Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **8**, Number **3**, August **2023**

E-RepLoss: A New Regression Loss Function Based On Repulsion Loss For Crowded Pedestrian Detection

Xinxin Zhou*

School of Computer Science
Northeast Electric Power University
Jilin, Jilin, China
zxx51@qq.com

Xuanyu Meng

School of Computer Science
Northeast Electric Power University
Jilin, Jilin, China
2202000707@neepu.edu.cn

Long Zhang

School of Computer Science
Northeast Electric Power University
Jilin, Jilin, China
m15055591464@163.com

*Corresponding author: Xinxin Zhou

Received October 3, 2022, revised December 5, 2022, accepted February 14, 2023.

Abstract. *In the crowded scenario, the pedestrians often gather together, which leads to occluding between pedestrians. The occlusion affects the performance of detecting human. Specifically, the predicted box of model may shift to surrounding objects or deleted by non-maximum suppression accidentally. In the article, a regression loss function is proposed for crowded scenarios, termed E-Reploss. In E-RepLoss, the repulsion loss function is constructed entirely based on EIoU loss function. It not only fits better with the model based on anchor, but also has all the properties of the repulsion loss function. It can both enhance the localization of the bounding boxes and keep a distance between surrounding object's bounding boxes and ground-truth boxes. This paper experiments with CrowdHuman benchmark in order to verify performance of the proposed strategy. The results of experiment indicate that this loss function reduces false positive and missed detection rate effectively in crowded scenes. Meanwhile, it maintains the better generalization.*
Keywords: Pedestrian detection, Crowded scenarios, EIoU, Regression loss function

1. **Introduction.** As an important part of object detection, human detection can provide important technical support for application scenarios such as automatic driving, intelligent surveillance, robot path planning and person re-identification [1, 2]. Moreover, it is widely applied in multi-target tracking, human pose estimation and person search [3, 4, 5, 6, 7]. Therefore, human detection brought out extensive attention of scholars in the field of industry and academia. However, there is a serious occlusion between pedestrians in the crowded scenes such as subways, shopping malls and scenic spots [8]. The general lack of features and high similarity between pedestrians cause the detector unable to extract

the effective feature of pedestrians. The predicted box is unable to accurately fit each pedestrian or is deleted by non-maximum suppression, resulting in missed detection and false positive. As a consequence, it is the challenge for the current human detection to decrease the rate of false positive and missed detection and enhance the precision of human detection and location.

Recently, deep convolutional neural network [9, 10, 11, 12, 13, 14] has developed rapidly. Using a single detector can achieve better effect and significantly improve human detection's performance. As an important part of network model, loss function is utilized to calculate error between real label and predicted output of model. It can be continuously optimized with the help of optimization function. The value of loss function can be utilized to judge the fitting of the model to data. If loss function is smaller, fitting degree is better. Conversely, the fitting degree will be worse. The training process of convolutional neural network is the process of continuously adjusting the weight and bias of the model through the loss function and optimization algorithm. Some research works have concentrated on the regression loss function of bounding box. Researches mainly focus on how to make the predicted boxes of the target fit its ground-truth box as much as possible and solve the problem of occlusion on human detection. Some classical approach is available in the literature [15] and literature [16]. In literature [15], an aggregation loss function (AggLoss) inspired by group behavior is introduced to strengthen the bounding box's location and reduce false detection. Meanwhile, a PORoI Pooling for partial occlusion perception is proposed to replace ROI pooling in Fast RCNN. This model can integrate the prior structure information and visibility prediction of human body into the network to deal with occlusion. To describe the uncertain prediction of bounding box, the bounding box and the ground-truth box are modeled as Gaussian distribution and Dirac distribution in reference [16]. Then the regression loss of the bounding box is called as KL divergence of predicted boxes distribution and ground-truth boxes distribution, and KL loss is proposed. Although the above strategy can solve occlusion patterns, it does not have a better breakthrough in convergence speed and precision of regression.

The predicted box cannot accurately fit each pedestrian or is deleted by non-maximum suppression. A regression loss function E-RepLoss suitable for crowded scenes with more serious occlusion is proposed to solve the problems. E-RepLoss not only makes the bounding box better cover its targets, but also can make the predicted box far keep a distance with the ground-truth boxes and adjacent pedestrians' predicted boxes. While it is more suitable for the CrowdDet based on anchor.

In this paper, the main contributions are summarized as follows:

(1) The reasons for high false positive and missed detection of human detection in crowded scenes are analyzed;

(2) An improved regression loss function E-RepLoss is proposed from the perspective of model loss function. E-RepLoss constructs each component of RepLoss loss function based on EIoU loss function. This loss function can strengthen the location of bounding box. Moreover, it maintains a distance from the ground-truth box and the bounding box of other pedestrians. Generally, the loss function is more suitable to CrowdDet based on anchor. The model's detection effect is improved;

(3) Experiments are designed to prove effectiveness of E-RepLoss. Firstly, based on E-RepLoss loss function, CrowdDet model is trained and tested on CrowdHuman dataset. Secondly, the loss function of CrowdDet model is replaced by EIoU loss function, RepLoss and other IoU variant loss construction RepLoss respectively. These loss functions are utilized to compare the performance.

Other parts of the article are arranged as follows: Section 2 summarizes the relevant work. The proposed method and common regression loss functions are introduced in

section 3. In Section 4, the improved loss function is described in detail. Section 5 introduces the relevant details of the experiment. The experimental result is analyzed in Section 6. Finally, in section 7, the paper makes conclusions and states the future work.

2. **Related Work.**

2.1. **Pedestrian Detection.** Pedestrian detection, as a domain-specific application of target detection, can provide important technical support for some realistic scenarios. Traditional pedestrian detection methods focus on improving the detection accuracy in terms of feature extraction and classification. For example, ACF [17], LDCF [18], and Checkerboard [19] use various filters and sliding window strategies on integral channel features (IDF) [20] to localize each target. In recent years, with the advent of large-scale datasets, deep learning, migration learning techniques, and the dramatic increase in GPU hardware computing power, convolutional neural network-based target detection methods have once again been rapidly developed, showing great potential for dominance in the field of pedestrian detection. In the literature [21, 22], features from deep neural networks rather than hand-crafted features are fed into an enhanced decision forest. The literature [23] proposed a network trained for multiple tasks to further improve the detection performance. The introduction of convolutional neural networks has improved the performance of pedestrian detection algorithms, but the occlusion problem is still a major difficulty in pedestrian detection [24, 25, 26]. In literature [24], they discuss different kinds of occlusions and propose an occluded scene text enhancing network to improve recognition performance. The network is based on generative adversarial networks, and they design accretion blocks to help the network generate the occluded image regions. In literature [25], due to the disadvantages of feature point center weighting, multiparticle template matching, and Kalman filter trajectory prediction algorithms in different cases, some algorithms with higher robustness and stability are developed to solve the occlusion problem. In literature [26], to handle the occlusion issue, they proposes a simple but effective pedestrian detector who designs a novel bi-center prediction mechanism (namely Bi-Center) based on the anchor-free network.

2.2. **Regression loss functions.** With the increasing representational power of convolutional neural networks in recent years, good detection results can be achieved using a single detector. Some research work has focused on the design of regression loss functions for the bounding boxes of pedestrian detection models, and this type of research is concerned with how to make the predicted bounding box of the target fit its corresponding ground truth as closely as possible.The general regression loss functions L_1, L_2, and Smooth_{L_1} [27] based on IoU and its variants [28, 29, 30]. The Smooth_{L_1} loss function first appeared in the Fast RCNN model of target detection. Then, a series of loss functions GIoU [28], DIoU and CIoU [30] have been proposed to improve the regression accuracy. They have made good contributions to improve model's detection accuracy, but the above loss functions still have the problems of low precision and slow convergence in the crowded scenes. The Focal and Efficient IoU Loss proposed in the literature [29] can accurately measure the differences of overlap area, centroid and edge length in the bounding box regression. It further improves the network convergence speed and accuracy of target localization. A repulsion loss function (RepLoss) added the penalty term is introduced in reference [31] to improve localization precision of pedestrians. The loss function makes the predicted box as close to the ground-truth box as possible and keeps away from ground-truth box and predicted box of other targets. This loss function is a classical method to deal with occlusion in human detection.

3. **The common regression loss functions.** The improved regression loss function E-RepLoss is based on Reploss and EIoU loss functions. First of all, the common regression loss functions are introduced.

The bounding box loss function used in CrowdDet model is calculated as follows:

$$\text{Smooth}_{L_1} = \begin{cases} \frac{(\beta x)^2}{2}, & |x| < \frac{1}{\beta^2} \\ |x| - \frac{0.5}{\beta^2}, & \text{otherwise} \end{cases} \tag{1}$$

Where β is smoothing coefficient. x represents input data. As can be seen from Equation (1), the smaller the loss is, the smaller the gradient is. It is conducive to convergence; When the loss is large, the gradient is 1. The model training is relatively stable. However, when the Smooth_{L_1} calculates the regression loss, the correlation of the bounding box's coordinates is ignored. It results in a variety of overlapping ways between the bounding boxes which have the same value of loss. But the bounding box with the highest degree of overlap is the most reasonable.

The EIoU loss function directly uses width and height to calculate the width and height penalty term, which has faster convergence speed. The calculation method is shown in Equation (2).

$$L_{\text{EIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{\text{gt}})}{c^2} + \frac{\rho^2(w, w^{\text{gt}})}{C_w^2} + \frac{\rho^2(h, h^{\text{gt}})}{C_h^2} \tag{2}$$

Where L_{EIoU} represents EIoU loss function. b^{gt} and b are ground-truth and box predicted box. C_h and are the height and width of the minimum outer rectangle of the two bounding boxes with the same loss. c is the diagonal length of the smallest enclosing box covering the two boxes. $\rho(\cdot)$ represents Euclidean distance. h, w, h^{gt} and w^{gt} represent the height and width of the predicted box and ground-truth box.

Another loss function is Reploss. Reploss loss function is an effective method in dealing with occlusion. The loss function is inspired by the attraction of dissimilar magnets and the repulsion of same magnets. It requires that the predicted box needs to fit its targets and be far away from predicted boxes of other surrounding targets. Moreover, the predicted box should keep a distance with ground-truth box. The Reploss includes L_{Attr}, L_{RepGT} and L_{RepBox} The calculation method is shown in Equation (3).

$$L = L_{\text{Attr}} + \alpha \cdot L_{\text{RepGT}} + \beta \cdot L_{\text{RepBox}} \tag{3}$$

Where α and β are equilibrium coefficient. L_{Attr} is the attraction term which requires a predicted box to approach its designated target. L_{RepBox} and L_{RepGT} represent the repulsion term.

L_{Attr} is the attraction term which requires a predicted box to approach its designated target. The Equation is shown as follows:

$$L_{\text{Attr}} = \frac{\sum_{p \in p_+} \text{Smooth}_{L_1}(B^p, G^P_{\text{Attr}})}{|P_+|} \tag{4}$$

Where p is the proposal bounding box. $|P_+|$ is the set of these candidate boxes. B^p and G^P_{Attr} represent the predicted box and ground-truth box of a pedestrian respectively. L_{RepGT} represents that predicted box B^p of a pedestrian keeps a distance with the ground-truth box of other pedestrians who are "closest" to the bounding box. It is calculated as follows:

$$L_{\text{RepGT}} = \frac{\sum_{p \in p_+} \text{Smooth}_{L_n}(\text{IoG}(B^p, G^P_{\text{Rep}}))}{|P_+|} \tag{5}$$

Where Smooth_{L_n} is a smoothing function. The calculation method is shown in Equation (6). represents the ratio of the intersection of B^p and G^P_{Rep} to G^P_{Rep}. Compared with IoU, only molecules can be changed for IoG, which is more suitable for loss. Otherwise, Smooth_{L_n} is used instead of Smooth_{L_1} in L_{RepGT}. The reason is that the punishment is influenced by the intersection of B^p and G^P_{Rep}. When the intersection is greater, the punishment of is large. The definition of G^P_{Rep} is shown in Equation (7).

$$\text{Smooth}_{L_n} = \begin{cases} -\ln(1-x) & x \leq \sigma \\ \dfrac{x-\sigma}{1-\sigma} - \ln(1-\sigma) & x > \sigma \end{cases} \tag{6}$$

Where σ is sensitivity hyperparameter of function.

$$G^P_{\text{Rep}} = \arg \max_{G \in \varsigma \setminus \{G^P_{\text{Attr}}\}} \text{IoU}(G, P) \tag{7}$$

Where ς represents the set of ground-truth box of all pedestrian. G and P are ground-truth box and predicted box. G^P_{Rep} is the ground-truth box when the intersection ratio of P with all ground-truth box is the second largest.

L_{RepBox} represents that the predicted boxes of different pedestrians should keep as far away as possible. The specific method is to divide the set of positive candidate boxes P_+ into mutually exclusive subsets according to the ground-truth box of each pedestrian. Then the predicted boxes of each subset are randomly selected for loss calculation. The formula of L_{RepBox} is shown as follows:

$$L_{\text{RepBox}} = \frac{\sum_{i \neq j} \text{Smooth}_{L_n}(\text{IoU}(B^{P_i}, B^{P_j}))}{\sum_{i \neq j} g[\text{IoU}(B^{P_i}, B^{P_j}) > 0] + \varepsilon} \tag{8}$$

Here, B^{P_i} and B^{P_j} represent the predicted box of different pedestrians. $g(\cdot)$ represents identity function. ε is a constant and it is not 0. In L_{RepBox}, the denominator indicates that there must be overlap between predicted boxes involved in loss calculation.

4. **The Improved Regression Loss Function Prepare.**

4.1. **Problems of human detection in crowded scenes.**

In crowded scenes, the occlusion between pedestrians is more common. The serious occlusion can cause the following problems.

(1) The serious occlusion leads to a high overlap between the optimal predicted box of an occluded target and that of other targets. It leads to the deletion of the optimal predicted box in the post-processing stage and results in missed detection.

(2) The occlusion between targets introduces a lot of interference information. And a large offset exists between the predicted box with the target.

(3) The model will retain more bounding boxes because the large number of pedestrians need to be identified in crowded scenes. Consequently, it is more dependent on the calculation of intersection over union. The loss function Smooth_{L_1} and intersection over union calculation in CrowdDet method are not suitable, which will ignore the correlation of four coordinates for the bounding box. It results in that regression direction of the

bounding box is not unique. Thus, it is not conducive to the regression for bounding box, resulting in the reduction of detection accuracy.

4.2. **The improved regression loss function E-RepLoss.** A new bounding box regression loss function E-RepLoss is proposed to enhance performance of the model for occluded objects.

The improved loss function E-RepLoss includes $L_{\text{E-Attr}}$, $L_{\text{E-RepGT}}$ and $L_{\text{E-RepBox}}$. The calculation formula is shown in Equation (9). E-RepLoss is completely based on the EIoU loss function to construct each component of loss function. It makes E-RepLoss have the respective advantages of the two functions at the same time.

$$L_{\text{RepLoss}} = L_{\text{E-Attr}} - \alpha \cdot L_{\text{E-RepGT}} - \beta \cdot L_{\text{E-RepBox}} \tag{9}$$

Where $L_{\text{E-Attr}}$ is the attraction term. $L_{\text{E-RepGT}}$ and $L_{\text{E-RepBox}}$ represent the repulsion term. α and β are the equilibrium coefficients.

(1) Attraction Term. The attraction term $L_{\text{E-Attr}}$ is to make the predicted boxes of the target as close to its ground-truth box as possible. The loss function is also called attraction loss of the ground-truth box to its predicted box. This function can optimize and adjust the bounding box from three aspects: overlapping area, distance of center point and the ratio of height and width. The calculation Equation of the attraction term $L_{\text{E-Attr}}$ is as follows:

$$L_{\text{E-Attr}} = 1 - \text{IOU} + \frac{\rho^2(b, b^{\text{gt}}_{\text{Attr}})}{c^2} + \frac{\rho^2(w, w^{\text{gt}}_{\text{Attr}})}{C_w^2} + \frac{\rho^2(h, h^{\text{gt}}_{\text{Attr}})}{C_h^2} \tag{10}$$

(2) Repulsion Term (E-RepGT).

For the repulsion term, bounding box's center distance and the ratio of the width and height are utilized to constrain the bounding box. Pedestrian's predicted box and the bounding box of the adjacent target are kept at a distance. The predicted box and the ground-truth box of the adjacent target is shown in Equation (11). $L_{\text{E-RepGT}}$ is decreasing. Therefore, the opposite number is taken in the overall loss.

$$L_{\text{E-RepGT}} = \frac{\rho^2(b, b^{\text{gt}}_{\text{Rep}})}{c^2} + \frac{\rho^2(w, w^{\text{gt}}_{\text{Rep}})}{C_w^2} + \frac{\rho^2(h, h^{\text{gt}}_{\text{Rep}})}{C_h^2} \tag{11}$$

Where $b^{\text{gt}}_{\text{Rep}}$ is the ground-truth box with which it has the largest IoU region except its designated target. $w^{\text{gt}}_{\text{Rep}}$ and $h^{\text{gt}}_{\text{Rep}}$ are its corresponding width and height.

(3) Repulsion Term (E-RepBox).

$L_{\text{E-RepBox}}$ indicates that the predicted boxes of different pedestrians should be as far as possible. The participation in the calculation is the predicted box of pedestrian and that of adjacent pedestrian. The Equation is as follows:

$$L_{\text{E-RepBox}} = \frac{\rho^2_{i\neq j}(b^i, b^j)}{c^2} + \frac{\rho^2_{i\neq j}(w^i, w^j)}{C_w^2} + \frac{\rho^2_{i\neq j}(h^i, h^j)}{C_h^2} \tag{12}$$

Where w^i and b^j denote the predicted boxes of different pedestrians. h and w are the height and width of the predicted boxes.

5. Experiment.

5.1. **Dataset.**

(1) CrowdHuman Dataset.

In 2018, CrowdHuman dataset [32] was published to promote technological development for human detection in crowded scenes. The images of the dataset are collected in a variety of scenes. A total of 24370 images is divided into the training set, test set and validation set. The number of these sets are 15000, 5000 and 4370 images respectively. Each image in CrowdHuman contains about 23 pedestrians on average, and the number of pedestrians in a single image is much more than that in other classic datasets of human detection. From the statistics of occlusion, the targets with occlusion degree less than 30%, or between 30% and 70% constitute 29.89% and 32.13% of entire CrowdHuman dataset, while the ratios for CityPersons which contains 6.47 pedestrians per image on average are 46.79% and 24.19%. It can be seen that the occlusion in CrowdHuman is more serious, and all kinds of occlusion are relatively balanced. The dataset has the characteristics of large amount of data and rich annotations. It can better evaluate the detector's performance in crowded scenes.

(2) COCO Dataset.

We use COCO dataset to test the adaptability of proposed approach for general object detection. For the large-scale dataset COCO, it has 80 categories. The validation set and training set of COCO are the same as that of COCO 2017.

5.2. **Evaluation Metrics.** Average precision (AP), missed detection rate (MR^{-2}) and Jaccard index (JI) are used to evaluate the network model's performance.

(1) Average Precision (AP).

AP is average precision, which refers to the area under the P-R curve formed by taking the recall as abscissa and precision as ordinate. The Equation is as follows:

$$AP = \int_0^1 p(r)dr \tag{13}$$

Where r is the independent variable which represents the recall. $p(\cdot)$ is the function to represent the precision rate.

(2) Missed Detection Rate ($\mathbf{MR}^{-2}$).

The rate of missed detection MR^{-2} is an important indicator for evaluating detectors of human detection. It represents false positives per image (FPPI). MR^{-2} only focuses on the recall when FPPI is less. It is quite sensitive to false detection samples. For the false detection with high confidence, it is more sensitive. When the value of MR^{-2} is smaller, the performance of detector is better. FPPI is defines as:

$$FPPI = \frac{FN}{Number} \tag{14}$$

Here, $Number$ is the number of images in test set. FN(False Negatives) is false negative sample.

(3) Jaccard Index (JI).

An index JI that measures the human detection model in crowded scenes is proposed, when the CrowdHuman dataset is published. It is utilized to evaluate the degree of overlap between ground-truth box and prediction set. The larger JI indicates better performance. The definition Equation is as follows:

$$JI(D,G) = \frac{|IoUMach(D,G)|}{|D| + |G| - |IoUMatch(D,G)|} \tag{15}$$

Where D is the set of detection box. G denotes the set of ground-truth box. IoUMatch represents matching algorithm.

5.3. **Implementation Details.** The human detection method of CrowdDet [33] is used as the benchmark model of the experiment. The model composition and parameter settings refer to the relevant settings in CrowdDet. The backbone is ResNet-50 pretrained on ImageNet. In the Region Proposal Networks (RPNs), the size of the anchor box should be set from low to high as $\{32^{-2}, 64^{-2}, 128^{-2}, 256^{-2}, 512^{-2}\}$, while the aspect ratios are $\{1:1, 1:2, 1:3\}$. Therefore, there are 15 anchor boxes of different sizes and proportions in the pyramid structure. To avoid the two quantization errors caused by the RoIPooling, RoIPooling is replaced by ROI align.

The model is trained for 30 epochs in total. The training phase is carried out on 1080Ti GPU. There are 2 images in the mini-batch. A total of 7500 iterations are included in an epoch. In the Region Proposal Networks, the positive and negative ratio of anchor used for training is 1:1, and the total number is 256; The positive and negative ratio of regions of interest (ROIs) used for classification and regression is 1:1, with a total of 512. The optimization algorithm chooses the mini-batch gradient descent. The value of momentum and weight decay are 0.9 and 0.0004. The initial learning rate is set to 0.00125. It decreased to 0.00025 and 0.00003 at the 24th and 27th epochs respectively.

In the initial experiment, equilibrium coefficient α and β in E-Reploss are 0.5 and 0.5. In addition, when IOU variant loss function replaces the Smooth_{L_1} loss function, there is a equilibrium coefficient λ between classification and regression loss in the Region Proposal Networks, as shown in Equation (16). In this experiment, the value of λ is set to 6.

$$L = L_{\text{cls}} + \lambda \cdot L_{\text{reg}} \tag{16}$$

Where L_{cls} and L_{reg} represents category loss and regression loss of bounding box respectively.

6. Experiment and Analysis of Result.

6.1. **Overall results.** For the two-stage model, S1 and S2 in the table represent the first stage and the second stage. $\surd$ indicates that the method is used in this stage. A null value (-) indicates that the regression loss function (Smooth_{L_1}) has not been changed in the baseline.

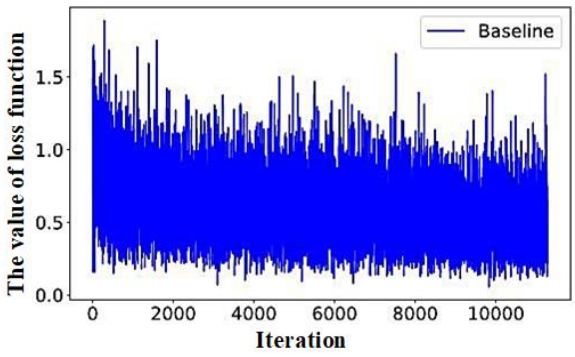
Table 1, shows that when E-Reploss appears in both stages of the model or only in the second stage, the test result MR^{-2} increases by 1.5% and 2.0% compared with the baseline. The model's performance is significantly lower than that of baseline. When E-Reploss only appeared in the first stage of the baseline, the model achieved the best test results which achieved 40.6%, 90.3% and 82.9% test results on MR^{-2}, AP and JI. In the results, MR^{-2} and JI were 0.8% and 0.2% higher than the baseline, and AP was at the same level as the baseline. The above experimental results indicate that E-RepLoss effectively reduces false detection and missed detection. Moreover, comparing (2) (3) and (4) in Table 1, it can be concluded that E-RepLoss can enhance the performance of the model in the first stage, while it will damage the performance in the second stage.

In order to look for the reasons of the above experimental results, this paper visualizes the values of the regression loss function in Table 1 (1) and (4). The results are shown in Figure 1, The ordinate represents loss function's value, and the abscissa represents iteration. Comparing the visualization results of the loss function in model's different stages, it indicates that E-RepLoss has a large loss compared with the loss function in the baseline in the first phase. While in the second phase, the difference between two results

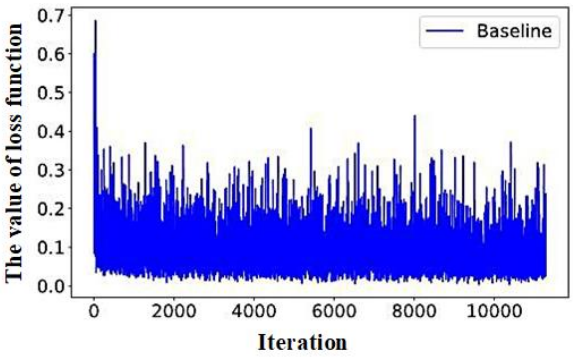
TABLE 1. Comparison of experimental results between the baseline and the baseline based on E-Reploss

Methods	S1	S2	$MR^{-2}/\%$	AP/%	JI/%	$\Delta MR^{-2}/\%$	$\Delta AP/\%$	$\Delta JI/\%$
(1) Smooth l1	√	√	41.4	90.3	82.7	-	-	-
(2) E-RepLoss	√	√	42.9	89.2	80.9	−1.5	−1.1	−1.8
(3) E-RepLoss	-	√	43.4	89.2	80.5	−2.0	−1.1	−2.2
(4) E-RepLoss	√	-	**40.6**	90.3	**82.9**	**+0.8**	0.0	**+0.2**

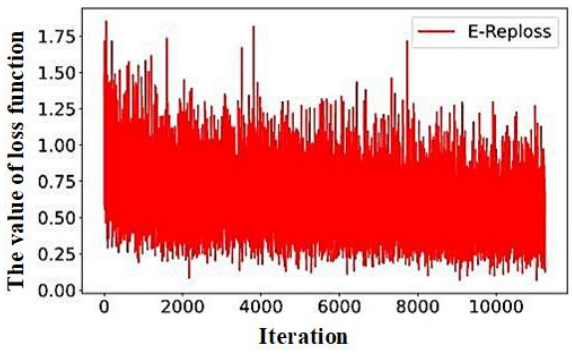
is not significant because both loss functions are Smooth_{L_1}. From the above visualization results, the larger value of loss function in E-RepLoss enables the model to update the parameters with relatively large gradient in the process of back-propagation. It makes the model have a strong learning ability. But the repulsion term may over-repel the surrounding targets. Using the loss function Smooth_{L_1} in the second stage can gradually make the deviated bounding box return. However, if E-RepLoss is still used in the second stage, the stability of the model will be destroyed.



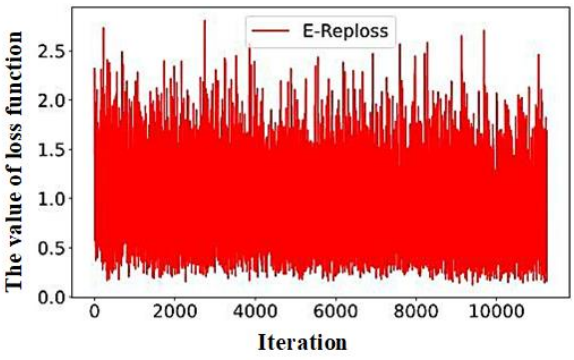
Smooth_{L_1} (S1)



Smooth_{L_1} (S2)



E-RepLoss (S1)



Smooth_{L_1} (S2)

FIGURE 1. Visual results of the benchmark model based on different loss functions

6.2. **Ablation study.** In order to explain the effects of EIoU loss function, RepLoss, and each of the parameters involved in the E-RepLoss, separate experiments are conducted for each component in this paper.

The performance of EIoU loss function and RepLoss in the first and second stages of the model are shown in Table 2, The table indicates that the model's performance is lower than that of the baseline when EIoU loss function is used in any stage of the model. The

indicator MR^{-2} of model is only lower than that of the baseline in the first stage. Due to the different number of anchors, the loss function based on the IOU variant performs better in the single-stage model than the two-stage model. This paper analyzes why the effect of EIoU loss function is poor. The reason is that the increase in the number of candidate boxes leads to the decline of performance.

TABLE 2. Comparison of experimental results between the baseline and the baseline based

Methods	S1	S2	MR^{-2}/%	AP/%	JI/%	ΔMR^{-2}/%	ΔAP/%	ΔJI/%
Smooth l1	√	√	41.4	90.3	82.7	-	-	-
EIoU Loss	√	√	44.2	88.3	78.6	−2.8	−2.0	−4.1
EIoU Loss	√	-	43.8	89.9	79.4	−2.4	−0.4	−3.3
EIoU Loss	-	√	44.4	88.2	78.7	−3.0	−2.1	−4.0
RepLoss	√	√	43.5	89.0	81.0	−2.1	−1.3	−1.6
RepLoss	-	√	43.9	88.7	80.2	−2.5	−1.6	−2.5
RepLoss	√	-	**41.1**	90.1	**82.8**	**+0.3**	−0.2	**+0.1**

To study the impact of equilibrium coefficient α and β on the results, relevant experiments have been done in this part. Table 3, shows experimental results of the equilibrium coefficients α and β at different values in Equation (9). In standard RepLoss, the performance of the model is the best when α and β are 0.5. Table 3 indicates that best results are obtained when the values of equilibrium coefficients α and β are 0.3 and 0.7 in the E-RepLoss respectively. The values of β become larger as expected, because the hyperparameter β corresponds to the loss $L_{\text{E-RepBox}}$. The offset problem of the bounding box is prevalent in the crowded scene.

TABLE 3. Impact of parameters α and β in E-Reploss for the model

α	β	MR^{-2}/%	AP/%	JI/%
0.3	0.7	**40.5**	**90.4**	**82.9**
0.5	0.5	40.6	90.3	**82.9**
0.7	**0.3**	40.9	90.2	82.6

Relevant experiments are carried out to analyze the equilibrium problem of classification and regression loss in Equation (12). Table 4, is the results of experiment. The value of λ is 2.5. The value is taken in the standard EIoU loss function. Table 4, indicates that the best performance of model is achieved when λ is 6.

TABLE 4. Impact of regression loss weight λ for the model

λ	MR^{-2}/%	AP/%	JI/%
1	40.9	90.2	82.5
2	40.8	90.2	82.7
6	**40.5**	**90.4**	**82.9**
12	41.1	90.2	82.6

In this paper, other common IoU variants loss is combined with RepLoss in addition to the EIoU loss function. The relevant experiments are conducted for the combined loss function. All experimental parameters have the same setting. Table 5, is the results of

experiment. In Table 5, (3) to (6) indicate the loss functions in which GIoU, DIoU, CIoU and EIoU loss functions are combined with RepLoss respectively. G-RepLoss is calculated by replacing IoU in RepLoss by GIoU, while D-RepLoss and C-RepLoss are designed in the way of E-RepLoss.

From Table 5, except for C-RepLoss, all indicators of this loss function have been improved. G-RepLoss achieved the same results as RepLoss on the index MR^{-2}. It is analyzed that the reason for the degradation of GIoU to IoU is overlaps between bounding boxes. Moreover, the performance of CIoU loss function is between DIoU and EIoU loss function in the field of general target detection. This paper predicts that the result of C-RepLoss should be between D-RepLoss and E-RepLoss, but the experimental results do not meet the expectations.

Table 5. The improved RepLoss based on IOU and its variant loss function

Methods	MR^{-2}/%	AP/%	JI/%
(1) Baseline	41.4	90.3	82.7
(2) RepLoss	41.1	90.1	82.8
(3) G-RepLoss(GIoU)	41.1	90.3	82.8
(4) D-RepLoss(DIoU)	40.8	90.4	83.0
(5) C-RepLoss(CIoU)	43.4	89.6	79.4
(6) E-RepLoss	40.5	90.4	82.9

6.3. **Generalization analysis.** In order to study the generalization ability of E-RepLoss, this paper carries out relevant experiments on the open source tools mmdetection (pytorch1.7.1, cuda10.1). E-RepLoss is combined with general models of human detection. The experiment is conducted on the conventional object detection dataset COCO2017. Table 6, show the experiment results. The table indicate that the performance of the two-stage object detection model Faster-RCNN combined with E-RepLoss and the single-stage object detection model RetinaNet has been improved.

Table 6. The performance of E-RepLoss on COCO dataset

Methonds	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster-RCNN	Resnet-50	37.5	58.4	40.8	21.6	41.3	48.3
Faster-RCNN + E-RepLoss	Resnet-50	**38.1**	**59.1**	**41.4**	**22.4**	**42.1**	**49.0**
RetinaNet	Resnet-50	36.5	55.6	39.2	20.6	40.1	47.8
RetinaNet + E-RepLoss	Resnet-50	**37.2**	**56.1**	**40.0**	**21.5**	**40.8**	**48.4**

The above experimental results show that the proposed method E-RepLoss has achieved improvement of performance on various indicators. It comprehensively illustrates the effectiveness of this method.

7. **Conclusions.** In crowded scenes, human detection has the problems of bounding box offset, false deletion and the lack of adaptation between the loss function and the model mechanism. An improved loss function E-Reploss is proposed from the perspective of the model's loss function to deal with the problems. The loss function is composed of attraction loss function and repulsion loss function. It strengthens the learning of the network model to the occluded target in the crowded scene, and improves model's recognition ability. The results of experiment indicate that the loss function has been

improved on all the indicators. As for MR^{-2} and JI, they obtained a 0.8% improvement and 0.2% improvement respectively. Secondly, in the ablation experiment, the effects of various functions and parameters constituting the E-Reploss loss function were explored. Except for the EIoU loss function, the loss of other IoU variants were combined with RepLoss to analyze the results of each loss function.

In the next work, some research directions are worth to be concerned. For example, this method is extended to other detection models and is used to solve the practical problem that pedestrians occlude each other in the dense crowds.

Funding Statement: This work was supported by Jilin Science and Technology Innovation Development Program Projects (No. 20190302202).

REFERENCES

[1] X. K. Xu, Y. Ma, X. Qian, "Scale-aware Efficient Det: real-time pedestrian detection algorithm for automated driving," *Journal of Image and Graphics*, vol. 26, no. 1, pp. 93-100, 2021.

[2] Q. Mou, L. Wei, C. Wang, D. Luo, and C. Gao, "Unsupervised Domain-adaptive Scene-specific Pedestrian Detection for Static Video Surveillance," *Pattern Recognition,* vol. 118, no. 9, 108038, 2021.

[3] X. Wu, D. Sahoo, and S. Hoi, "Recent Advances in Deep Learning for Object Detection," *Neurocomputing*, vol. 396, pp. 39-64, 2020.

[4] D.-T. Nguyen, W.-Q. Li., and P. O. Ogunbona, "Human detection from images and videos: a survey," *Pattern Recognition*, vol. 51, pp. 148-175, 2016.

[5] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How Far are We from Solving Pedestrian Detection?" *IEEE Conference on Computer Vision & Pattern Recognition IEEE,* pp. 1259-1267, 2016.

[6] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, pp. 436-444, 2015.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *IEEE Computer Society IEEE Computer Society,* 2013. [Online]. Available: https://doc.taixueshu.com/foreign/arXiv13112524.html, 2013.

[8] C. Zhou, J. Yuan, "Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection," *Asian Conference on Computer Vision. Springer, Cham,* pp. 305-320, 2016.

[9] 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,* pp. 481-498, 2017.

[10] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 10781-10790, 2020.

[11] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv:* 2004.10934, 2020.

[12] Y. Tian, L. Ping, X. Wang, and X. Tang, "Deep Learning Strong Parts for Pedestrian Detection," *IEEE International Conference on Computer Vision IEEE,* pp.1904-1912, 2016.

[13] J. Redmon, A. Farhadi, "YOLO9000: better, faster, stronger," *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6517-6525, 2017.

[14] C. Zhu, Y. He, M. Savvides, "Feature selective anchor-free module for single-shot object detection," *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 840-849, 2019.

[15] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd," *European Conference on Computer Vision (ECCV),* pp. 637-653, 2018.

[16] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding Box Regression With Uncertainty for Accurate Object Detection," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE,* pp. 2883-2892, 2019.

[17] P. Doll´ar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, no. 8, pp. 1532-1545, 2014.

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

[19] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp.1751-1760, 2015.

[20] P. Doll´ar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," *British Machine Vision Conference, BMVC 2009,* London, UK, September 7-10, 2009. Proceedings DBLP.

[21] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" *European Conference on Computer Vision,* pp. 443-457, 2016.

[22] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," *In Proceedings of the IEEE International Conference on Computer Vision,* pp. 82–90, 2015.

[23] 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),* pp. 1-7, 2017.

[24] Y.-X. Gong, Z.-Q. Zhang, G.-Z. Duan, Z. Ma, and M. Xie, "AccNet: occluded scene text enhancing network with accretion blocks," *Machine Vision and Applications,* vol. 34, 1, 2023.

[25] S.-L. Zhang, K.-X. Zheng, and H.-Y. Sun, "Analysis of the Occlusion Interference Problem in Target Tracking," *Mathematical Problems in Engineering,* vol. 2022, 4605111, 2022.

[26] Q. Li, Y.-Q. Bi, R.-S. Cai, and J. Li, "Occluded pedestrian detection through bi-center prediction in anchor-free network," *Neurocomputing,* vol. 507, pp. 199-207, 2022.

[27] R. Girshick, "Fast R-CNN," *IEEE International Conference on Computer Vision,* pp. 1440-1448, 2015.

[28] H. Rezatofighi, N. Tsoi, J.-Y. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE,* pp. 658-666, 2019.

[29] Y.-F. Zhang, W.-Q. Ren, Z. Zhang, Z. Jia, L. Wang, and T.-N. Tan, "Focal and Efficient IOU Loss for Accurate Bounding Box Regression," *arXiv preprint arXiv:* 2101.08158, 2021.

[30] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," *AAAI Conference on Artificial Intelligence,* vol. 34, no. 07, pp. 12993-13000, 2020.

[31] X. Wang, T. Xiao, Y. Jiang, S. Shuai, and C. Shen, "Repulsion Loss: Detecting Pedestrians in a Crowd," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE,* pp. 7774-7783, 2018.

[32] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "CrowdHuman: A Benchmark for Detecting Human in a Crowd," *arXiv preprint arXiv:* 1805.00123, 2018.

[33] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in Crowded Scenes: One Proposal, Multiple Predictions," *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 12211-12220, 2020.