



Article

A Sample Balance-Based Regression Module for Object Detection in Construction Sites

Xiaoyu Wang¹, Hengyou Wang¹ , Changlun Zhang^{1,*} , Qiang He¹ and Lianzhi Huo²

¹ School of Science, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; wwangxiaoyuu@163.com (X.W.); wanghengyou@bucea.edu.cn (H.W.); heqiang@bucea.edu.cn (Q.H.)

² Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; huolz@aircas.ac.cn

* Correspondence: zclun@bucea.edu.cn; Tel.: +86-139-1055-8268

Abstract: Object detection plays an important role in safety monitoring, quality control, and productivity management at construction sites. Currently, the dominant method for detection is deep neural networks (DNNs), and the state-of-the-art object detectors rely on a bounding box regression (BBR) module to localize objects. However, the detection results suffer from a bounding box redundancy problem, which is caused by inaccurate BBR. In this paper, we propose an improvement of the object detection regression module for the bounding box redundancy problem. The inaccuracy of BBR in the detection results is caused by the imbalance between the hard and easy samples in the BBR process, i.e., the number of easy samples with small regression errors is much smaller than the hard samples. Therefore, the strategy of balancing hard and easy samples is introduced into the EIOU (Efficient Intersection over Union) loss and FocalL1 regression loss function, respectively, and the two are combined as the new regression loss function, namely EFocalL1-SEIOU (Efficient FocalL1-Segmented Efficient Intersection over Union) loss. Finally, the proposed EFocalL1-SEIOU loss is evaluated on four different DNN-based detectors based on the MOCS (Moving Objects in Construction Sites) dataset in construction sites. The experimental results show that the EFocalL1-SEIOU loss improves the detection ability of objects on different detectors at construction sites.

Keywords: object detection; loss function; construction sites



Citation: Wang, X.; Wang, H.; Zhang, C.; He, Q.; Huo, L. A Sample Balance-Based Regression Module for Object Detection in Construction Sites. *Appl. Sci.* **2022**, *12*, 6752. <https://doi.org/10.3390/app12136752>

Academic Editor: Krzysztof Koszela

Received: 25 May 2022

Accepted: 1 July 2022

Published: 3 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In construction projects, inefficiencies are often caused by the occurrence of unknown accidents [1]. To improve productivity, the construction industry has continuously strengthened safety management and engineering supervision to reduce casualties and improve construction efficiency during construction. Taking China as an example, there were 34,600 accidents in 2021, of which 26,300 died. Construction industry casualties are consistently high across all industries, and more than half of them should have been preventable. The traditional methods of monitoring construction operations are through job sampling, personnel testing, etc. However, these methods require many human, material and financial resources [2]. Therefore, the application of modern information and communication technology is of great significance in preventing safety accidents and improving the efficiency of construction operations.

Computer vision techniques have aroused wide interest in academia and industry, such as smart glasses systems [3], automatic fire detection and notification [4], and construction worker safety inspection [5]. Many applications have been developed, such as collision risk prevention [6], ground engineering equipment activity analysis [7], safety helmet wearing detection [8], noncertified work detection [9], and construction activities identification [10]. For these applications, it is a basic requirement to accurately detect and correctly identify workers and equipment through digital images in construction sites [5–10]. However, the detectors used in previous studies [5] are mainly based on the

form of hand-crafted features and sliding windows traversing the whole image; also, the algorithm runs slowly, with insufficient accuracy and without generalization.

In recent years, DNNs have shown superior performance in object detection. Compared with the previously artificially designed detectors, the detector based on DNNs realizes feature extraction by adaptive convolution, which improves the accuracy and speed of the algorithm and is more robust in object detection. The deep learning object detection method consists of three parts. First, the backbone [11–15] part of the network is responsible for extracting the necessary feature information from the input data. The quality of feature information extraction is crucial for subsequent object detection. The second part is the neck [16–20], which is followed by the backbone to better fuse and extract feature maps. The last part is the detection head, which is responsible for obtaining the object category and position from the extracted features. Deep neural network methods based on deep learning can be roughly divided into two categories. One is the regression-based one-stage method [21,22]. The other is a two-stage method [23].

Despite these different detection frameworks, BBR is a critical step in predicting rectangular boxes for localizing target objects. In terms of metrics to measure the localization of object detection, there are various loss functions for object detection, such as focal loss [24], class-balanced loss [25], balanced loss for classification and BBR [26], and gradient flow balancing loss [27]. Nevertheless, rectangular BBR loss is still the most popular loss function approach.

The loss functions are evolving through continuous innovation. Initially, L1 loss was used, but it is difficult to converge when the error is small in the later stages of training. The derivative of L2 loss used in R-CNN (regional convolutional neural network) [28] and SPP-Net (spatial pyramid pooling network) [29] is x , but it is unstable against outliers. SmoothL1 loss [30] combines L1 loss and L2 loss to perfectly avoid their drawbacks and is applied in Fast R-CNN [30] and Faster R-CNN [23] but does not have scale invariance. YOLOv1 (You Only Look Once) [31] proposed that the square root of the difference of a bounding box can alleviate the scale sensitivity of the loss function. The $2 - wh$ loss used in YOLOv3 [22] also reduces the effect of scale on regression accuracy. The Dynamic SmoothL1 loss in Dynamic R-CNN [32] dynamically controls the shape of the loss function, thus gradually focusing on high-quality anchor boxes. BalancedL1 loss in Libra R-CNN [26] similarly focuses on increasing the weight of the easy sample gradient but does not control the gradient of outliers. FocalL1 [33] increases the gradient for high-quality samples and decreases the gradient for low-quality samples. However, any example of better regression will directly improve the quality of the final predicted boxes and should not overly suppress the contribution of low-quality samples. However, the l_n -norm losses do not pay attention to the intrinsic connection in the four variables (x, y, w, h) , and for this problem, IOU (Intersection over Union) [34] loss was proposed and achieved excellent results. Since then, GIOU (Generalized IOU) [35], DIOU (Distance IOU) [36], CIOU (Complete IOU) [36], and EIOU (Efficient IOU) [33] have been proposed to address the weaknesses of the IOU loss function for specific problems, resulting in faster convergence and better performance. PIOUS (Pixels-IOU) [37] added angles to assist the IOU loss function. Alpha-IOU [38] is an existing uniform idempotent based on the IOU loss. The above research proves the importance of the loss function in the regression process of object detection. However, the current research on the loss function still has some shortcomings, and the detector based on anchors generally has a problem of imbalance between hard and easy samples.

Currently, object detection algorithms ensure that moving objects on construction sites can be detected by detectors. However, the bounding box redundancy problem in object detection is shown in Figure 1, especially for objects with large stretches. The reason for the redundant bounding box problem is the imbalance in the number of hard and easy samples in the regression process. Therefore, the hard-to-regress samples in the existing regression modules provide a larger contribution, resulting in bounding box redundancy. The hard-versus-easy sample balance strategy can improve the easy sample regression

contribution and reduce the hard sample regression contribution. To this end, we address this problem by incorporating a hard–easy sample balancing strategy into the BBR module.



Figure 1. Bounding box redundancy problem for object detection at a construction site.

In this paper, we propose a loss function for the object detection BBR module, which adopts a hard-versus-easy sample regression strategy. The design is based on the bounding box redundancy problem, and the regression module is studied to improve the accuracy of object detection. The proposed loss functions include three types: (1) A Segmented Efficient IOU (SEIOU) loss function is proposed, which uses a piecewise function to perform piecewise regression on the penalty term of the loss function. (2) A regression loss function focusing on the balance of hard and easy samples is designed to change the contribution of hard and easy samples in the model optimization process to improve the model accuracy. (3) Finally, the two methods are combined into a new BBR loss function EfocalL1-SEIOU loss to achieve more accurate object detection. The evaluation on several advanced object detection models, including Faster R-CNN [23], Mask R-CNN [39], Cascade R-CNN [40], and YOLOF [41], uses the construction site dataset MOCS (Moving Objects in Construction Sites) [42]. The results show that an improvement in the accuracy of significance is achieved, which illustrates the generalizability of this proposed loss function.

The contributions of this paper can be summarized as follows:

1. Considering the redundancy problem of bounding boxes in the construction site, it is caused by the imbalance of hard and easy samples in the BBR process. The strategy of balancing hard and easy samples is introduced into the IOU-based loss to help the BBR as much as possible by segmenting the hard and easy samples.
2. The strategy of balancing hard and easy samples is also introduced into the l_n -norm loss by controlling the regression gradient to obtain better regression results.
3. Compared with the previous object detection loss function, this loss function introduces the hard-versus-easy sample balancing strategy and combines the IOU-based loss and the l_n -norm loss as a new loss function to obtain better performance.

The remainder of the paper is organized as follows. Section 2 introduces the recent literature on the hard-versus-easy sample balance problem and object detection in construction sites. In Section 3, we describe the limitations of IOU-based losses and the generation and analysis of SEIOU, EfocalL1 loss, and EfocalL1-SEIOU. In Section 4, we evaluate the performance of EfocalL1-SEIOU loss on four different advanced object detectors and perform some ablation experiments. In Section 5, we describe the limitations and discussion. In Section 6, we draw some conclusions.

2. Related Work

In this section, we briefly survey the related work of the problem of the imbalance of hard and easy samples and object detection in construction sites.

2.1. Imbalance of Hard and Easy Sample

There are two common problems with anchor-based detectors: positive and negative sample imbalance and hard and easy sample imbalance. Easy samples are samples with small loss. Hard samples are samples with large loss. For classification problems, the two above are included. For the regression problem, only the hard and easy sample imbalance problems are included.

For the classification problem, OHEM (Online Hard Example Mining) [43] considers that there are many simple negative samples; thus, the hard negative samples should be mined. However, OHEM is sensitive to noise. In Focal loss [24], the contribution of positive samples is increased, and the contribution of negative samples is decreased. Since GHM (Gradient Harmonizing Mechanism) [27] considers hard samples to be outliers, it reduces the weight of hard samples. For regression problems, GHM decreases the hard sample weight because it considers hard samples as outliers. However, it does not reduce the easy sample weight, because any example of better regression will directly improve the quality of the final predicted boxes. Both BalancedL1 loss [26] and FocalL1 loss [33] emphasize improving the gradient of the easy sample gradient, while the latter also reduces the hard sample gradient. However, compared to the SmoothL1 loss [30], the easy sample gradient does not increase significantly, while the hard sample gradient decreases significantly.

2.2. Object Detection in Construction Sites

Compared to general object detection, few research has been made on object detection in construction sites. Roberts et al. performed object detection for cranes to monitor crane-related safety hazards [44]. Kim et al. used YOLOv3 to detect workers and heavy machinery to prevent being attacked by hazards [6]. Roberts et al. trained a DNN-based detector on an advanced infrastructure management dataset [45] containing heavy machinery for five different types of construction work to analyze the productivity of excavators and trucks [7]. As the collected datasets tend to be small in size, the generalization performance of the algorithms is reduced. With the generation of large-scale datasets, a simple approach is to directly apply mainstream object detection algorithms to datasets under construction sites to improve detection accuracy and thus carry out various applications.

3. Materials and Methods

3.1. Experiments Details

3.1.1. Dataset and Evaluation Metric

The current dataset suitable for DNNs is MOCS [42], which contains 174 construction sites covering various projects. The dataset of moving objects in construction sites contains 41,668 images in 13 categories. A benchmark containing different DNN-based detectors was made using the MOCS dataset. The results show that all the trained detectors can detect the objects on the construction site.

Construction site images have more semantic information and complex texture information; thus, we conducted simulated experiments with the MOCS dataset to investigate the advantages of SEIOU and EfocalL1 loss and the importance of the balance between hard and easy samples, respectively. We present the experimental results of object detection with the dataset MOCS in construction sites. We use the MOCS training set for training and show the ablation studies of the MOCS validation set. COCO-style Average Precision (AP) is chosen as the evaluation index.

3.1.2. Implementation Details

For fair comparisons, all experiments are implemented with PyTorch [46]. The backbones used in the experiments are publicly available. For all experiments on the MOCS

dataset, we use ResNet-50 backbone and run 12 epochs. We train detectors with a GPU (2 images per GPU), adopting the stochastic gradient descent (SGD) optimizer. The default weight of BBR is set to 1 based on l_n -norm losses and 10 for IOU-based losses.

3.1.3. Network Architecture

There are roughly two types of deep learning object detection network architectures. One is a two-stage method based on region proposal, which performs two classifications and regressions, as shown in Figure 2. First, the candidate regions are generated and screened, the features of the region of interest (ROI) in the image are obtained, and then, the neural network is trained for classification. Another class of regression-based one-stage methods regresses the target in a single convolutional neural network, as shown in Figure 3. Despite the different detection frameworks, BBR is a critical step in locating target objects. As shown in the figure, the red dotted box in the two architecture diagrams is the placement of our proposed regression loss function.

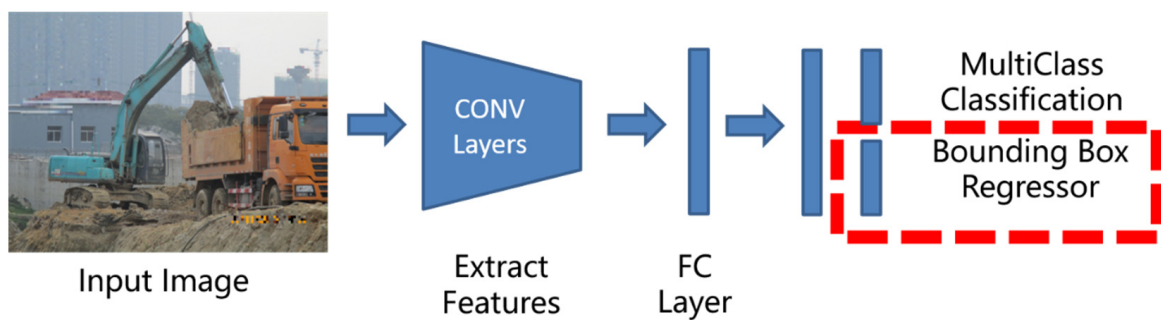


Figure 2. Network architecture of a one-stage detector.

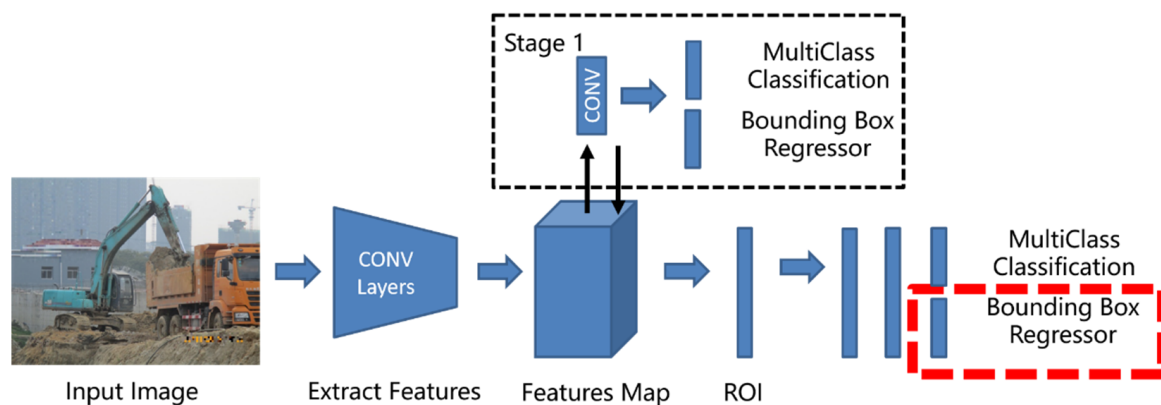


Figure 3. Network architecture of a two-stage detector.

3.2. Limitations of IOU-Based Losses

In this subsection, we analyze the flaws of four IOU-based loss functions, i.e., the IOU [34], GIOU [35], CIOU [36], and EIOU [33] loss.

3.2.1. Limitations of IOU Loss

The IOU loss [34] for measuring similarity between two bounding boxes B , B^{gt} is attained by:

$$L_{IOU} = 1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|}, \tag{1}$$

which has some good properties, such as symmetry and scale insensitivity. However, it cannot reflect the closeness of B and B^{gt} . Therefore, the direction of gradient descent cannot be provided for optimization. Second, the IOU loss does not reflect how the two boxes intersect, i.e., how good the overlap is.

3.2.2. Limitations of Generalized IOU Loss

To solve the drawback of IOU loss when the boxes overlap, GIOU [35] is proposed as follows.

$$L_{GIOU} = 1 - \frac{|B \cap B^{st}|}{|B \cup B^{st}|} + \frac{|C - (B \cup B^{st})|}{|C|}, \tag{2}$$

C is the closure of B and B^{st} . GIOU mitigates the problem of gradient disappearance in the non-overlapping case by penalizing the term and reflects the goodness of the overlap of the two boxes. However, in order to reduce the loss function, the regression process will first choose the optimized penalty term to reduce the closure and then choose the optimized IOU term. There is a problem that the area of the prediction frame is mistakenly increased.

3.2.3. Limitations of Complete IOU Loss

To solve the problem of incorrectly increasing the area of the predicted box, CIOU [36] is proposed as in Equation (3).

$$L_{CIOU} = 1 - \frac{|B \cap B^{st}|}{|B \cup B^{st}|} + \frac{\rho^2(B, B^{st})}{C^2} + \alpha v, \tag{3}$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{st}}{h^{st}} - \arctan \frac{w}{h} \right)^2, \tag{4}$$

$$\alpha = \frac{v}{(1 - IOU) + v}, \tag{5}$$

where ρ^2 is the Euclidean distance between the centroids of the two boxes, and C is the diagonal of the closure of B and B^{st} . v is used to measure the width-to-height ratio, and α is the equilibrium factor. w and h are the width and height of the two boxes. CIOU loss adds a scaling coefficient penalty term to solve the problem of incorrectly increasing the area of the prediction box. However, there are still problems with the scale factor introduced by the CIOU loss. Width and height cannot be increased and decreased at the same time during the optimization process.

3.2.4. Limitations of Efficient IOU Loss

To solve the problem of optimizing the penalty term in aspect ratio, EIOU [33] is proposed for any two bounding boxes B and B^{st} , as in Equation (6).

$$L_{EIOU} = 1 - \frac{|B \cap B^{st}|}{|B \cup B^{st}|} + \frac{\rho^2(B, B^{st})}{C^2} + \frac{\rho^2(w, w^{st})}{C_w^2} + \frac{\rho^2(h, h^{st})}{C_h^2}, \tag{6}$$

C is the diagonal of the closure of B and B^{st} , C_w and C_h are the width and height of the closure. EIOU proposes to optimize the width and height of the box, respectively. When the two bounding boxes keep approaching, the parallel set will keep decreasing. To make the loss decrease, the closure should keep increasing, which is contradictory.

3.3. EfocalL1-SEIOU Loss for BBR

As mentioned in FocalL1 loss [33], the problem of hard and easy sample imbalance also exists in the BBR problem. That is, the number of high-quality and easily regressive bounding boxes with small regression loss is far less than that of low-quality and hard-to-regress bounding boxes with large regression loss. In this subsection, we first propose SEIOU, and then, we propose EfocalL1 loss based on FocalL1 loss. In addition, we combine EfocalL1 loss with SEIOU to improve the performance of the BBR loss function.

3.3.1. Segmented Efficient Intersection over Union Loss

In this method, the hard–easy sample balance strategy is introduced into the EIOU by segmenting the loss function. As shown in Figure 4, green is the EIOU method, and both red and green are our method. For any two bounding boxes B and B^{gt} defined as follows:

$$L_{SEIOU} = \begin{cases} 1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} + \frac{\rho^2(B, B^{gt})}{C_1^2} + \frac{\rho^2(w, w^{gt})}{C_{w1}^2} + \frac{\rho^2(h, h^{gt})}{C_{h1}^2}, & IOU \leq 0.5 \\ 1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} + \frac{\rho^2(B, B^{gt})}{C_2^2} + \frac{\rho^2(w, w^{gt})}{C_{w2}^2} + \frac{\rho^2(h, h^{gt})}{C_{h2}^2}, & IOU > 0.5 \end{cases} \quad (7)$$

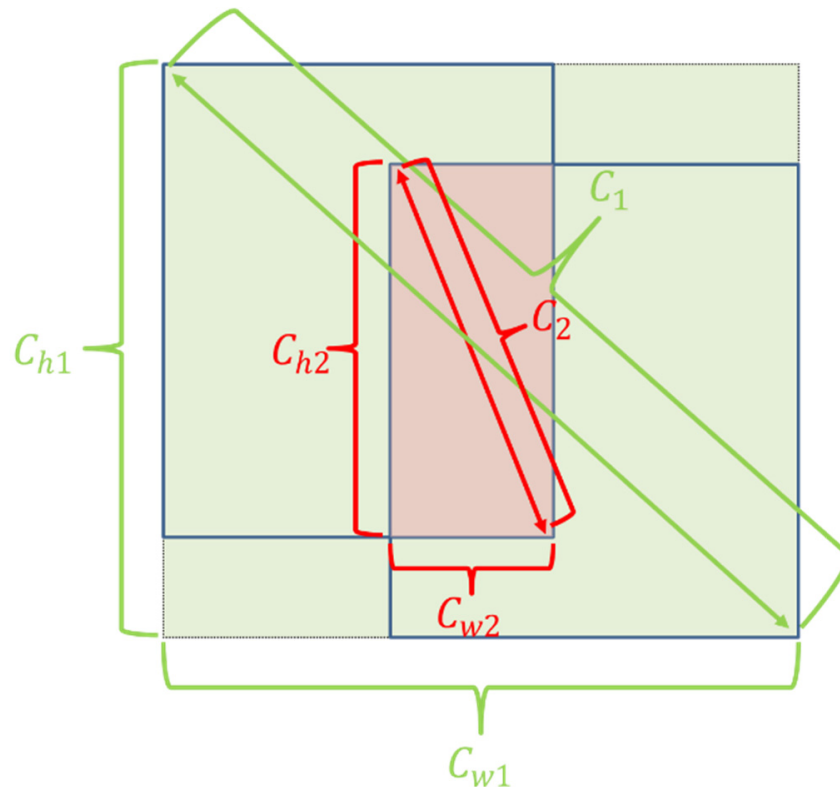


Figure 4. Differences between EIOU (green) and our method (both green and red).

C_1 is the diagonal of the closure of B and B^{gt} , C_{w1} and C_{h1} are the width and height of the closure. C_2 is the diagonal of the intersection, C_{w2} and C_{h2} are the width and height of the intersection. Similar to the definition of IOU-based losses, $iou = 0.5$ is used as the cut-off point to divide the samples with regression loss greater than 0.5 as easy samples, and those with regression loss less than 0.5 as hard samples. Among them, the intersection is used as the denominator of penalty terms for easy regression samples. As the predicted box keeps getting closer to the target box, the intersection keeps increasing, and the loss function keeps decreasing, which is the direction we expect. For the hard regression samples, we still use the closure as the denominator of the penalty term. When the two boxes keep getting closer, the closure is decreasing and the loss function is changing erratically. Therefore, the loss function regression is focused on the easy regression samples. This ensures the speed and accuracy of EIOU loss convergence while focusing more on the regression effect of high-quality bounding boxes, which makes the regression better.

3.3.2. Efficient FocalL1 Loss

The FocalL1 loss [33] emphasizes increasing the gradient of the easy sample regression and decreasing the gradient of the hard sample at the same time. However, compared with BalancedL1 loss [26], the effect of increasing the easy sample gradient is slightly less, but the hard sample gradient is reduced substantially. The outliers are harmful to the training,

but the hard samples in the initial training are not necessarily outliers. It is mentioned in GHM [27] that better prediction of any example in the regression problem will directly improve the quality of the final bounding box. Therefore, it is crucial to increase the easy sample gradient and decrease the hard sample gradient appropriately. In this method, by increasing the gradient of easy samples and reducing the gradient of hard samples, the balance strategy of hard and easy samples is introduced into the loss FocalL1 loss.

Similar to the definition of l_n -norm losses, $x = 1$ is used as the dividing point, the samples with regression loss less than 1 are divided into easy samples, and the samples with regression loss greater than 1 are divided into hard samples. According to the idea that the gradient value increases relatively in the area with small error and decreases relatively in the area with large error, we design a function curve of the BBR gradient magnitude. As β increases, the gradient of hard samples will be further suppressed, and the gradient of the easy samples will be further suppressed, as shown in Figure 5a, which is not what we expect. Therefore, the change in the gradient of the easy sample is controlled by the parameter α . As α increases, the gradient of easy samples will further increase, as shown in Figure 5b. The final gradient loss function is written in the following form.

$$f(x) \begin{cases} -\alpha x \ln(\beta x^2), & x < 1 \\ -\alpha x \ln(\beta), & \text{otherwise} \end{cases} \quad (8)$$

where x is the difference between the predicted box and the target box.

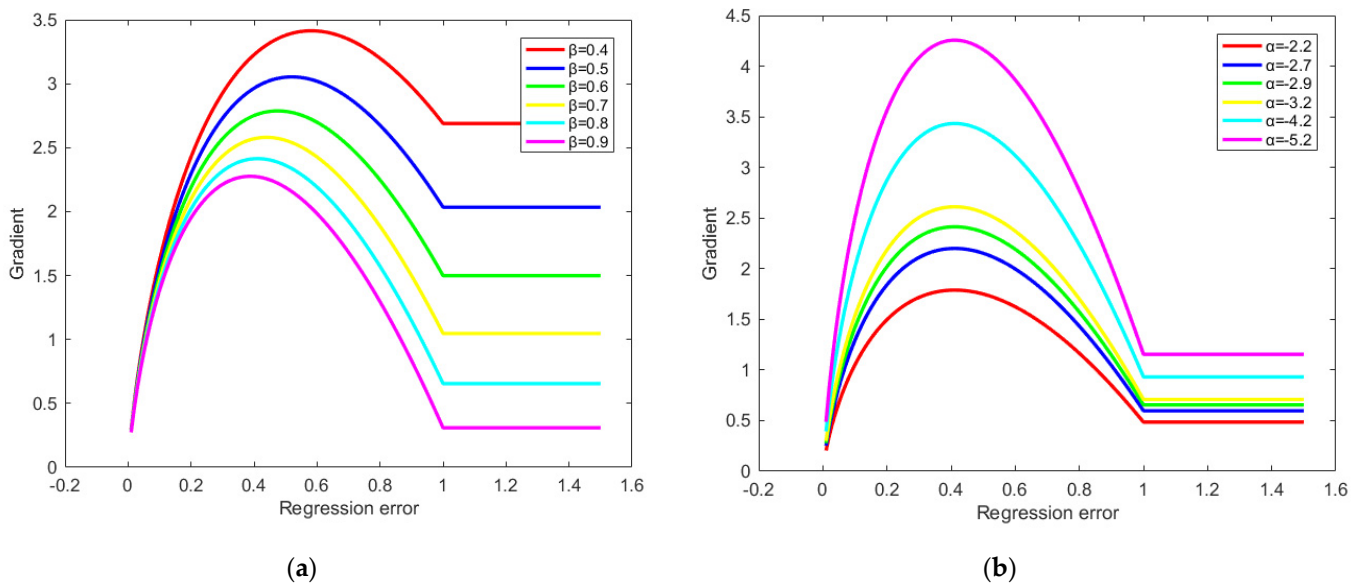


Figure 5. Possible gradient curves: (a) use β to control curve shape; (b) use α to control curve shape.

By integrating the gradient formulation above, we can obtain the EfocalL1 loss for BBR,

$$F(x) = \begin{cases} -\frac{\alpha}{2} x^2 \ln(\beta x^2) + \frac{\alpha}{2} x^2, & x < 1 \\ C, & \text{otherwise} \end{cases} \quad (9)$$

where C is a constant value. To ensure that $F(x)$ in Equation (9) is continuous at $x = 1$, we have $C = -\alpha \ln(\beta)x - \alpha \ln(\beta) + \frac{\alpha}{2} [\ln(\beta) - 1]$. As shown in Figure 6, the easy sample contribution can be increased substantially while the hard sample contribution can be decreased.

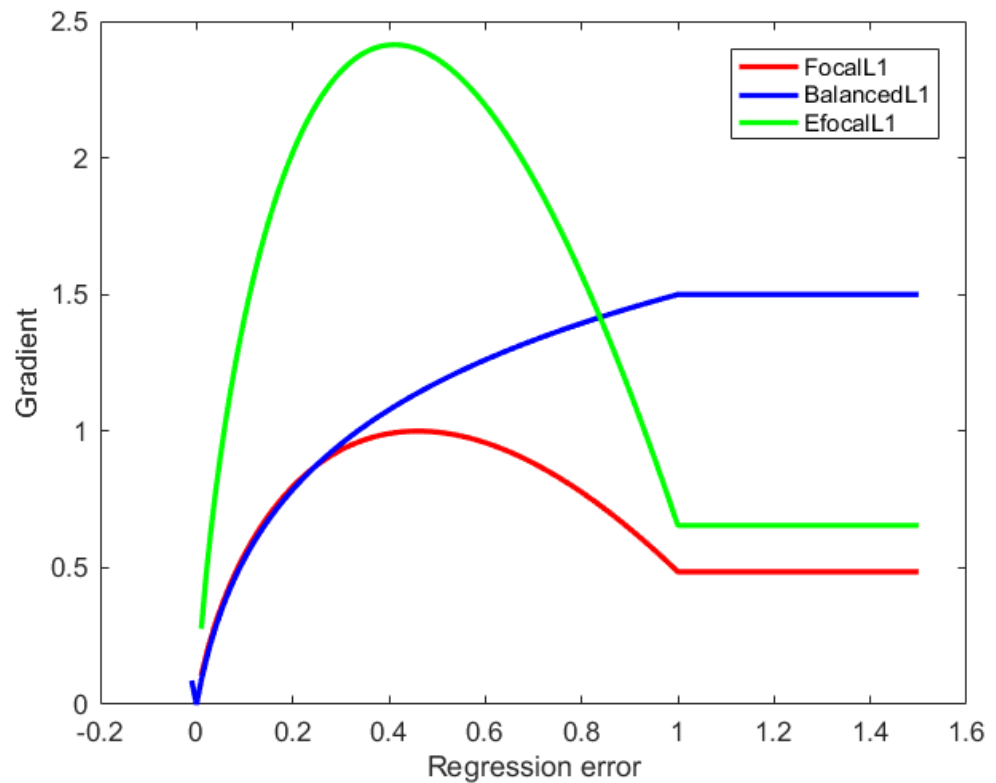


Figure 6. Curves for the gradient of FocalL1 loss, BalancedL1 loss, and our EfocalL1 loss for BBR.

3.3.3. EfocalL1-SEIOU Loss

In this method, to combine the advantages of the two types of loss functions, the two types of loss functions are nested. To enable the SEIOU loss regression process to increase the easy sample regression gradient and relatively decrease the hard sample regression gradient, one can naturally consider replacing x in Equation (9) with the SEIOU loss. However, we observe that the above combination does not work well. As L_{SEIOU} approaches zero, $\frac{\partial L_{Efocal-seiou}}{\partial SEIOU}$ approaches zero, and therefore, the global gradient also approaches zero, i.e., the weight influence of samples with a small loss of L_{SEIOU} is weakened. To tackle this problem, we use the value of IOU to reweight the SEIOU loss and obtain high regression accuracy by paying more attention to the high IOU target. We obtain EfocalL1-SEIOU loss as follows.

$$L_{Efocal-seiou} = IOU^\gamma F(SEIOU), \tag{10}$$

γ is a parameter to control the degree of inhibition of outliers.

4. Results

4.1. Ablation Studies on SEIOU Loss

We first conducted experiments with the IOU, GIOU, CIOU, EIOU, and SEIOU losses in Table 1. Bounding box loss weights control the balance between classification loss and regression loss in object detection, and we set the weight for BBR to 10.0 for fair comparisons. In addition, we chose $iou = 0.5$ as the segmentation because it is mentioned in the review [47] that $iou < 0.5$ as the false positives. We also verified the segmentation point, as shown in Figure 7a; the performance of IOU at $[0.3, 0.5]$ is maintained at 0.479. When the segmentation points are taken less at than 0.3 or more than 0.5, the performance decreases; thus, the segmentation point is set to $iou = 0.5$. The experimental results show that for other IOU-based losses, the method has better performance.

Table 1. Overall ablation studies on MOCS.

Method	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Baseline	0.458	0.730	0.501	0.171	0.364	0.564
IOU	0.476	0.728	0.520	0.179	0.373	0.590
GIOU	0.476	0.729	0.521	0.174	0.376	0.589
CIOU	0.476	0.731	0.519	0.178	0.376	0.589
EIOU	0.472	0.723	0.514	0.167	0.374	0.585
EfocalL1	0.478	0.724	0.522	0.174	0.372	0.597
SEIOU	0.479	0.736	0.522	0.174	0.376	0.594
EfocalL1-SEIOU	0.482	0.728	0.524	0.189	0.374	0.594

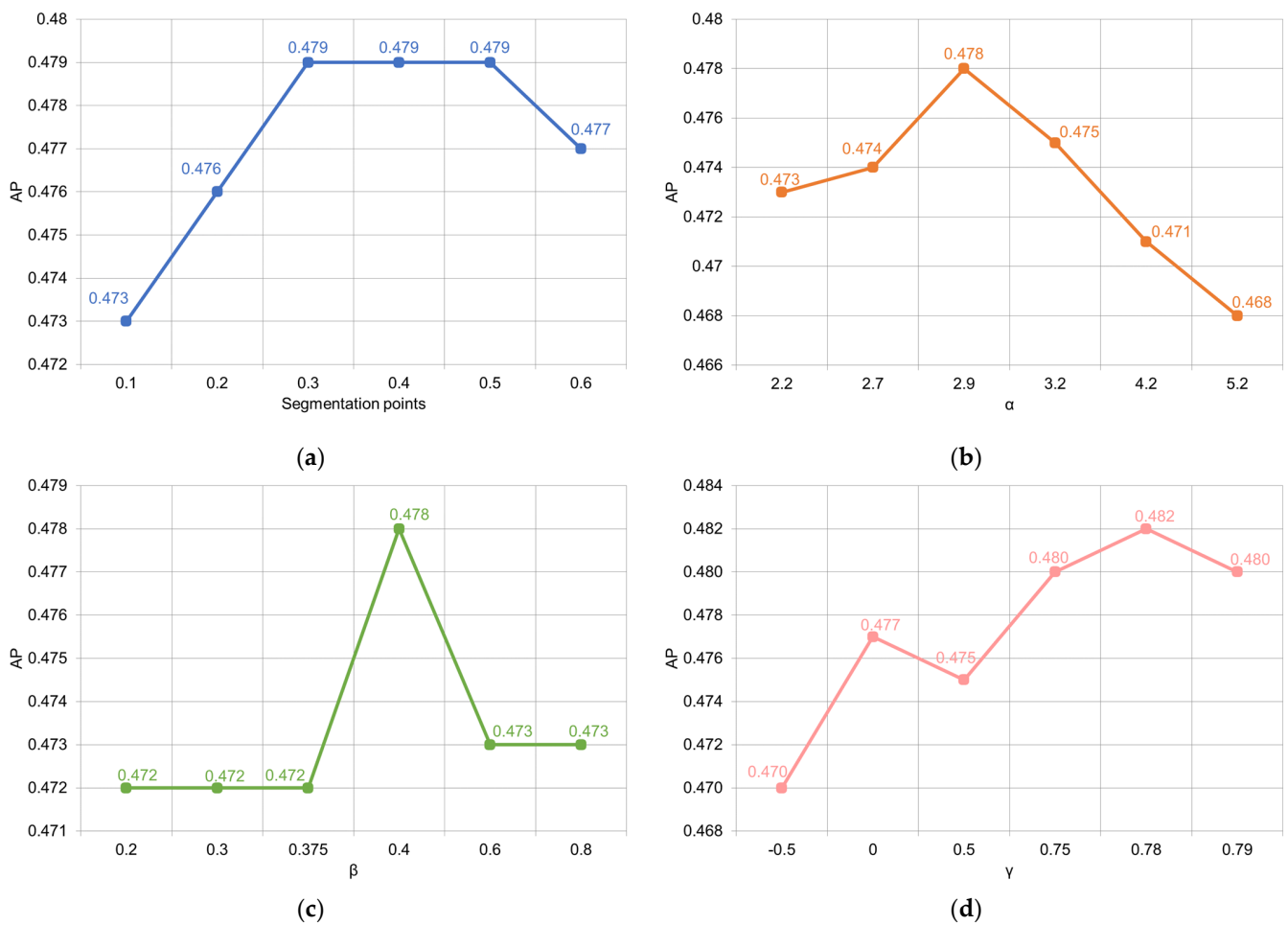


Figure 7. Performance of methods with different value of parameters. (a) SEIOU loss with different segmentation points; (b) EfocalL1 loss with different α ; (c) EfocalL1 loss with different β ; (d) EfocalL1-SEIOU loss with different γ .

4.2. Ablation Studies on EfocalL1 Loss

As shown in Figure 7b,c, we test the EfocalL1 loss for different α and β . Setting a smaller α further increases the gradient of the easy samples, and setting a larger β further suppresses the gradient of the hard samples. Finally, we find that the optimal equilibrium AP value is 0.478 when $\alpha = -2.75$ and $\beta = 0.4$, which is 0.9 higher than the FocalL1 loss [33]. We also compare the previous work BalancedL1 loss [26], which improved the AP by 0.5. These experimental results show that the EfocalL1 loss makes the model perform better.

4.3. Ablation Studies on EfoCaLL1-SEIOU Loss

To illustrate the improvements brought by the different parameters for reweighting the SEIOU loss, we compare the results of the different adjustments here. We first find that replacing x in Equation (9) with SEIOU leads to a reduction in the gradient of the easy samples and therefore does not improve the performance of this loss function. Then, we weight the EfoCaLL1 loss by IOU to obtain the EfoCaLL1-SEIOU loss. As shown in Figure 7d, we can obtain an improvement in performance. As γ increases, the performance of the loss function first increases and then decreases. We cannot suppress the gradient of hard samples extremely because the validity of hard samples still exists in the BBR process. We find that the best results are achieved when setting $\gamma = 0.78$ and using it as the default value for subsequent experiments.

4.4. Overall Ablation Studies

To demonstrate the effectiveness of each method, we performed overall ablation studies, as shown in Table 2. The SEIOU loss improved AP from 0.472 to 0.479 compared to the EIOU loss, and the EfoCaLL1 loss compared to the FocalL1 loss improved the AP by 0.9. Applying the SEIOU loss as x directly to Equation (9) is not satisfactory. The EfoCaLL1-SEIOU loss of Equation (10) is reasonable, improving the AP of baseline by 2.4. The visual effect is shown in Figure 8, where the first column is the original image, the second column is the result of Faster R-CNN using baseline, and the third column is the effect of Faster R-CNN using EfoCaLL1-SEIOU loss. From the comparison, our method can reduce the redundancy of bounding box in construction sites, which verifies the effectiveness of the EfoCaLL1-SEIOU loss function.

Table 2. Ablation studies of the EfoCaLL1 loss on MOCS.

Setting	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Baseline	0.458	0.730	0.501	0.171	0.364	0.564
BalancedL1	0.473	0.723	0.520	0.171	0.368	0.587
FocalL1	0.469	0.729	0.515	0.179	0.371	0.580
EfoCaLL1	0.478	0.724	0.522	0.174	0.372	0.597



Figure 8. Comparison chart of object detection results.

4.5. Incorporations with State-of-the-Arts

In this subsection, we evaluate the EfoCAL1-SEIOU loss by incorporating it into popular object detectors, including Mask R-CNN [39], Cascade R-CNN [40], and YOLOF [41]. Figure 9 shows the AP trends of the baselines of the corresponding detectors compared to our method during training. Our method significantly outperforms the corresponding baseline when the network finally converges stably. The results in Table 3 show that training these models by the EfoCAL1-SEIOU loss can consistently improve their performance compared to their own regression losses. Compared to other models, the improvements of Mask R-CNN and Cascade R-CNN are relatively small. There are two reasons for this. First, Mask R-CNN and Cascade R-CNN are both based on l_n -norm loss with carefully adjusted parameters. These parameters may not apply to the proposed EfoCAL1-SEIOU loss. Second, both Mask R-CNN and Cascade R-CNN have the process of adjusting the weighting for the respective losses. Mask loss is added in Mask R-CNN correspondingly, and iterative BBR is used in Cascade R-CNN. Although these adjustment weighting methods limit the effect brought by EfoCAL1-SEIOU loss, they confirm the necessity of hard and easy sample balancing. In addition, the improved effect in YOLOF illustrates that there is also more room for improvement in the one-stage network for the hard and easy sample imbalance problem.

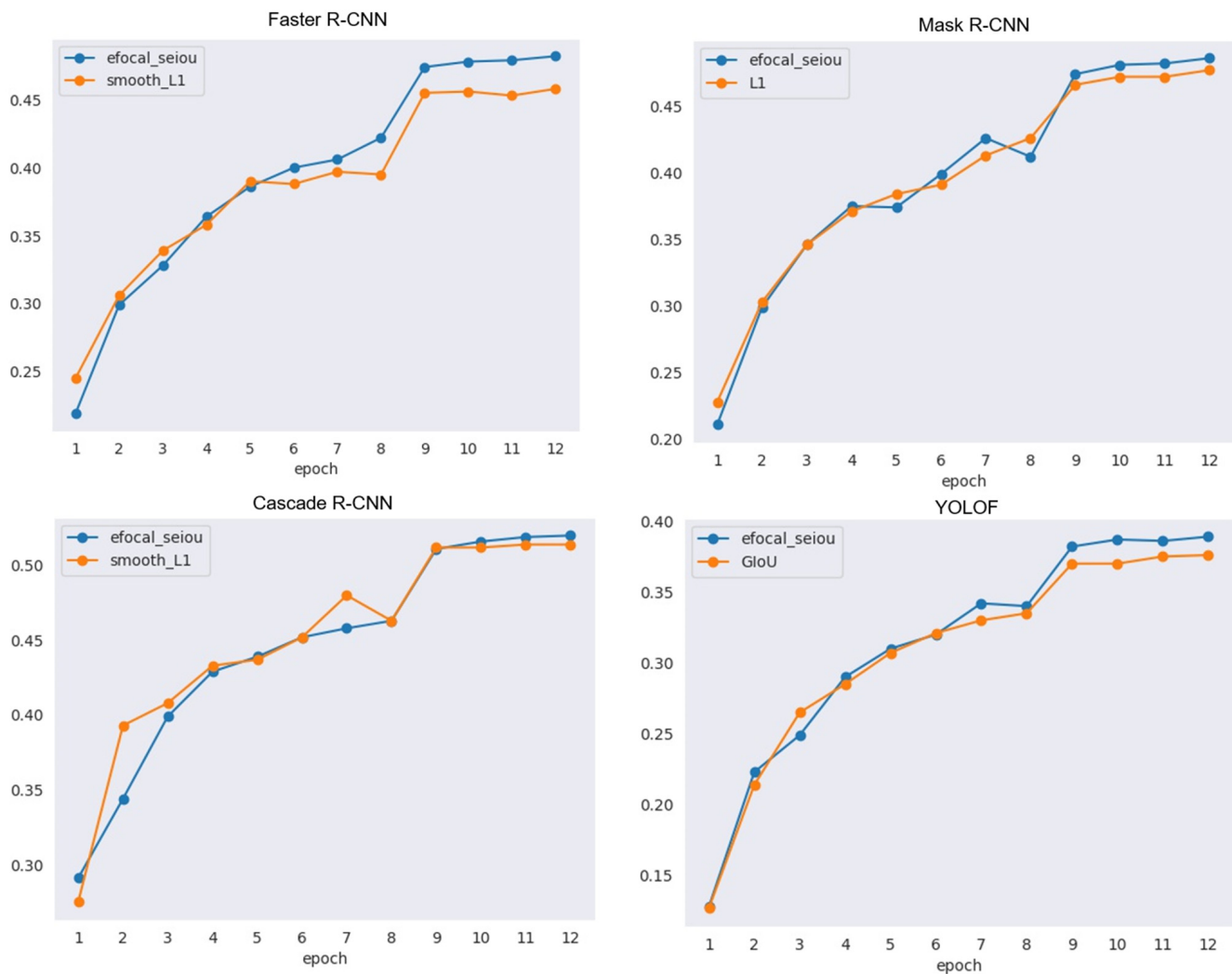


Figure 9. Comparison chart of AP trend.

Table 3. The performance when incorporating the EfoCaL1-SEIOU loss with different SOTA models.

Method	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Faster R-CNN	ResNet-50-FPN	0.458	0.730	0.501	0.171	0.364	0.564
Faster R-CNN *	ResNet-50-FPN	0.482	0.728	0.524	0.189	0.374	0.594
Mask R-CNN	ResNet-50-FPN	0.477	0.727	0.527	0.176	0.374	0.593
Mask R-CNN *	ResNet-50-FPN	0.486	0.727	0.526	0.172	0.385	0.602
Cascade R-CNN	ResNet-50-FPN	0.514	0.736	0.561	0.176	0.396	0.632
Cascade R-CNN *	ResNet-50-FPN	0.520	0.733	0.566	0.184	0.400	0.636
YOLOF	ResNet-50	0.376	0.620	0.397	0.073	0.232	0.524
YOLOF *	ResNet-50	0.389	0.622	0.409	0.061	0.254	0.541

* indicates using the EfoCaL1-SEIOU loss instead of their original losses.

5. Limitation and Discussion

Despite the above achievements, the proposed loss function still suffers from certain shortcomings. Regarding the detection of workers, since there is often large equipment on construction sites, the pixel value of workers with large equipment accounts for a smaller proportion; thus, it is often missed, as shown in Figure 8. These limitations mainly focus on images obtained from surveillance perspectives at the height of the building. In addition, when there is a high degree of overlap between workers or machines on a construction site, there is still a problem of bounding box redundancy. Therefore, there is still some room for improvement in the detection effect. To this end, for the problem of unbalanced hard and easy samples, we hope to explore bounding box evaluation metrics that apply the hard and easy sample balance strategy to apply to the classification loss function, positive and negative sample allocation, and non-maximum suppression modules.

6. Conclusions

In this paper, we analyzed the bounding box redundancy problem of object detection in construction sites and found that there is still some room for improvement in the loss function of the regression module by introducing a balanced strategy of hard and easy samples. First, the existing loss functions all have certain defects that hinder the direction of the BBR. Second, the existing studies ignore the importance of hard and easy sample balance. As a result, the gradient of hard sample regression is too large, which limits the performance of BBR. Based on the above two problems, we propose the EfoCaL1-SEIOU loss to address the shortcomings of the existing loss function and balance the gradient of the hard and easy sample regression. Extensive experiments on the MOCS dataset show that EfoCaL1-SEIOU loss brings some improvements on many advanced object detectors, solves practical problems, and can be applied to various applications. In the future, we hope to design a low-light enhancement module and its loss function to perform low-light target detection tasks for building operation behaviors in scenes such as nighttime and tunnels.

Author Contributions: Conceptualization, X.W. and H.W.; methodology, X.W. and C.Z.; software, H.W.; validation, X.W., H.W. and C.Z.; investigation, X.W.; resources, Q.H.; data curation, Q.H. and L.H.; writing—original draft preparation, X.W.; writing—review and editing, X.W. and H.W.; visualization, X.W., C.Z. and L.H.; supervision, C.Z. and Q.H.; project administration, X.W. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (no. 62072024 and 41971396), the Projects of Beijing Advanced Innovation Center for Future Urban Design (no. UDC2019033324 and UDC2017033322), R&D Program of Beijing Municipal Education Commission (KM202210016002), and the Fundamental Research Funds for Municipal Universities of Beijing University of Civil Engineering and Architecture (no. X20084 and ZF17061).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset generated and analyzed during the current study can be obtained at http://www.anlab340.com/Archives/IndexArctype/index/t_id/17.html (accessed on 30 November 2020).

Acknowledgments: The authors are thankful to all the personnel who either provided technical support or helped with data collection. We also acknowledge all the reviewers for their useful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sacks, R.; Radosavljevic, M.; Barak, R. Requirements for building information modeling based lean production management systems for construction. *Autom. Constr.* **2010**, *19*, 641–655. [[CrossRef](#)]
2. Su, Y.Y.; Liu, L.Y. Real-time tracking and analysis of construction operations. In Proceedings of the 2007 ASCE/CIB Construction Research Congress, Grand Bahama Island, Bahamas, 6–8 May 2007.
3. Mukhiddinov, M.; Cho, J. Smart Glass System Using Deep Learning for the Blind and Visually Impaired. *Electronics* **2021**, *10*, 2756. [[CrossRef](#)]
4. Mukhiddinov, M.; Abdusalomov, A.B.; Cho, J. Automatic Fire Detection and Notification System Based on Improved YOLOv4 for the Blind and Visually Impaired. *Sensors* **2022**, *22*, 3307. [[CrossRef](#)] [[PubMed](#)]
5. Park, M.W.; Elsafty, N.; Zhu, Z.H. Hardhat-wearing detection for enhancing on-site safety of construction workers. *J. Constr. Eng. Manag.* **2015**, *141*, 04015024. [[CrossRef](#)]
6. Kim, D.; Liu, M.Y.; Lee, S.H.; Kamat, V.R. Remote proximity monitoring between mobile construction resources using camera-mounted UAVs. *Autom. Constr.* **2019**, *99*, 168–182. [[CrossRef](#)]
7. Roberts, D.; Golparvar-Fard, M. End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level. *Autom. Constr.* **2019**, *105*, 102811. [[CrossRef](#)]
8. Fang, Q.; Li, H.; Luo, X.C.; Ding, L.Y.; Luo, H.B.; Rose, T.M.; An, W.P. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* **2018**, *85*, 1–9. [[CrossRef](#)]
9. Fang, Q.; Li, H.; Luo, X.C.; Ding, L.Y.; Rose, T.M.; An, W.P.; Yu, Y.T. A deep learning-based method for detecting non-certified work on construction sites. *Adv. Eng. Inform.* **2018**, *35*, 56–68. [[CrossRef](#)]
10. Luo, X.C.; Li, H.; Cao, D.P.; Dai, F.; Seo, J.; Lee, S.H. Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. *J. Comput. Civ. Eng.* **2018**, *32*, 04018012. [[CrossRef](#)]
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *NeurIPS* **2012**, *25*, 1106–1114. [[CrossRef](#)]
12. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
13. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7 June 2015; pp. 1–9.
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 770–778.
16. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, Honolulu, HI, USA, 16–21 July 2017; pp. 2117–2125.
17. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
18. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12595–12604.
19. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 7036–7045.
20. Xu, H.; Yao, L.; Zhang, W.; Liang, X.; Li, Z. Auto-fpn: Automatic network architecture adaptation for object detection beyond classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 6649–6658.
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computervision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
22. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
23. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
24. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.M.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Shenzhen, China, 22–29 October 2017; pp. 2999–3007.

25. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 9260–9269.
26. Pang, J.M.; Chen, K.; Shi, J.P.; Feng, H.J.; Ouyang, W.L.; Lin, D.H. Libra R-CNN: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 821–830.
27. Li, B.; Liu, Y.; Wang, X. Gradient harmonized single-stage detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 24 January–1 February 2019; Volume 33, pp. 8577–8584.
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
29. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [[CrossRef](#)] [[PubMed](#)]
30. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1440–1448.
31. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 779–788.
32. Zhang, H.K.; Chang, H.; Ma, B.P.; Wang, N.Y.; Chen, X.L. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 260–275.
33. Zhang, Y.F.; Ren, W.Q.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T.N. Focal and efficient IOU loss for accurate bounding box regression. *arXiv* **2021**, arXiv:2101.08158.
34. Yu, J.H.; Jiang, Y.N.; Wang, Z.Y.; Cao, Z.M.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, New York, NY, USA, 15–19 October 2016; pp. 516–520.
35. Rezatofighi, H.; Tsoi, N.; Gwak, J.Y.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
36. Zheng, Z.H.; Wang, P.; Liu, W.; Li, J.Z.; Ye, R.G.; Ren, D.W. Distance-iou loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–16 February 2020; pp. 12993–13000.
37. Chen, Z.M.; Chen, K.A.; Lin, W.Y.; See, J.; Yu, H.; Ke, Y.; Yang, C. Piou loss: Towards accurate oriented object detection in complex environments. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 195–211.
38. He, J.B.; Erfani, S.; Ma, X.J.; Bailey, J.; Chi, Y.; Hua, X.S. Alpha-IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression. *Proc. Adv. Neural Inf. Process. Syst.* **2021**, *34*, 20230–20242.
39. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Shenzhen, China, 22–29 October 2017; pp. 2980–2988.
40. Cai, Z.W.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
41. Chen, Q.; Wang, Y.M.; Yang, T.; Zhang, X.Y.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 13 November 2021; pp. 13034–13043.
42. An, X.H.; Zhou, L.; Liu, Z.G.; Wang, C.Z.; Li, P.F.; Li, Z.W. Dataset and benchmark for detecting moving objects in construction sites. *Autom. Constr.* **2021**, *122*, 103482.
43. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 761–769.
44. Roberts, D.; Bretl, T.; Golparvar-Fard, M. Detecting and classifying cranes using camera-equipped UAVs for monitoring crane-related safety hazards. *J. Comput. Civ. Eng.* **2017**, *2017*, 442–449.
45. Kim, H.; Kim, H.; Hong, Y.W.; Byun, H. Detecting construction equipment using a region-based fully convolutional network and transfer learning. *J. Comput. Civ. Eng.* **2018**, *32*, 04017082. [[CrossRef](#)]
46. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.M.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 3–9 December 2017.
47. Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance problems in object detection: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3388–3415. [[CrossRef](#)] [[PubMed](#)]