remote sensing

MDPI

Article

An Efficient Rep-Style Gaussian–Wasserstein Network: Improved UAV Infrared Small Object Detection for Urban Road Surveillance and Safety

Tuerniyazi Aibibu [1,2], Jinhui Lan [1,2,*], Yiliang Zeng [1,2], Weijian Lu [3] and Naiwei Gu [3]

[1] Department of Instrument Science and Technology, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2] Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
[3] Beijing Institute of Space Launch Technology, Beijing 100076, China
* Correspondence: lanjh@ustb.edu.cn

Abstract: Owing to the significant application potential of unmanned aerial vehicles (UAVs) and infrared imaging technologies, researchers from different fields have conducted numerous experiments on aerial infrared image processing. To continuously detect small road objects 24 h/day, this study proposes an efficient Rep-style Gaussian–Wasserstein network (ERGW-net) for small road object detection in infrared aerial images. This method aims to resolve problems of small object size, low contrast, few object features, and occlusions. The ERGW-net adopts the advantages of ResNet, Inception net, and YOLOv8 networks to improve object detection efficiency and accuracy by improving the structure of the backbone, neck, and loss function. The ERGW-net was tested on a DroneVehicle dataset with a large sample size and the HIT-UAV dataset with a relatively small sample size. The results show that the detection accuracy of different road targets (e.g., pedestrians, cars, buses, and trucks) is greater than 80%, which is higher than the existing methods.

Keywords: deep learning; aerial image; infrared image; detection; road object

Citation: Aibibu, T.; Lan, J.; Zeng, Y.; Lu, W.; Gu, N. An Efficient Rep-Style Gaussian–Wasserstein Network: Improved UAV Infrared Small Object Detection for Urban Road Surveillance and Safety. *Remote Sens.* **2024**, *16*, 25. https://doi.org/10.3390/rs16010025

Academic Editor: Arturo Sanchez-Azofeifa

Received: 22 October 2023
Revised: 24 November 2023
Accepted: 8 December 2023
Published: 20 December 2023

1. Introduction

With the rapid development of information processing [1] and integrated circuit [2,3] technology, the cost of advanced scientific and technological applications (e.g., unmanned aerial vehicle (UAV) remote sensing [4–7] and deep learning networks [8–10]) is decreasing, leading to many notable social and industrial improvements. These improvements enable scientists to develop a variety of new products. For example, researchers have classified crops with UAV remote sensing images and deep learning algorithms to provide decision support for farmers' planting and reduce crop management costs [11]. Some researchers have applied drone-based remote sensing image processing for river characterization and analysis to prevent or respond to unexpected flooding events [12]. Some researchers have detected cracks on roads using UAV-based remote sensing images and convolutional neural networks to analyze road damage [13]. Some researchers have observed wildlife activities with UAV-based visible infrared remote sensing images to analyze information such as population sizes, feeding sites, and migration directions [14]. It can be seen that UAV-based remote sensing image processing has high application value, especially infrared remote sensing image processing completed after the availability of low-cost infrared UAV remote sensing equipment. High-performance and lower-cost infrared thermal imaging has many advantages. First, object features can be obtained 24 h/day because there is no need for external light sources during infrared imaging. Second, collectors can now penetrate fog and provide anti-interference abilities [15]. Compared with fixed infrared thermal imaging, UAV remote-sensing platforms have better flexibility and perform more complex

imaging tasks, obtaining multidimensional imaging characteristics, even at night and in poor weather, as can be seen in Figure 1. With the plummeting costs of aerial infrared imaging tools, there is a great opportunity to make solutions more efficient with superior performance. Although infrared aerial images have many advantages, they do not have color and lack relatively fine texture information compared with visible images; therefore, detecting small targets from infrared images is a big challenge, and how to detect infrared small targets quickly and accurately has become a hot topic in the field of infrared remote sensing. Presently, detection methods use either traditional artificial intelligence processes or deep learning models. Notably, deep learning models provide dramatic improvements in precision and accuracy with their iterative model training with dynamic multi-class image features. Consequently, deep learning-based target detection techniques for UAV infrared images were brought into the field of urban road surveillance and safety.

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

Figure 1. Advantages of infrared images compared with visual images. (**a**) Visual image at night, (**b**) infrared image at night, (**c**) visual image in fog, and (**d**) infrared image in fog.

In the field of infrared image processing, researchers have performed a lot of work for road surveillance and traffic target detection in recent years. The limited computing power of early systems restricted artificial intelligence methods to image preprocessing, with feature detection relying largely on parametric algorithms. For example, Iwasaki et al. [16] later proposed a vehicle recognition algorithm based on infrared images for use in inclement weather that extracts Haar-like features and pixel information and inputs them into a cascaded set of multistage classifiers. Although the method can detect road congestion, the target class can only account for positive and negative samples, which greatly limits its practical utility.

Notably, traditional artificial intelligence methods require relatively small amounts of computation, but a large number of parameters are required for training, and detection accuracy is low. More recently, with the rapid improvements in and lowering costs of semiconductor technologies, it has become possible to build sophisticated deep learning models. Consequently, in 2012, Krizhevsky et al. [17] won the ImageNet image competition using the AlexNet method for target detection, outperforming traditional support vector machine options and opening a new era of machine learning. Zhu et al. [18] proposed a deep learning vehicle recognition model that uses a pretrained YOLOv3 network to detect targets using a transfer learning approach; however, the method can only detect one target class at a time and cannot manage multi-class target detection. Ren et al. [19], from the Nanjing University of Science and Technology, proposed a super-resolution infrared small-target recognition model that uses a generative adversarial network to detect super-resolution small targets with high accuracy; however, the method requires a vast number of computations and is inefficient. Alhammadi et al. [20] proposed a transfer learning method with relatively high accuracy, but it requires infrared images collected using vehicle-mounted thermal cameras. Hence, it is not suitable for handling aerial target prediction. Zhang et al. [21] developed an aerial method that provides good feature extraction and processing but only provides single-class detection results. Bhadoriya et al. [22] proposed a method designed to handle low-visibility conditions using multiple long-wave infrared imagery to overcome low-visibility restrictions while providing decision support for intelligent driving; however, it requires multiple thermal imaging sensors and data fusion at very high computational costs. Tichý́d et al. [23] analyzed the thermal characteristics of road vehicle objects and used

detailed features for vehicle detection and identification, but the data must be manually processed, which is practically prohibitive.

In summary, researchers have made many advancements in road target recognition using infrared images; however, current methods still struggle with multi-class detection, high computational costs, and narrow application areas. To overcome these limitations, this paper proposes an efficient Rep-style Gaussian–Wasserstein network (ERGW-net) based on YOLOv8 for small road object detection in infrared aerial images. The ERGW-net leverages state-of-the-art networks (i.e., Resnet and InceptionNet), adds an improved loss function, and lays the foundation for a vast array of improvements. Using experiments, we verify its improved effectiveness on a DroneVehicle dataset [24] and the famous HIT-UAV dataset [25]. The main contributions of our method can be summarized as follows:

- By redesigning and improving the backbone and neck, network parameters are reduced, and target detection accuracy is improved.
- A new loss function is proposed. Aiming to address the drawbacks of the existing loss function in small target recognition, we propose the loss function L_{GWPIoU} to improve target detection accuracy.
- To the best of our knowledge, this is the first time that up to five small target detection categories are considered using only UAV infrared images.

2. Materials and Methods

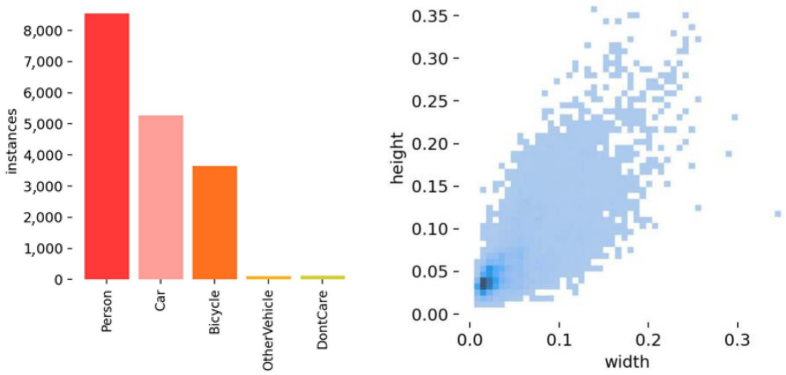
In order to better design the ERGW-net for small object detection in UAV-based infrared remote sensing images, we need to understand the characteristics of UAV remote sensing infrared images and prepare a large number of UAV infrared remote sensing images. Our private dataset has the disadvantages of a single shooting scene and inaccessibility for subsequent comparison studies by other researchers, so we selected two public datasets such as HIT-AUV and DroneVehicle.

2.1. Datasets

The two datasets contain different target classes and numbers of instances, so their structure and characterization are introduced separately.

2.1.1. HIT-UAV

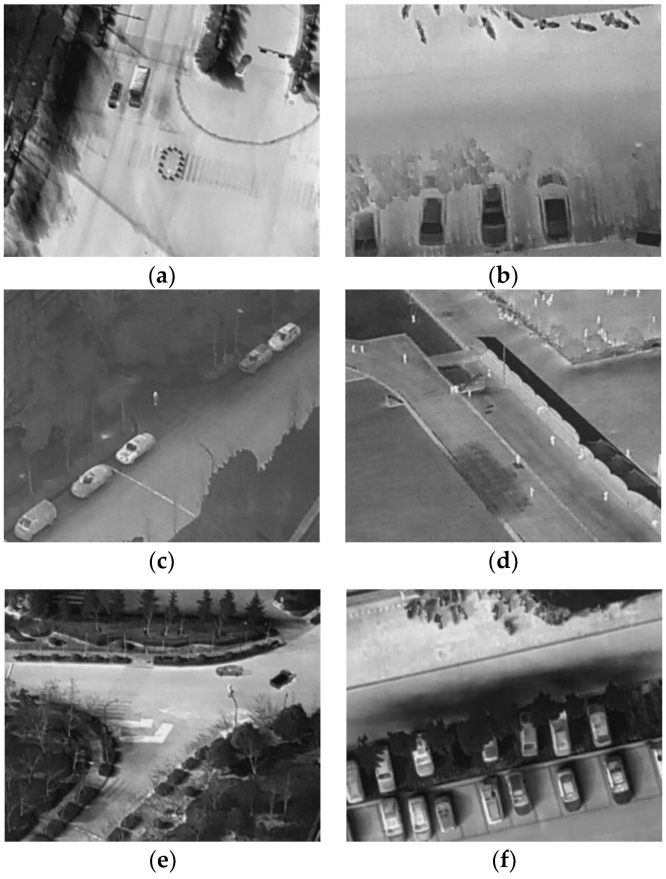
The HIT-UAV dataset is a collection of UAV infrared remote sensing images provided by Zhang et al. [25]. A drone DJI Matrice M210 V2 is used to obtain the dataset, and a DJI Zenmuse XT2 infrared camera is mounted on the drone, which has a resolution of 640 $\times$ 512 pixels and a 25 mm lens. The dataset contains 2898 infrared images taken above roads, parking lots, and schools. Because it contains several targets from various complex scenes, it is more suitable for verifying the robustness of detection algorithms than datasets with multiple targets from a single scene. Furthermore, the infrared images in this dataset contain several imaging influencing factors information, such as different flight altitudes and different shooting angles. The variations in altitude from 60 m to 130 m and shooting angles different from 30° to 90° play an important role in the diversity of target imaging results in the dataset. This study divided the data into training, validation, and testing sets, and person, car, bicycle, other, and "DontCare" classifications were labeled. The number and size distribution of target instances used for training is described in Figure 2, where the total number of instances from different classes is shown in Figure 2a. The horizontal and vertical axes in Figure 2b indicate the normalized target size relative to the whole image size. The horizontal scale represents the ratio of target width and image width; the vertical scale represents the ratio of target height and image height. Thus, both the x and y axes in Figure 2b have no units. In Figure 2b, different squares represent instances of different sizes, and the more overlapping instances of the same size are, the deeper color of their corresponding regions.



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

Figure 2. Distribution of object instances in the HIT-UAV dataset. (**a**) Instance number distribution and (**b**) instance size distribution.

According to Figure 2a, the number of person samples is the largest, and the numbers of other vehicle and DontCare samples are the smallest. Figure 2b shows that the resolution of most of the instances is less than two-tenths of the resolution of the infrared image, and even some of the instances have one-tenth of the resolution of the original image, which is because there is the largest number of people instances in the dataset. To better understand the relationship between instance resolution and infrared image resolution, it is necessary to view several typical sample infrared images from the dataset. Sample images of different road objects are shown in Figure 3.



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

(**c**) (**d**)

(**e**) (**f**)

Figure 3. Sample HIT-UAV images containing different objects: (**a**) cars and other vehicles on a road, (**b**) cars and bicycles in a parking lot. (**c**) cars and people on a roadside, (**d**) people in a playground, (**e**) cars and other vehicles at a crossroad, and (**f**) cars in a parking lot.

It can be seen in Figure 3 that the HIT-UAV dataset contains a variety of people, cars, bicycles, and other vehicles in complex scenarios, where people and bicycle targets are much smaller than cars and other vehicles. Regardless of the varying sizes between the different classes, it can be seen that they occupy few pixels in the infrared image, most of which belong to the category of small objects. Therefore, this dataset is suitable for testing the performance of our model.

2.1.2. DroneVehicle

The DroneVehicle dataset is a collection of UAV remote sensing images provided by Zhu et al. [24]. To produce this dataset, they used the DJI M200 UAV platform with a Zenmuse XT 2 longwave infrared camera, which uses a VOx uncooled imaging sensor with a resolution of 640 × 512, and the UAV platform was also equipped with a CMOS visual camera. So, this dataset provides visible and infrared dual-mode image data that can be used for the automatic detection, classification, and localization of road vehicles. Its data volume is relatively large, containing 28,439 pairs of visible and infrared images in the daytime and darkness (with and without lights), five types of targets: cars, trucks, buses, vans, and freight cars, and different scenarios such as parking lots, urban roads, overpasses and other types of parking lots. It has different shooting angles and different flight altitudes from 80 m to 120 m. We selected infrared images from the dataset and converted its oriented bounding-box (OBB) labels [26] into horizontal bounding box (HBB) labels [27] because only the rotating target labels were provided in the original dataset. There may be one or more target instances in an infrared image, so it is necessary to understand the distribution of different target instances in the dataset. The distribution of the target instances used for training is shown in Figure 4.

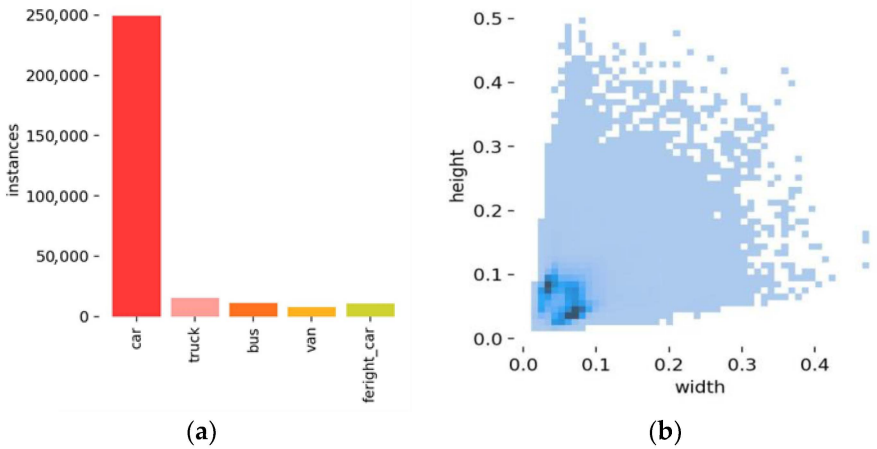


Figure 4. Distribution of target instances in the DroneVehicle dataset: (**a**) instance number distribution and (**b**) instance size distribution.

Figure 4a shows that, within the dataset, cars have the largest number of target instances, and vans had the smallest. In terms of instance size, most instances are less than one-tenth the overall size of images in terms of length and width. Some example pictures are shown in Figure 5.

As can be seen in Figure 5, the DroneVehicle dataset includes five types of road target instances with relatively small sizes in different scenarios, making it very suitable for evaluating the performance of our model.

After analyzing the structural characteristics of the UAV remote sensing dataset, we found that the road objects in infrared remote sensing images are not only smaller but also have lower contrast. Therefore, we pre-processed the datasets using the Adaptive Contrast Enhancement [28] algorithm to improve the contrast of the images and then designed the detection model ERGW-net that can detect small road objects in infrared remote sensing images.

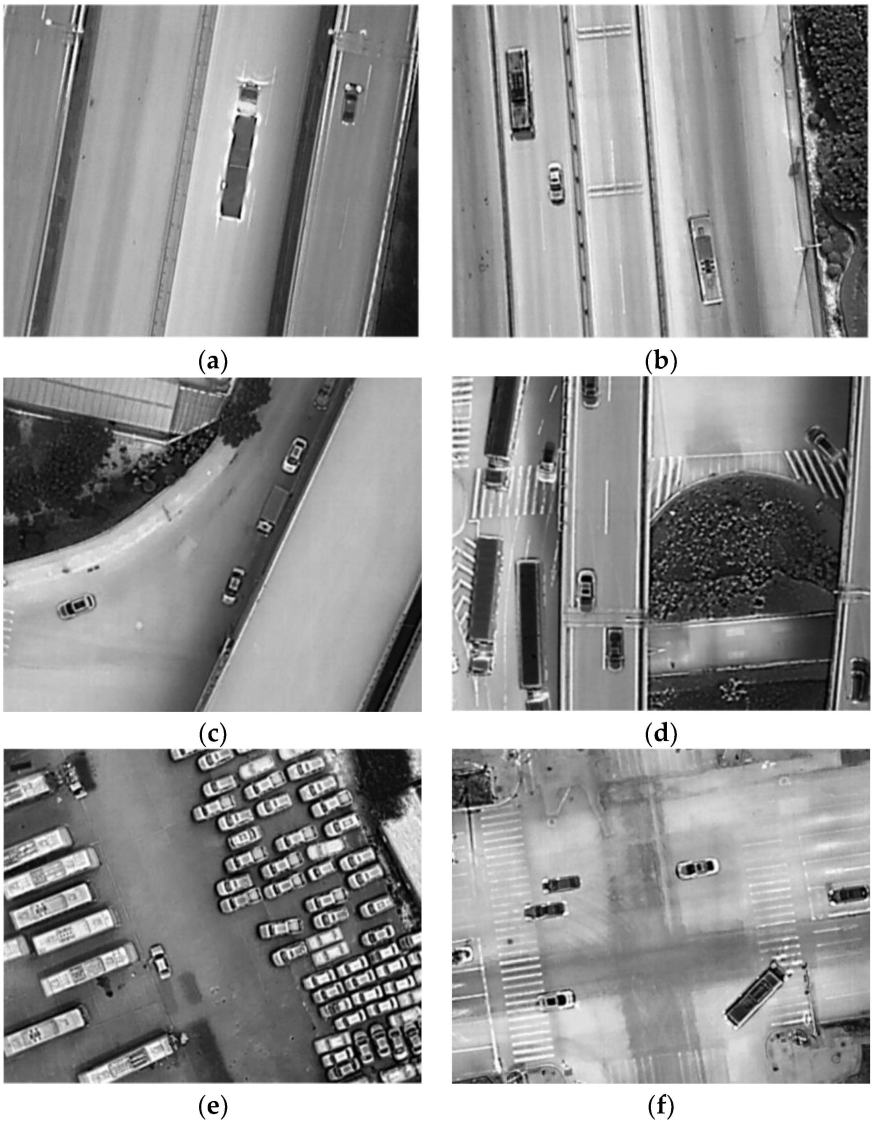


Figure 5. Sample DroneVehicle images containing road objects: (**a**) freight cars at night, (**b**) cars and buses on a road, (**c**) different vehicles on an overpass, (**d**) cars and freight cars on an overpass, (**e**) cars and buses in a parking lot, and (**f**) cars and buses at a crossroad.

2.2. Methods

The ERGW-net consists of a backbone, neck, and head. The backbone leverages modified CSPDarknet53 [29] with a new block called iRepblock, which combines the advantages of InceptionNet [30] and ResNet [31] to improve feature acquisition while decreasing computational demands. The neck fuses and categorizes the infrared image features, and a new loss function is provided at the head to improve the network's ability to process small road objects from aerial infrared images. The overall structure of the ERGW-net is shown in Figure 6.

The backbone, neck, and head structures of the ERGW-net are given on the right side in Figure 6, and the left half gives a schematic of the main module's basic structure from the right-side part, where some different colored modules are from YOLOv8. In this paper, we propose the iRepblock of backbone, the ERC block of neck, and the loss function of the head.

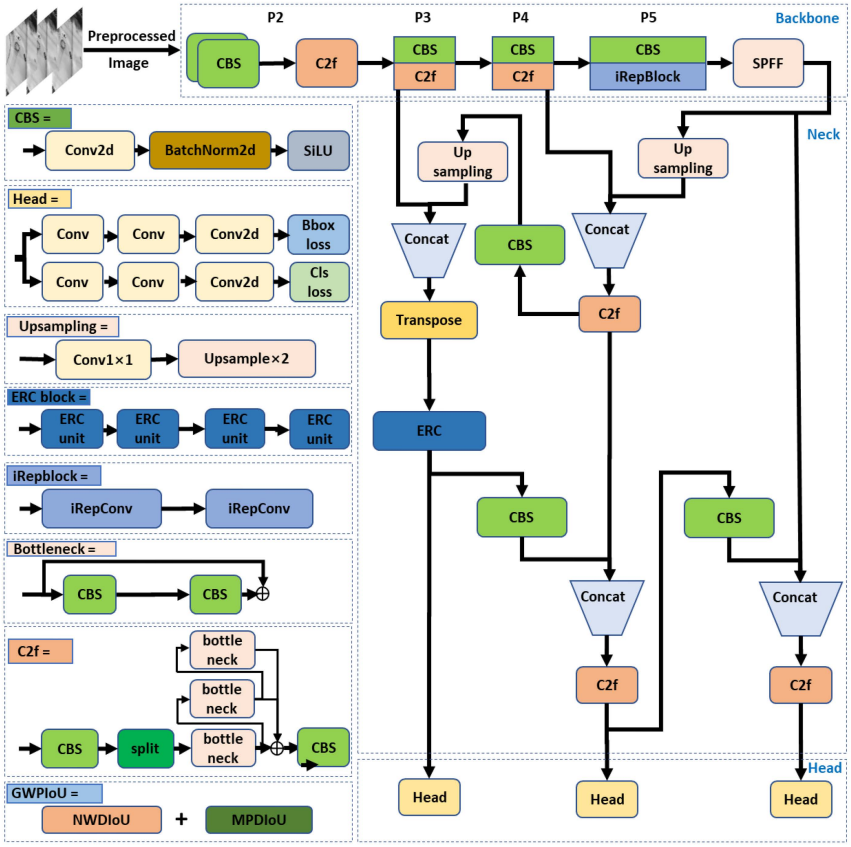


Figure 6. Overall ERGWnet structure.

2.2.1. Backbone

The role of the backbone is to extract features from images. To improve the overall performance of the backbone, we propose a new Rep-style backbone structure based on modified CSPDarknet53 [29] from YOLOv8. In other words, we provide a Rep-style capability that supports the modified CSPDarknet53 by orchestrating ResNet, InceptionNet, and efficient RepVGG ConvNet capabilities [32]. Our Rep-style method reduces the parameter dimension with its improved RepConV (iRepConV) structure and increases the efficiency of feature extraction with its improved Repblock (iRepblock) module using several iRepConV structures. To understand the orchestration process, it is necessary to first understand the architectures of iRepConV and iRepblock, as shown in Figure 7.

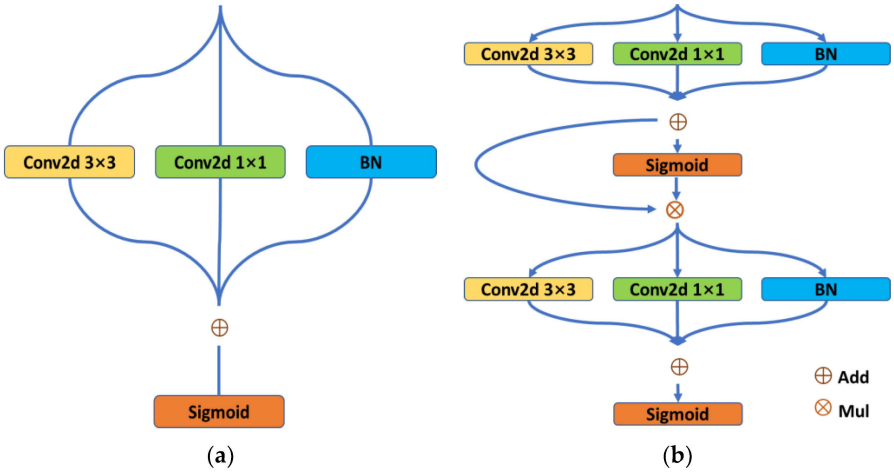


Figure 7. Improved RepConV and Repblock: (**a**) improved RepConV and (**b**) improved Repblock.

As shown in Figure 7a, there are several types of neural network modules such as batch normalization (BN) [33] and 1 × 1 Conv2d. To solve the problem of gradient vanishing, we imitate a ResNet by transmitting inputs directly to the output layer with the addition of a BN module. A 1 × 1 Conv2d is used for dimension reduction and rectified linear activation, which greatly reduces computational costs. Several iRepConVs comprise an iRepblock, and the number of iRepConVs, n, varies according to the task. In this case, $n = 2$. With these preparations, we change the original structure of the modified CSPDarknet53 and use iRepblock to reduce the number of parameters, without decreasing the performance, to obtain a new backbone structure named Rep-style network. In this way, we are able to obtain the most useful features from infrared images where the target is not obvious. The specific structure is shown in Figure 8.

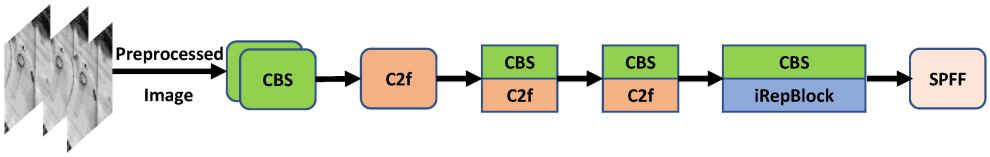


Figure 8. Structure of the improved darknet.

In Figure 8, it can be seen that the Rep-style network retains the CBS structure and SPFF structure of the modified CSPDarknet53, thus retaining the original advantages. iRepblock is used in the latter half of feature extraction to achieve the purpose of reducing the parameters and improving the performance.

2.2.2. Neck

To improve the multiscale feature fusion efficiency of the network and achieve a balance between network parameters and detection accuracy, this paper designs a new neck structure with Efficient-RepConV (ERC) based on YOLOv8. The ERC improves network's feature fusion capabilities by leveraging the advantages of ResNet and InceptionNet with combining transpose, conv, and upsampling operations for better road small object recognition (Figure 9).

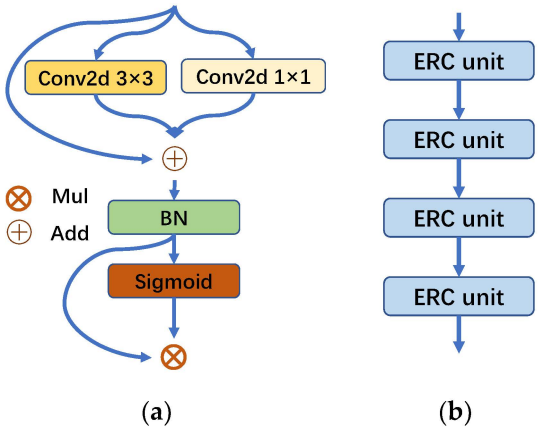


Figure 9. Structure of the improved neck and ERC blocks: (**a**) ERC unit and (**b**) ERC block.

As illustrated in Figure 9a, the ERC residual network avoids the vanishing gradient problem by directly passing inputs to the output, and a 1 × 1 Conv2d supplies dimensional reduction and rectified linear activation. Subsequently, BN and Sigmoid activation [34] are used to obtain the ERC unit. Each unit is the basis for constructing a transposition-based ERC block (Figure 9b), and the number depends on the training situation. In this study, four ERC units constitute a block that leverages the upsampling and C2f structure of YOLOv8. As a result, multiscale feature fusion improves with detection accuracy. Considering the

mutual constraints between parameter reduction and detection accuracy, we put ERCblock on the feature-output side of the neck, and the specific structure of the improved neck is shown in Figure 10.

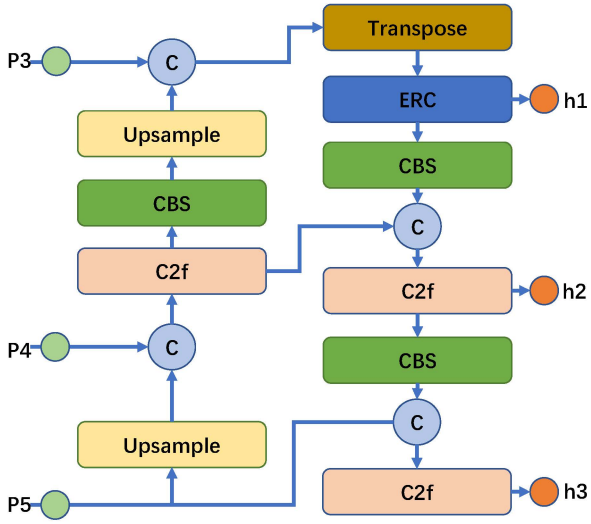


Figure 10. Structure of the improved neck.

According to Figure 10, it can be seen that the ERC block takes advantage of the positional advantages in the neck structure, not only outputting feature information to head1 but also affecting head2 and head3 with the feature information outputted to the CBS. This optimization of the neck structure plays an important role in processing infrared images with low contrast and few features. The effect of the improved neck on the detection of road small objects is given in the ablation experiment section.

2.2.3. Loss Function of the Head

Due to the small sizes of most road objects found in UAV aerial images, general deep learning detectors have low accuracy, which is exacerbated by the absence of color and texture features in infrared images. To solve this problem, we redesigned the loss function of the head and constructed its intersection over union (IoU) calculation method. Traditional IoU calculation methods are less useful, especially with the presence of overlaps and occlusions [35]. In the traditional IoU calculation, when the detection area of a small target changes slightly relative to the ground truth, the calculated IoU and the loss function change greatly, which reduces the detection accuracy, so we provide a novel Gaussian–Wasserstein Points (GWPIoU) calculation method based on Wasserstein [35] and minimum points distances [36].

The GWPIoU method consists of NWDIoU and MPDIoU, where NWDIoU is used to solve the problem of detecting small targets with a statistical approach, and the problem of overlapping targets is solved using MPDIoU. To make it easier to understand the relationship between traditional IoUs, NWDIoUs, and MPDIoUs, we give the corresponding schematic diagrams in Figure 11. The NWDIoU [29] calculation is as follows:

$$\mathrm{NWD}(N_a, N_b) = \exp\left(-\frac{\sqrt{W_2^2(N_a, N_b)}}{c}\right), \tag{1}$$

where $W_2^2(N_a, N_b)$ is the Gaussian distribution distance matrix and c is a constant that is generally related to the dataset.

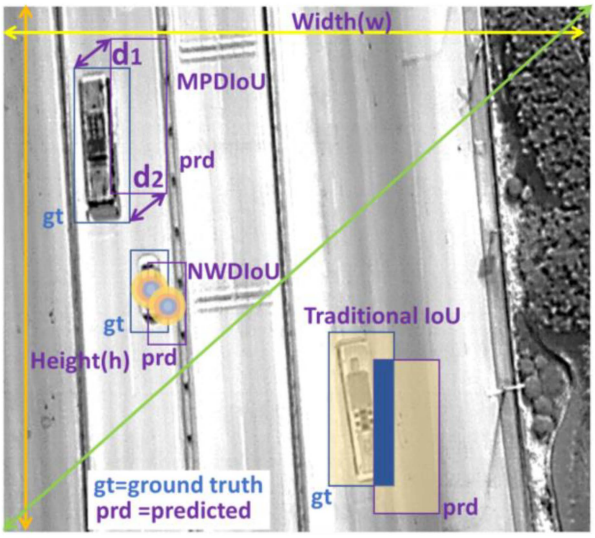


Figure 11. Schematic diagram of traditional IoU, NWDIoU, and MPDIoU.

The MPDIoU calculation [36] increases the detection rate of occluded or overlapped small objects as they relate to the coordinates of the prediction area, ground truth region, and image length and width. This is calculated as follows:

$$\mathrm{MPDIoU} = \frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \tag{2}$$

where A is the ground truth region, B is the prediction region, and d_1 and d_2 are the distances corresponding to the upper-left and lower-right corners of regions A and B, respectively, which are calculated as follows:

$$\begin{aligned} d_1^2 &= \left(x_1^B - x_1^A\right)^2 + \left(y_1^B - y_1^A\right)^2 \\ d_2^2 &= \left(x_2^B - x_2^A\right)^2 + \left(y_2^B - y_2^A\right)^2. \end{aligned} \tag{3}$$

The GWPIoU formula is as follows:

$$\mathrm{GWPIoU} = \alpha \mathrm{GWIoU} + \beta \mathrm{MDPIoU}, \tag{4}$$

where α is a constant whose value range is (0–1), which is chosen based on the dataset. The smaller an object, the larger the value of α (0.9 in this paper). Similarly, β is a constant with a value of 0.1 in this paper. According to the loss function theory, L_{GWPIoU} is calculated as follows:

$$L_{GWPIoU} = 1 - \mathrm{GWPIoU}. \tag{5}$$

Section 3 relates the results and analyses of all these components.

3. Results

The computer used for the experiment was an Intel Core i5-13400 with 32 GB memory and an NVIDIA GeForce RTX 3060 GPU. The validation experiments were divided into comparison and ablation types, where the comparisons were used to verify the effectiveness of the proposed algorithm compared with others, and the ablation tests were conducted to verify the effectiveness of individual modules.

To verify the effectiveness of the ERGW-net in different scenarios, this paper carries out algorithm validation experiments using the DroneVehicle and HIT-UAV datasets. In order to reduce the negative impact of low-contrast infrared images on object detection, we preprocessed the datasets for contrast enhancement using the Adaptive Contrast En-

hancement [28] method. The preprocessed infrared images with high contrast were more suitable for subsequent experiments.

3.1. Evaluation Metrics

In this section, to evaluate the road object detection capability of the ERGW-net, the following variables are introduced and defined.

3.1.1. Precision

Precision measures the number of true positives in a detection result and is specified by the following formula [37]:

$$P = \frac{TP}{TP + FP} \tag{6}$$

where P is the precision of road object detection, TP is the number of true positives in the detected set, and FP is the number of false positives.

3.1.2. Average Precision (AP)

The detection accuracy of our algorithm for a particular class of road targets was measured using AP [38], calculated by averaging the overall precision of detection. The mathematical expression is as follows:

$$AP = \frac{1}{m}\sum_{i}^{m} P_i, \tag{7}$$

where AP is the average detection accuracy of a particular class of road targets, m is a positive sample among n road samples belonging to the same class, and P_i is the probability of the detection precision for each positive sample.

3.1.3. Mean Average Precision (mAP)

mAP measures the overall detection precision of different classes in a dataset. AP measures the detection precision of one class, and there are generally several classes in the dataset. Thus, mAP has the following expression:

$$mAP = \frac{1}{c}\sum_{j}^{c} AP_j \tag{8}$$

where c is the number of road object classes in the dataset, and AP_j is the average detection accuracy of one of the classes.

3.1.4. mAP_{50}

Bounding box regression and object position estimation tasks are parts of the larger road object detection effort, and the precision of the bounding box is measured using the IoU. The mAP value when IoU = 0.5 is referred to as mAP_{50}, which is what we used to measure algorithmic effectiveness.

3.2. Comparative Experiments

In this study, the performance of the proposed algorithm was compared with that of other target detection algorithms, including Faster R-CNN [39], YOLOv5 [40], YOLOv7 [41], and YOLOv8 [42]. These experiments verified the superiority of ERGW-net over other algorithms for road small target detection.

3.2.1. Results on the HIT-UAV Dataset

To verify the detection performance datasets with a small number of samples, the ERGW-net and the other algorithms were tested on the HIT-UAV dataset. We trained the different algorithms on the dataset for 300 epochs and obtained the corresponding training

results. The mAP_{50} variation curves of the different algorithms from the training phase are illustrated in Figure 12.

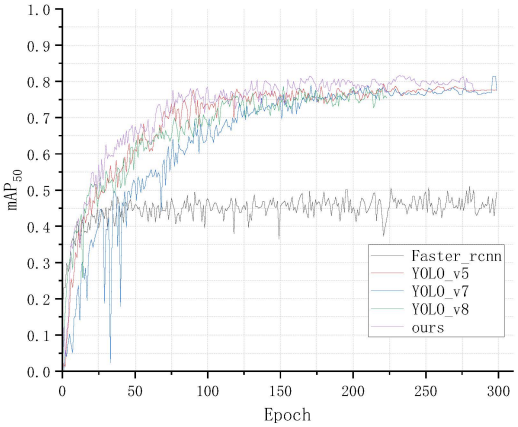


Figure 12. mAP_{50} variation curves of different algorithms during training on the HIT-UAV dataset.

According to the variation in the curves of different algorithms in Figure 12, the corresponding curve of our proposed algorithm trends above the others, indicating that it has better training performance.

The results of the comparative experiments are listed in Table 1. According to Table 1, compared with other target detection algorithms, the ERGW-net had the highest mPA_{50}, reaching more than 80% on the HIT-UAV dataset. When we analyze the AP_{50} of different classes, we find that all the classes have the highest accuracy except BC. This is because e-bicycles have different types and sizes in real life, and their infrared image features vary a lot. At the same time, there are not many corresponding improvements in the detection network, which causes the AP_{50} of BC to not be the highest. Visualized images of the detection results are shown in Figure 13.

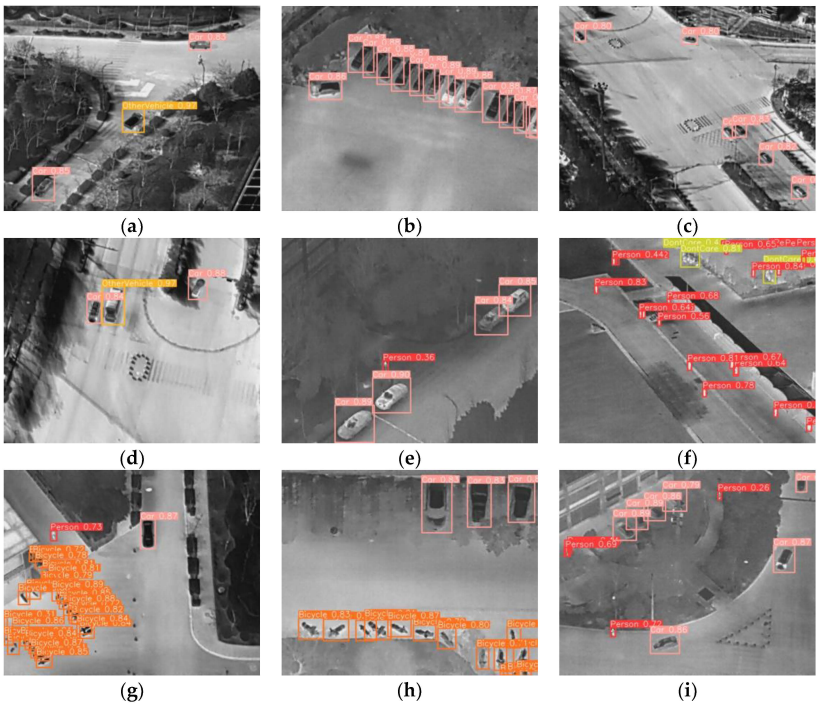


Figure 13. Detected HIT-UAV targets: (**a**) cars and another vehicle, (**b**) cars, (**c**) cars, (**d**) other vehicle, (**e**) a person and cars, (**f**) people and DontCare, (**g**) a person, a car, and bicycles, and (**h**) cars and bicycles, and (**i**) people and cars.

Table 1. Results of the comparative experiment on the HIT-UAV dataset.

Different Methods, Classes	Person	Car	Bicycle	Other Vehicles	DontCare	mAP50
Faster RCNN	20.5	80.4	51	35.6	40	45.5
YOLOV5	91.2	98	85.9	77.1	37.1	78
YOLOV7	90.4	97.5	92.1	78.1	26.3	77.5
YOLOV8	90.4	97.9	87.9	71.7	29.7	75.5
Ours	91.2	98	88.9	83.4	45.7	81.5

According to Figure 13, it can be seen that our proposed algorithm does a great job when detecting different targets in a variety of scenarios.

3.2.2. Results on the DroneVehicle Dataset

When faced HIT-UAV with small sample sizes, some of the detection features of the ERGW-net cannot be reflected in the experimental results. Hence, to obtain more detailed performance, we need a dataset with a large sample size: the DroneVehicle dataset.

Using the DroneVehicle dataset, the mAP_{50} variation curves of the different algorithms during training are shown in Figure 14.

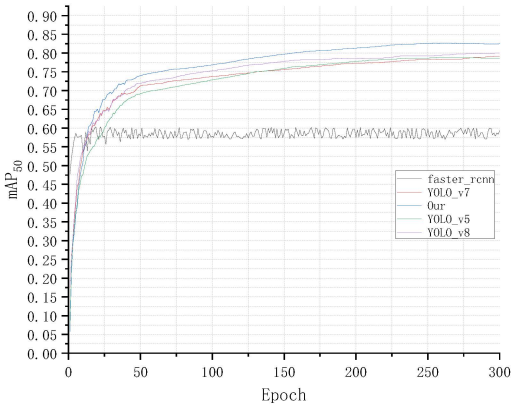


Figure 14. mAP_{50} variation curves of different algorithms during training on the DroneVehicle dataset.

Based on the variation in the curves of the different algorithms in Figure 14, the experimental results are listed in Table 2.

Table 2. Results of the comparative experiment on the DroneVehicle dataset.

Different Methods, Classes	Car	Truck	Bus	Van	Freight Car	mAP50
Faster RCNN	80.4	53.2	73.5	47.5	46.9	60.3
YOLOV5	96	73	94.1	59.3	71.4	78.8
YOLOV7	96.7	73.4	95	64.4	70.0	79.8
YOLOV8	96.5	73.6	94.8	64.4	72.5	80.4
Ours	96.9	77.9	96.1	66.8	74.8	82.5

According to Table 2, it can be seen that the proposed algorithm obtains superior mAP_{50} scores on the DroneVehicle dataset compared with the other algorithms, reaching higher than 80%. Some of the visual detection results are shown in Figure 15.

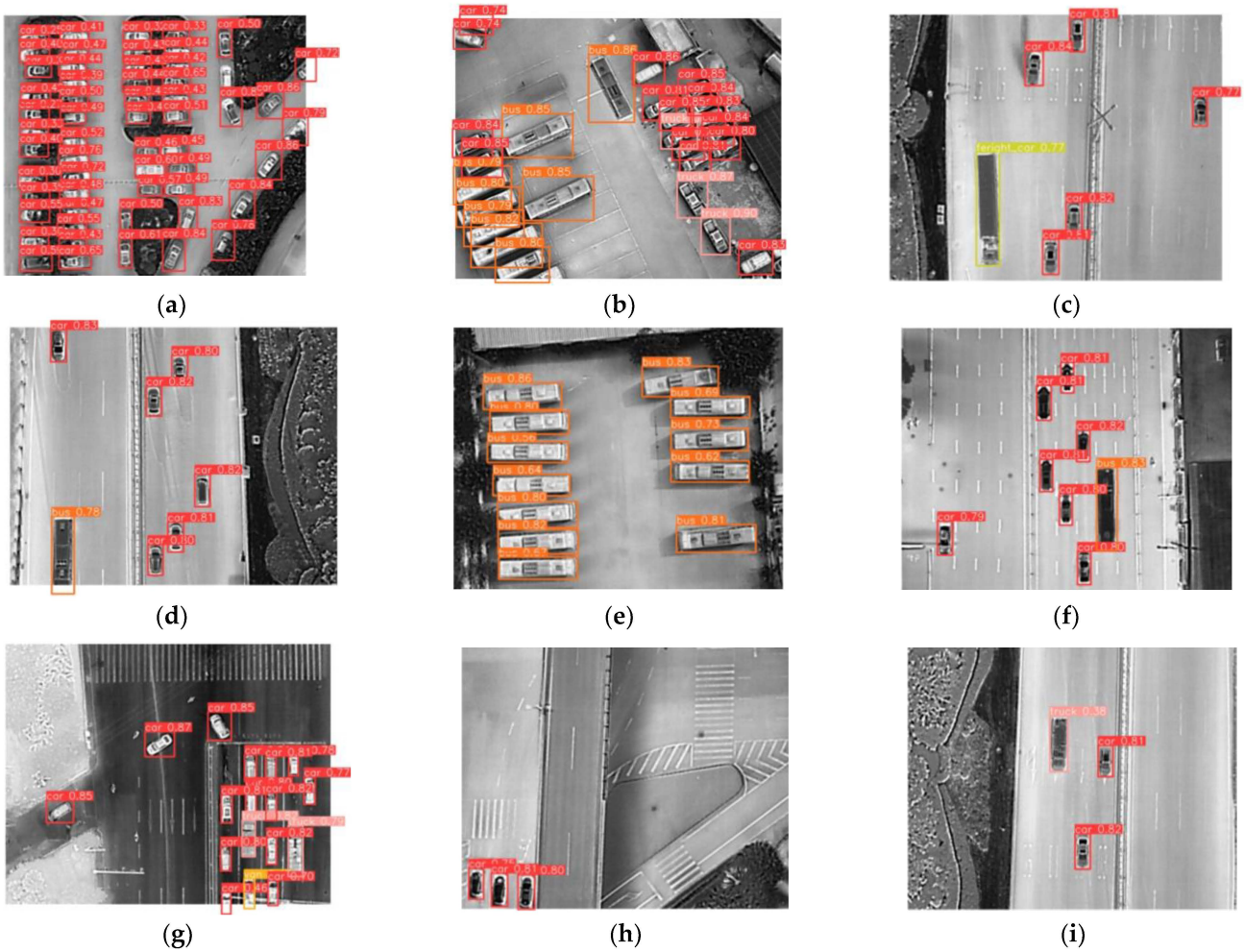


Figure 15. Detected DroneVehicle targets: (**a**) stopped cars, (**b**) cars, buses, and trucks, (**c**) cars and a freight car, (**d**) cars and a bus, (**e**) buses in a parking lot, (**f**) cars and a bus on urban highway, (**g**) cars, a van, and trucks, and (**h**) cars on overpasses, and (**i**) cars and a truck.

From Figure 15, it can be observed that the proposed algorithm accurately detects multiple road targets in different environments.

3.3. Ablation Experiment

Next, we report on the effectiveness of the different modules comprising the ERGW-net for road target detection. Because the ERGW-net contains three new modules (i.e., the L_{GWPIoU} loss function, ERC block, and Repblock), one module at a time was discarded to check for differences in the results (Table 3).

Table 3. Results of the ablation experiment on the DroneVehicle dataset.

iRepblock	ERC Block	L_{GWPIoU}	mAP_{50} on DroneVehicle
	✓	✓	80.1
✓		✓	79.2
✓	✓		78.4
✓	✓	✓	82.5

✓means ERGW-net contains this module, blank means it does not.

Table 3 shows that the mAP_{50} of the ERGW-net on the DroneVehicle dataset decreased from 82.5% to 78.4% after removing the L_{GWPIoU} loss function. Similarly, it dropped from

82.5% to 79.2% after removing the ERC module and from 82.5% to 80.1% after removing iRepblock. Therefore, the L_{GWPIoU} loss function contributes the most to small target detection accuracy improvement, whereas iRepblock contributes the least.

4. Discussion

According to the experimental results, it can be seen that the algorithm proposed in this paper has good performance on the HIT-UAV and DroneVehicle datasets. The following summary of the discussion points is provided:

- During model training, the degree of fluctuation in the training curves on the HIT-UAV dataset was larger than that on the DroneVehicle dataset, indicating that the large sample dataset was more suitable for training.
- The mAP_{50} score of the proposed algorithm on both datasets was greater than 80%, but it still has space for improvement; hence, we plan to improve the algorithm in future work.
- To understand which bounding boxes were used to make predictions, in this paper, we used class activation maps (CAM) [43–45] to help overcome the black-box rationale of deep learning models. The CAM of different classes from the DroneVehicle datasets are shown in Figure 16. The redder color indicates the higher classification contribution and the bluer color represents the lower classification contribution

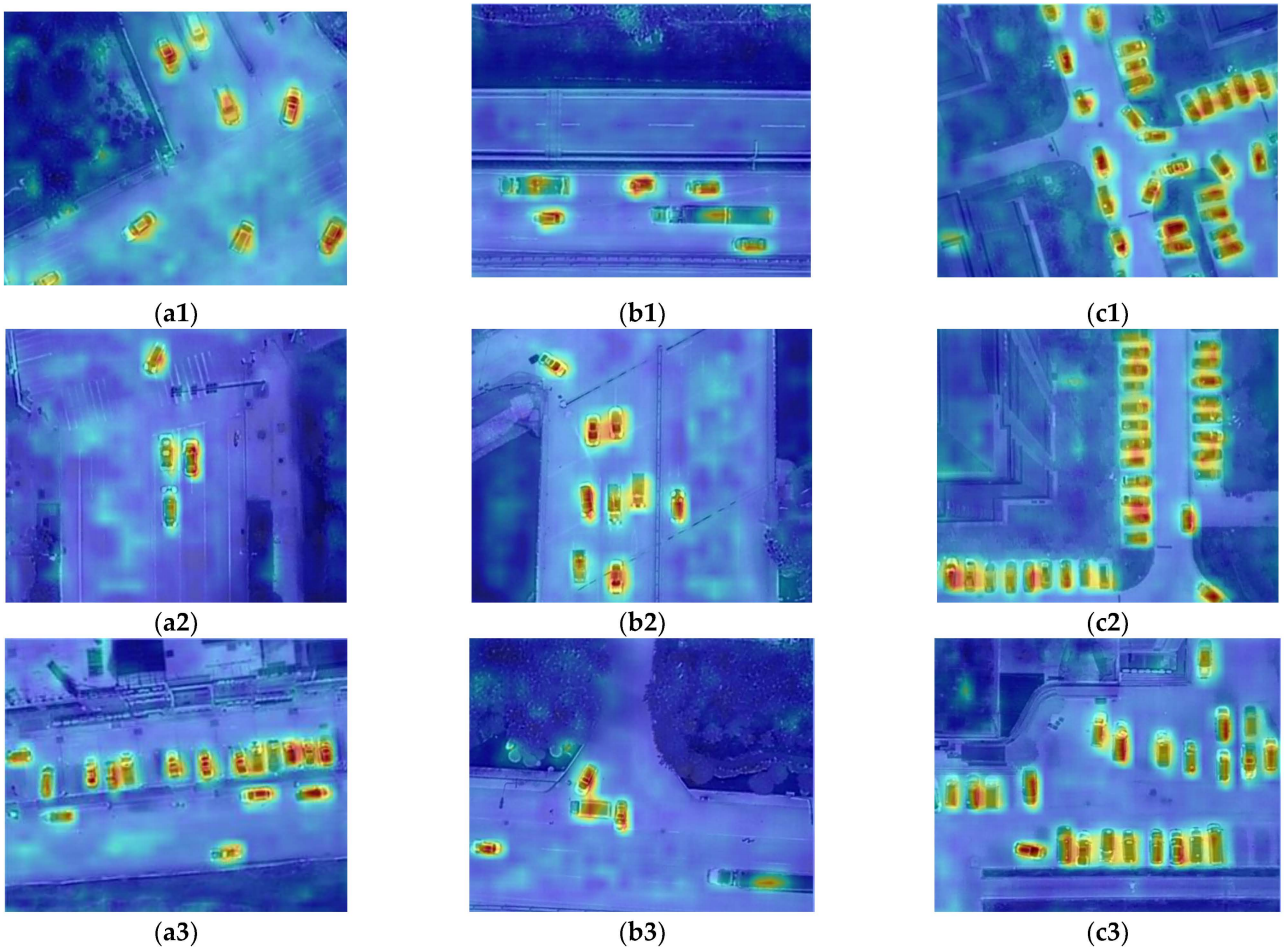


Figure 16. CAMs of different classes from the DroneVehicle datasets: (**a1**,**a2**,**a3**) moving cars on the road, (**b1**,**b2**,**b3**) cars, vans and freight cars, (**c1**,**c2**,**c3**) stopped cars.

- It can be seen from the CAM results on the DroneVehicle dataset that the ERGW-net achieves accurate localization results for small road targets from different infrared aerial images and categorizes them into different classes with high efficiency.

- In order to comparatively analyze the CAM visualization results of ERGW-net on the different datasets, Figure 17 shows the CAM visualization results of the algorithm on the HIT-UAV dataset for different classes.

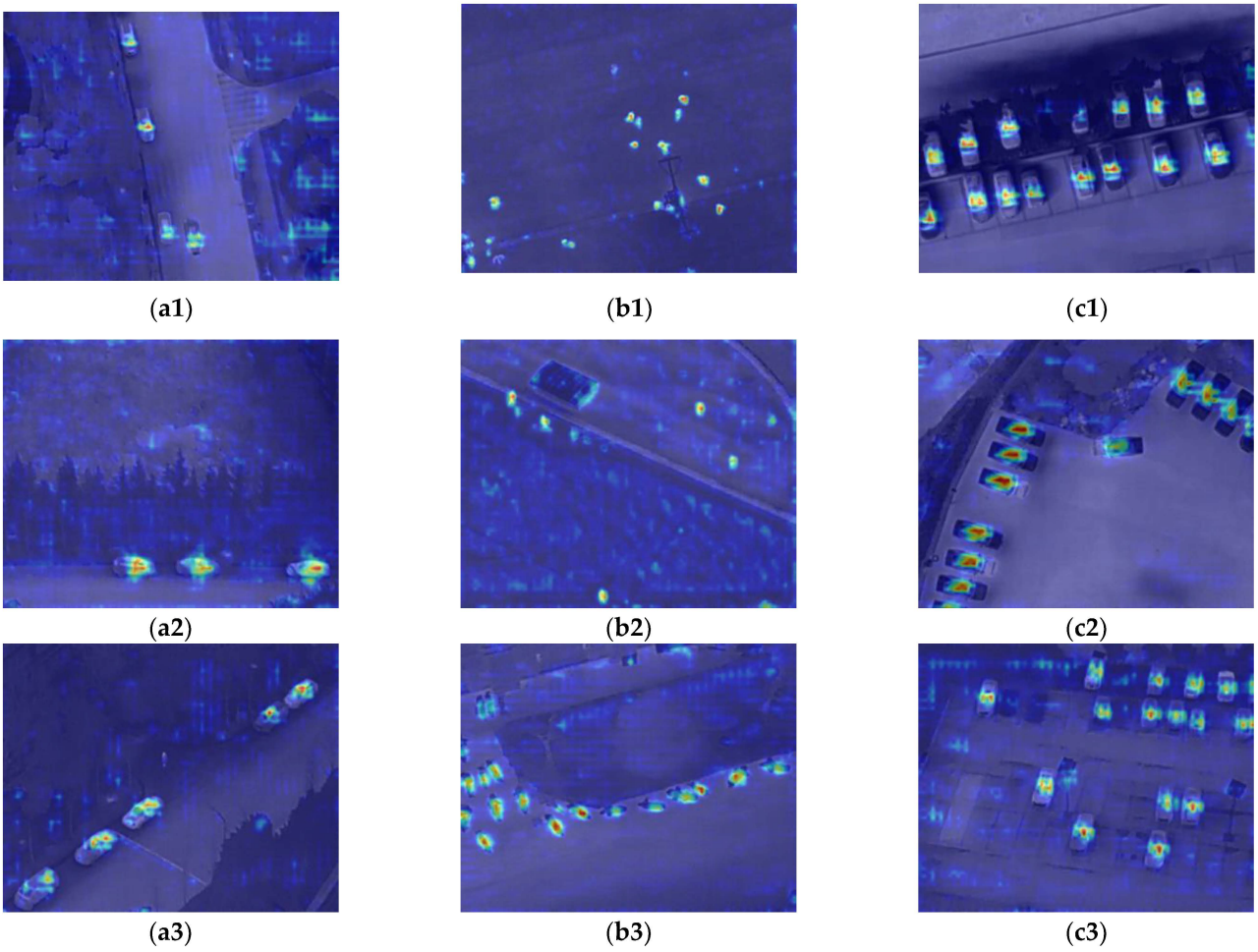
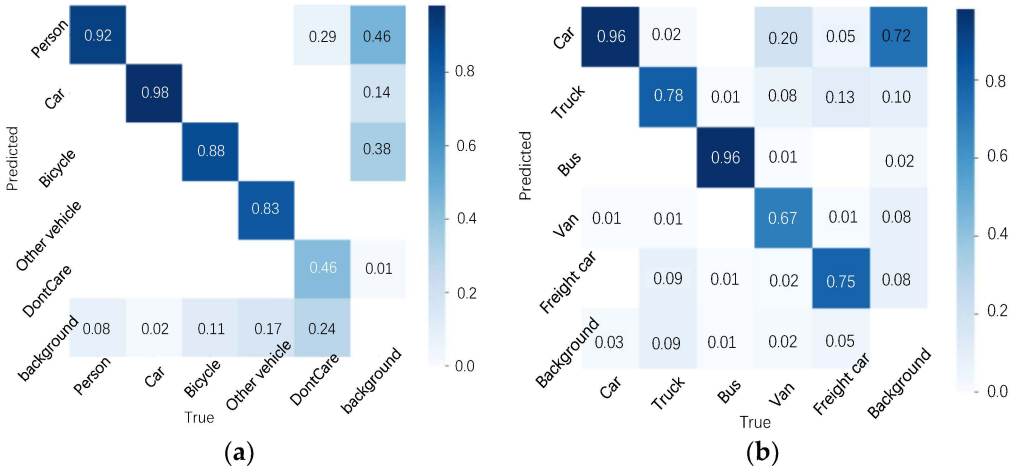


Figure 17. CAMs of different classes from the HIT-UAV datasets: (**a1**,**a2**,**a3**) cars on the road, (**b1**,**b2**,**b3**) people and bicycles, (**c1**,**c2**,**c3**) cars in parking lot.

- Figure 17 illustrates that the algorithm's CAM results on the HIT-UAV dataset are not as good as those on the DroneVehicle dataset. This is because the number of samples is small and the feature information of small targets such as bicycles and people is not obvious in the HIT-UAV dataset.
- In order to understand the false recognition rate of the ERGW-net between different classes during target detection, the confusion matrix of the algorithm on different datasets is given in Figure 18.
- Figure 18a shows that there is a higher probability that some background objects such as streetlights or small thermal targets are recognized as people, which is up to 0.46.
- Because there was no corresponding object called "DontCare" in the real world, and the number of relevant samples was very small in the HIT-UAV dataset, the mAP_{50} of "DontCare" is quite low.
- Because some of the traffic facilities and small houses have similar imaging features to car in the infrared image, it can be seen in Figure 18b that there is a higher probability that some of the objects in the background will be detected as car.
- According to Figure 18b, it is known that sometimes a truck is recognized as a freight_car; meanwhile, a freight_car is also detected as a truck. This is because a truck and a freight_car have relatively similar features except for the shape ratio.

- Compared with similar research, the number of classes detected using our method based on UAV infrared images is more than that of the other methods, and the proposed loss function in this method has some value for subsequent small object detection research.



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

Figure 18. Normalized confusion matrix of the ERGW-net in different datasets: (**a**) HIT-UAV and (**b**) the DroneVehicle datasets.

5. Conclusions

To realize the potential value of UAVs and thermal infrared technology to provide continuous road-target detection 24 h/day, this paper proposed the ERGW-net. Leveraging the advantages of InceptionNet, ResNet, and YOLOv8, our novel high-efficiency backbone, neck, and head constructions leveraged L_{GWPIoU} loss function to improve the overall detection accuracy of small road targets. Comparison experiments were carried out on the latest HIT-UAV and DroneVehicle datasets to measure the performance of our model and compare it with other state-of-the-art methods. All were used to classify road targets into a variety of classes (e.g., small cars, lorries, pedestrians, bicycles, and buses) and to produce visualizations of the results using CAMs. The mAP_{50} accuracy results of our model exceeded 80%, which is higher than any previous method. The effects of each part of the proposed model were analyzed with an ablation experiment, showing that the loss function contributed the most.

Author Contributions: T.A. and J.L. are co-first authors with equal contributions. Conceptualization, T.A. and J.L.; methodology, T.A.; validation, Y.Z.; writing, T.A. and Y.Z.; investigation, W.L. and N.G.; supervision, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the 14th Five-Year Plan Funding of China, grant number 50916040401, and in part by the Fundamental Research Program, grant number 514010503-201.

Data Availability Statement: The HIT-AUV dataset mentioned in this paper is openly and freely available at https://pegasus.ac.cn/ (accessed 24 September 2023). The drone vehicle dataset used in this study is freely available at https://github.com/VisDrone/DroneVehicle/blob/master/README.md (accessed 24 September 2023).

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

References

1. Cheng, J.R.; Gen, M. Accelerating genetic algorithms with GPU computing: A selective overview. *Comput. Ind. Eng.* **2019**, *128*, 514–525. [CrossRef]
2. Pennisi, S. The Integrated Circuit Industry at a Crossroads: Threats and Opportunities. *Chips* **2022**, *1*, 150–171. [CrossRef]
3. Hao, Y.; Xiang, S.; Han, G.; Zhang, J.; Ma, X.; Zhu, Z.; Guo, X.; Zhang, Y.; Han, Y.; Song, Z.; et al. Recent progress of integrated circuits and optoelectronic chips. *Sci. China Inf. Sci.* **2021**, *64*, 201401. [CrossRef]

4. Lee, C.Y.; Lin, H.J.; Yeh, M.Y.; Ling, J. Effective Remote Sensing from the Internet of Drones through Flying Control with Lightweight Multitask Learning. *Appl. Sci.* **2022**, *12*, 4657. [CrossRef]
5. Ecke, S.; Dempewolf, J.; Frey, J.; Schwaller, A.; Endres, E.; Klemmt, H.J.; Tiede, D.; Seifert, T. UAV-Based Forest Health Monitoring: A Systematic Review. *Remote Sens.* **2022**, *14*, 3205. [CrossRef]
6. Zhang, J.Z.; Guo, W.; Zhou, B.; Okin, G.S. Drone-Based Remote Sensing for Research on Wind Erosion in Drylands: Possible Applications. *Remote Sens.* **2021**, *13*, 283. [CrossRef]
7. Wavrek, M.T.; Carr, E.; Jean-Philippe, S.; McKinney, M.L. Drone remote sensing in urban forest management: A case study. *Urban For. Urban Green.* **2023**, *86*, 127978. [CrossRef]
8. Wang, X.T.; Pan, Z.J.; Gao, H.; He, N.X.; Gao, T.G. An efficient model for real-time wildfire detection in complex scenarios based on multi-head attention mechanism. *J. Real Time Image Process.* **2023**, *20*, 4. [CrossRef]
9. Liu, H.M.; Jin, F.; Zeng, H.; Pu, H.Y.; Fan, B. Image Enhancement Guided Object Detection in Visually Degraded Scenes. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [CrossRef]
10. Zhang, T. Target Detection for Motion Images Using the Improved YOLO Algorithm. *J. Database Manag.* **2023**, *34*, 3. [CrossRef]
11. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A.M. Deep learning techniques to classify agricultural crops through UAV imagery: A review. *Neural Comput. Appl.* **2022**, *34*, 9511–9536. [CrossRef]
12. La Salandra, M.; Colacicco, R.; Dellino, P.; Capolongo, D. An Effective Approach for Automatic River Features Extraction Using High-Resolution UAV Imagery. *Drones* **2023**, *7*, 70. [CrossRef]
13. Fakhri, S.A.; Satari Abrovi, M.; Zakeri, H.; Safdarinezhad, A.; Fakhri, S.A. Pavement crack detection through a deep-learned asymmetric encoder-decoder convolutional neural network. *Int. J. Pavement Eng.* **2023**, *24*, 2255359. [CrossRef]
14. Perz, R.; Wronowski, K.; Domanski, R.; Dąbrowski, I. Case study of detection and monitoring of wildlife by UAVs equipped with RGB camera and TIR camera. *Aircr. Eng. Aerosp. Technol.* **2023**, *95*, 1461–1469. [CrossRef]
15. Zhang, J.Y.; Rao, Y. A Target Recognition Method Based on Multiview Infrared Images. *Sci. Program.* **2022**, *2022*, 1358586.
16. Iwasaki, Y.; Kawata, S. A Robust Method for Detecting Vehicle Positions and Their Movements Even in Bad Weather Using Infrared Thermal Images. In *Technological Developments in Education and Automation*; Springer: Dordrecht, The Netherlands, 2010; pp. 213–217.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 1; Curran Associates Inc.: Lake Tahoe, NV, USA, 2012; pp. 1097–1105.
18. Zhang, X.; Zhu, X. Vehicle Detection in the Aerial Infrared Images via an Improved Yolov3 Network. In Proceedings of the IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 19–21 July 2019; pp. 372–376.
19. Ren, K.; Gao, Y.; Wan, M.; Gu, G.; Chen, Q. Infrared small target detection via region super resolution generative adversarial network. *Appl. Intell.* **2022**, *52*, 11725–11737. [CrossRef]
20. Alhammadi, S.A.; Alhameli, S.A.; Almaazmi, F.A.; Almazrouei, B.H.; Almessabi, H.A.; Abu-Kheil, Y. Thermal-Based Vehicle Detection System using Deep Transfer Learning under Extreme Weather Conditions. In Proceedings of the 8th International Conference on Information Technology Trends (ITT), Dubai, United Arab Emirates, 25–26 May 2022; pp. 119–123.
21. Zhang, X.X.; Zhu, X. Moving vehicle detection in aerial infrared image sequences via fast image registration and improved YOLOv3 network. *Int. J. Remote Sens.* **2020**, *41*, 4312–4335. [CrossRef]
22. Bhadoriya, A.S.; Vegamoor, V.; Rathinam, S. Vehicle Detection and Tracking Using Thermal Cameras in Adverse Visibility Conditions. *Sensors* **2022**, *22*, 4567. [CrossRef]
23. Tichý, T.; Švorc, D.; Růžička, M.; Bělinová, Z. Thermal Feature Detection of Vehicle Categories in the Urban Area. *Sustainability* **2021**, *13*, 6873. [CrossRef]
24. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
25. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection. *Sci. Data* **2023**, *10*, 227. [CrossRef] [PubMed]
26. Li, Z.H.; Hou, B.; Wu, Z.T.; Ren, B.; Ren, Z.L.; Jiao, L.C. Gaussian Synthesis for High-Precision Location in Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5619612. [CrossRef]
27. Wen, L.; Cheng, Y.; Fang, Y.; Li, X. A comprehensive survey of oriented object detection in remote sensing images. *Expert Syst. Appl.* **2023**, *224*, 119960. [CrossRef]
28. Liu, C.; Sui, X.; Kuang, X.; Liu, Y.; Gu, G.; Chen, Q. Adaptive Contrast Enhancement for Infrared Images Based on the Neighborhood Conditional Histogram. *Remote Sens.* **2019**, *11*, 1381. [CrossRef]
29. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef] [PubMed]
30. Szegedy, C.; Wei, L.; Yangqing, J.; 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 (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
31. 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.

32. Weng, K.; Chu, X.; Xu, X.; Huang, J.; Wei, X. EfficientRep: An Efficient Repvgg-style ConvNets with Hardware-aware Neural Network Design. *arXiv* **2023**, arXiv:2302.00386.
33. Sergey, I.; Christian, S. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 37, pp. 448–456.
34. Dubey, S.R.; Singh, S.K.; Chaudhuri, B.B. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* **2022**, *503*, 92–108. [CrossRef]
35. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.-S. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 79–93. [CrossRef]
36. Ma, S.; Xu, Y. MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. *arXiv* **2023**, arXiv:2307.07662.
37. Lan, J.H.; Zhang, C.; Lu, W.J.; Gu, N.W. Spatial-Transformer and Cross-Scale Fusion Network (STCS-Net) for Small Object Detection in Remote Sensing Images. *J. Indian Soc. Remote Sens.* **2023**, *51*, 1427–1439. [CrossRef]
38. Padilla, R.; Netto, S.L.; Silva, E.A.B.d. A Survey on Performance Metrics for Object-Detection Algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 1–3 July 2020; pp. 237–242.
39. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
40. Mahaur, B.; Mishra, K.K. Small-object detection based on YOLOv5 in autonomous driving systems. *Pattern Recognit. Lett.* **2023**, *168*, 115–122. [CrossRef]
41. Hussain, M. YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection. *Machines* **2023**, *11*, 677. [CrossRef]
42. Kim, J.H.; Kim, N.; Won, C.S. High-Speed Drone Detection Based On Yolo-V8. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–2.
43. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
44. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
45. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.