our experiments involved CityPersons are conducted on the *reasonable* train/validation sets for training and testing, respectively. For evaluation, the log miss rate is averaged over the false positive per image (FPPI) range of $[10^{-2},\ 10^0]$ (MR^{-2}) is used (lower is better).

Detector. Our baseline detector is the commonly used Faster R-CNN [24] detector modified for pedestrian detection, generally following the settings in Zhang *et al.* [31] and Mao *et al.* [21]. The difference between our implementation and theirs is that we replace the VGG-16 backbone with the faster and lighter ResNet-50 [12] network. It is worth noting that ResNet is rarely used in pedestrian detection, since the down-sampling rate at convolution layers is too large for the network to detect and localize small pedestrians. To handle this, we use dilated convolution and the final feature map is $1/8$ of input size. The ResNet-based detector achieves $14.6\,\mathrm{MR}^{-2}$ on the validation set, which is sightly better than the reported result ($15.4\,\mathrm{MR}^{-2}$) in [33].
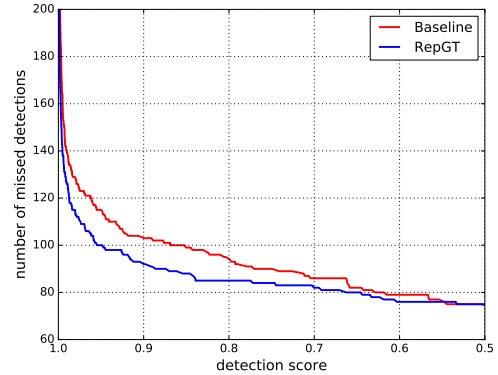
(a)
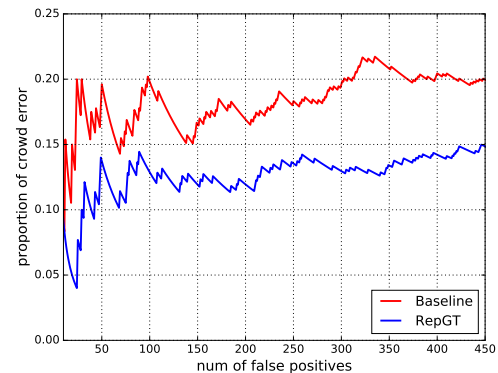
(b)

Figure 3. Errors analysis of our baseline and RepGT. (a) The number of missed detections in reasonable-crowd subset under different detection scores. (b) The proportion of false positives caused by crowd occlusion of all false positives. RepGT Loss effectively reduces missed detections and false positives caused by crowd occlusion.

3.2. Analysis on Failure Cases

Missed Detections. With the results of the baseline detector, we first analyze missed detections caused by crowd occlusion. Since the bounding box annotation of the visible part of each pedestrian is provided in CityPersons, the occlusion can be calculated as $occ \triangleq 1 - \frac{area(BBox_{visible})}{area(BBox)}$. We define a ground-truth pedestrian whose $occ \geq 0.1$ as an occlusion case, and one whose $occ \geq 0.1$ and $IoU \geq 0.1$ with any other annotated pedestrian as a crowd occlusion case. By definition, from the total $1,579$ non-ignored pedestrian annotations in the reasonable validation set, two subsets are extracted: the *reasonable-occ* subset, consisting of 810 occlusion cases (51.3%) and the *reasonable-crowd* subset, consisting of 479 crowd occlusion cases (30.3%). Obviously the reasonable-crowd subset is also a subset of reasonable-occ subset.

In Figure 2, we report the numbers of missed detections and MR^{-2} on the reasonable, reasonable-occ and

reasonable-crowd subsets. We observe that the performance drops significantly from 14.6 MR^{-2} on the reasonable set to 18.6 MR^{-2} on the reasonable-occ subset; of all missed detections at 20, 100, and 500 false positives, occlusion makes up approximately 60%, indicating that it is a main factor which harms the performance of the baseline detector. Of missed detections in the reasonable-occ subset, the proportion of crowd occlusion stands at nearly 60%, making it a main obstacle for addressing occlusion issues in pedestrian detection. Moreover, the miss rate on the reasonable-crowd subset (19.1) is even higher than the reasonable-occ subset (18.6), indicating that crowd occlusion is an even harder problem than inter-class occlusions; when we lower the threshold from 100 to 500 false positives, the portion of missed detections caused by crowd occlusion becomes larger (from 60.7% to 69.2%). It implies that missed detections caused by crowd occlusion are hard to be rescued by lowering the threshold.

In Figure 3(a), the red line shows how many ground-truth pedestrians are missed in the reasonable-crowd subset with different detection scores. As in real-world applications, only predicted bounding boxes with high confidence will be considered, the large number of missed detections on the top of the curve implies we are far from saturation for real-world applications.

False Positives. We also analyze how many false positives are caused by crowd occlusion. We cluster all false positives into three categories: background, localization and crowd error. A background error occurs when a predicted bounding box has $IoU < 0.1$ with any ground-truth pedestrian, while a localization error has $IoU \geq 0.1$ with only one ground-truth pedestrian. Crowd errors are those who have $IoU \geq 0.1$ with at least two ground-truth pedestrians.

After that we count the number of crowd errors and calculate its proportion of all false positives. The red line in Figure 3(b) shows that crowd errors contribute to a relative large proportion (about 20%) of all false positives. Through visualization in Figure 4, we observe that the crowd errors usually occur when a predict box shifts slightly or dramatically to neighboring non-target ground-truth objects, or bounds the union of several overlapping ground-truth objects together. Moreover, the crowd errors usually have relatively high confidences thus leading to top-ranked false positives. It indicates that to improve the robustness of detectors to crowd scenes, more discriminative loss is needed when performing bounding box regression. More visualization examples can be found in supplementary material.

Conclusion. The analysis on failure cases validates our observation: pedestrian detectors are surprisingly tainted by crowd occlusion, as it constitutes the majority of missed detections and results in more false positives by increasing the difficulty in localization. To address these issues, in Sec-
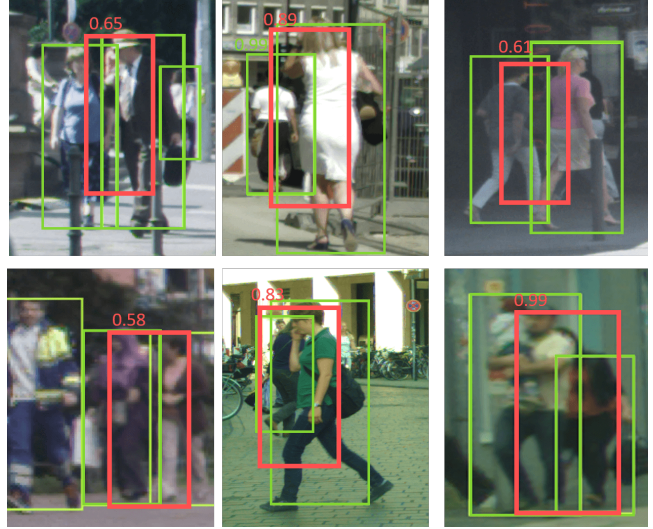

Figure 4. The visualization examples of the crowd errors. Green boxes are correct predicted bounding boxes, while red boxes are false positives caused by crowd occlusion. The confidence scores outputted by detectors are also attached. The errors usually occur when a predict box shifts slightly or dramatically to neighboring ground-truth object (*e.g.*, top-right one), or bounds the union of several overlapping ground-truth objects (*e.g.*, bottom-right one).

tion 4, the repulsion loss is proposed to improve the robustness of pedestrian detectors to crowd scenes.

4. Repulsion Loss

In this section we introduce the repulsion loss to address the crowd occlusion problem in detection. Inspired by the characteristics of magnet, *i.e.*, *magnets attract and repel*, the Repulsion Loss is made up of three components, defined as:

$$L = L_{\mathrm{Attr}} + \alpha * L_{\mathrm{RepGT}} + \beta * L_{\mathrm{RepBox}}, \qquad (1)$$

where L_{Attr} is the *attraction* term which requires a predicted box to approach its designated target, while L_{RepGT} and L_{RepBox} are the *repulsion* terms which require a predicted box to keep away from other surrounding ground-truth objects and other predicted boxes with different designated targets, respectively. Coefficients α and β act as the weights to balance auxiliary losses.

For simplicity we consider only two-class detection in the following, assuming all ground-truth objects are from the same category. Let $P = (l_P, t_P, w_P, h_P)$ and $G = (l_G, t_G, w_G, h_G)$ be the proposal bounding box and ground-truth bounding box which are represented by their coordinates of left-top points as well as their widths and heights, respectively. $\mathcal{P}_+ = \{P\}$ is the set of all positive proposals (those who have a high IoU (*e.g.*, $IoU \geq 0.5$) with at least one ground-truth box are regarded as positive samples, while negative samples otherwise), and $\mathcal{G} = \{G\}$ is the set of all ground-truth boxes in one image.

Attraction Term. With the objective to narrow the gap between predicted boxes and ground-truth boxes measured by some kind of distance metrics[1], *e.g.*, Euclidean distance [10], Smooth_{L1} distance [9] or IoU [29], attraction loss has been commonly adopted in existing bounding box regression techniques. To make a fair comparison, in this paper we adopt Smooth_{L1} distance for the attraction term as in [21, 33]. We set smooth parameter in Smooth_{L1} as 2. Given a proposal $P \in \mathcal{P}_+$, we assign the ground-truth box who has the maximum IoU as its designated target: $G^P_{Attr} = \arg\max_{G \in \mathcal{G}} IoU(G, P)$. B^P is the predicted box regressed from proposal P. Then the attraction loss could be calculated as:

$$L_{\mathrm{Attr}} = \frac{\sum_{P \in \mathcal{P}_+} \mathrm{Smooth}_{L1}(B^P, G^P_{Attr})}{|\mathcal{P}_+|}. \quad (2)$$

Repulsion Term (RepGT). The RepGT Loss is designed to repel a proposal from its neighboring ground-truth objects which are not its target. Given a proposal $P \in \mathcal{P}_+$, its repulsion ground-truth object is defined as the ground-truth object with which it has the largest IoU region except its designated target:

$$G^P_{Rep} = \arg\max_{G \in \mathcal{G}\backslash\{G^P_{Attr}\}} IoU(G, P). \quad (3)$$

Inspired by IoU Loss in [29], the RepGT Loss is calculated to penalize the overlap between B^P and G^P_{Rep}. The overlap between B^P and G^P_{Rep} is defined by Intersection over *Ground-truth* (IoG): $IoG(B, G) \triangleq \frac{area(B \cap G)}{area(G)}$. As $IoG(B, G) \in [0, 1]$, we define RepGT Loss as:

$$L_{\mathrm{RepGT}} = \frac{\sum_{P \in \mathcal{P}_+} \mathrm{Smooth}_{ln}\left(IoG(B^P, G^P_{Rep})\right)}{|\mathcal{P}_+|}, \quad (4)$$

where

$$\mathrm{Smooth}_{ln} = \begin{cases} -\ln(1-x) & x \le \sigma \\ \dfrac{x-\sigma}{1-\sigma} - \ln(1-\sigma) & x > \sigma \end{cases} \quad (5)$$

is a smoothed $\ln$ function which is continuously differentiable in $(0, 1)$, and $\sigma \in [0, 1)$ is the smooth parameter to adjust the sensitiveness of the repulsion loss to the outliers. Figure 5 shows its curve with different σ. From Eqn. 4 and Eqn. 5 we can see that the more a proposal tends to overlap with a non-target ground-truth object, a larger penalty will be added to the bounding box regressor by the RepGT Loss. In this way, the RepGT Loss could effectively stop a predicted bounding box from shifting to its neighboring objects which are not its target.

---

[1]Here the distance is simply a measurement of difference of two bounding boxes. It may not satisfy triangle inequality.
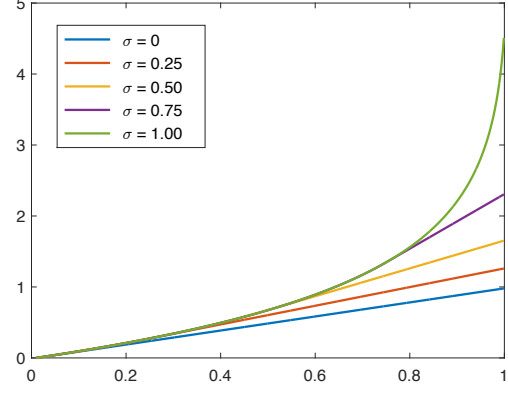

Figure 5. The curves of Smooth_{ln} under different smooth parameter σ. The smaller σ is, the less sensitive loss is to the outliers.

Repulsion Term (RepBox). NMS is a necessary post-processing step in most detection frameworks to merge the primary predicted bounding boxes which are supposed to bound the same object. However, the detection results will be affected significantly by NMS especially for the crowd cases. To make the detector less sensitive to NMS, we further propose the RepBox Loss whose objective is to repel each proposal from others with different designated targets. We divide the proposal set $\mathcal{P}_+$ into $|\mathcal{G}|$ mutually disjoint subsets based on the target of each proposal: $\mathcal{P}_+ = \mathcal{P}_1 \cap \mathcal{P}_2 \cap \ldots \cap \mathcal{P}_{|\mathcal{G}|}$. Then for two proposals randomly sampled from two different subsets, $P_i \in \mathcal{P}_i$ and $P_j \in \mathcal{P}_j$ where $i, j = 1, 2, \ldots, |\mathcal{G}|$ and $i \ne j$, we expect that the overlap of predicted box B^{P_i} and B^{P_j} will be as small as possible. Therefore, the RepBox Loss is calculated as:

$$L_{\mathrm{RepBox}} = \frac{\sum_{i \ne j} \mathrm{Smooth}_{ln}\left(IoU(B^{P_i}, B^{P_j})\right)}{\sum_{i \ne j} \mathbb{1}[IoU(B^{P_i}, B^{P_j}) > 0] + \epsilon}, \quad (6)$$

where $\mathbb{1}$ is the identity function and ϵ is a small constant in case divided by zero. From Eqn. 6 we can see that to minimize the RepBox Loss, the IoU region between two predicted boxes with different designated targets needs to be small. That means, the RepBox Loss is able to reduce the probability that the predicted bounding boxes with different regression targets are merged into one after NMS, which makes the detector more robust to the crowd scenes.

4.1. Discussion

Distance Metric. It is worth noting that we choose the IoG or IoU rather than Smooth_{L1} metric to measure the distance between two bounding boxes in the repulsion term. The reason is that the values of IoG and IoU are bounded in range $[0, 1]$ while Smooth_{L1} metric is boundless, *i.e.*, if we use Smooth_{L1} metric in the repulsion term, in the RepGT Loss for example, it will require the predicted box to keep away from its repulsion ground-truth object as far as possible. On

the contrary, IoG criteria only requires the predicted box to minimize the overlap with its repulsion ground-truth object, which better fits our motivation.

In addition, IoG is adopted in RepGT Loss rather than IoU because, in the IoU-based loss, the bounding box regressor may learn to minimize the loss by simply enlarging the bounding box size to increase the denominator $area(B^P \cup G^P_{Rep})$. Therefore, we choose IoG whose denominator is a constant for a particular ground-truth object to minimize the overlap $area(B^P \cap G^P_{Rep})$ directly.

Smooth Parameter σ. Compared to [29] which directly uses $-\ln(IoU)$ as loss function, we introduce a smoothed ln function Smooth_{ln} and a smooth parameter σ in both RepGT Loss and RepBox Loss. As shown in Figure 5, we can adjust the sensitiveness of the repulsion loss to the outliers (the pair of boxes with large overlap) by the smooth parameter σ. Since the predicted boxes are much denser than the ground-truth boxes, a pair of two predicted boxes are more likely to have a larger overlap than a pair of one predicted box and one ground-truth box. It means that there will be more outliers in RepBox than in RepGT. So, intuitively, RepBox Loss should be less-sensitive to outliers (with small σ) than RepGT Loss. More detailed studies about the smooth parameter σ as well as the auxiliary loss weights α and β are provided Section 5.2.

5. Experiments

The experiment section is organized as follows: we first introduce the basic experiment settings as well as the implementation details of repulsion loss in Section 5.1; then the proposed RepGT Loss and RepBox Loss are evaluated and analyzed on the CityPersons [33] benchmark respectively in Section 5.2; finally, in Section 5.3, the detector with repulsion loss is compared with the state-of-the-art methods side-by-side on both CityPersons [33] and Caltech-USA [7].

5.1. Experiment Settings

Datasets. Besides the CityPersons [33] benchmark introduced in Section 3, we also carry out experiments on the Caltech-USA dataset [7]. As one of several predominant datasets and benchmarks for pedestrian detection, Caltech-USA has witnessed inspiring progress in this field. A total of 2.5-hour video is divided into training and testing subsets with $42,500$ frames and $4,024$ frames respectively. In [31], Zhang *et al.* provide refined annotations, in which training data are refined automatically while testing data are meticulously re-annotated by human annotators. We conduct all experiments related to Caltech-USA on the new annotations unless otherwise stated.

Training Details. Our framework is implemented on our self-built fast and flexible deep learning platform. We train

	MR^{-2}			Improvement		
σ	0	0.5	1.0	0	0.5	1.0
RepGT	14.3	14.5	**13.7**	+0.3	+0.1	**+0.9**
RepBox	**13.7**	14.2	14.3	**+0.9**	+0.4	+0.3

Table 1. The MR^{-2} of RepGT and RepBox Losses and their improvements with different smooth parameters σ on the validation set of CityPersons.

α (RepGT)	0.3	0.4	0.5	0.6	0.7
β (RepBox)	0.7	0.6	0.5	0.4	0.3
MR^{-2}	13.9	13.9	**13.2**	13.3	14.1

Table 2. We balance the RepGT and RepBox Losses by adjusting the weights α and β. Empirically, $\alpha = 0.5$ and $\beta = 0.5$ yields the best performance. The results are obtained on CityPersons validation subset.

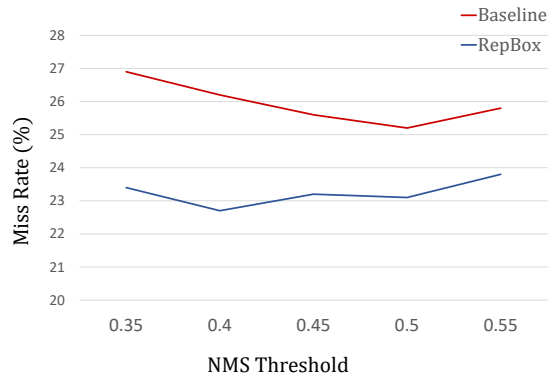Figure 6. Results with RepBox Loss across various NMS thresholds at $\mathrm{FPPI} = 10^{-2}$. The curve of RepBox is smoother than that of baseline, indicating it is less sensitive to the NMS threshold.

the network for $80k$ iterations and $160k$ iterations, with the base learning rate set to 0.016 and decreased by a factor of 10 after the first $60k$ and $120k$ iterations for CityPersons and Caltech-USA, respectively. The Stochastic Gradient Descent (SGD) solver is adopted to optimize the network on 4 GPUs. A mini-batch involves 1 image per GPU. Weight decay and momentum are set to 0.0001 and 0.9. Multi-scale training/testing are not applied to ensure fair comparisons with previous methods. For Caltech-USA, we use the 10x set ($\sim 42k$ frames) for training. Online Hard Example Mining (OHEM) [25] is used to accelerate convergence.

5.2. Ablation Study

RepGT Loss. In Table 1, we report the results of RepGT Loss with different parameter σ for Smooth_{ln} loss. When set σ as 1.0, adding RepGT Loss yields the best performance of $13.7 \ \mathrm{MR}^{-2}$ in terms of reasonable evaluation setup. It outperforms the baseline with an improvement of $0.9 \ \mathrm{MR}^{-2}$. Setting $\sigma = 1$ that means we directly sum over

Method	+RepGT	+RepBox	+Segmentation	Scale	Reasonable	Heavy	Partial	Bare
Zhang *et al.* [33]				×1	15.4	55.0	18.9	9.3
			✓	×1	14.8	-	-	-
				×1.3	12.8	-	-	-
Baseline				×1	14.6	60.6	18.6	7.9
RepLoss	✓			×1	13.7	57.5	17.3	7.2
		✓		×1	13.7	59.1	17.2	7.8
	✓	✓		×1	13.2	56.9	16.8	7.6
	✓	✓		×1.3	11.6	55.3	14.8	7.0
	✓	✓		×1.5	10.9	52.9	13.4	6.3

Table 3. Pedestrian detection results using RepLoss evaluated on the CityPersons [33]. Models are trained on train set and tested on validation set. We use ResNet-50 as our back-bone architecture. The best 3 results are highlighted in red, blue and green, respectively.

$-\ln(1 - IoG)$ with no smooth at all, similar to the loss function used in IoU Loss [29].

We also provide comparisons on missed detections and false positives between RepGT and baseline. In Figure 3(a), adding RepGT Loss effectively decreases the number of missed detections in the reasonable-crowd subset. The curve of RepGT is consistently lower than that of baseline when the threshold of detection score is rather high, but two curves agree when the score is at 0.5. The saturation points of curves are both at ~ 0.9, also a commonly used threshold for real applications, where we reduce the quantity of missed detections by relatively 10%. In Figure 3(b), false positives produced by RepGT Loss due to crowd occlusion cover less proportion than the baseline detector. This demonstrates that RepGT Loss is effective on reducing missed detections and false positives in crowd scenes.

RepBox Loss. For RepBox Loss, we experiment with a different smooth parameter σ, reported in the fourth line of Table 1. When setting σ as 0, RepBox Loss yields the best performance of 13.7 MR^{-2}, on par with RepGT with $\sigma = 1.0$. Setting σ as 0 means we completely smooth a $\ln$ function into a linear function and sum over IoU. We conjure that RepBox Loss tends to have more outliers than RepGT Loss since predicted boxes are much denser than ground-truth boxes.

As mentioned in Section 1, detectors in crowd scenes are sensitive to the NMS threshold. A high NMS threshold may lead to more false positives, while a low NMS threshold may lead to more missed detections. In Figure 6 we show our results with RepBox Loss across various NMS thresholds at FPPI $= 10^{-2}$. In general, the performance of detector with RepBox Loss is smoother than baseline. It is worth noting that at the NMS threshold of 0.35, the gap between baseline and RepBox is 3.5 points, indicating that the latter is less sensitive to NMS threshold. Through visualization in Figure 7, there are fewer predictions lying in between two adjacent ground-truths of RepBox, which

Method	Reasonable	
	IoU=0.5	IoU=0.75
Zhang *et al.* [33]	5.8	30.6
Mao *et al.* [21]	5.5	43.4
Zhang *et al.* [33]*	5.1	25.8
Baseline	5.6	28.7
+RepGT	5.0	27.1
+RepBox	5.3	26.2
+RepGT & RepBox	5.0	26.3
+RepGT & RepBox*	4.0	23.0

Table 4. Results on Calech-USA test set (reasonable), evaluated on the new annotations [31]. On a strong baseline, we further improve the state-of-the-art to a remarkable 4.0 MR^{-2} under 0.5 IoU threshold. The consistent gain when increasing IoU threshold to 0.75 demonstrates effectiveness of repulsion loss. *: indicates pre-training network using CityPersons dataset.

is desirable in crowd scenes. More examples are shown in supplementary material.

Balance of RepGT and RepBox The introduced RepGT and RepBox Loss help detectors do better in crowd scenes when added alone, but we have yet studied how to balance these two losses. Table 2 shows our results with different settings of α and β. Empirically, $\alpha = 0.5$ and $\beta = 0.5$ yields the best performance.

5.3. Comparisons with State-of-the-art Methods

To demonstrate our effectiveness under different occlusion levels, we divide the reasonable subset (occlusion $\leq$ 35%) into the *reasonable-partial* subset ($10\% <$ occlusion $\leq 35\%$), denoted as Partial subset, and the *reasonable-bare* subset (occlusion $\leq 10\%$), denoted as Bare subset. For annotations whose occlusion is above 35% (not in the reasonable set), we denote them as Heavy subset. Table 3 summarizes our results on CityPersons. In general, RepGT Loss and RepBox Loss show improvement across all evaluation

(a) Input	(b) Baseline	(c) +RepBox

Figure 7. Visualized comparison of predicted bounding boxes before NMS of baseline and RepBox. In the results of RepBox, there are fewer predictions lying in between two adjacent ground-truths, which is desirable in crowd scenes. More examples are shown in supplementary material.
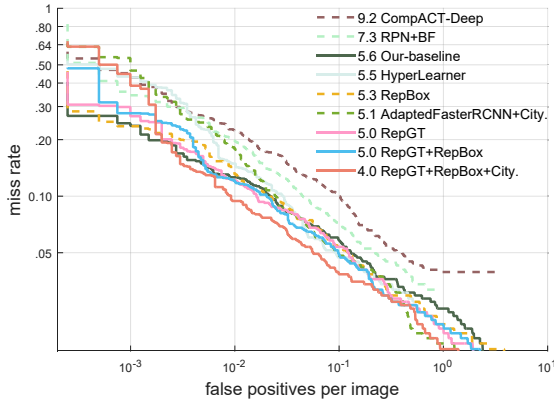

Figure 8. Comparisons with state-of-the-art methods on the new Caltech test subset.

subsets. Combined together, our proposed repulsion loss achieves 13.2 MR^{-2}, which is an absolute 1.4-point improvement over our baseline. In terms of different occlusion levels, performance with RepLoss on the Heavy subset is boosted by a remarkably large margin of 3.7 points, and on the Partial subset by a relatively smaller margin of 1.8 points, while causing non-obvious improvement on the Bare subset. It is in accordance with our intention that RepLoss is specifically designed to address the occlusion problem.

We also evaluate RepLoss on new Caltech-USA dataset. Results are shown in Table 4. On a strong reference, RepLoss achieves MR^{-2} of 5.0 at .5 IoU matching threshold and 26.3 at .75 IoU matching threshold. The consistent and even larger gain when increasing IoU threshold demonstrates the ability of our framework to handle occlusion problem, for it that occlusion is known for its tendency of being more sensitive at a higher matching threshold. Result curves are shown in Figure 8.

6. Extensions: General Object Detection

Our RepLoss is a generic loss function for object detection in crowd scenes and can be used in applications other

Method	mAP	mAP on Crowd
Faster R-CNN [12]	76.4	-
Faster R-CNN (*ReIm*)	79.5	38.7
+ RepGT	**79.8**	**40.8**

Table 5. General object detection results evaluated on PASCAL VOC 2007 [8] benchmark. *ReIm* is our re-implemented Faster R-CNN. Crowd subset contains ground-truth objects who has overlaps above 0.1 IoU region with at least another ground-truth object of the same category. Our RepGT Loss outperforms baseline by 2.1 mAP on crowd subset.

than pedestrian detection. In this section, we apply the repulsion loss to general object detection.

We conduct our experiments on the PASCAL VOC dataset [8], a common evaluation benchmark for general object detection. This dataset consists of over 20 object categories. Standard evaluation metric for VOC dataset is mean Average Precision (mAP) over all categories. We adopt the vanilla Faster R-CNN [24] framework, using ImageNet-pretrained ResNet-101 [12] as the backbone. The NMS threshold is set as 0.3. The model is trained on the train and validation subsets of PASCAL VOC 2007 and PASCAL VOC 2012, and is evaluated on the test subset of PASCAL VOC 2007. Our re-implemented baseline is better than original one by 3.4 mAP.

Results are shown in Table 5. The gain over the entire dataset is not significant. Nevertheless, when evaluated on the crowd subset (objects have intra-class IoU greater than 0.1), RepLoss outperforms the baseline by 2.1 mAP. These results demonstrate that our method is generic and can be extended to general object detection.

7. Conclusion

In this paper, we have carefully designed the repulsion loss (RepLoss) for pedestrian detection, which improves detection performance, particularly in crowd scenes. The main motivation of the repulsion loss is that the attraction-by-target loss alone may not be sufficient for training an optimal detector, and repulsion-by-surrounding can be very beneficial.

To implement the repulsion energy, we have introduced two types of repulsion losses. We have achieved the best reported performance on two popular datasets: Caltech and CityPersons. Significantly, our result on CityPersons without using pixel annotation outperforms the previously best result [33] that uses pixel annotation by about 2%. Detailed experimental comparison have demonstrated the value of the proposed RepLoss, which improves detection accuracy by a large margin in occlusion scenarios. Results on generic object detection (PASCAL VOC) further show its usefulness. We expect wide application of the proposed loss in many other object detection tasks.

References

[1] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision*, pages 354–370. Springer, 2016. 1

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 3

[3] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 1, 2

[4] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011. 2

[5] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. 1, 2

[6] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. 2009. 1, 2

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009. 1, 2, 6

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 8

[9] R. Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 5

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2, 5

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 8

[13] J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[14] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4073–4082, 2015. 1, 2

[15] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 2

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[17] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 2017. 2

[18] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2359–2367, 2017. 2

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[21] J. Mao, T. Xiao, Y. Jiang, and Z. Cao. What can help pedestrian detection? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 7

[22] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved detection. *arXiv preprint arXiv:1406.1134*, 2014. 2

[23] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3258–3265. IEEE, 2012. 2

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 1, 2, 3, 8

[25] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016. 6

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[27] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1904–1912, 2015. 2

[28] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *Proceedings of the IEEE international conference on computer vision*, pages 82–90, 2015. 2

[29] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang. Unitbox: An advanced object detection network. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 516–520. ACM, 2016. 2, 5, 6, 7

[30] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *European Conference on Computer Vision*, pages 443–457. Springer, 2016. 2

[31] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2016. 1, 3, 6, 7

[32] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1751–1760. IEEE, 2015. 2

[33] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 6, 7, 8

[34] C. Zhou and J. Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3486–3495, 2017. 2

References

[1] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision*, pages 354–370. Springer, 2016. 1

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 3

[3] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 1, 2

[4] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011. 2

[5] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. 1, 2

[6] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. 2009. 1, 2

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009. 1, 2, 6

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 8

[9] R. Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 5

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2, 5

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 8

[13] J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[14] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4073–4082, 2015. 1, 2

[15] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 2

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[17] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 2017. 2

[18] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2359–2367, 2017. 2

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[21] J. Mao, T. Xiao, Y. Jiang, and Z. Cao. What can help pedestrian detection? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 7

[22] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved detection. *arXiv preprint arXiv:1406.1134*, 2014. 2

[23] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3258–3265. IEEE, 2012. 2

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 1, 2, 3, 8

[25] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016. 6

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[27] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1904–1912, 2015. 2

[28] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *Proceedings of the IEEE international conference on computer vision*, pages 82–90, 2015. 2

[29] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang. Unitbox: An advanced object detection network. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 516–520. ACM, 2016. 2, 5, 6, 7

[30] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *European Conference on Computer Vision*, pages 443–457. Springer, 2016. 2

[31] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2016. 1, 3, 6, 7

[32] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1751–1760. IEEE, 2015. 2

[33] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 6, 7, 8

[34] C. Zhou and J. Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3486–3495, 2017. 2

Supplementary material

A. The examples of missed detections and false positives

In Figure 9 we show more examples of missed detections and false positives before and after applying RepLoss. The blue bounding boxes represent false positives, and the red ones represent the missed detections. The examples above the grey dashed line are to demonstrate the effectiveness of proposed RepLoss on *eliminating false positives*, while those bolow the grey dashed line are to demonstrate the effectiveness of proposed RepLoss on *detecting more missed pedestrians.*

B. More examples on CityPersons In Figure 10, we demonstrate more examples on challenging CityPersons dataset. Green bounding boxes are predicted pedestrians whose score (in range $[0, 1.0]$) is at a relatively high threshold (greater than 0.8 in this case).

Figure 9. Comparison of baseline and RepLoss. The blue bounding boxes represent false positives, and the red ones represent the missed detections. On two sides of the grey dashed line, samples on the first row of each side are predictions of our baseline, while samples on the second row of each side are the predictions after adding the RepLoss.

Figure 10. More examples on CityPersons dataset. Green bounding boxes are predicted pedestrians whose score ([0, 1.0]) is greater than 0.8.