electronics

Article

Rotating Object Detection for Cranes in Transmission Line Scenarios

Lingzhi Xia [1], Songyuan Cao [2], Yang Cheng [1], Lei Niu [1], Jun Zhang [3] and Hua Bao [3,*]

[1] State Grid Anhui Electric Power Research Institute, Hefei 230022, China; xialz0@ah.sgcc.com.cn (L.X.)
[2] State Grid Anhui Electric Power Co., Ltd., Hefei 230022, China
[3] School of Artificial Intelligence, Anhui University, 111 Jiulong Road, Hefei 230601, China; junzhang@ahu.edu.cn
* Correspondence: baohua@ahu.edu.cn

Abstract: Cranes are pivotal heavy equipment used in the construction of transmission line scenarios. Accurately identifying these cranes and monitoring their status is pressing. The rapid development of computer vision brings new ideas to solve these challenges. Since cranes have a high aspect ratio, conventional horizontal bounding boxes contain a large number of redundant objects, which deteriorates the accuracy of object detection. In this study, we use a rotating target detection paradigm to detect cranes. We propose the YOLOv8-Crane model, where YOLOv8 serves as a detection network for rotating targets, and we incorporate Transformers in the backbone to improve global context modeling. The Kullback–Leibler divergence (KLD) with excellent scale invariance is used as a loss function to measure the distance between predicted and true distribution. Finally, we validate the superiority of YOLOv8-Crane on 1405 real-scene data collected by ourselves. Our approach demonstrates a significant improvement in crane detection and offers a new solution for enhancing safety monitoring.

Keywords: rotating object detection; transmission line scenarios; YOLOv8; Kullback–Leibler divergence

check for updates

Citation: Xia, L.; Cao, S.; Cheng, Y.; Niu, L.; Zhang, J.; Bao, H. Rotating Object Detection for Cranes in Transmission Line Scenarios. *Electronics* **2023**, *12*, 5046. https://doi.org/10.3390/electronics12245046

Academic Editor: KC Santosh

Received: 30 October 2023
Revised: 8 December 2023
Accepted: 13 December 2023
Published: 18 December 2023

1. Introduction

Transmission line scenarios are characterized by diverse environments, complex scenarios, and potential safety hazards [1,2]. Specifically, equipment with large boom spans, such as cranes and cement pumps, are typically constructed in environments including urban roads, fields, mountains, and viaducts. Due to the large size of these pieces of equipment, the unstable center of gravity caused them to overturn, resulting in many injuries and property damage [3].

Object detection is one of the mainstream tasks in computer vision [4,5]. The main task is to accurately localize the object of interest by providing bounding boxes of targets. The framework of target detection is broadly divided into one-stage algorithms and two-stage algorithms, and they mark the targets using horizontal rectangular boxes. R-CNN [6], Fast R-CNN [7], and Faster R-CNN [8] are classical two-stage detection models that first extract region proposals and then correct them to obtain detection results. Although they have high detection accuracy, the inference is slow, and it is difficult to meet the real-time requirements. The two-stage models treat detection as a classification problem, whereas the one-stage models treat detection as a regression task. The YOLO [9–11] and SSD [12] families are typical one-stage models that offer higher detection accuracy with improved real-time performance. Unlike the two-stage model, these one-stage models streamline the detection process, treating it as a direct regression task without the need for intermediate steps like region proposal extraction. Therefore, one-stage models obtain a balance between accuracy and speed, making them more suitable for applications that demand real-time processing. That is the reason we chose YOLO as the backbone of our network framework. The YOLO families offer several advantages in terms of speed, efficiency, simplicity, and suitability for real-time applications when compared to traditional two-stage object detection methods. These advantages have made YOLO a popular choice in real applications. In January 2023, Ultralytics released YOLOv8, the latest version of the YOLO family. The framework of YOLOv8 still incorporates the three core components of the YOLO family: backbone, neck, and detection head. The backbone extracts low- and high-level features from input images. Then, the neck fully integrates the features' output from the backbone through top-down and bottom-up cross-layer connections. The detection head uses a structure that separates detection and classification and determines positive and negative samples based on classification and regression scores.

At present, object detection has been widely used for hazard detection in construction scenarios. Lu et al. [3] use a swin transformer as a backbone network for crawler crane detection to recognize images acquired by an unmanned aerial vehicle and provide drivers with risk information about hazardous work zones. Chian et al. [13] employ CenterNet to dynamically identify possible areas for crane load falls. All of the above methods use horizontal bounding boxes to label objects. However, the boom of a construction vehicle has a certain tilt angle and a high aspect ratio. Therefore, the use of horizontal bounding boxes will contain many irrelevant backgrounds and objects, resulting in a decrease in recognition accuracy.

Rotating object detection is mainly for the detection of objects with a certain rotation angle [14–16]. It works to make bounding boxes outline targets according to the contour of the objects and reduces the area difference between the target area and the bounding box area. The use of rotating object detection is a good solution when the target scale is proportionally large, dense, and has a complex background. Figure 1 illustrates the horizontal and rotational detection paradigm for cranes. As shown in Figure 1a, in addition to the cranes, the horizontal bounding box includes background information such as trees, fences, and buildings. However, the rotating bounding boxes contain only the crane without other interfering objects. Therefore, we use rotating object detection to recognize cranes in the transmission line scene.

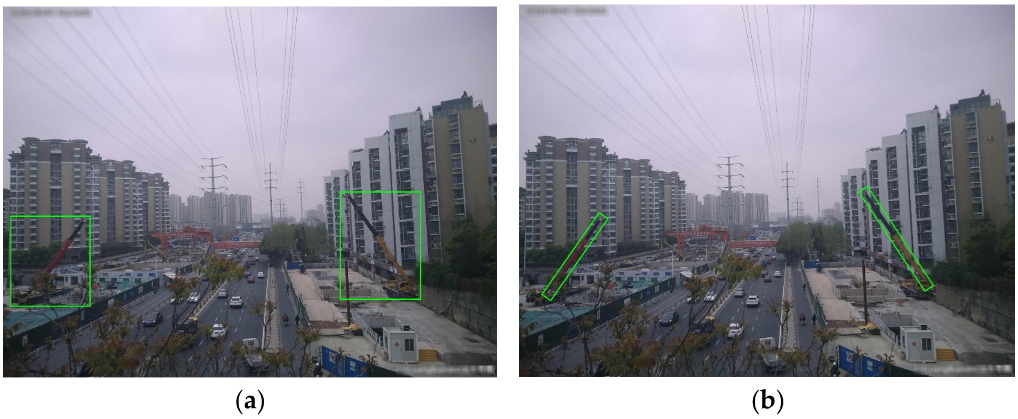
(**a**) (**b**)

Figure 1. (**a**,**b**) are the horizontal and rotational detection paradigms, respectively.

The main contributions of this study can be summarized as follows:

(1) We propose the YOLOv8-Crane model, specifically designed for crane detection in transmission line scenarios. In this model, a Transformer [17] is integrated into the backbone architecture to capture global context information, enhancing the model's understanding of the overall scene. More importantly, the study innovatively employs the Kullback–Leibler divergence (KLD) [18] as a loss function for rotating object detection. This strategic choice significantly improves the detection accuracy of cranes with high aspect ratios.

(2) We collect 1405 images of cranes in transmission line scenarios. The dataset encompasses diverse settings, including a variety of environments, such as urban neighborhoods, fields, riverbanks, daytime, and night-time. To enhance the generalization

ability of the detection network, gamma and contrast transformations are applied as data-augmentation techniques, simulating luminance variations. Additionally, the dataset is annotated with both horizontal and rotated bounding boxes to facilitate effective training.

(3) Extensive experiments were conducted to demonstrate the superiority of the YOLOv8-Crane model. Comparative analyses demonstrate that YOLOv8-Crane outperforms several state-of-the-art models, including Faster R-CNN [8], YOLOv6 [19], YOLOv7 [20], R3Det [14], R4Det [21], and SCRDet [22]. Ablation studies provide insights into the impact of the introduced KLD loss function, revealing its significant role in improving detection performance. Furthermore, this study shows that KLD outperforms alternative rotating object losses, such as smooth L1 [7] and Gaussian Wasserstein distance (GWD) [23].

(4) The rest of the article is structured as follows. Section 2 reviews rotating target detection. Section 3 describes the proposed YOLOv8-Crane model. Experimental results are presented in Section 4. Finally, the conclusion is summarized in Section 5.

2. Related Work

Unlike horizontal target detection, the key issue in rotating target detection is to transform the horizontal bounding box to the rotating case [18,22–24]. Yang et al. [22] propose SCRDet that fuses spatial and channel attention mechanisms with effective anchor sampling and design an improved smooth L1 loss for solving the regression problem of rotating bounding box. In subsequent work [14], they propose R3Det and design a feature refinement module to improve detection performance by obtaining more accurate features. R4Det adds a feedback mechanism to the recursive feature pyramid network (FPN) to improve the accuracy of rotating target detection [21]. Yang et al. [23] convert an arbitrary rotational bounding box $\mathcal{B}(x, y, w, h, \theta)$ into a two-dimensional (2D) Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and then the distance between the two Gaussian distributions is calculated as the final loss. The conversion process is calculated as

$$\boldsymbol{\mu} = (x, y)^\top \tag{1}$$

$$\begin{aligned}
\boldsymbol{\Sigma}^{1/2} &= \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^\top \\
&= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \frac{w}{2} & 0 \\ 0 & \frac{h}{2} \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\
&= \begin{pmatrix} \frac{w}{2}\cos^2\theta + \frac{h}{2}\sin^2\theta & \frac{w-h}{2}\cos\theta\sin\theta \\ \frac{w-h}{2}\cos\theta\sin\theta & \frac{w}{2}\sin^2\theta + \frac{h}{2}\cos^2\theta \end{pmatrix}
\end{aligned} \tag{2}$$

where $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\Sigma}$ is the covariance, $\mathbf{R}$ is the rotation matrix, and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. The GWD is used to measures the probability of $\mathbf{X}_p \sim \mathcal{N}_p\left(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p\right)$ and $\mathbf{X}_t \sim \mathcal{N}_t(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$:

$$\mathbf{D}_w(\mathcal{N}_p, \mathcal{N}_t)^2 = \underbrace{\left\|\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\right\|_2^2}_{\text{center distance}} + \underbrace{\mathbf{Tr}\left(\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_t - 2\left(\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}_p^{1/2}\right)^{1/2}\right)}_{\text{coupling terms about } h_p,\, w_p \text{ and } \theta_p} \tag{3}$$

where $\mathbf{D}_w$ is the GWD. The GWD consists of the distance between the center points (x, y)

and the coupling terms about h, w, and θ. Equation (3) can be further simplified for the horizontal detection box ($\theta = 0$):

$$\begin{aligned}\mathbf{D}_w^h(\mathcal{N}_p,\mathcal{N}_t)^2 &= \left\|\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\right\|_2^2 + \left\|\boldsymbol{\Sigma}_p^{1/2} - \boldsymbol{\Sigma}_t^{1/2}\right\|_F^2 \\ &= (x_p - x_t)^2 + (y_p - y_t)^2 + \left[(w_p - w_t)^2 + (h_p - h_t)^2\right]/4 \\ &= l_2\text{-norm}(\Delta x, \Delta y, \Delta w/2, \Delta h/2)\end{aligned} \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm. The GWD is not scale-invariant. However, the KLD has excellent scale invariance and is used as a loss function to measure the distance between the predicted and true distribution. Therefore, we use a new method based on the KLD.

3. Materials and Methods

3.1. Kullback–Leibler Divergence (KLD)

A rotating bounding box is converted into a 2D Gaussian distribution, and then KLD is calculated between the Gaussian distributions as the regression loss [18]. The KLD has excellent scale invariance and can be used as a loss function to measure the distance between the predicted distribution and the true distribution, allowing models to accomplish high-accuracy rotating object detection. The KLD between two 2D Gaussians is defined as

$$\mathbf{D}_{kl}(\mathcal{N}_p\mathcal{N}_t) = \underbrace{\frac{1}{2}\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\right)^\top \boldsymbol{\Sigma}_t^{-1}\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\right)}_{\text{term about } x_p \text{ and } y_p} + \underbrace{\frac{1}{2}\mathbf{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_p\right) + \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_t|}{|\boldsymbol{\Sigma}_p|}}_{\text{coupling tems about } h_p,\, w_p \text{ and } \theta_p} - 1 \tag{5}$$

$$\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\right)^\top \boldsymbol{\Sigma}_t^{-1}\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\right) = \frac{4(\Delta x\cos\theta_t + \Delta y\sin\theta_t)^2}{w_t^2} + \frac{4(\Delta y\cos\theta_t - \Delta x\sin\theta_t)^2}{h_t^2} \tag{6}$$

$$\mathbf{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_p\right) = \frac{h_p^2}{w_t^2}\sin^2\Delta\theta + \frac{w_p^2}{h_t^2}\sin^2\Delta\theta + \frac{h_p^2}{h_t^2}\cos^2\Delta\theta + \frac{w_p^2}{w_t^2}\cos^2\Delta\theta \tag{7}$$

$$\ln\frac{|\boldsymbol{\Sigma}_t|}{|\boldsymbol{\Sigma}_p|} = \ln\frac{h_t^2}{h_p^2} + \ln\frac{w_t^2}{w_p^2} \tag{8}$$

where $\Delta x = x_p - x_t$, $\Delta y = y_p - y_t$, and $\Delta\theta = \theta_p - \theta_t$.

In addition, the KLD and its derivatives can dynamically modify parameter gradients based on target characteristics [18]. It can adjust the gradient weights of the angular parameters according to the aspect ratio, which is crucial for improving the detection accuracy.

3.2. YOLOv8-Crane

In January 2023, Ultralytics released YOLOv8, the latest upgrade to the YOLO family. We propose the YOLOv8-Crane model based on YOLOv8. As shown in Figure 2, the network architecture of YOLOv8-Crane consists of input, backbone, neck, detection head, and output. There are two main improvements to YOLOv8-Crane. First, we add Transformers to the backbone for global context modeling, which helps to identify small crane targets. In addition, we use rotating object detection. Specifically, we replace the complete Intersection over Union (CIoU) [25] and distribution focal loss (DFL) [26] of YOLOv8 with KLD loss. Compared to horizontal bounding boxes, rotating bounding boxes reduce redundant background objects, especially when detected objects are characterized by a high aspect ratio. Therefore, combining Transformer and KLD, YOLOv8-Crane has better detection performance.

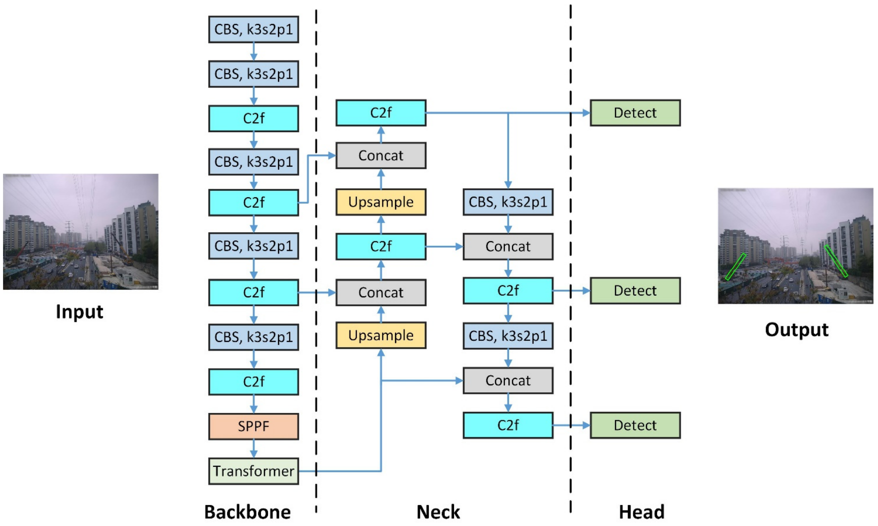

Figure 2. Architecture of YOLOv8-Crane, including input, backbone, neck, detection head, and output. For the CBS, "k" is the kernel size, "s" is the stride, and "p" is the padding, where "k3s2p1" means that the hyperparameters k, s, and p are set to 3, 2, and 1, respectively.

The role of the backbone as the network that extracts features is to pass feature information from extracted images to subsequent networks. The backbone model for YOLOv8-Crane is DarkNet-53 [11]. We add a Transformer at the end of the backbone. In this way, the backbone consists mainly of CBS (Conv + BatchNorm + SiLU), C2f (cross-stage partial network bottleneck with 2 convolutions), SPPF (spatial pyramid pooling-fas), and Transformer. CBS module is used to assist in feature extraction, while the SPPF module enhances the feature expression ability of the backbone. The C2f module in YOLO plays a critical role in contextual understanding and feature fusion, ultimately leading to improved object-detection performance. As shown in Figure 3a, the CBS module includes convolution, batch normalization [27], and the sigmoid-weighted linear unit (SiLU) [28]. In Figure 3b, the C2f structure is derived from the extended efficient layer aggregation network (ELAN) [20] of YOLOv7. C2f enhances the feature-fusion capability of convolutional neural networks and improves the inference speed of the backbone. In addition, C2f enriches the gradient flow via more branching cross-layer connections. In Figure 3c, SPPF is a fast spatial pyramid pooling that consists of two CBS modules and three max pooling layers. It enriches feature information via local and global feature fusion. Compared to convolution, the Transformer has a global receptive field, and its core component is the self-attention mechanism [17]. Self-attention is written as

$$\mathrm{Self}-\mathrm{attention}(Q,K,V)=\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{9}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and d_K is the dimension of K. The role of d_K is to improve training convergence and avoid vanishing gradients. As shown in Figure 3d, the Transformer splits hidden state vectors into multi-head self-attention (MHSA) mechanisms to form multiple sub-semantic spaces, allowing models to focus on information in different dimensional semantic spaces. MHSA is defined as

$$\mathrm{MHSA}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)W^O \tag{10}$$

$$\mathrm{head}_i=\mathrm{Self}-\mathrm{attention}\left(QW_i{}^Q,KW_i{}^K,VW_i{}^V\right) \tag{11}$$

In addition to the self-attention mechanism and MHSA, layer normalization [29] is used to stabilize training and accelerate convergence, and residual connectivity is used to improve information flow [30]. The multi-layer perceptron (MLP) consists of two linear layers and a Gaussian error linear units (GELUs) [31] activation function:

$$\mathrm{MLP}(x) = \mathrm{GELU}(xW_1 + b_1)W_2 + b_2 \tag{12}$$

where W denotes the weight, and b denotes the bias term.

The neck is an FPN structure where low- and high-level features are integrated. FPN is a type of neural network architecture, which is commonly used in many object-detection tasks. Objects may vary in size, and some objects might be small while others are large. FPN addresses this issue by creating a feature pyramid, a top-down architecture, where features from multiple convolutional layers of the backbone network at different resolutions are combined to provide a set of feature maps at multiple scales. This is performed to create a feature pyramid that maintains both semantic richness and spatial resolution. The detection heads are used for classification and bounding box regression, which consists of the CBS module and a 1 × 1 convolution (see Figure 3e). Cross-entropy loss function is employed for classification loss. KLD loss is utilized for bounding box regression.

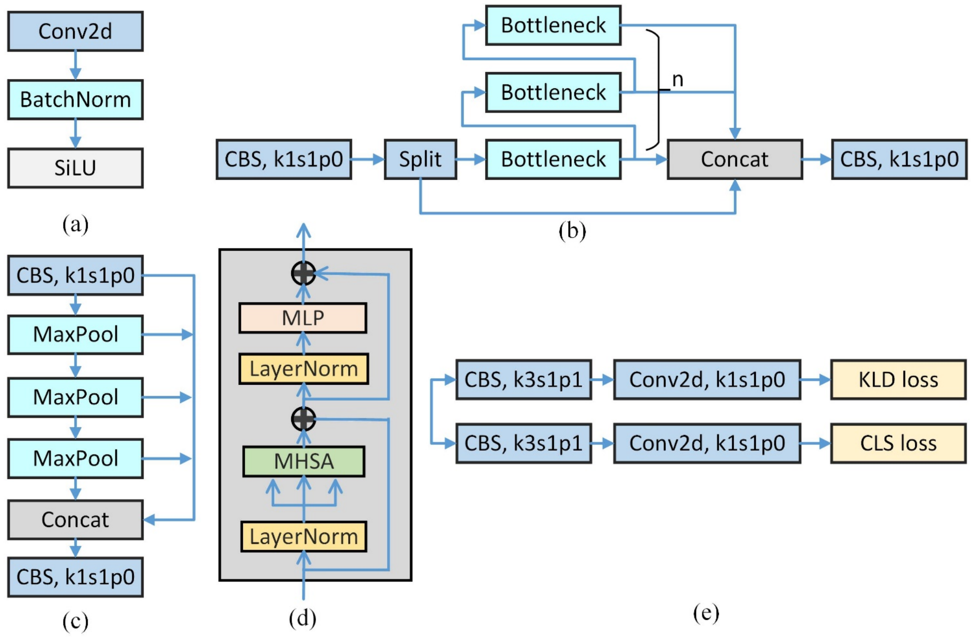

Figure 3. (**a**) CBS, (**b**) C2f, (**c**) SPPF, (**d**) Transformer, and (**e**) detection head.

4. Experiments

4.1. Dataset

We collected 1405 images of cranes in transmission line scenes. As shown in Figure 4, these images include a variety of environments such as urban neighborhoods, fields, riverbanks, daytime, and night-time. These images are divided into training/validation/test sets according to 8:1:1. Thus, the training/validation/test sets have 1125, 140, and 140 images, respectively. Data augmentation is a technique commonly used in machine learning, particularly in computer vision tasks such as object detection. It involves applying various transformations to the existing training dataset to artificially increase its size. In this work, the augmented data are used in this work to train the proposed model, improving the model's generalization and robustness. Among these techniques, we use mosaic [32], mixup [33], inversion, and rotation for data augmentation. Mosaic augmentation involves combining four images into a single mosaic image. It takes four random images and combines them into a single image. Mosaic achieves this by resizing each of the four images, stitching them together, and then taking a random cutout of the stitched images to obtain the final Mosaic image. Mixup augmentation blends two or more images together by taking a weighted linear combination of pixel values from two randomly selected images.

MixUp encourages the model to be more robust and generalize better by learning from combinations of different examples. Inversion and Rotation Augmentation are very simple: Inversion involves flipping images horizontally or vertically to create mirror images. This augmentation technique helps the model become more invariant to changes in pixel intensity or color, making it more robust to variations in lighting conditions. Rotation applies random rotations to images by a specified angle. This helps the model become more invariant to different orientations of objects in the images and improves its ability to recognize objects from various viewpoints. In addition, we also use gamma contrast transformations to simulate lighting variations, which helps improve detection performance at night.

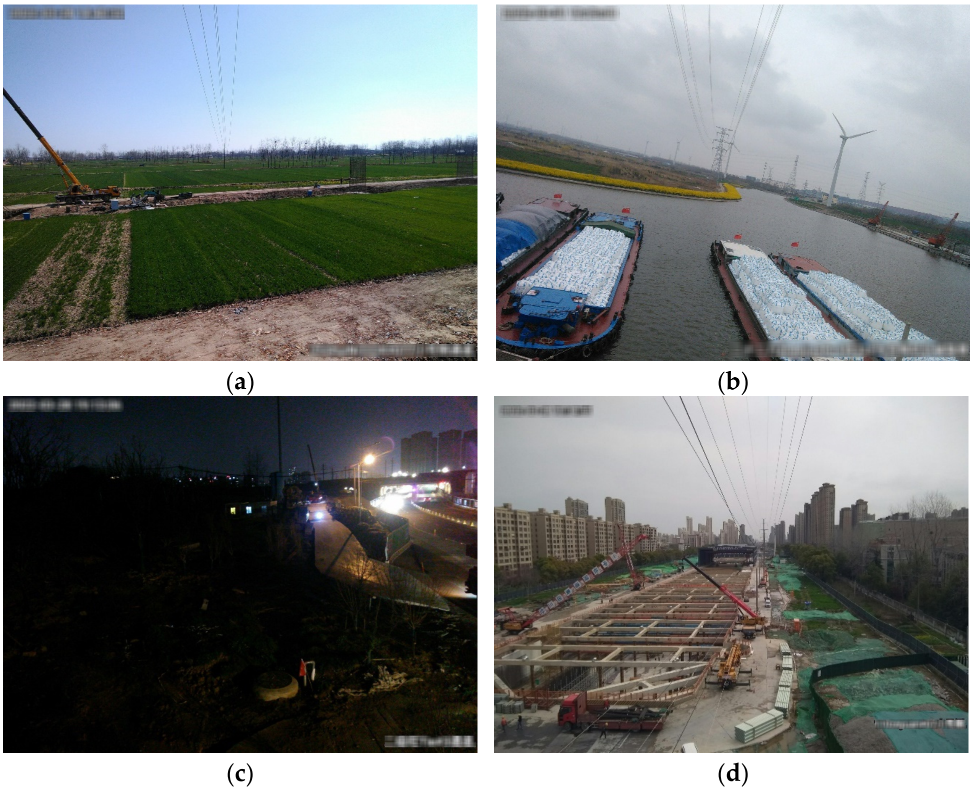
(**a**) (**b**) (**c**) (**d**)

Figure 4. Examples of images from collected data. (**a**) Fields, (**b**) riverbanks, (**c**) night-time, and (**d**) urban neighborhoods.

4.2. Experimental Setup

The input images are scaled to 640 × 640 pixels. The Stochastic gradient descent (SGD) with momentum (SGDM) is used to train networks for 350 epochs. SGD is an optimization algorithm commonly used in machine learning and deep learning for training models. SGDM stands for Stochastic Gradient Descent with Momentum. It is an extension of the standard SGD optimization algorithm, with the addition of a momentum term. In general, it adds a momentum term to the update rule for the model's parameters. This momentum term helps accelerate convergence and dampen oscillations during training. SGDM is particularly effective in training deep neural networks because it helps the optimization process converge faster and more reliably compared to standard SGD. Its initial learning rate is set to 0.01, and weight decay is 5×10^{-4}. The batch size is set to 8. Cosine annealing [34] is used to accelerate model training, and the decay period is set to 50. Although cosine annealing deteriorates performance during the restart phase, it is beneficial to performance in the long run [34]. The number of Transformer layers for YOLOv8-Crane is set to 2, each layer has an input dimension of 128, and the MHSA is set to 8. The experimental platform is an Intel i7-8700 CPU, 16GB of RAM, and a GTX 1080Ti GPU. The deep learning framework is PyTorch.

4.3. Evaluation Metric

The mean average precision (mAP) is used as an evaluation metric for object detection, where an IoU greater than 0.5 is denoted as mAP@50. Given a model-predicted box and a manually labeled ground-truth box, the IoU is written as

$$\mathrm{IoU} = \frac{\text{Area of Intersection of two boxes}}{\text{Area of Union of two boxes}} \tag{13}$$

A larger value of IoU indicates a more accurate predicted box. In addition, mAP is the mean of the average precision (AP) of all the object categories. AP measures the area under the precision–recall curve. The mAP value reflects the detection precision of models and is defined as

$$\mathrm{mAP} = \frac{1}{K}\sum_{i=1}^{K} AP_i \tag{14}$$

where K denotes the number of classes.

4.4. Experimental Results

In this subsection, we compare state-of-the-art models to demonstrate the superiority of YOLOv8-Crane. Table 1 shows the experimental results of the different models. Faster R-CNN, as a representative of the two-stage network, outperforms YOLOv6 and YOLOv7 in terms of detection performance. YOLOv8-Crane achieves the best detection results, in which mAP@50 and mAP@50-95 are 59.4% and 40.2%, respectively. In addition, Table 1 further indicates that KLD can significantly improve mAP results (from 54.1% to 59.4%).

Table 1. Quantitative experimental results of different models.

Model	mAP@50 (%)	mAP@50-95 (%)
Faster R-CNN [8]	55.3	34.7
YOLOv6 [19]	53.1	32.8
YOLOv7 [20]	52.2	31.5
YOLOv8-Crane (w/o KLD)	**54.1**	**34.4**
YOLOv8-Crane (w/KLD)	**59.4**	**40.2**

Figure 5 illustrates the qualitative experimental results. The object-detection models using horizontal bounding boxes all contain many redundant background objects. For example, the first column of images has buildings and iron poles in the horizontal bounding box. YOLOv6 even misidentifies the iron pole as a crane. YOLOv8-Crane with KLD fits better to the boom profile of cranes, which shows the great advantage of rotating target detection. For the fifth column of images, Faster R-CNN missed one of the cranes due to their proximity. For the night scene (fourth column), all models have poor detection performance. Faster R-CNN and YOLOv6 mistakenly detect the vertical iron pole in the distance as a crane. In addition, Faster R-CNN and YOLOv7 recognize shadows in a 45° oblique direction as cranes. In dim light, the shadow in a 45° oblique direction is similar to a black crane. Similarly, YOLOv8-Crane with KLD detects the distant iron pole as a crane, but it does not appear to be a missed detection. However, in daytime scenes, this linear structure is rarely misidentified as a crane—for example, the wind turbine in the third column of images. Therefore, we conjecture that using other methods to naturally convert daytime images to night-time images can improve crane detection at night instead of using gamma and contrast transformations.

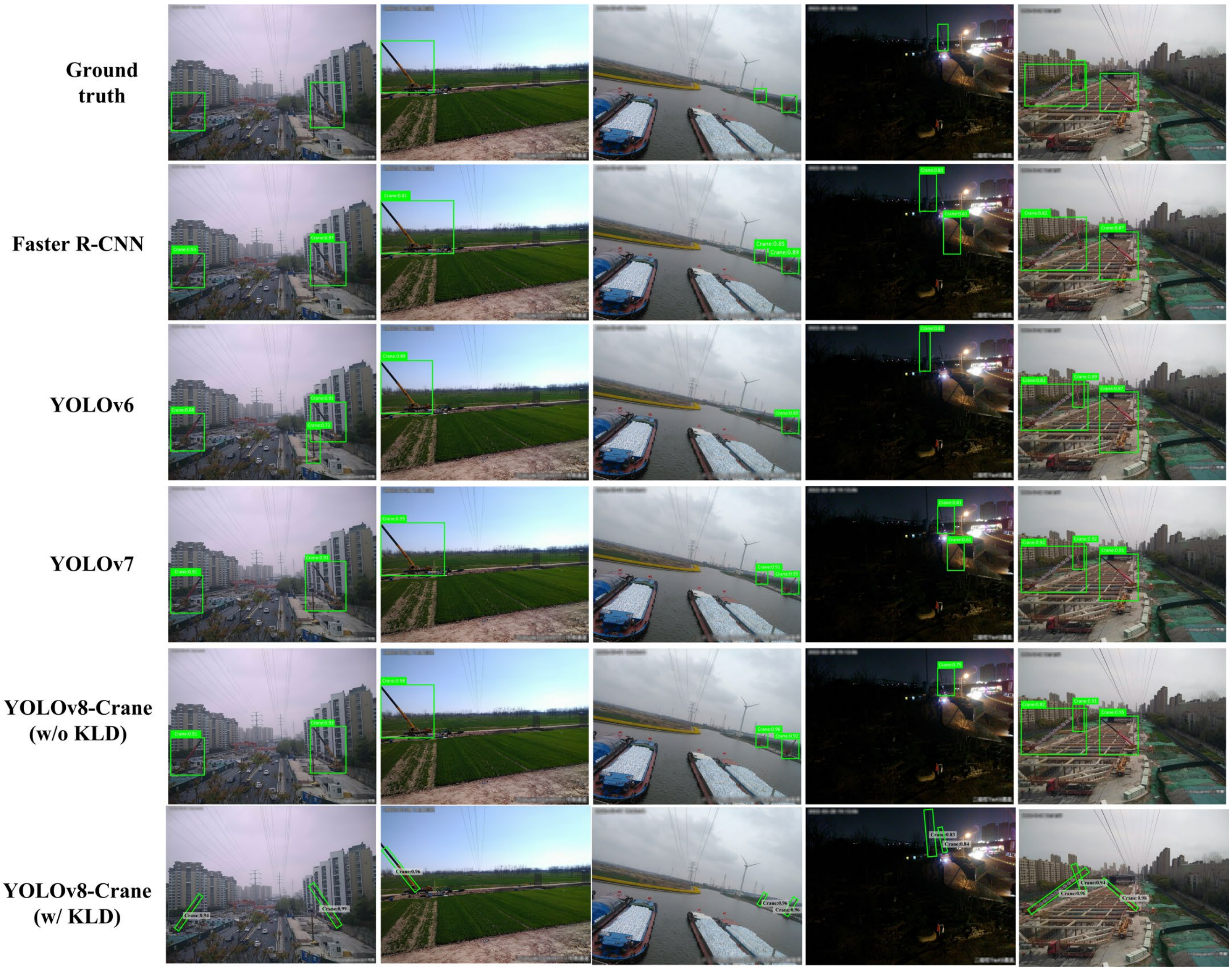

Figure 5. Detection results. Each row represents different test models, where the first row is the ground truth. Each column represents different test images.

Further, we employ different rotational object loss functions, including smooth L1 and GWD. The KLD and its derivatives can dynamically adjust the parameter gradient according to the properties of the object [18], which is crucial for improving object detection accuracy. As a result, the KLD-equipped YOLOv8-Crane outperforms smooth L1 and GWD. Specifically, in Table 2, the mAP@50 of KLD is 2.7% and 1.2% higher than that of smooth L1 and GWD, respectively. Figure 6 illustrates the detection results. For the fifth column of images, there is a significant deviation in the angle of the rotation bounding box of smooth L1, followed by GWD, and KLD can better fit the crane. For dim light scenarios (fourth column), both smooth L1 and GWD do not detect the crane correctly. KLD detects two cranes, but there is one misjudgment. Overall, qualitative and quantitative experimental results indicate that YOLOv8-Crane using KLD has better performance.

Table 2. Quantitative results for YOLOv8-Crane using different rotating objective loss functions.

Model	Loss	mAP@50 (%)	mAP@50-95 (%)
YOLOv8-Crane	Smooth L1 [7]	56.7	36.4
	GWD [23]	58.2	38.1
	KLD [18]	59.4	40.2

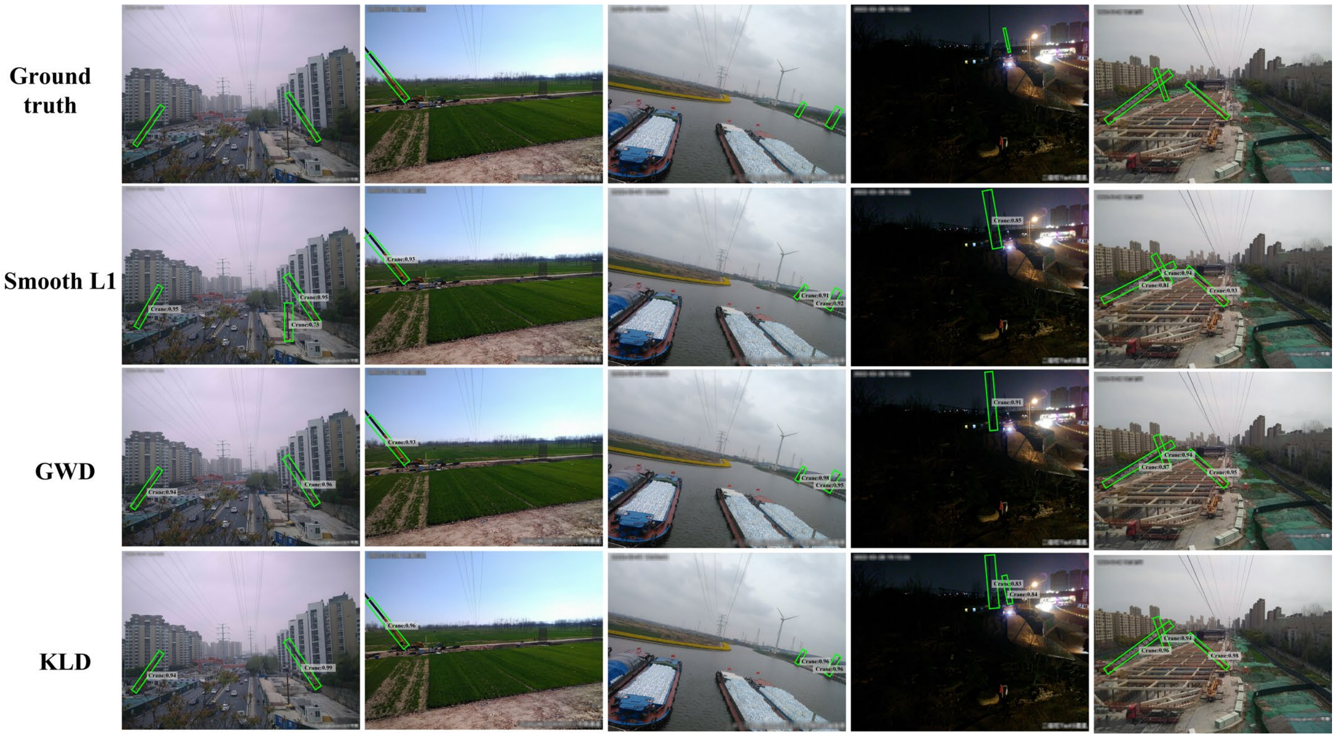

Figure 6. Experimental results of YOLOv8-Crane using various rotating object losses.

4.5. Comparison of Rotating Object-Detection Models

In this subsection, we compare other rotating target-detection models, including R3Det [14], R4Det [21], and SCRDet [22]. As shown in Table 3, YOLOv8-Crane achieves the highest mAP results. Specifically, the mAP@50 and mAP@50-95 of YOLOv8-Crane are 0.8% and 1.2% higher than that of R4Det, respectively. YOLOv8-Crane is also better than R3Det and SCRDet significantly. In addition, we found that the results for rotating objects are all higher than those for conventional object detection (see Table 1), which further confirms the superiority of rotating objects for crane detection.

Table 3. Experimental results of rotating target-detection models.

Model	mAP@50 (%)	mAP@50-95 (%)
R^3Det [14]	57.9	38.1
R^4Det [21]	58.6	39.0
SCRDet [22]	57.3	37.7
YOLOv8-Crane	59.4	40.2

4.6. Parameter Sensitivity

The number of Transformer layers and input dimensions of YOLOv8-Crane affect the ability to model global context. The setting of these two key hyperparameters needs to consider real-time performance and hardware costs. As shown in Figure 7a, as the number of Transformer layers increases, mAP@50 first increases rapidly and then enters the saturation phase when the number of layers is greater than 2. Increasing the number of layers can easily lead to overfitting, resulting in a deterioration trend in mAP@50. Hence, setting the number of Transformer layers to 2 balances efficiency and performance. Figure 7b illustrates the effect of input dimensions on detection performance. The performance growth rate starts to slow down when the input dimension reaches 128. Although consistently increasing the input dimension yields more performance gains, the computational amount of self-attention is the square of the input [17]. Increasing the input dimension increases computational consumption significantly. Therefore, we set this hyperparameter to 128 in our experiments.

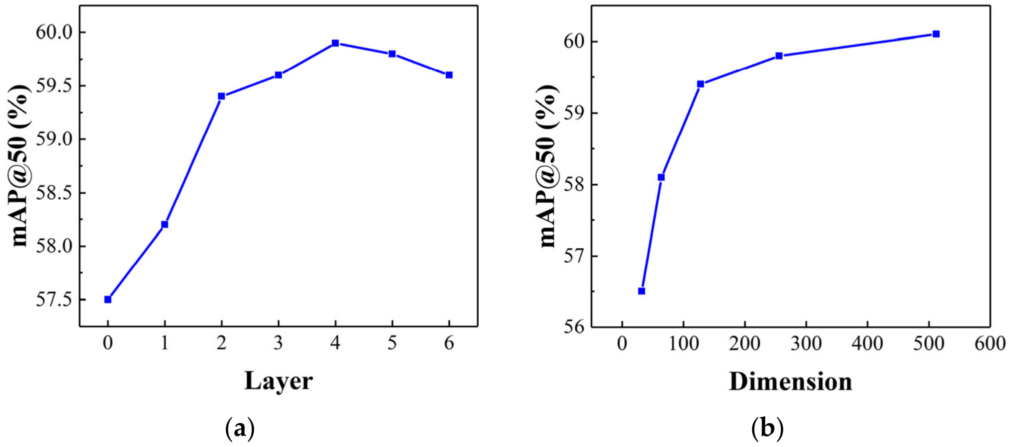

Figure 7. (**a**) Effect of the number of Transformer layers on mAP@50. (**b**) The mAP@50 results for different input dimensions of Transformer.

4.7. Ablation Study

In this subsection, ablation studies validate the effectiveness of Transformer and KLD. The results of the ablation experiments are shown in Table 4. YOLOv8-Crane is a standard YOLOv8 model when the Transformer and KLD are not used. Experimental results for YOLOv8 compete with YOLOv6 and YOLOv7, in which mAP@50 and mAP@50-95 are 52.3% and 32.1%, respectively. When the Transformer is added, the mAP metrics increase by approximately 2%. The Transformer improves global context modeling, which helps to detect cranes in complex backgrounds. KLD has a more significant improvement in performance, with mAP@50 growth from 52.3% to 57.5%. Rotating object detection is more advantageous than conventional object detection for cranes with high aspect ratios. When combined with the Transformer and KLD, YOLOv8-Crane achieves higher mAP, where mAP@50 and mAP@50-95 are 59.4% and 40.2%, respectively.

Table 4. Ablation studies on Transformer and KLD.

Model	Transformer	KLD	mAP@50 (%)	mAP@50-95 (%)
YOLOv8-Crane	×	×	52.3	32.1
	√	×	54.1	34.4
	×	√	57.5	37.1
	√	√	59.4	40.2

5. Conclusions

In this study, we have addressed the challenging problem of crane recognition in transmission line scenarios, proposing a novel and effective solution termed YOLOv8-Cranel. The proposed model integrates key advancements in object detection, specifically incorporating KLD-based rotating object detection and leveraging Transformer architecture for global context modeling. These techniques within the YOLOv8 detection framework yield a significant improvement in crane detection performance, particularly for cranes with high aspect ratios. The utilization of KLD-based rotating object detection is a noteworthy aspect of our approach. This method obviously enhances the detection accuracy of cranes characterized by high aspect ratios, a common occurrence in transmission line scenarios. By adapting the detection mechanism to account for the unique geometry of cranes, YOLOv8-Crane excels in accurately identifying and localizing these structures within the given environment. Furthermore, the integration of Transformer architecture plays an important role in enhancing global context modeling. Transformers are adept at capturing long-range dependencies in data, enabling the model to consider the relationships between various components of the input features. In the context of crane recognition, this facilitates a more complicated understanding of the scene, contributing to improved detection performance. The experimental results show the efficacy of YOLOv8-Crane, with the model achieving a remarkable 59.4% mAP@50 and 40.2% mAP@50-95. These metrics surpass the performance of established comparison models such as Faster R-CNN, YOLOv6, and YOLOv7. The higher mAP values underscore the superior accuracy and reliability of our proposed model in the context of crane detection. The implications of these results are significant, positioning YOLOv8-Crane as a benchmark method for the ongoing development of crane status monitoring and safe area detection.

Author Contributions: Conceptualization, L.X.; methodology, S.C.; algorithm L.X. and L.N.; experiemtal analysis, Y.C.; writing—original draft preparation, L.X.; writing—review and editing, J.Z.; supervision and project administration, H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by in part by the provincial natural science foundation of Anhui under Grant 1908085MF217, and in part by the Natural Science Research Project of Anhui Provincial Education Department under Grant 2023AH050085.

Data Availability Statement: For reasons of data confidentiality, the data cannot be public available temporarily.

Conflicts of Interest: Author Songyuan Cao was employed by the company State Grid Anhui Electric Power Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Belagoune, S.; Bali, N.; Bakdi, A.; Baadji, B.; Atif, K. Deep learning through LSTM classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems. *Measurement* **2021**, *177*, 109330. [CrossRef]
2. Deng, F.; Xie, Z.; Mao, W.; Li, B.; Shan, Y.; Wei, B.; Zeng, H. Research on edge intelligent recognition method oriented to transmission line insulator fault detection. *Int. J. Electr. Power Energy Syst.* **2022**, *139*, 108054. [CrossRef]
3. Lu, Y.; Qin, W.; Zhou, C.; Liu, Z. Automated detection of dangerous work zone for crawler crane guided by UAV images via Swin Transformer. *Autom. Constr.* **2023**, *147*, 104744. [CrossRef]
4. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef] [PubMed]
5. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [CrossRef] [PubMed]
7. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
10. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
11. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
12. 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 Computer Vision–ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
13. Chian, E.Y.T.; Goh, Y.M.; Tian, J.; Guo, B.H.W. Dynamic identification of crane load fall zone: A computer vision approach. *Saf. Sci.* **2022**, *156*, 105904. [CrossRef]
14. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; pp. 3163–3171.
15. Dai, L.; Chen, H.; Li, Y.; Kong, C.; Fan, Z.; Lu, J.; Chen, X. TARDet: Two-stage Anchor-free Rotating Object Detector in Aerial Images. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 4266–4274.
16. Feng, X.; Yao, X.; Cheng, G.; Han, J. Weakly Supervised Rotation-Invariant Aerial Object Detection Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14126–14135.

17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
18. Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; Yan, J. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18381–18394.
19. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
20. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
21. Sun, P.; Zheng, Y.; Zhou, Z.; Xu, W.; Ren, Q. R4 Det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. *Image Vis. Comput.* **2020**, *103*, 104036. [CrossRef]
22. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8231–8240.
23. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, virtual, 18–24 July 2021; pp. 11830–11841.
24. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; Xu, C. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11204–11213.
25. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
26. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
27. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
28. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **2018**, *107*, 3–11. [CrossRef] [PubMed]
29. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
30. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
32. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
33. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
34. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.