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 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.

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.

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

*The work was done when Xinlong Wang and Tete Xiao were interns at Megvii, Inc.

threshold leads to more missed detections. Such undesirable behaviors can harm 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 outperforms 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.

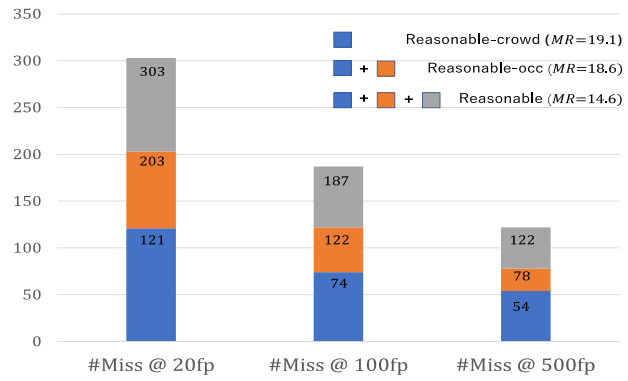


Figure 2. Missed detection numbers and MR^{-2} scores of our baseline on the reasonable, reasonable-occ, reasonable-crowd subsets. Of all missed detection in reasonable-occ subset, crowd occlusion accounts for $\sim$60%, making it a main obstacle for addressing occlusion issues.

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 $\sim 35,000$ persons with an additional $\sim 13,000$ 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].

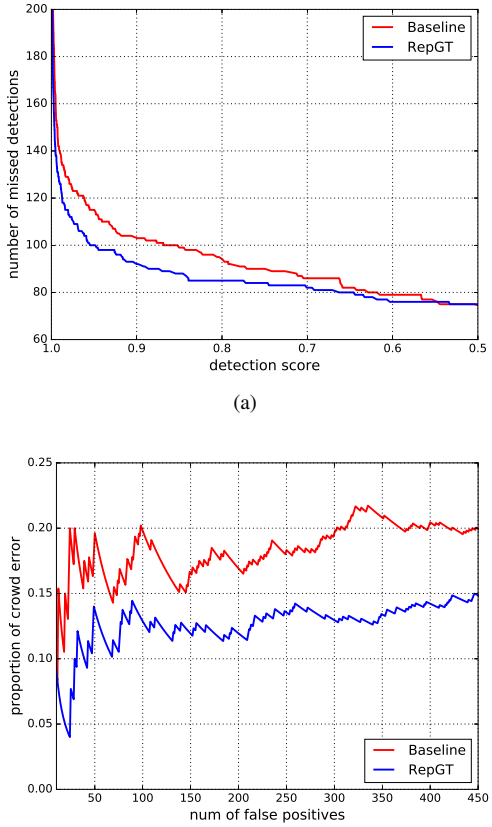


Figure 3. Errors analysis of our baseline and RepGT. (a) The number of missed detections in reasonable-crowd subset under different detection scores. (b) The proportion of false positives caused by crowd occlusion of all false positives. RepGT Loss effectively reduces missed detections and false positives 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 Section 4, the repulsion loss is proposed to improve the robustness of pedestrian detectors to crowd scenes.

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).

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:

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

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:

$$G^P_{Rep} = \underset{G \in \mathcal{G} \setminus \{G^P_{Attr}\}}{\arg\max} \; IoU(G, P). \quad (3)$$

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:

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

where

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

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.

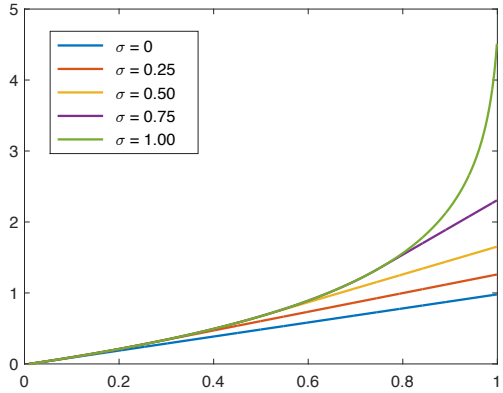


Figure 5. The curves of Smooth_{ln} under different smooth parameter σ. The smaller σ is, the less sensitive loss is to the outliers.

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 $i \neq j$, 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:

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

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 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.

	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.

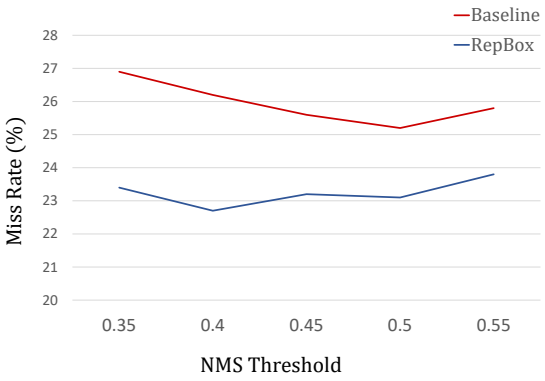


Figure 6. Results with RepBox Loss across various NMS thresholds at $\text{FPPI} = 10^{-2}$. The curve of RepBox is smoother than that of baseline, indicating it is less sensitive to the NMS threshold.

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 MR^{-2} in terms of reasonable evaluation setup. It outperforms the baseline with an improvement of 0.9 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 $\sim$ 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 is desirable in crowd scenes. More examples are shown in supplementary material.

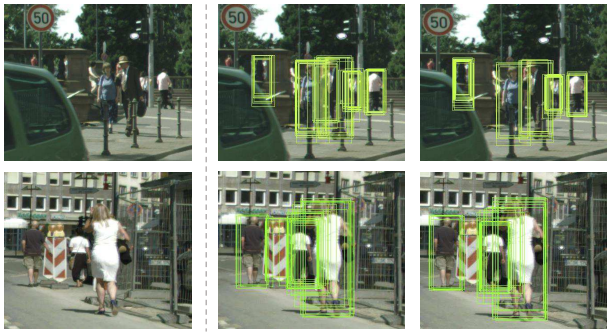
Method	Reasonable	
	IoU=0.5	IoU=0.75
Zhang *et al*. [33]	5.8	30.6
Mao *et al*. [21]	5.5	43.4
Zhang *et al*. [33]*	5.1	25.8
Baseline	5.6	28.7
+RepGT	5.0	27.1
+RepBox	5.3	26.2
+RepGT & RepBox	5.0	26.3
+RepGT & RepBox*	4.0	23.0

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.

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.

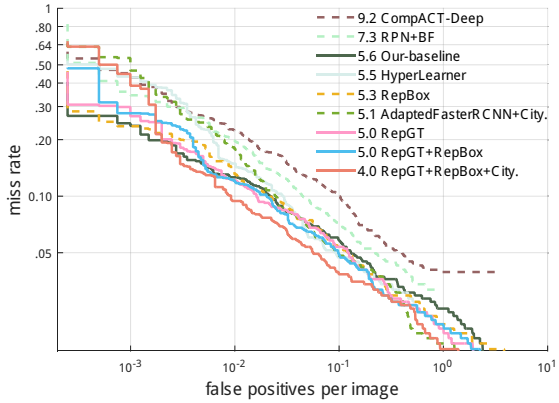


Figure 8. Comparisons with state-of-the-art methods on the new Caltech test subset.

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

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.

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

CVPR2018 Paper Translation

姓名：吴佳雨

学号：2015300155

班号：10011501

排斥损失：在人群中检测行人

Xinlong Wang Tete Xiao Yuning Jiang Shuai Shao Jian Sun Chunhua Shen

Tongji University
1452405wxl@tongji.edu.cn

Megvii, Inc.
jyn, shaoshuai, sunjian@megvii.com

Peking University
jasonhsiao97@pku.edu.cn

The University of Adelaide
chunhua.shen@adelaide.edu.au

摘要

由于在现实生活中，人群总是聚集在一起并且互相遮挡，所以从人群中检测单独的行人一直都是一个具有挑战性问题。在这篇文章中，我们将首先通过实验，探究目前最先进的行人探测器性能是如何在人流拥挤的情况下受到影响，从而提供对这一问题的见解。其次，我们将提出一种专门为人群场景的新型的边界框回归损失，称为排斥损失。这种排斥损失受两种因素影响：目标的吸引力和其他周围物体的排斥。排斥的情况通过阻止检测提案在周围对象中转移，从而保证一个更具有人群鲁棒性性的定位。我们通过排斥损失训练的探测器较目前最先进的探测器具有更优良的性能，尤其在人群拥挤的情况下，性能显著提高。

1. 引言/绪论

尽管今年来有显著的进步，但遮挡问题一直是目标检测中最具挑战的问题之一[10, 9, 24, 19, 1, 20, 11, 3]。一般来说，遮挡问题可以被分为两种：类间遮挡和类内遮挡。前者发生于一个物体被其他类别的物体遮挡的情况。而后者，也被成为密集遮挡，往往是指一个物体被同类物体所遮挡的情况。

在行人检测中[31, 14, 6, 5, 7, 21]，密集遮挡发生数量占遮挡情况的绝大多数。原因是在进行行人检测的应用场景中，例：视频监控和自动驾驶，行人们往往走在一起，相互遮挡。例如，在CityPerson的数据集中[33]，一共含有3157个行人

图1. 我们所提出的排斥损失的例证。排斥有两部分组成：目标的吸引力将提议的结果和特定目标的差距缩小；对周围物体的排斥使结果和非目标物体的差距扩大。

注解的验证子集中，48%的行人注释都有重叠，并且其交并比（IoU）高于0.1。此外，26.4%的行人和有不同注释的其他行人有着相当高的重叠，交并比高于0.3。如此高概率的密集遮挡对行人探测器的性能产生了十分严重的影响。

密集遮挡的主要影响是其大大提升了行人定位的难度。例如，当一名目标行人 T 和另一名行人 B 在图像中重叠，因两名行人具有相似的外观特征，探测器容易产生混淆。从而很有可能导致本应划分出 T 的预测框向 B 转移，给出不准确的定位。更糟糕的情况是，作为最初得到的检测结果，还需要进行非极大值抑制（NMS）的进一步的处理：原本从 T 移位的检测框被 B 所取代，造成检测错漏了行人 T。这也使检测器对于 NMS 的阈值十分敏感：过高的阈值导致更多的错误定位，阈值过低则导致更多的检测遗漏。这样的不佳的行为会对实时分割框

架产生不利的影响[11, 18]，因为实时分割需要准确的检测结果。因此，对于人群中单个行人的定位鲁棒性是行人检测最重要的工作之一。在当前最先进的检测框架中[9, 24, 3, 19]，将边界框回归方法应用于对象定位，其中，通过训练回归量以缩小用某种距离度量测量的候选检测框和实际目标框之间的差距（例，SmoothL1 或 IoU）。尽管如此，现有的方法只需要候选目标靠近特定目标，无需考虑周围物体。如图 1 所示，在标准边框回归损失方法中，当预测框转移至周围物体是，没有给其的额外的惩罚因子。这一观察结果让我们想到：在检测人群中的目标时，是否可以将周围物体计入考虑的范围之中？受磁铁特性的启发，例：磁铁的相吸与互斥，在这篇文章中，我们提出了一个全新的定位技术，称为排斥损失（RepLoss）。在排斥损失方法中，每个候选不仅需要靠近其指定的目标 T，同时还需远离其他的实际对象与指定目标不为 T 的其他候选目标。换句话说，排斥损失中的边框回归器受两个因素驱动：目标对其的吸引和周围物体对其的排斥。例如，如图 1 所示，当红色的边界框转移到 B 物体时，因其与周围的非目标物体产生重叠，会得到一个额外的惩罚因子。因此，排斥损失可以有效防止预测边界框转移至存在重叠的临近物体，从而使检测器在实际人流较大的情景中更具鲁棒性。我们的主要工作（贡献）如下：

- 我们首先通过实验研究人群拥挤状况对行人检测的影响。特别地，在 CityPersons 基准[33]中，我们定量分析了由于人群拥挤所引起的定位错误和漏检，这为人群拥挤而产生的遮挡问题提供了重要的见解。
- 提出了两种类型的排斥损失来解决人群拥挤时的遮挡问题，分别为 RepGT 损失和 RepBox 损失。RepGT 损失直接对预测框转移至其他物体的情况进行惩罚，RepBox 损失则要求各个不同指定目标的预测框相互远离，从而使检测结果对于 NMS 算法的敏感度下降。
- 通过使用提出的的排斥损失算法，对具有人群鲁棒性的行人探测器进行端到端训练，其训练结果优于目前最先进的在 CityPerson 和 Caltech-USA benchmarks [7]中所使用的方法。还应指出的是，使用排斥损失算法后，探测器对于拥挤情况的探测准确度有了显著提高，这也进一步强调了排斥损失的有效性。另外，我们在 PASCAL VOC [8]检测数据集上的实验显示，RepLoss 除了行人检测，对于一般物体的检测也存在优势。

2. 相关研究

目标定位 由于近年来卷积神经网络（CNNs）[16, 26, 12]的发展，目标检测有了非常显著的进步。其中，目标定位被普遍认为是R-CNN中将最初的候选目标与指定目标进行重定位的回归问题。一个线性回归模型按欧几里得坐标距离用候选目标和指定目标进行训练。在文献[9]中，Smooth_{L1}缺失算法提出用R-CNN中的欧式距离代替边界框回归。文献[24]提出了区域生成网络（RPN），通过两次边界框回归将预先定义的锚点转换为最终的检测框。Densebox [15]提出一个无锚点，全卷积的检测框架。文献[29]提出IoU 缺失算法，使实际边框与预测框之间的IoU最大化。我们还注意到一个由Desai *et al.* [4]提出的方法，利用对象只见那的（特征）吸引与排斥来获取不同对象类之间的空间排列，当然，这是通过全局模型解决对象分类的问题。在本次研究中，我们将展示在人群场景下，使用排斥缺失算法进行目标定位的有效性。

行人检测 行人检测是许多现实生活应用中最初也是最重要的一步。传统的行人探测器，例如ACF [5]、LDCF [22]和Checkerboard [32]，根据积分通道特征(IDF) [6]上的不同过滤器，使用滑动窗口策略来定位每一个目标。最近，一个基于CNN的探测器[17, 30, 21, 14, 28]展现出对于行人检测领域的巨大潜力。在文献[28, 30]中，来自深度神经网络的特征取代了手工标记的特征，输入进一个增强决策树内。文献[21]提出一个多任务训练网络，用于进一步提升检测性能。同时，在文献[23, 27, 34]中，一个基于部分的模型被用于处理被遮挡的行人。研究[13]致力于提高NMS的鲁棒性，但其最终需要依赖于额外的网络进行后续处理。事实上，在之前的研究中，鲜少有关注人群拥挤对检测的影响。

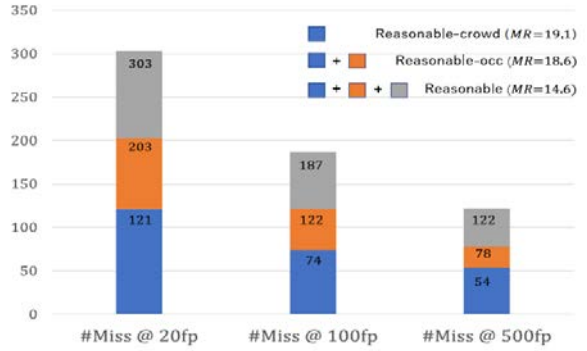


图2. reasonable, reasonable-occ, reasonable-crowd子集的基线漏检数与MR-2得分。在reasonable-occ子集的所有漏检中，人群遮挡占了~60%，是解决遮挡问题的主要阻碍。

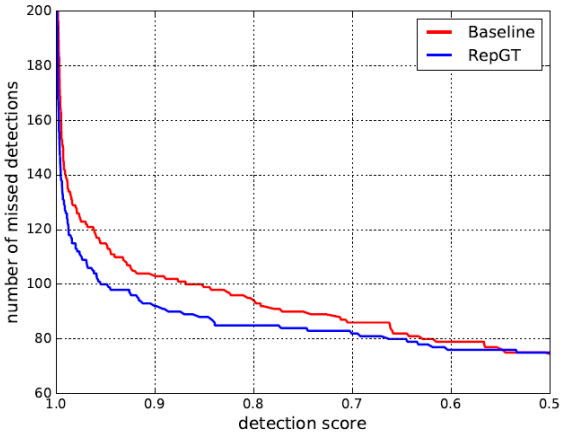
3. 人群拥挤产生的影响

为提供对于人群拥挤所产生的遮挡问题的见解，在这一部分，我们通过实验研究了人群拥挤对行人检测结果的影响程度。在深入研究分析之前，首先介绍一下我们所使用的数据集和基线探测器。

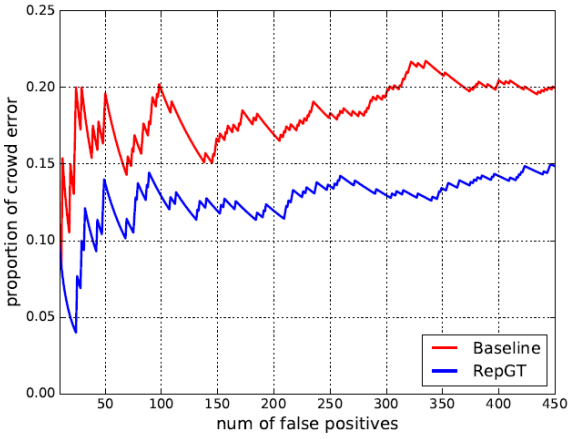
3.1 初步措施

数据集与评估指标 CityPersons [33]是一个在语义分割数据集CityScapes [2]上的新的行人检测数据集，拥有5000个在德国一些城市获取的图像。一共有~35000个人和~13000个忽略区域，提供了所有行人的边界框和可见部分的注释。我们所有关于CityPersons的用于训练和测试的实验都是在合理训练/验证集上分别进行的。为了评估，未命中率在使用$[10^{-2}, 10^{0}]$ (MR^{-2})区间下的每张图像的误检率范围内取平均值并取对数（越低越好）。

探测器 我们的基线探测器是普遍使用的修改为行人检测的R-CNN [24]探测器，遵循Zhang *et al.* [31]和Mao *et al.* [21]中的一般参数设定。我们与他们在执行上的不同之处在于我们使用更快更轻的ResNet-50 [12]网络替换了VGG-16骨干（网络）。需要指出的是，由于卷积层中的下采样率过大使神经网络无法用于小目标的行人检测和定位，ResNet很少用于行人检测。为解决这一问题，我们使用扩张卷积并且将最终的特征图变为输入大小的1/8。基于ResNet的探测器在验证集上达到了$14.6MR^{-2}$，远好于[33]报告中的结果($15.4\ MR^{-2}$)。



(a)



(b)

图3. 基线与RepGT的错误分析；(a) 在不同检测得分下的reasonable-crowd子集的漏检数；(b)由于人群拥挤所产生的错误定位占所有错误定位的比例，RepGT缺失算法有效减少了由于人群拥挤所造成的漏检与误定位.

3.2 失败案例的分析

漏检 根据基线探测器的结果，我们首先分析了由于人群拥挤所造成的漏检。由于CityPersons提供了每个行人的可见部分的边界框注释，结论可以用 $\text{occ} \triangleq 1 - \frac{area(B\ Box_{visible})}{area(B\ Box)}$ 进行计算。我们定义occ ≥ 0.1的标准比对行人为（普通）遮挡案例，而occ ≥ 0.1和与其他标注的行人的IoU ≥ 0.1的行人作为人群拥挤时的遮挡案例。按照定义，在合理验证集所有1570名未忽略的行人注释中提取出两个子集：由

810个（普通）遮挡案例(51.3%)组成的*reasonable-occ*子集和由479个人群拥挤时的遮挡案例(30.3%)构成的*reasonable-crowd*子集。显然，reasonable-crowd子集同时也是reasonable-occ子集的一个子集。

在图2中，我们报告了漏检的数量与reasonable、reasonable-occ 和 reasonable-crowd子集的MR^{-2}。我们观测到从reasonable集上的14.6 MR^{-2}到reasonable-occ子集上的18.6 MR^{-2}，性能有显著下降：一共漏检20100，误报500起。其中遮挡情况占大约60%，表明这是影响基线探测器性能的一个主要原因。在reasonable-occ子集的漏检中，人群拥挤产生的遮挡比例占了将近60%，是其成为解决行人检测中遮挡问题的主要障碍。此外，reasonable-crowd子集中的漏检率(19.1)远高于reasonable-occ子集的漏检率(18.6)，表明在类间遮挡情况中，人群拥挤产生的遮挡所造成的问题更严重；当我们降低100到500的误报阈值，由于人群拥挤产生的遮挡所造成的误报部分逐渐增大(从60.7%升至69.2%)。这表明有人群拥挤产生的遮挡所引起的误报很难通过降低阈值来解决。

在图3(a)中，红线表示在不同检测得分的reasonable-crowd自己中，标准比对的行人被漏检的数量。在现实生活中，只有具有高可信度的预测的边框会被纳入考虑，在曲线顶部的大量漏检意味着远不能满足实际生活中的应用。

误报（误检） 我们还分析了由于人群拥挤产生的遮挡造成的误报数量。我们将所有误报分为三类：背景(background)、定位(localization)、人群错误(crowd error)。背景错误发生在预测边框和任一标准比对行人之间的IoU < 0.1，定位错误则仅与一个行人之间的IoU ≥ 0.1，人群错误指至少与两个行人之间的IoU ≥ 0.1

在这之后，我们计算了人群错误发生的数量并计算了其占所有误报的比例。如图3(b)中所示，人群错误占所有误报中相当大的比例(大约20%)。即使如图4所示，我们发现人群错误通常发生在预测框略微或显著移动到临近的非指定的标准比对物体，或将几个重叠的标准比对物体结合在一起时的情况下。另外，人群错误通常具有相对较高的可信度，

图4. 关于人群错误的可视化例子. 绿色框是正确的预测框，而红色框表示因人群遮挡产生的错误定位框. 同时附加了探测器输出的可信度分数. 错误通常发生在预测框略微或显著移动到临近的非指定的标准比对物体(例，右上方)，或将几个重叠的标准比对物体结合在一起时的情况下(例，右下方)

从而导致最多的误报。这说明，为了提高探测器对人群场景的鲁棒性，在进行边框回归时，需要有更多的歧视性损失。更多可视化的例子可参阅补充材料。

结论 对于失败案例的分析验证了我们的观测：行人探测器非常容易受到人群遮挡的影响（污染），它通过增加定位难度造成了绝大多数的漏检和错误定位。为了解决这些问题，在第4部分，排斥损失将被用于提升行人探测器在人群拥挤情景下的鲁棒性。

4. 排斥损失

在这一部分，我们将介绍排斥损失算法，用于解决检测中的人群遮挡问题。受磁铁特性的启发，*例，磁体的相吸与互斥*，排斥损失算法由三个分量构成，定义为：

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

L_{Attr}表示要求预测框接近指定目标的吸引力的变量，L_{RepGT}和l_{RepBox}分别表示要求预测框远离周围其他的标准比对对象和其他预测框指定的目标对象的排斥变量。系数α和β为平衡辅助损失的权重。

为简化，我们仅考虑如下两类检测，假设所有的标准比对对象来自同一类。令$P = (l_P, t_P, w_P, h_P)$，$G = (l_G, t_G, w_G, h_G)$，分别表示由其左上角左标及其宽度和高度所表示候选边界框和实际边界框。$P_+ = \{P\}$表示所有正候选集（至少与一个标准比对框之间有高IoU(*例，$IoU \geq 0.5$*)被认为是正例子，反之则为负例子），$G = \{G\}$表示一张图像中所有的标准比对框。

吸引变量 由于目的是缩小预测框与实际边界框以某种距离度量1衡量的差距，*例，*欧式距离[10]，$Smooth_{L1}$距离[9] 或 IoU [29]，吸引损失已经在目前的边框回归技术中普遍使用。为做一个公平比较，在这篇文章中，我们采用$Smooth_{L1}$距离衡量吸引变量[21, 33]。我们将$Smooth_{L1}$中的Smooth变量设为2。设$P \in P_+$，令具有最大IoU的作为其指定目标的标准比对框：$G^P_{Atrr} = \arg max_{G \in g} IoU(G, P)$。$B^P$是从P回归的预测框。然后，吸引损失可用如下方法计算：

$$L_{Atrr} = \frac{\sum_{P \in P_+} Smooth_{L1}(B^P, G^P_{Atrr})}{|P_+|} \quad (2)$$

排斥参数(RepGT) RepGT损失被设计用于反驳非指定目标的临近对象的候选对象。设$P \in P_+$，其排斥的实际场景中的物体定义为标准比对对象，除指定目标外具有最大的IoU区域：

$$G^P_{Rep} = \underset{G \in g \setminus \{G^P_{Atrr}\}}{\arg max}\ IoU(G, P) \quad (3)$$

受文献[29]中IoU损失的启发，RepGT 损失用于计算B^P和G^P_{Rep}重叠时惩罚因子。B^P和G^P_{Rep}之间的重叠由IoG交点确定：$IoG(B, G) \triangleq \frac{area(B \cap G)}{area(G)}$。当$IoG(B, G) \in [0,1]$时，我们定义RepGT损失为：

$$L_{RepGT} = \frac{\sum_{P \in P_+} Smooth_{L1}(IoG(B^P, G^P_{Rep}))}{|P_+|} \quad (4)$$

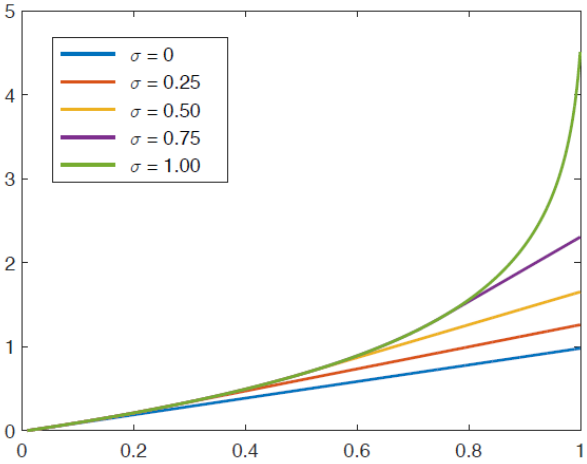


图5. Smoothln在不同smooth参数σ下的曲线。σ越小，对异常值的敏感损失越小.

其中

$$Smooth_{L1} = \begin{cases} -\ln(1-x) & x \leq \sigma \\ \frac{x-\sigma}{1-\sigma} - \ln(1-\sigma) & x > \sigma \end{cases} \quad (5)$$

这是一个smooth的自然对数函数，在(0,1)上连续可分，$\sigma \in [0,1)$是smooth参数用于调整对异常值的排斥损失的敏感度。图5显示了公式4和公式5中不同σ值时的曲线，可以看出，候选越倾向于和非目标标准比对对象重合，RepGT损失就会对边框回归增加越高的惩罚因数。通过种方法，RepGT损失可以有效地防止预测框向不是其目标的临近对象偏移。

排斥参数 (RepBox). NMS以大多数检测框架将最初预计绑定同一对象的的预测框合并的所必须的后期处理步骤。然而，在人群拥挤产生的遮挡情况下，NMS将严重影响检测结果。为了令探测器对于NMS不过于敏感，我们进一步提出了RepBox损失，其目的为排斥具有不同指定目标的各个候选。我们根据各个候选的目标将候选集P_+分为$|g|$个互不相交的子集：$P_+ = P_1 \cap P_2 \cap \ldots \cap P_{|g|}$ 然后，对于两个子集中随机选出的候选，$P_i \in \boldsymbol{P}_i$和$P_j \in \boldsymbol{P}_j$，其中$i, j = 1,2, \ldots, |g|$ 且 $i \neq j$。我们希望预测框B^{P_i}的B^{P_j}和尽可能小。因此，B^{P_i}和B^{P_j}按如下方法计算：

$$L_{RepBox} = \frac{\sum_{i \neq j} Smooth_{L1}(IoG(B^{P_i}, B^{P_j}))}{\sum_{i \neq j} \mathbb{I}[IoU(B^{P_i}, B^{P_j}) > 0] + \epsilon} \quad (6)$$

其中$\mathbb{I}$是本身函数，ϵ是一个小的常量防止被分为0个案例。根据公式6，我们可以看出，为缩小RepBox Loss，需要缩小两个具有不同特定目标的预测框之间的IoU区域。这意味着RepBox Loss能够降低不同特定目标的预测边框在NMS后合并的可能性，使得探测器在人群拥挤的情况下具有更高的鲁棒性。

4.1 讨论

距离度量 值得一提的是，我们选择了IoG或IoU测量两个边界框的排斥参数的距离，而非$Smooth_{L1}$度量方法。原因是因为IoG和IoU的值位于[0,1]，而$Smooth_{L1}$的值无限定区间，*例*，如果我们使用$Smooth_{L1}$度量方法测量排斥参数，在RepGT损失下，它要求预测框与其排斥的实际物体距离尽可能远。相反，IoG标准只需要预测框将其与排斥的实际对象之间的重叠最小化即可，这更符合我们的要求。

另外，RepGT损失中使用IoG而不是IoU，因为基于IoU的损失算法，边界框回归量可以通过放大边界框大小增加分母：$\text{area}(B^P \cap G^P_{Rep})$，学习如何将损失最小化。因此，我们选择分母为某个特定标准比对对象的常数的IoG，以直接将重叠区域$\text{area}(B^P \cap G^P_{Rep})$最小化。

Smooth 参数σ 相比于直接使用−ln(IoU)作为损失函数的文献[29]，我们将介绍同时用于RepGT 损失和RepBox 损失的Smooth的自然对数函数Smoothln和Smooth参数σ。如图5所示，通过Smooth参数σ，我们可以调整排斥损失对异常值（有大面积重合的一对边界框）的敏感度。由于预测边框比实际的标准比对边框更密集，两个预测框之间的重叠部分更容易比预测框与实际边框之间的重叠部分面积更大。这意味着，在RepBox会比RepGT有更多的异常值。因此，直观地说，RepBox损失相对于RepGT损失，对异常值敏感度不高（σ较小）。关于Smooth参数σ和辅助损失权重α与β更加详细的研究请见5.2节

	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

表1. RepGT和RepBox损失的MR−2与在CityPersons验证集上不同Smooth参数σ所产生的改进

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

表2. 我们通过调整权重α与β平衡RepGT和RepBox损失.根据经验，α = 0.5，β = 0.5时为最佳性能. 结果由CityPersons验证子集上获得

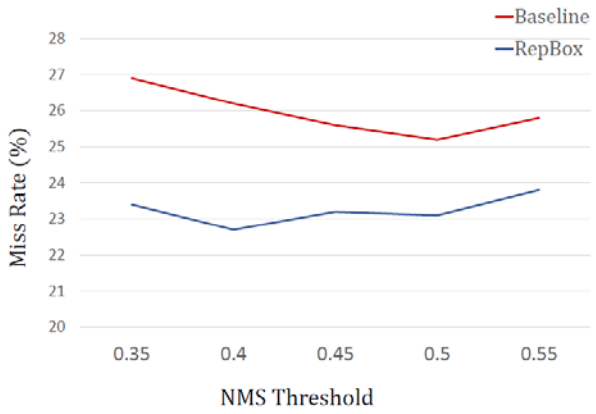


图6. 在FPPI = 10−2时，不同NMS阈值的RepBox损失结果. RepBox的曲线较基准线更为平滑，可说明其对于NMS阈值的敏感度较小

5. 测试

测试部分的组织如下：我们首先在5.1节介绍基本的测试设置和排斥损失的实施细节；然后在5.2节在CityPersons [33]的基础上分别分析评估提出的RepGT损失和RepBox损失；最后，在5.3节，将使用排斥损失的探测器与CityPersons [33]和Caltech-USA [7]中最先进的检测方法进行比较。

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

表3. 使用RepLoss在CityPersons[33]上评估行人检测结果。模型在训练集上训练并在验证集上进行测试.我们使用ResNet-50作为骨干架构。3个最好的结构分别用红色、蓝色和绿色标注

5.1测试设置

数据集 除了在第3节介绍的CityPersons [33]基准，我们还在Caltech-USA dataset [7]上进行了实验。作为行人检测的几个主要数据集和基准之一，Caltech-USA见证了这一领域令人振奋的进展。总共2.5小时的视频分为训练和测试子集，分别为42,500帧和4,024帧。文献[31], Zhang *et al*提供了精确的注释，其训练数据可自动细化且测试数据均精心的由人工标注。除另外说明，均为使用Caltech-USA数据集新标记的注释所进行的测试。

测试细节 我们的框架是通过自行建立的快速灵活的深度学习平台上实现的。我们对网络进行了80k迭代和160k迭代训练，基本学习率设置为0.016，并分别在CityPersons和Caltech-USA上的第一次60k与120k次迭代后减少了10倍。采用随机梯度下降(SGD)求解器优化4个GPU上网络。权重衰减和动量设置为0.0001和0.9。未应用多规模训练/测试以保证与之前方法的公平比较。对于Caltech-USA（数据集），我们使用10x集(~42k帧)进行训练。使用在线硬例挖掘(OHWM) [25]来加速收敛。

5.2模型简化测试

RepGT损失 在表1中，我们报告了RepGT损失在不同σ参数下的$Smooth_{ln}$损失。当使σ为1.0时，就合理的评估设置而言，将RepGT损失设为13.7 MR−2可

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

表4. Calech-USA 测试集(reasonable)结果，按新注释评估[31]. 在强大的基线上，我们进一步极爱那个目前最先进的4.0 MR−2改进到0.5IoU阈值以下。当增加IoU阈值时的持续增长说明排斥损失的有效性*：使用CityPersons数据集进行网络的预训练

获得最佳性能。其性能优于0.9 MR^{-2}改善后的基线。令σ为1意味着直接将计算$-\ln(1-LoG)$而不进行平滑，类似于使用IoU损失的损失函数[29]。我们还提供了RepGT和基线之间的漏检和误报对比。在图3(a)中，添加RepGT损失可以显著减少在reasonable-crowd子集中漏检的数量。在检测得分阈值相当高的情况下，RepGT曲线始终低于基线，但曲线在0.5处重合。曲线的饱和点均位于~0.9，是实

图7. 在NMS之前的基线与RepBox的预测框的可视化比较. 在RepBox的结果中，两个相邻的实际标准对比之间存在较少的预测，这在人群检测中是被期望的. 更多例子详见补充材料

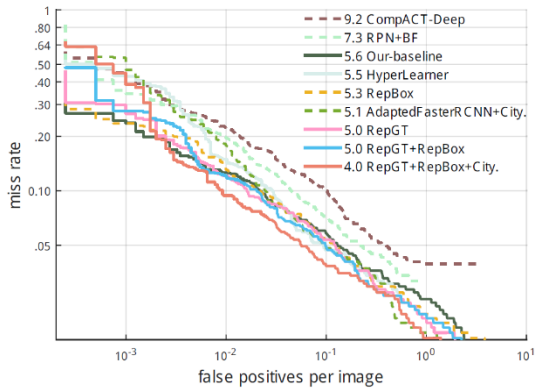


图8. 在新Caltech测试集下与目前最先进方法的比较

际应用中常用的阈值，在此点上，漏检数量减少了10%。在图3(b)中，应用RepGT损失的情况下，由于人群拥挤造成的误报比例低于基线探测器。这表明RepGT损失在人群场景下可有效减少漏检和误报。

RepBox损失 对于RepBox损失，我们用不同的Smooth参数σ进行实验，并将结果报告在表1的第四行。当σ设为0时，RepBox损失产生了13.7 MR^{-2}的最佳性能，与RepGT的σ设为1.0时的性能相近。将σ设为0意味着我们将ln函数完全平滑为线性函数并

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

表5.在PASCAL VOC 2007 [8]基准上进行的普通目标检测评估. ReIm是重置的更快的RCNN. 人群子集包含标准比对对象，在0.1 IoU区域内至少与另一相同类别的标准比对对象重叠. 我们的RepGT损失在人群子集上优于基线2.1 mAP。

对IoU进行求和。由于预测框密度远大于实际边界框，我们猜测RepBox损失比RepGT损失倾向于有更多的异常值。正如第1节中所提到的，在人群场景中，探测器对于NMS的阈值十分敏感。高NMS阈值会导致更多误报，而低NMS阈值会导致更多的漏检。在图6中，我展示了在FPPI = 10^{-2}时，在不同NMS阈值上应用RepBox损失的结果。一般来说，使用RepBox损失的探测器比基线更为平滑。值得指出的是，早NMS阈值为0.35时，基线与RepBox之间的差距为3.5分，表明后者对于NMS阈值更不敏感。尽管如图7所示，在RepBox的两个相邻实际对象之间存在较少的预测，但这在人群场景中仍是可取的。更多的例子请见补充材料。

平衡RepGT和RepBox 之前已经说明了单独应用RepGT和RepBox可以帮助探测器在人群场景下的检测，但我们还未研究如何平衡二者。表2展示了我们通过设置不同的α和β的值所得到的结果。根据经验，α=0.5，β=0.5时可获得最佳性能。

5.3. 与当前最先进方法比较

为证明我们在不同遮挡程度下的有效性，我们将reasonable子集(遮挡 ≤35%)划分成表示部分子集的*reasonable-partial*子集(10% < 遮挡 ≤35%)和表示裸露子集的*reasonable-bare*子集(遮挡 ≤10%)。对于遮挡大于35%的注释(不在reasonable集中)，我

们用Heavy子集表示。表3总结了在CityPersons的结果。一般来说，RepGT损失和RepBox损失可以展示对所有评估子集的改进。结合两者的排斥损失可以达到13.2 MR^{-2}，比基线提升了1.4个百分点。就不同的遮挡程度而言，使用排斥损失的Heavy子集得到了3.7个百分点的显著提升，而对于Partial子集提升了相对较少的1.8个百分点，Bare子集未能得到提升。这是由于根据我们的目的，RepLoss是专门用于解决遮挡问题的。我们还在Caltech-USA数据集上对RepLoss进行测试。结果展示在表4中。在一个强有力的参考下，RepLoss达到了在.5的IoU 匹配阈值下5.0MR^{-2}，在.75的IoU 匹配阈值下26.3的结果。当增加IoU阈值时，已知遮挡会造成对于匹配阈值的敏感度增加，一致且甚至更大的增益表明了我们的框架处理遮挡问题的能力。结果曲线由图8展示。

6. 扩展：一般物体检测

我们的RepLoss是一个在人群场景中一个通用的物体检测损失函数，可以用于除行人检测以外的应用。在这一部分，我们将排斥损失应用于一般物体的检测。　　我们在一般物体检测通用的评估基准，PASCAL VOC数据集[8]上进行测试。这个数据集由20多个对象类别组成。VOC数据集的标准评估指标为所有类别的平均精度(mAP)。我们采用vanilla Faster R-CNN [24]框架，使用ImageNet-pretrained ResNet-101 [12]作为骨干。NMS阈值设为0.3。模型在PASCAL VOC 2007 和 PASCAL VOC 2012的验证集上进行训练，并在PASCAL VOC 2007的测试集上评估。我们重新设置的基线比原始基线好3.4 mAP。

结果显示在表5中。整个数据集的增长并不显著。尽管如此，在人群子集上进行评估时（对象中有类内IoU大于0.1），RepLoss的性能优于2.1 mAP。这些结果表明我们的方法是通用的，可以扩展到一般对象检测。

7. 结论

在这篇文章中，我们仔细设计了用于行人检测的排斥损失(RepLoss)，大大提升了检测性能，尤其是在人群拥挤的情况下。排斥损失的主要因素是仅用对目标的吸引损失不足以训练出一个最佳的探测器，而考虑周围的排斥因素会有所助益

为实现排斥，我们介绍了两种排斥损失的种类。我们在两大最通用的数据集Caltech和CityPersons上取得了最优的性能报告。值得注意的是，我们在不使用像素注释的CityPersons上的结果优于之前使用像素注释的最佳结果[33] 约2%。详细的实验比较证明了所提出的RepLoss的价值，它在人群遮挡场景中可以大大提高检测精度。检测（PASCAL VOC）的结果进一步显示其多效性。我们期望在许多其他对象的检测任务中能广泛应用我们所提出的损失算法。

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