Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 8683226, 12 pages
https://doi.org/10.1155/2021/8683226

Research Article

Vehicle Detection in Remote Sensing Image Based on Machine Vision

Liming Zhou[1,2] **Chang Zheng**[1,2] **Haoxin Yan**[1,2] **Xianyu Zuo**[1,2] **Baojun Qiao**[1,2] **Bing Zhou**[1,2] **Minghu Fan**[1,2] **and Yang Liu**[1,2,3]

[1]*Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, Henan, China*
[2]*School of Computer and Information Engineering, Henan University, Kaifeng, Henan, China*
[3]*Henan Engineering Laboratory of Spatial Information Processing, Henan University, Kaifeng, Henan, China*

Correspondence should be addressed to Bing Zhou; zhoub1631@163.com

Received 5 April 2021; Revised 23 June 2021; Accepted 30 July 2021; Published 10 August 2021

Academic Editor: Anastasios D. Doulamis

Target detection in remote sensing images is very challenging research. Followed by the recent development of deep learning, the target detection algorithm has obtained large and fast growth. However, in the application of remote sensing images, due to the small target, wide range, small texture, and complex background, the existing target detection methods cannot achieve people's hope. In this paper, a target detection algorithm named IR-PANet for remote sensing images of an automobile is proposed. In the backbone network CSPDarknet53, SPP is used to strengthen the learning content. Then, IR-PANet is used as the neck network. After the upper sampling, depthwise separable convolution is used to greatly avoid the lack of small target feature information in the convolution of the shallow network and increase the semantic information in the high-level network. Finally, Gamma correction is used to preprocess the image before image training, which effectively reduces the interference of shadow and other factors on training. The experiment proves that the method has a better effect on small targets obscured by shadows and under the color similar to the background of the picture, and the accuracy is significantly improved based on the original algorithm.

1. Introduction

Remote sensing target detection is to mark the object of interest in remote sensing images and then forecast the type and location of this targets. In the traditional detection dataset, the target is concentrated, while the aviation dataset is not, and the object strength in the aviation image usually appears in arbitrary orientation, which depends on the perspective of the Earth vision platform [1].

Object detection is the process of detecting instances of semantic objects of a certain class (such as humans, airplanes, or birds) in digital images and video [2]. Small target detection has always been a hot and difficult area in target detection tasks. The study of remote sensing imagery has vital applications in military, disaster control, environmental management, and transport planning [3–6]. And vehicles in remote sensing images as a special category, whether civilian, military, or transportation, have an important meaning and at the same time are more challenging.

Firstly, the small target in the target detection task is usually a target that is less than 30 pixels in the image whereas vehicle targets in remotely sensed images are usually below 20 pixels or even 10 pixels. Secondly, the class of vehicle in remotely sensed images is often subject to weather and environmental images such as atmospheric occlusion, shadow occlusion, and building occlusion and other factors, for example, different overhead views, different sizes of vehicle targets in the same image, similar colors between vehicles and their surroundings, and so on. It resulting in poor detection accuracy of car targets.

For the past few years, followed by the rapid science development of deep learning, great progress in target

detection methods has been made. Among them, the popular target detection methods based on the convolutional neural network [7–12] can be divided into one stage [7–9, 13–18] and two stages [8, 19–24]. One stage, such as YOLO [9], SSD [7], DSSD [13], YOLOv2 [17], YOLOv3 [18], and Retina Net [16] and so on. The one-stage algorithm has a very fast speed in the detection process, and it has a widespread application in each kind of scenes which needs high efficiency. In the traditional target dataset, the one-stage algorithm in the process of detection always has a fast speed and high accuracy. However, when applied to optical remote sensing datasets, its accuracy will be greatly reduced. The reasons for this phenomenon are as follows: (1) optical remote sensing image is a bird's-eye view; thus, its height and shooting angles are uncertain, and (2) there are many small targets of less than 30 pixels in aerial images, but the one-stage method, such as SSD, does not perform well for small targets.

The two-stage algorithms, such as RCNN [20], SPP net [24], fast RCNN [21], faster RCNN [19], and Mask R-CNN [23], divide the detection task into two subtasks: recognition and localization. Compared with one-stage algorithms, the accuracy of the two-stage algorithm is higher, but their training time is often longer, and the detection speed is relatively slow.

Since the YOLO series algorithm was put forward, its speed has been widely affirmed, but the accuracy is relatively poor. The latest YOLOv4 algorithm has greatly improved in accuracy and has been widely used in many fields. The backbone network of the YOLOv4 algorithm improves the darknet53 network proposed in the YOLOv3 algorithm by adding CSP[25] module, and SPP enhanced network [24] is added after the backbone network. PANet [26] is used for the neck network of the algorithm, and YOLO head is still used as the detector. In the PANet structure [26], the algorithm divides the feature map into three scales according to the scale size, namely,19×19, 38×38, and 76×76 for detection. Upsampling is carried out at the two adjacent scales, and then downsampling is performed after several convolutions. Although the YOLOv4 algorithm has achieved good results so far, there is still a lot of room for improvement in its detection effect in the targets of remote sensing images, especially in the car class.

In this paper, we propose the IR-PANet algorithm for targets in shadow occlusion and targets with a similar color to their surroundings in remote sensing images. The detector replaces convolution with inverted residual [27] based on PANet, which can increase the depth of the model, enrich the semantic information of the algorithm and increase its detection accuracy. In the training, CSPDarknet53 is used as the backbone network. To test the capability of the network, we use the HRSSD [28] dataset to pretrain our model. Considering that vehicles in remote sensing images are easily obscured by shadows, we apply GAMMA correction to the image for brightening the image and removing noise. The main contributions of this study are as follows:

(1) Given the particularity of remote sensing images, an IR-PANet is proposed and applied to the YOLOv4 method. The recognition ability of the model for small targets and occluded targets is increased. In this method, we replace the original convolution with inverted residual after upsampling, and the improved algorithm is shown in Figure 1. In our method, the network becomes deeper without much increase in computation, which greatly deepens the robustness of the network and increases the network semantic information.

(2) The original image is preprocessed. Before training, Gamma correction is used to preprocess the image, which reduces the noise in the original image, brightens the shadow part of the image, and improves the recognition rate of the shadow-covered target. The processed image is shown in Figure 2.

(3) YOLOv4 algorithm has a low recognition rate for targets with complex backgrounds. In this paper, we compare two classes of the HRSSD dataset, vehicle class and ship class. Among them, the vehicle data and background are complex and have shadow, building occlusion, incomplete semantic information, dense target, and so on. However, the background of ship data is simple and the target is single. Proved after the experiment, the accuracy of the original algorithm in the vehicle dataset is not high, while the accuracy of the IR-PANet algorithm in the vehicle dataset has been significantly improved after training.

The rest of this paper is organized as follows: Section 2 introduces related research on vehicle detection in common images and remote sensing images. Section 3 describes our proposed model of remote sensing image target detection. Section 4 describes the experimental design and experimental details. Section 5 describes the experimental results and analysis. Finally, we conclude this study in section 6.

2. Related Work

While existing algorithms have achieved good results in detection tasks, much work has been done to improve them for the characteristics of vehicle targets. For example, Jun et al. [29] proposed a vehicle detection model YOLOv2_-vehicle based on the YOLOv2 algorithm and obtained an average accuracy of 94.78% on the Beijing Institute of Technology (BIT) Vehicle validation dataset. Hu et al. [30] proposed a scale-insensitive convolutional neural network (SINet) for fast detection of vehicles with large scale variance. Nguyen [31] proposed an improved fast R-CNN based framework for fast vehicle detection. Xu et al. [32] proposed a Side Fusion FPN algorithm for application in the Resnet-86 backbone network to detect vehicles and pedestrians on the road.

Deep learning continues to have a wide range of applications in remote sensing image target detection tasks. For instance, Deng et al. [33] proposed a unified and effective method for simultaneous detection of multiple classes of targets in large scale change remote sensing images by using multiscale object proposal network (MS-OPN) and an accurate object detection network (AODN) for target classification and detection and achieved higher accuracy in

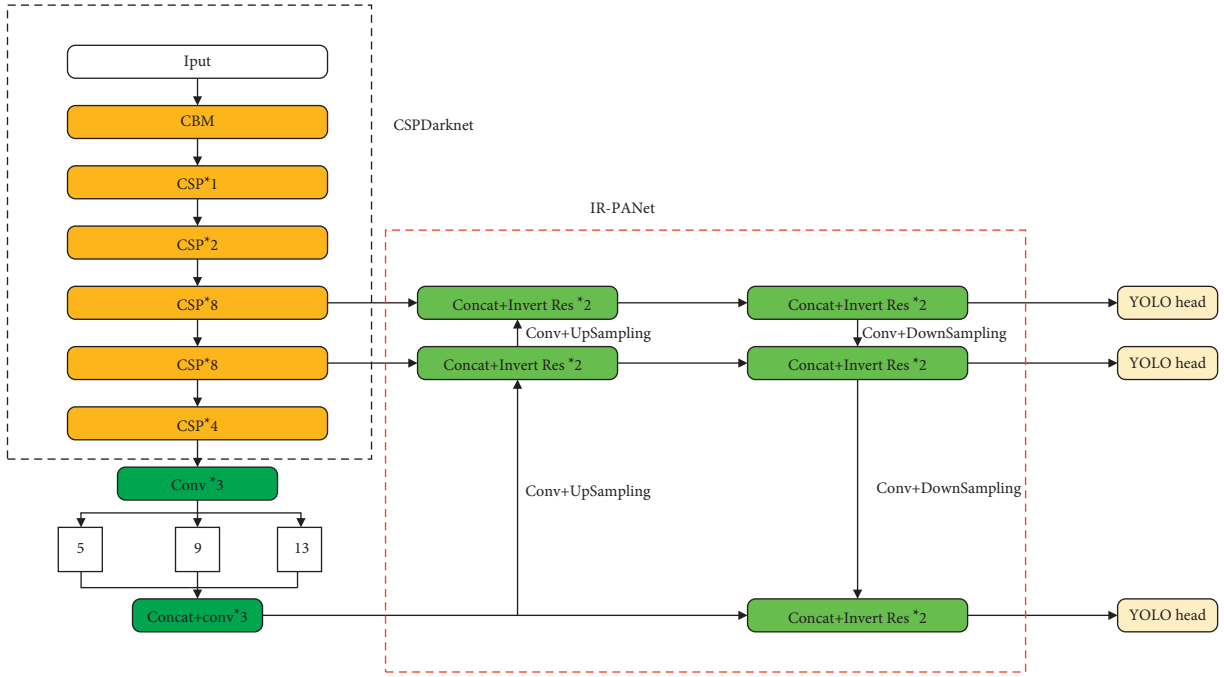


Figure 1: The network structure of IR-PANet with CSPDarknet.

datasets such as NWPU VHR-10. Zhang et al. [34] proposed a dual multiscale feature pyramid network (DM-FPN) by combining strong semantics, low-resolution features, and weak semantics using the inherent multiscale pyramidal features of remote sensing images. Eten et al. [35] improved a network for raw resolution detection of remote sensing data based on YOLOv2 and concluded that objects as small as 5 pixels in size can still be located at high resolution. Ming et al. [36] introduced the CFC-NET detection network, which optimizes the single-stage detector in terms of feature representation, anchor refinement, and training sample selection.

For vehicle-based targets in remote sensing images, Tang et al. [6] used an R-CNN network-based approach for real-time monitoring of remote sensing images of vehicle targets. Gao et al. [37] proposed a novel detection model, DE-CycleGAN, to enhance weak targets and achieve accurate remote sensing image detection. Zhang et al. [38] proposed a three-step local proposal method (LRP) for the detection of live vehicles in satellite video. Shi et al. [39] proposed a single-stage and anchorless detection method to detect oriented vehicles in high-resolution aerial images by linking coarse and fine feature maps output from different stages of the residual network through a feature pyramid fusion strategy.

3. Materials and Methods

In this section, we first introduce the used backbone network CSPDarknet and, then, the structure of the inverted residual block and its advantages. Then, the next section introduces our proposed IR-PANet network structure. Finally, we detail the image preprocessing method, GAMMA correction.

3.1. Backbone Network. The backbone network used in this paper is CSPDarknet [40]. YOLO-based target detection algorithms say that detection problems are defined as single regression problems; that is, a single neural network can achieve probability prediction of multiple enclosing boxes and classes for an entire image through a positive operation. Based on YOLOv3, CSPDarknet53 introduces the CSP module based on Darknet53. Before each downsampled, it will carry out semantic fusion with the top-level information to increase network depth and enrich the network. Each residual module consists of convolution layers of 1×1 and 3×3 and a quick connection. Convolution structure is composed of convolution layer, batch normalization, and mish [41] activation function.

CSPDarknet53 attributes the problem to repeated gradient information in network optimization and focuses on the variability of the gradient by integrating feature mapping from the beginning and end of the network control phase [25]. The structure of the CSP module is shown in Figure 3.

3.2. The Module of Invert Residual. In the remote sensing image, the content is complex. All these lead to the loss of semantic information in the image during the convolution process. Objects like vehicles are easily affected by clouds, shadows, buildings, and other factors. The principle of invert residual is to replace convolution with depthwise separable convolution, deeply separate information, fuse channels through multilayer $1 * 1$ convolution, and finally restore to the target size, in which each block contains an input, the middle is several bottlenecks, and then it is the expansion. However, the bottleneck contains all the necessary information, and the extension layer is only used as the implementation details of the neighborhood tensor nonlinear

Figure 2: The car category in the remote sensing dataset. Where targets marked by yellow circles are vehicle targets obscured by trees, target marked by the white circle is vehicle target in shadow, and target marked by the red circle is vehicle target whose color is similar to that of the surrounding environment.

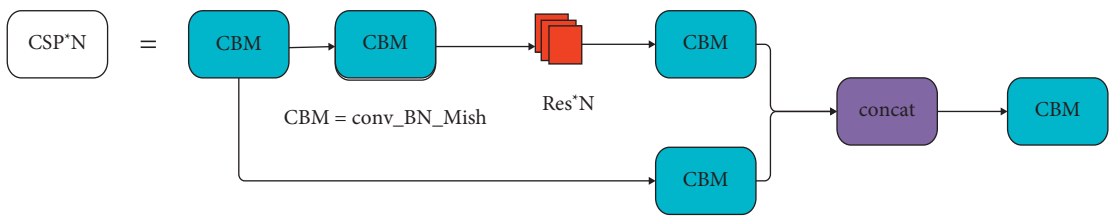


Figure 3: The network structure of the CSP module.

transformation. We directly use the shortcut between the bottlenecks [27].

In the invert residual module, we still use LeakyRelU as our activation function, and the expression is as follows:

$$Leaky\ Re\ LU\ (x) = \begin{cases} x, & x \geq 0, \\ ax, & x \leq 0, \end{cases} \in R \qquad (1)$$

where x is the input information. In invert residual, the total number of multilayer additions required for blocks is $(m/s) \times (n/s) \times (tc)$ with the input size of $m \times n$, t is the expansion factor, and c is the core size. As shown in Table 1, compared with a convolution, the amount of computation is greatly reduced. The input and output parameters for the calculation of standard convolution and inverse residual are shown in Table 1.

The interface diagram of the inverted residual module is shown in Figure 4.

3.3. The Network Structure of IR-PANet. The high-level network has a strong sense of the whole image, while other neurons are more likely to be activated by local texture and pattern [26], which indicates that it is necessary to increase the top-down path to spread semantic features to enhance the robustness of the network.

PANet network first follows the definition of FPN, which has upsampling from bottom to top, enriching the feature information of the upper network detection layer, while CSPDarknet53 detection layer generates three prediction layers, which are 19×19, 38×38, and 76×76, respectively. The semantic information of the upper network is convoluted and then downsampled to the deep network, which can also enrich the semantic information of the network. This greatly improves the FPN network's ability to detect small targets. We replace the cubic convolution (1 × 1and 3 × 3 as a group) in PANet with twice inverse residual, and the output of the upper network is downsampled. Remote sensing images contain many objects with little textural information and low counterbalance. The original model is difficult to recognize complex features. The IR-PANet increases the depth of the model and reduces its calculation time, which has a positive effect on the learning process of the network. The IR-PANet network structure is shown in Figure 2.

As can be seen from Figure 2, IR-PANet performs target detection by merging feature maps of different sizes. The network upsamples each layer of the feature pyramid from top to bottom and then performs two invert residual horizontally, followed by downsampling and semantic fusion layer by layer. Besides, the output of the network contains more information about the target's location in the deep layer and the network outputs contain less information on the location of the target in the complex network. In our algorithm, we use 19×19, 38×38, and 76×76 feature maps to construct feature pyramids for large, medium, and small target detection. Prediction at multiple scales makes the algorithm have greater sensitivity to small targets and the targets in a complex environment and significantly improves its detection ability.

3.4. Gamma Correction. Before we send the image into the deep learning algorithm, we first preprocess the image. As shown in Figure 5, remote sensing images are easy to receive weather, angle, building occlusion, shadow occlusion, and other factors, which affect the accuracy of our algorithm. By gamma correction, adjusting the contrast of the image, and enhancing local details, we can reduce the shadow and other

Table 1: Comparison of standard convolution and inverse residual.

Input	Operator	Output
$m \times n \times c$	$1 * 1$ convolution	$m \times n \times (tc)$
$m \times n \times (tc)$	$3 * 3$ invert residual	$(m/s) \times (n/s) \times (tc)$

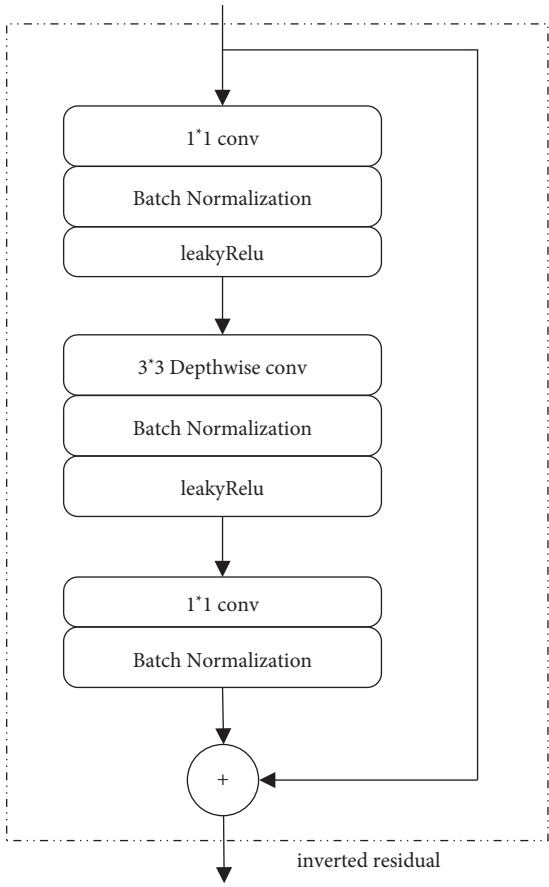


Figure 4: The network structure of the inverted residual module.

factors of the image. According to [42], the formula based on Gamma correction is

$$F(I) = I^{\gamma}, \tag{2}$$

where γ is an important adaptive parameter designed according to the principle of image chromaticity and uniform illumination intensity. As can be seen from Figure 4, when γ is smaller than 1, the effect is that the image almost dims, and when equaling 1, the image is not enhanced. Again, when γ is higher than 1, the effect of the image shows artificial facts.

4. Experiments Design

4.1. Datasets and the Evaluation Metric

4.1.1. The Description of Dataset. We use the HRSSD dataset of Xi'an optical and Precision Machinery Research Institute of Chinese Academy of Sciences to verify our proposed method and select automobiles category in the dataset to evaluate. In addition, we still tested the ship class in this dataset to check the robustness of the algorithm. There are 1188 train sets, 1186 validation sets, and 2382 test sets in the automobile category and 950 train sets, 948 validation sets, and 1988 test sets in the ship category. If you want to find out more information about the dataset, please visit https://github.com/CrazyStoneonRoad/TGRS-HRRSD-Dataset.

4.1.2. Evaluation Index. In this experiment, loss curve value and average precision (AP) were used for evaluating our method. The change of curve loss can reflect the error. The change in the loss curve may reflect the error between the predicted and actual results. Therefore, the faster the decrease in the loss curve and the smaller the value of the loss, the better the result of the training model. AP refers to the proportion of correct box selection in the prediction box. The higher the proportion is, the more correct box selection targets are, which indicates that the following result of the training model is better.

The loss function we use is DIoU [43]. Compared with IOU [44], DIoU directly minimizes the distance between two middle points, and the prediction box is closer to the target box. DIoU can be defined as follows:

$$\mathfrak{R}_{DIou} = 1 - \mathrm{IoU} + \frac{\rho^2\left(a, a^{gt}\right)}{b^2}, \tag{3}$$

where a and a^{gt} are the center of the forecasting box and the target box, ρ is the Euclidean distance, and b is the diagonal length of the shortest closed box that covers both boxes.

In multiclass target detection, mean accuracy (map) is being widely used for the evaluation index. This mAP is equal to the average of the AP values of all the categories. Its expression is as follows:

$$\begin{aligned} \mathrm{precision}_j &= \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j}, \\ \mathrm{recall}_j &= \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FN}_j}, \\ \mathrm{AP} &= \int_0^1 \mathrm{precision}_j\left(r_j\right) d\left(r_j\right), \\ m\mathrm{AP} &= \frac{1}{n}\sum_{j=1}^{n} AP_j. \end{aligned} \tag{4}$$

If the IOU of the detected box is greater than 0.5 with that of the box labeled in the actual dataset, the detected box is considered as the true position. For class J, TP_j, FP_j, and FN_j are the figures for true positions, false positions, and false negatives respectively. n indicates the number of HRSSD dataset classes, and $k = 2$. r_j is the class of recall. However, the accuracy and recall are contradictory. As the

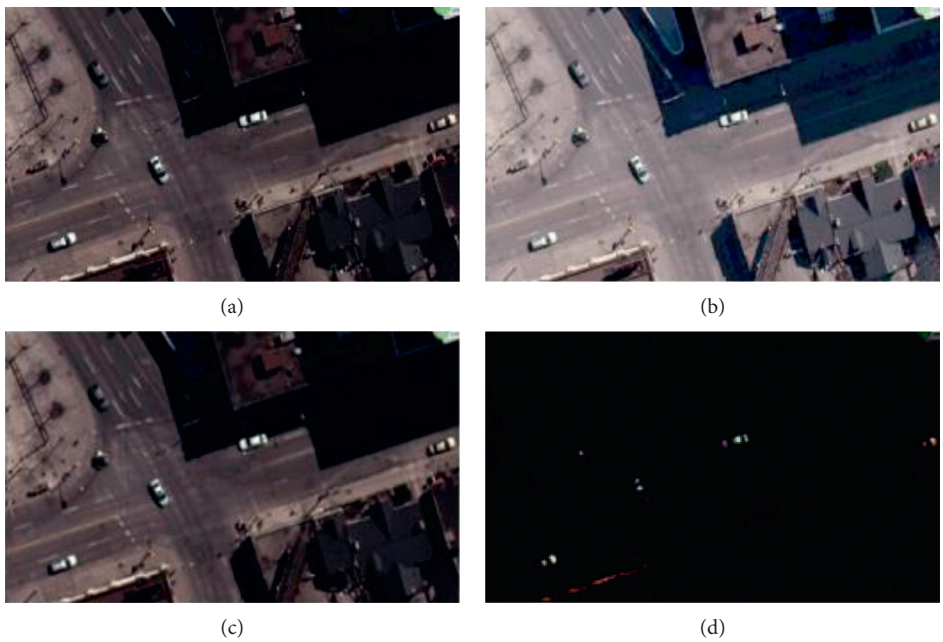


Figure 5: Result of the effect of gamma correction: (a) original; (b) $\gamma < 1$; (c) $\gamma = 1$; and (d) $\gamma > 1$.

number of recalls increases, the accuracy of recall also decreases. Therefore, combining the accuracy with recall, we use mAP to evaluate our algorithm.

4.2. Experimental Details. In this paper, the algorithm language is C language, the operating system is Ubuntu 16.04, GPU is Quadro P4000 8GB×2, and the hardware platform is Intel (R) Xeon (R) silver 4114, CPU @ 2.2 GHz. In the process of training, the momentum is set to 0.949, and the model is optimized by asynchronous small-batch stochastic gradient descent. The initial learning rate is set to 0.0013. When the figure for training iterations is 4800 and 5400, the learning rate is adjusted to 0.00013 and 0.000013, respectively. We have iterated 6000 times of this algorithm.

In this section of the experiment, the input size of the image changes randomly in the experimental training with a multiple of 32 (the range of change is 320×320 to 608×608 pixels). The batch size is 64. Meanwhile, the training data can be increased by adjusting saturation, exposure, and tone.

5. Experimental Result

Figure 6 is the convergence curve of the loss value of our network framework when training the ship dataset. The horizontal axis represents the figure for training iterations, the maximum iteration is 6000, and the longitudinal axis represents the loss and accuracy. From Figure 6, it is to recognize that when the number of iterations is 500, it tends to be stable. Then, we can see that the loss fluctuation interval of IR-PANet is always lower than that of YOLOv4, and its scatter is denser. Then, as can be seen in Table 2, the number of TP detected by IR-PANet is much higher than that of YOLOv4 and FN is also much lower than that of YOLOv4 while the value of precision is lower due to higher FP.

There are 2382 test targets in the automobile dataset and 1988 test targets in the ship dataset. Among them, in the automobile dataset, the target is generally small, the distribution range is dense, the target features are occluded, and the environment is complex, resulting in the relatively high loss value of this method, but we can also see that the value of loss tends to be stable around 1000 generations. For different networks, the test results of the two classes in the HRRSD dataset are shown in Table 3 and Figure 7.

In this paper, after 1188 remote sensing images are tested, it is found that the accuracy of vehicles in remote sensing images is very high, and it also has a good recognition effect in the areas with dense vehicles and shaded areas. The detection effect and accuracy are shown in Figure 8, and the contrast area is marked with a yellow box. From Figures 8(a) and 8(b), it can be found that IR-PANet detects more dense and occluded vehicle targets than IR-PANet. In Figure 8(c), the target location of missed detection is dense and shadowed with the accuracy of 78% and 84% in Figure 8(d). In IR-PANet, the accuracy of the two targets detected by YOLOv4 in these target vehicles is also increased from 100% and 99% to 99% and 94%, respectively. This indicates that IR-PANet is higher than YOLOv4 in this complex environment in detection accuracy.

In the ship dataset, as a result of the large pixel size and obvious features of the target, the map of IR-PANet is similar to YOLOv4. From Figure 9(a), the left ship is affected by shadow and missed in the algorithm. In Figure 9(b), the image after Gamma correction preprocessing is completely detected. This clearly shows that after Gamma correction, the algorithm is easier to learn useful feature information.

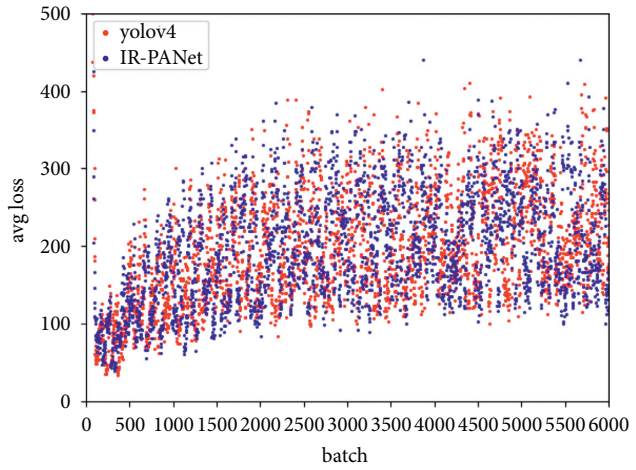


Figure 6: Loss comparison chart of YOLOv4 and IR-PANet.

Table 2: Comparison of TP (the number of actual positive classes predicted as positive classes), FP (the number of actual negative classes predicted as positive classes), FN (the number of actual positive classes predicted as negative classes), *F*1-score, recall, and precision of YOLOv4 and IR-PANet in vehicle classes

	YOLOv4	IR-PANet
TP	2029	**2284**
FP	123	404
FN	294	**39**
*F*1-score	0.91	0.91
Recall	0.87	0.98
Precision	0.94	0.85

Table 3: Results of Faster RCNN [45], CACL Faster RCNN [45], YOLOv4, IR-PANet, and IR-PANet with GAMMA in the dataset.

	Vehicle	Ship	mAP	FPS
Faster RCNN	0.84	0.885	0.8625	—
CACL Faster RCNN	0.869	0.885	0.877	—
YOLOv4	0.9090	0.9768	0.9429	19
IR-PANet	0.9823	0.9773	0.9798	15.2
IR-PANet with GAMMA	**0.9835**	**0.9791**	**0.9813**	15.2

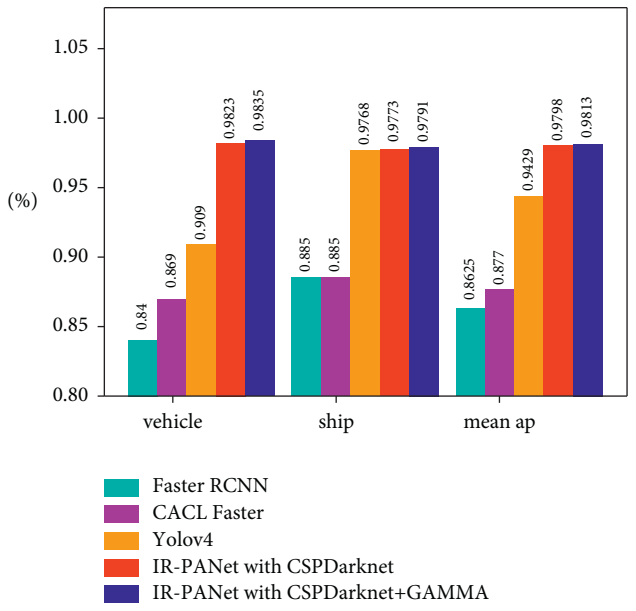


Figure 7: Bar chart comparison of different methods of AP and map in automobile and ship datasets.

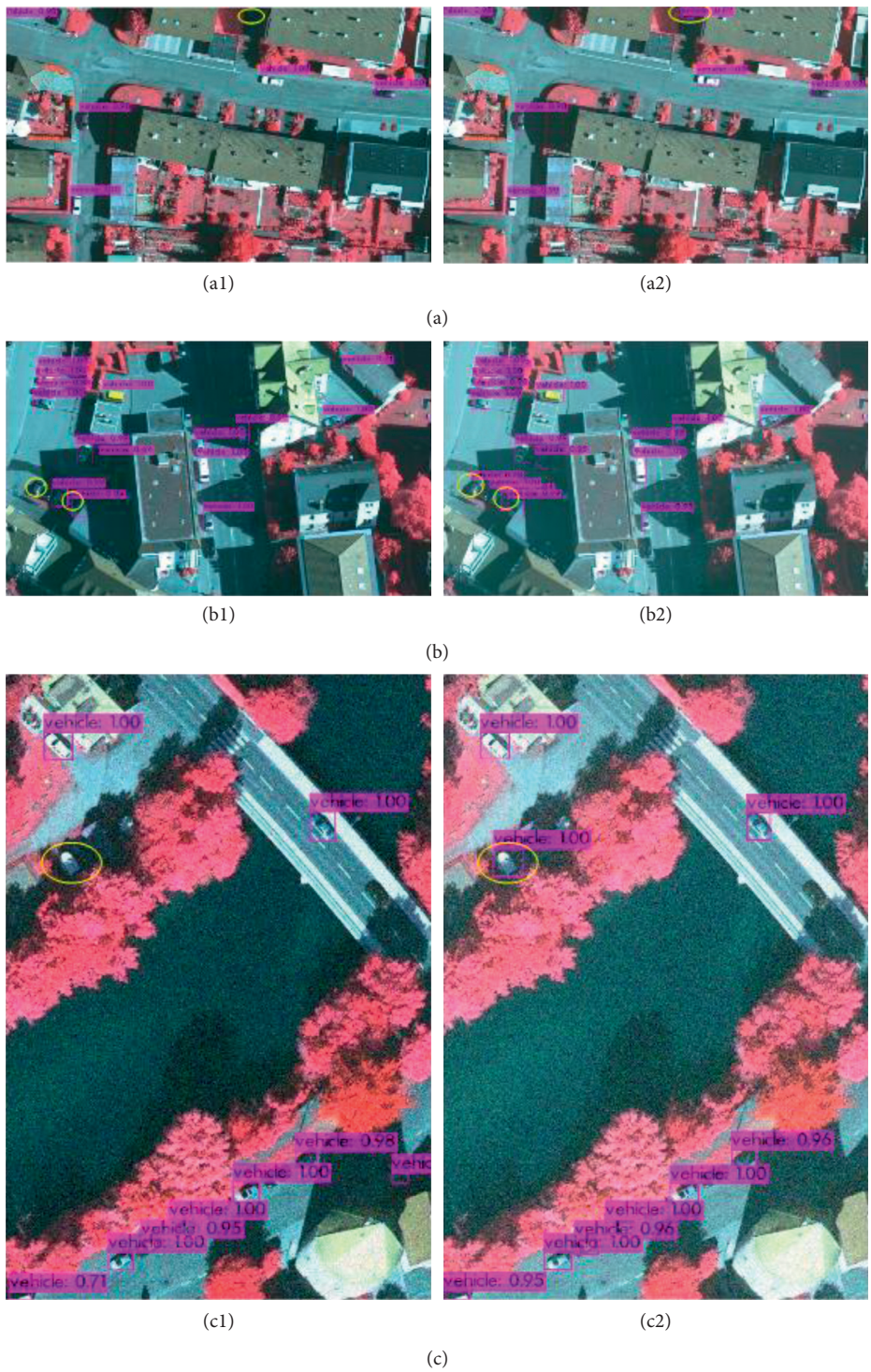
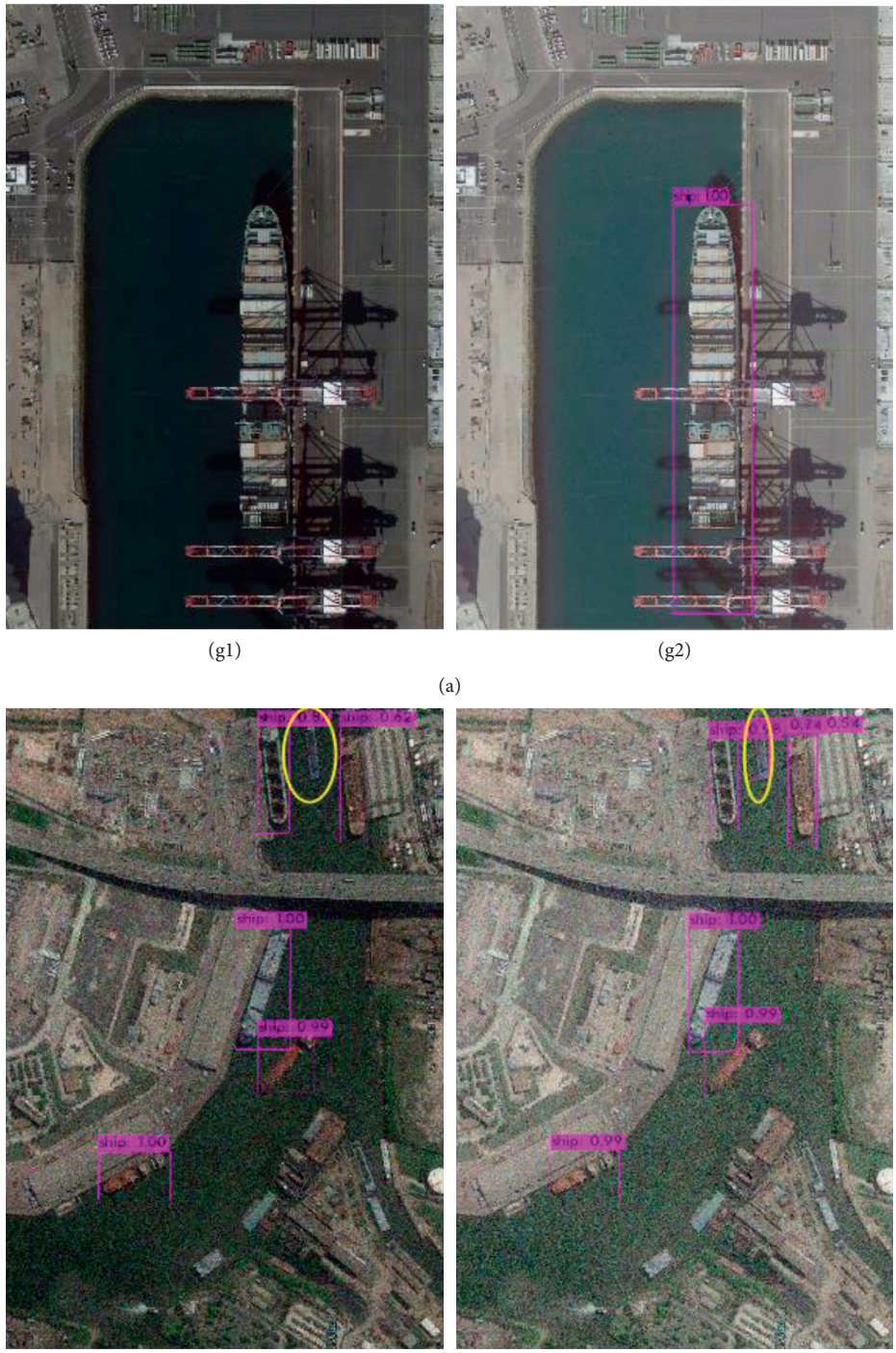


Figure 8: Detection results in the shadow-obscured background, where (a, c, e) are the detection results of the YOLOv4 algorithm and (b, d, f) are the detection results of IR-PANet.

In the following, we will show the comparison result graphs of YOLOv4 and the proposed algorithm in terms of targets in a shadow-obscured environment and dark target miss detection.

Target shadow occlusion is a difficult problem for target detection in remote sensing images, and as can be seen in Figure 8, the proposed algorithm works better in this scene. It is also clear from Figures 8(e) and 8(f) that the detection of targets in shadow is more accurate.

The dark target features in remote sensing images are close to the shadows as well as the picture background, which is also easy to miss and error. It can be seen from Figure 10 that IR-PANet has slightly better accuracy than YOLOv4 in identifying dark targets, while Figures 10(e) and



(g1) (g2)

(a)

(h1) (h2)

(b)

Figure 9: Comparison results of the YOLOv4 algorithm with the proposed algorithm in the ship class.

10(f) demonstrate that the proposed algorithm is more accurate in detecting shadows clearly.

In the ship class, the target is less affected by the environment; as can be seen in Table 3, the map of IR-PANet and YOLOv4 algorithm is close. And from Figure 9, the ship target in g1 receives shadow occlusion, while the image is clearer after image processing, and our algorithm is better recognized. In Figures 9(c) and 9(d), the detection effect of

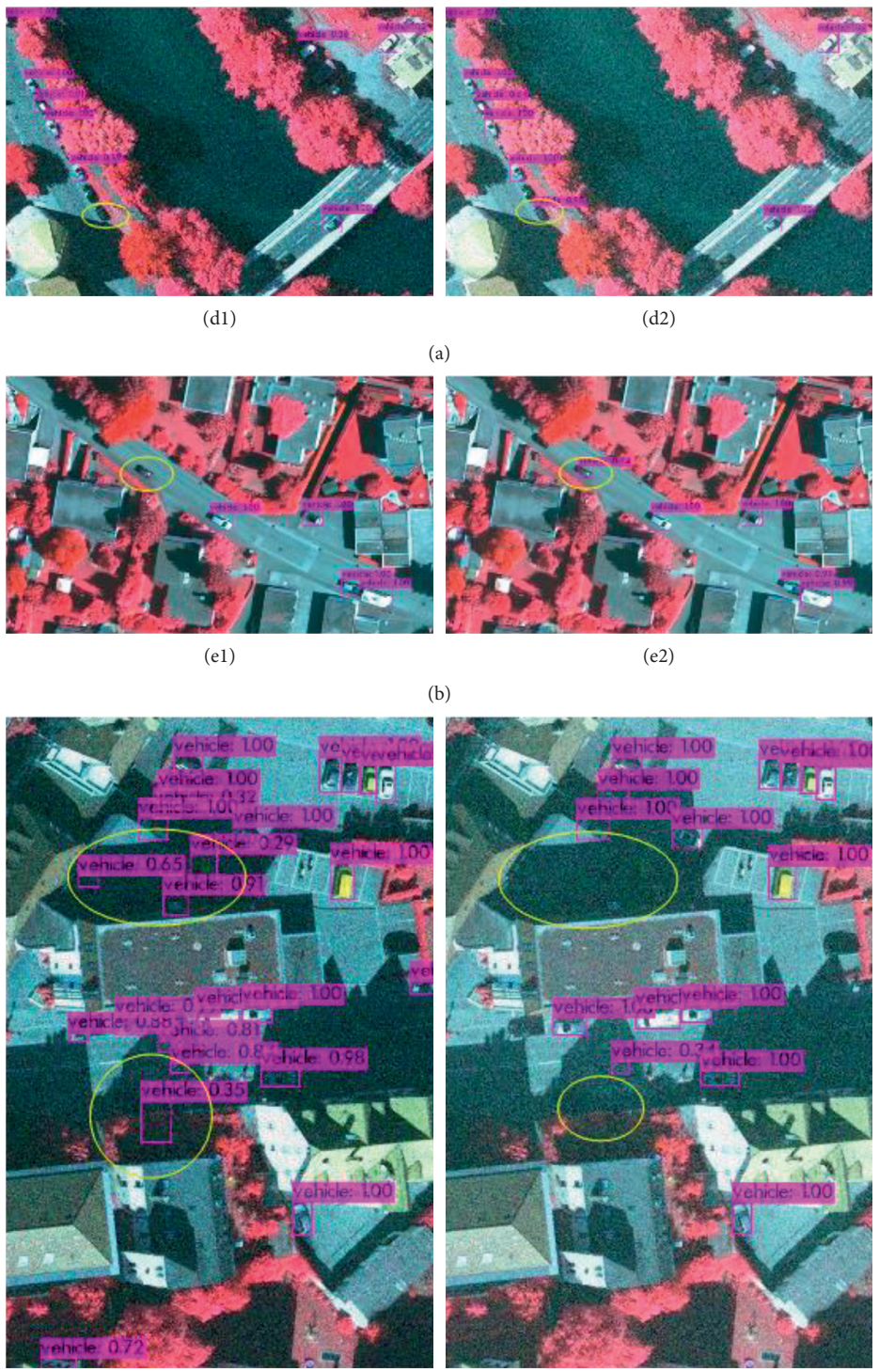


Figure 10: The results of dark target detection are compared, where (a, c, e) are the detection results of YOLOv4 and (b, d, f) are the detection results of IR-PANet, where the comparison results of (e and f) can see that YOLOv4 is easier to identify the shadow block as a target, while IR-PANet is more accurate.

the target color is similar to the background color which can also be better shown. This fully illustrates that IR-PANet has strong robustness.

6. Conclusions

We propose a new CNN structure, IR-PANet, to identify vehicle targets in optical remote sensing imagery. The network is applied in the CSPDarknet backbone. In the detection layer, we improve the network by making it deeper and larger. Before training, the images are preprocessed with GAMMA correction and the appropriate number of anchor boxes is reset to provide better target detection performance in difficult environments such as shadows. In the experiments, IR-PANet which is running under Quadro P4000 obtained the achievement of $ap = 98.35\%$ in the vehicle class, improving the accuracy by 7.45% with a loss of 3.8 fps detection speed, outperforming other neural networks. However, the experimental results indicated that the number of FP (the number of actual negative classes predicted as positive classes) was higher for IR-PANet. This is an area for improvement in our future research. In addition, experiments have shown that the performance of IR-PANet is demonstrated in the ship class. By evaluating the optimization of the proposed network models for small objectives (cars and ships) in the context of two different complexes, this study is relevant to the research of aerial remote sensing image detection technology.

Data Availability

The data are available at https://github.com/CrazyStoneonRoad/TGRS-HRRSD-Dataset.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by grants from the National Basic Research Program of China (Grant no. 2019YFE0126600); the Major Project of Science and Technology of Henan Province (Grant no. 201400210300); the Key Scientific and Technological Project of Henan Province (Grant no. 212102210496); the Key Research and Promotion Projects of Henan Province (Grant nos. 212102210393, 202102110121, 202102210368, and 19210221009), and Kaifeng Science and Technology Development Plan (Grant no. 2002001).

References

[1] G. S. Xia, X. Bai, and J. Ding, "DOTA: a large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, Salt Lake City, UT, USA, December 2018.

[2] A. Voulodimos, "Deep learning for computer vision: a brief review," *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 7068349, 13 pages, 2018.

[3] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 1–11, 2016.

[4] C. Kamusoko, "Importance of remote sensing and land change modeling for urbanization studies," in *Urban Development in Asia and Africa*Springer, Singapore, 2017.

[5] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and P. Halvorsen, "Social media and satellites," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 2837–2875, 2019.

[6] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, no. 2, 2017.

[7] W. Liu, D. Anguelov, and D. Erhan, "SSD: Single shot multibox detector," in *Proceedings of the 2016 European Conference on Computer Vision*, pp. 21–37, Amsterdam, Netherlands, August 2016.

[8] T.-Y. Lin, P. Dollár, and R. Girshick, "Feature pyramid networks for object detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, November 2017.

[9] J. Redmon, S. Divvala, and R. Girshick, "You only look once: unified, real-time object detection,," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[10] S. Liu and D. Huang, "Receptivefield block net for accurate and fast object detection," in *Proceedings of the 2018 European Conference on Computer Vision*, pp. 385–400, Munich, Germany, September 2018.

[11] Z. Li, P. Chao, and Y. Gang, "Detnet: a backbone network for object detection," 2018, https://arxiv.org/abs/1804.06215.

[12] Z. Cai, Q. Fan, and R. S. Fe Ris, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proceedings of the 2016 European Conference on Computer Vision*, pp. 354–370, Amsterdam, Netherland, June 2016.

[13] C. Y. Fu, W. Liu, and A. Ranga, "Deconvolutional Single Shot Detector," 2017, https://arxiv.org/abs/1701.06659.

[14] L. Zheng, C. Fu, and Y. Zhao, "Extend the shallow part of single shot multibox detector via convolutional neural network," 2018, https://arxiv.org/abs/1801.05918.

[15] Z. Li and F. Zhou, "FSSD: feature fusion single shot multibox detector," 2017, https://arxiv.org/abs/1712.00960.

[16] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pp. 2980–2988, Venice, Italy, October 2017.

[17] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, July 2017.

[18] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018, https://arxiv.org/abs/1804.02767.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-Cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the 2015 Advances in Neural Information Processing Systems*, pp. 91–99, Montreal, Canada, December 2015.

[20] R. Girshick, J. Donahue, and T. Darrell, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[21] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pp. 1440–1448, Santiago, Chile, December 2015.

[22] J. Dai, Y. Li, and K. He, "R-Fcn: Object detection via region-based fully convolutional networks," in *Proceedings of the 2016 Advances in Neural Information Processing Systems*, pp. 379–387, Red Hook; NY, USA, December 2016.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-cnn,," in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pp. 2961–2969, Venice, Italy, October 2017.

[24] K. He, X. Zhang, and S. Ren, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[25] C.-Y. Wang, "CSPNet: a new backbone that can enhance learning capability of CNN," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580, Seattle, WA, USA, June 2020.

[26] S. Liu, L. Qi, and H. Qin, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759–8768, Salt Lake City, UT, USA, June 2018.

[27] M. Sandler, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June 2018.

[28] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 5535–5548, 2019.

[29] J. Sang, "An improved YOLOv2 for vehicle detection," *Sensors*, vol. 18, no. 12, 2018.

[30] X. Hu, "SINet: a scale-insensitive convolutional neural network for fast vehicle detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1010–1019, 2018.

[31] H. Nguyen, "Improving faster R-CNN framework for fast vehicle detection," *Mathematical Problems in Engineering*, vol. 2019, no. 3, 11 pages, Article ID 3808064, 2019.

[32] C. Xu, G. Wang, S. Yan et al., "Fast vehicle and pedestrian detection using improved Mask R-CNN," *Mathematical Problems in Engineering*, vol. 2020, Article ID 5761414, 15 pages, 2020.

[33] Z. Deng, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 3–22, 2018.

[34] X. Zhang, K. Zhu, and G. Chen, "Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network," *Remote Sensing*, vol. 11, no. 7, p. 755, 2019.

[35] A. V. Etten, "You only look twice: rapid multi-scale object detection in satellite imagery," 2018, https://arxiv.org/abs/1805.09512.

[36] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: a critical feature capturing network for arbitrary-oriented object detection in remote sensing images," https://arxiv.org/abs/2101.06849.

[37] P. Gao, T. Tian, L. Li, J. Ma, and J. Tian, "DE-CycleGAN: An object enhancement network for weak vehicle detection in satellite images," *In IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3403–3414, 2021.

[38] J. Zhang, X. Jia, and J. Hu, "Local region proposing for frame-based vehicle detection in satellite videos," *Remote Sensing*, vol. 11, no. 20, pp. 2072–4292, 2019.

[39] F. Shi, T. Zhang, and T. Zhang, "Orientation-aware vehicle detection in aerial images via an anchor-free object detection approach," *In IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5221–5233, 2021.

[40] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[41] D. Misra, "Mish: a self regularized nonmonotonic neural activation function," 2019, https://arxiv.org/abs/1908.08681.

[42] S. Kansal and R. K. Tripathi, "Adaptive gamma correction for contrast enhancement of remote sensing images," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 25241–25258, 2019.

[43] Z. Zheng, P. Wang, and W. Liu, "Distance-IoU loss: faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artifificial Intelligence*, February 2020.

[44] J. Yu, "UnitBox: an advanced object detection network," in *Proceedings of the 24th ACM International Conference on Multimedia*, New York, NY, USA, October 2016.

[45] X. Lu, Y. Zhang, and Y. Yuan, "Gated and axis-concentrated localization network for remote sensing object detection," *In IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 179–192, 2020.