remote sensing

MDPI

Article

Oil Well Detection under Occlusion in Remote Sensing Images Using the Improved YOLOv5 Model

Yu Zhang [1], Lu Bai [2], Zhibao Wang [1,3,*], Meng Fan [4], Anna Jurek-Loughrey [2], Yuqi Zhang [5], Ying Zhang [4], Man Zhao [6] and Liangfu Chen [4,7]

1 School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China
2 School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT9 6SB, UK
3 Bohai-Rim Energy Research Institute, Northeast Petroleum University, Qinhuangdao 066004, China
4 State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China
5 Department of Applied Statistics, College of Science, Purdue University, West Lafayette, IN 47907, USA
6 School of Communication and Electronic Engineering, Qiqihaer University, Qiqihaer 161003, China
7 University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wangzhibao@nepu.edu.cn

Abstract: Oil wells play an important role in the extraction of oil and gas, and their future potential extends beyond oil and gas exploitation to include the development of geothermal resources for sustainable power generation. Identifying and detecting oil wells are of paramount importance given the crucial role of oil well distribution in energy planning. In recent years, significant progress has been made in detecting single oil well objects, with recognition accuracy exceeding 90%. However, there are still remaining challenges, particularly with regard to small-scale objects, varying viewing angles, and complex occlusions within the domain of oil well detection. In this work, we created our own dataset, which included 722 images containing 3749 oil well objects in Daqing, Huatugou, Changqing oil field areas in China, and California in the USA. Within this dataset, 2165 objects were unoccluded, 617 were moderately occluded, and 967 objects were severely occluded. To address the challenges in detecting oil wells in complex occlusion scenarios, we propose the YOLOv5s-seg CAM NWD network for object detection and instance segmentation. The experimental results show that our proposed model outperforms YOLOv5 with F1 improvements of 5.4%, 11.6%, and 23.1% observed for unoccluded, moderately occluded, and severely occluded scenarios, respectively.

Keywords: oil well; object detection; instance segmentation; remote sensing; occlusion; YOLOv5

check for updates

Citation: Zhang, Y.; Bai, L.; Wang, Z.; Fan, M.; Jurek-Loughrey, A.; Zhang, Y.; Zhang, Y.; Zhao, M.; Chen, L. Oil Well Detection under Occlusion in Remote Sensing Images Using the Improved YOLOv5 Model. *Remote Sens.* **2023**, *15*, 5788. https://doi.org/10.3390/rs15245788

Academic Editors: Andrea Garzelli and Amin Beiranvand Pour

Received: 24 October 2023
Revised: 7 December 2023
Accepted: 12 December 2023
Published: 18 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background

Oil and gas are linked to the sustainable, stable, and prosperous development of the national economy and represent a vital component of people's livelihoods, according to the *BP Statistical Yearbook of World Energy 2022* [1]. At present, in response to industry challenges brought by energy conservation, emission reduction, and carbon reduction, it is very important to improve resource utilization efficiency and carry out the overall allocation of oil and gas resources to cope with emergency situations effectively. Oil wells serve as an important facility for the extraction of crude oil; the distribution and quantity of oil wells offer valuable insights into the status of oil and gas resources. By utilizing remote sensing monitoring, we can capture the dynamics of global oil energy exploitation and assess the production of oil in various regions.

With the development of deep learning methods and the emergence of high-resolution satellite remote sensing images, remote sensing object detection and instance segmentation have become one of the current research hotspots [2–4]. High-resolution remote sensing

images have richer detailed features of ground objects, spatial structures, and topological relationships, and are mostly used to detect different types of ground objects, such as aircraft, ships, roads, and buildings. Yu et al. [5] proposed a rotation-invariant method, a multi-scale rotation-invariant Hough forest with embedded scale factors, where orientation information is trained to cast rotation-invariant votes for estimating airplane centroids; this method is capable of performing well in accurately and correctly detecting arbitrarily orientated and varying-sized airplanes. In order to solve the limitations of traditional ship detection, such as complex application scenarios and intensive object detection, Yang et al. [6] proposed a framework called rotation dense feature pyramid networks (R-DFPNs) by building high-level semantic feature maps for all scales by means of dense connections, designing a rotation anchor strategy to predict the minimum circumscribed rectangle of the object. And they also proposed multi-scale ROI Align, which is more suitable for ship detection tasks. Zao et al. [7] used a richer U-Net model incorporating the detailed recovery of decoding networks through an enhanced detail recovery structure (EDRS). The implementation of the edge-focused loss function, which prioritizes pixels nearer to the edges, leads to increased precision in road detection outcomes. In view of the small distance between buildings, strong aggregation, and serious mutual occlusion, Han et al. [8] conducted three types of remote sensing image preprocessing. The training data were refined through a combination of threshold segmentation and fuzzy clustering techniques, which involved shadow removal and image enhancement through noise addition and flipping.

High-resolution remote sensing satellites provide the necessary basis for spatial data, attribute data, remote sensing data, and various other data types. They can be applied to the production and operation of oil and gas fields and the intelligent monitoring of remote sensing in oil and gas fields. They can enhance monitoring efficiency, reduce manual monitoring costs, contribute to environmental protection, and aid in informed decision making, leading to sustainable development and economic advantages. Currently, research in the field of oil-related studies utilizing remote sensing methods primarily focuses on oil spill detection and the identification of objects like oil tanks and well sites. However, there are very limited studies on oil well detection. Remote sensing images combined with deep learning can quickly monitor oil wells. However, in remote sensing images, image quality may be affected by factors such as weather, cloud cover, sensor resolution, shooting angle, and shadow, which make it difficult to extract features of small-scale objects like oil wells. In addition, the terrain significantly impacts the accuracy of oil well detection, leading to a high false-alarm rate. Potential obstructions like buildings and trees surrounding the oil wells can also lead to instances of missed oil well detection.

While the object interpretation of remote sensing images is becoming more and more obvious, occlusion detection has always been a difficult point in computer vision. Yu et al. [9] proposed a real-time face detector based on the one-stage detector YOLOv5, named YOLO-FaceV2. The attention network SEAM block and repulsion loss were used to solve the problem of face occlusion. Du et al. [10] proposed the FA-YOLO, which significantly improved detection efficiency on 318 infrared occluded vehicle images from the VIVID—infrared dataset. Since the shape and texture information of an object is seriously affected by occlusion, it is difficult to detect an oil well effectively under occlusion in an actual real-world scenario. Therefore, enhancing oil well detection in remote sensing images is important.

This paper studies automatic oil well detection methods based on deep learning in remote sensing images under occlusion. Currently, mainstream object detection methods can be divided into two categories, namely two-stage and one-stage detection algorithms. With the development of deep learning, one-stage methods have gradually become mainstream as they do not require the generation of proposal boxes and has higher efficiency. In this work, we adopt the YOLOv5 as the detection model. YOLOv5 provides four model networks: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. The network YOLOv5s, which has the smallest depth and the smallest width of feature map, is used. Given the absence of certain features in occluded objects, object detection algorithms can be affected by these

missing characteristics. Thus, we employ the YOLOv5s-seg algorithm, which takes into account pixel-wise classification, for oil well object detection and instance segmentation. Compared with YOLOv5, the use of the YOLOv5s-seg algorithm results in an increase of 6.6% in average precision (AP) and a 6.23% boost in F1 score. In order to further improve the detection accuracy of occluded oil wells, YOLOv5s-seg CAM NWD is used in combination with the CONTEXT_AUGMENTATION_MODULE (CAM) method and normalized Wasserstein distance (NWD). Compared with that of the YOLOv5 model, the AP and F1 are improved by 11.7% and 10.72%, respectively. In scenarios without occlusion, with moderate occlusion and under severe occlusion, F1 score shows significant improvements of 5.4%, 11.6% and 23.1%, respectively.

The main contributions of this paper are as follows:

1. We construct the first segmentation dataset of oil wells in remote sensing images, with many occluded objects. The dataset can be used as a reference for evaluating remote sensing image instance segmentation under occlusion.
2. We propose combining the CONTEXT_AUGMENTATION_MODULE and Normalized Weighted Distance methods and demonstrate that it improves the accuracy of oil well detection under different occlusion scenarios.

1.2. Related Work

1.2.1. Research on Object Detection and Instance Segmentation Algorithms Based on Deep Learning

Object detection is one of the core tasks of computer vision, whose purpose is to obtain the location and category of an object. Convolutional neural networks are one of the most used methods for object detection due to their ability to extract and learn features from image data effectively. Addressing the difficulty of locating small objects in faster regions, a convolutional neural network-based method for multi-class objects in remote sensing images with large scale changes was proposed by Deng et al. [11]. Li et al. [12] introduced a feature attention object detection framework, which uses channel and pixel attention to enhance object-related representation and reduce background information when fusing multi-scale visual features of a backbone network. Xiao et al. [13] proposed a feature pyramid network, which combines context enhancement and feature refinement to supplement context information and prevent small objects from being overwhelmed by conflicting information. Sun et al. [14] propose a part-based convolutional neural network (PBNet) for the detection of composite objects in remote sensing images, such as sewage treatment plants, golf courses, airports, and other objects with neither a fixed size nor fixed shape. Mahmoud et al. [15] used the adaptive mask Region-based Convolutional Network (mask-RCNN) to detect various classes of objects in remote sensing images and overcome the problems of object scale change, small size, large density, and small amount of annotation in remote sensing images.

Instance segmentation is a technology that simultaneously solves the problem of object detection and semantic segmentation [16], achieves classification at the pixel level, locates different instances, and can obtain very rich and refined object information. Instance segmentation methods are divided into two types: two-stage and one-stage. Two-stage instance segmentation is divided into top–down methods based on object detection and bottom–up methods based on semantic segmentation. One-stage instance segmentation can expel the restrictions within the detection frame, which is the future research trend. In order to solve the challenges of scale changes and low contrast in remote sensing images, Liu Y et al. [17] proposed a context aggregation network (CATNet) by incorporating the dense feature pyramid network (DenseFPN) in the feature domain, the Spatial Context Pyramid (SCP) in the spatial domain and the Hierarchical Region of Interest Extractor (HRoIE) in the instance domain to aggregate the global visual context, respectively. Lin et al. [18] proposed a face detection and segmentation method based on improved Mask R-CNN, called G-Mask, which integrates face detection and segmentation into a framework to obtain more granular face information, and can segment each face from the background image.

1.2.2. Research on Oil and Gas and Remote Sensing

Oil plays an important role in advancing modern society. At present, research in the domain of oil-related studies employing remote sensing techniques predominantly emphasizes oil spill detection and the identification of objects such as oil tanks and well sites. Wang et al. [19] proposed a BO-DRNET model for oil spill detection on SAR images, with ResNet-18 as the backbone network of DeepLabv3+ and Bayesian optimization (BO) to optimize model hyperparameters. Zhu et al. [20] used the SSD algorithm combined with the Hough transform to identify the circular features at the top of industrial storage tanks in order to reduce false alarms for the detection of industrial storage tanks at the city level. Wu et al. [21] proposed a YOLOX-TR network in order to solve the problem of dense oil tank detection due to overlapping, contour blurriness and geometric distortion. A transformer encoder can obtain the area of interest of the oil tank and enhance the feature representation. RepVGG can reparametrize the multi-branch trunk to improve the classification accuracy. He et al. [22] proposed oil well site extraction using the OWS Mask R-CNN model by adding a semantic segmentation branch to Mask R-CNN to make the whole network focus on the relationship between the route objects near the well sites and the well sites. Considering that the number and geographical location of oil wells are important for energy resource planning, Wang et al. [23–25] built the first oil well object detection dataset and obtained high accuracy using the most advanced deep learning models. YOLOv4 with sliding slice and discarded edge was proposed by Shi et al. [26], which effectively solves the problem of repeated detection and inaccurate positioning in large-scale and high-resolution oil well detection.

1.2.3. Transfer Learning

In deep learning, the method of taking a pre-trained model as the starting point of a new model from the perspective of similarity is called transfer learning [27]. Yosinski et al. [28] experimentally quantified the generality versus specificity of neurons in each layer of a deep convolutional neural network. It was found that even features transferred from distant tasks are better than random weights and that this is a universally useful technique for improving the performance of deep neural networks even if significant fine-tuning is required on a new task. Ruan et al. [29] used the disturbance label information in a large-scale face database for transfer learning and were able to effectively extract multiple disturbing factors from facial expression images. Ma et al. [30] introduced the label transfer learning paradigm to decouple known and unknown features, promote unknown learning, and adjust the learning process through other strategies to achieve a balance between unknown learning and known learning. One of the challenges that needs to be considered when using transfer learning is to avoid negative transfer. It is necessary to pay attention to whether the relationship between the original task and the target task is close, and whether the transfer method can make good use of the relationship between the tasks, to achieve some improvement in the target task. Song and Yang [31] proposed a GSCCTL model based on clustering and transfer learning, which was tested on UCMerced, AID, and NWPU-RESISC45 remote sensing datasets. It was found to be suitable for semi-supervised scene classification. Alem and Kumar [32] proposed the transfer learning (TL) method, which is widely used for land cover or land use (LCLU) classification in remote sensing images.

2. Methods

2.1. The Network Structure of YOLOv5s-Seg

YOLOv5 is mainly composed of an input terminal, backbone, neck, and prediction components. An adaptive anchor box calculation and adaptive picture scaling method was adopted. The preset anchor box scale was inputted into the network, and the obtained predicted bounding box was compared with the ground truth bounding box to update the network parameters. Mosaic data enhancement was used to splice images according to random scaling, random cropping, and random arrangement, which greatly enriched the background of objects to be detected. The focus structure was used to improve computing

power without losing information. The backbone was composed of CBL, CSP and SPPF modules. Some improvements were made to the structure of FPN+PAN, using the CSP2_X structure to strengthen the capability of network feature fusion. The YOLOv5s-seg model introduces instance segmentation within object detection. In order to improve the performance of the model in the task of oil well object detection, we adopted a model-training strategy based on transfer learning. We used the officially provided weight file YOLOv5s.pt pre-trained on the COCO dataset as the starting point for model initialization. The COCO (Microsoft Common Objects in Context) object detection dataset [33] is a large-scale dataset with rich context information, multi-tasking, diversity, and high-quality annotations, which is widely used in various tasks in computer vision research, especially for object detection and image understanding tasks. In the process of transfer learning, we first loaded the pre-training weights into the YOLOv5s-seg network. The purpose of this was to use the general image features learned on the COCO dataset to speed up the model's learning of the feature representation related to oil wells. Meanwhile, using the pre-training weights on the large-scale dataset could help our model resist the overfitting of the network and improve the generalization of the model. The transfer learning pipeline and the network structure are shown in Figures 1 and 2, respectively.

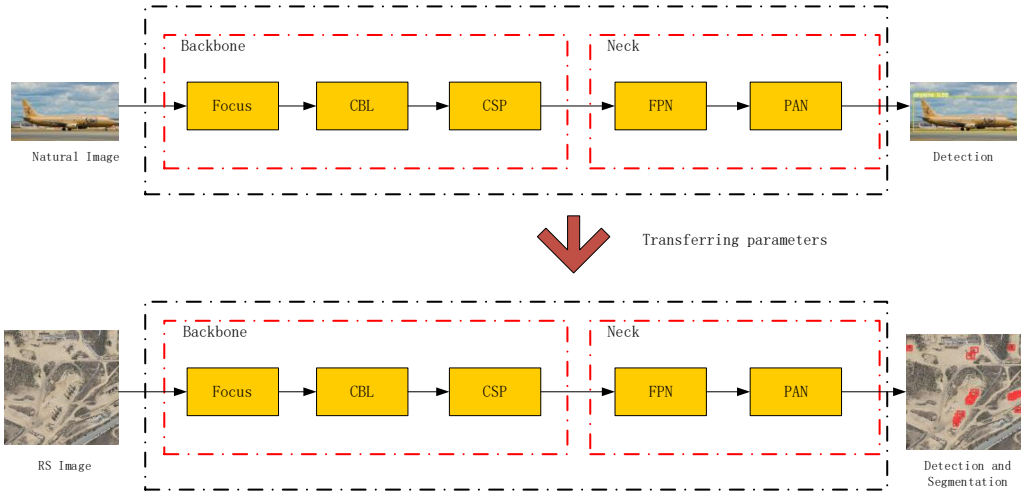


Figure 1. Transfer learning pipeline structure.

2.2. CONTEXT AUGMENTATION MODULE Structure

Rich semantic information from the surrounding context, such as the environment around the oil well, plays a significant role in object analysis. It enables the detector to gain a deeper understanding of the object's context, thereby enhancing the oil well recognition capability. The proposed context module by Yu and Koltun [34] proposed a context module, which is a network module that uses extended convolution to aggregate multi-scale context information without the loss of resolution and contributes to dense prediction. In order to avoid the loss of spatial detail and positioning accuracy, a high-resolution network was introduced by Zhang et al. [35], and local context was aggregated by introducing adaptive spatial pooling. Compared with the baseline HRNet, the proposed architecture has an advantage of 0.47% in OA and 0.59% in the average F1 score on the Potsdam dataset and 0.67% in OA and 0.96% in the average F1 score on the Vaihingen dataset. Therefore, in the task of oil well object detection, which has a rich semantic environment of surrounding context, the CONTEXT_AUGMENTATION_MODULE [13] enables the detection model to understand the context of the object more deeply, further improving the ability to identify partially occluded oil wells. A CAM module is directly connected in series behind the backbone of the YOLOv5s-seg model. The CAM module uses different expansion convolution rates of 1, 3 and 5, respectively, to obtain the context information of different receptor fields, thus achieving the purpose of enriching the context information of FPN. These spatial features are convolved and fused with three different outward expansions as

inputs to the neck part of the YOLOv5s-seg network. There are three strategies for feature fusion, namely weighted fusion, concatenation fusion, and adaptive fusion. Adaptive fusion is more suitable for large and medium objects, and concatenation fusion has the greatest advantage for small objects. In this work, concatenation fusion was adopted to directly add feature maps into space and channel dimensions, therefore enriching the overall feature representation. The CAM fusion structure is shown in Figure 3.

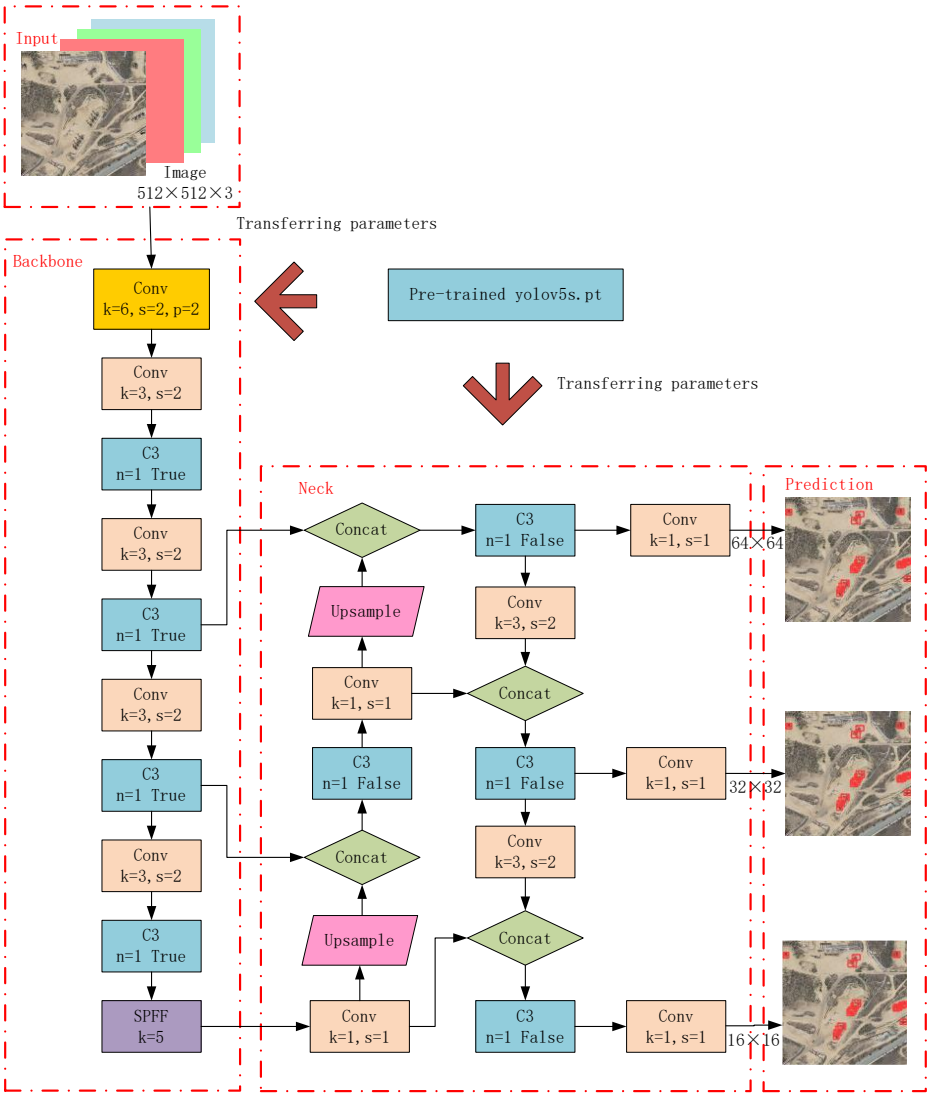


Figure 2. YOLOv5s-seg structure.

2.3. Normalized Weighted Distance

Small objects exist in large numbers in real-world scenarios; in terms of pixel size, the COCO dataset defines small targets as resolutions less than 32 × 32 pixels. In terms of relative area, defined as the median ratio of the bounding box area to the image area, it is between 0.08% and 0.58% [36]. Oil wells, being relatively small objects, provide minimal information, particularly in cases of occlusion, which poses challenges for the network in learning distinctive features and can lead to detection errors. As the sensitivity of IoU to objects of different scales varies greatly, as shown in Figure 4, for oil wells, a small position deviation will lead to huge changes in IoU. Therefore, Normalized Weighted Distance (NWD) was used to measure the similarity, and it is worth noting that NWD is not sensitive to scale [37]. Since there are always some background pixels in the bounding box, the foreground pixels are concentrated in the middle. When calculating the distance between the prediction box and the object box, the bounding box was modeled as a 2D Gaussian

distribution. The similarity between the prediction box and the object box is the scalar value obtained by the normalization of the weighted summation distance of the Gaussian distribution, which is used as the loss function of the object detection. The distribution can be measured regardless of whether there is overlap between the two surrounding boxes. W_2^2 is the distance between two bounding boxes. Suppose the two boundary boxes are (cx_a, cy_a, w_a, h_a) and (cx_b, cy_b, w_b, h_b), and the formula W_2^2 is defined by Equation (1). The normalized distance formula is defined by Equation (2), Na, Nb is the Gaussian distribution modeled by bounding boxes A and B, and C is a constant related to the dataset. The loss of NWD is calculated as per Equation (3):

$$W_2^2(Na, Nb) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2} \right]^T \right) \right\|_2^2, \tag{1}$$

$$NWD(Na, Nb) = \exp\left(-\frac{\sqrt{W_2^2(Na, Nb)}}{C} \right), \tag{2}$$

$$L_{NWD} = 1 - NWD(Na, Nb) \tag{3}$$

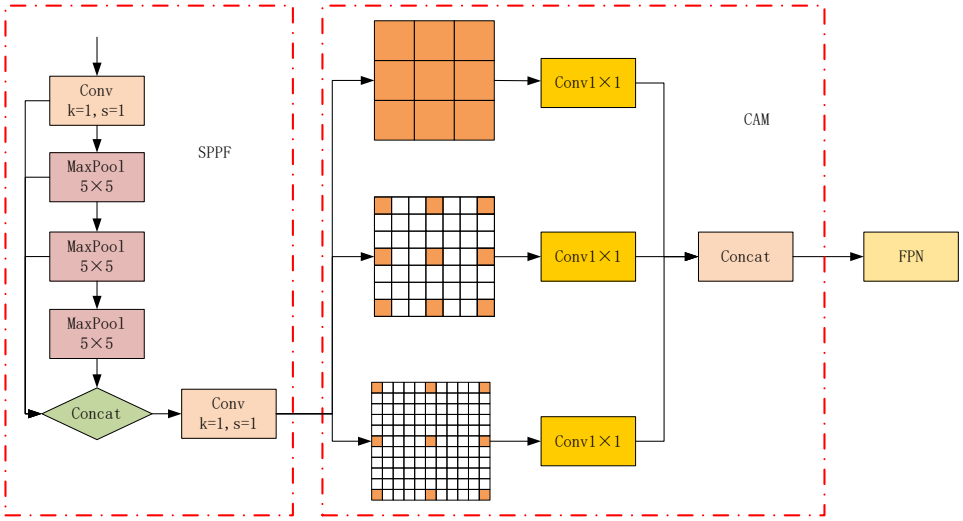


Figure 3. CAM fusion structure.

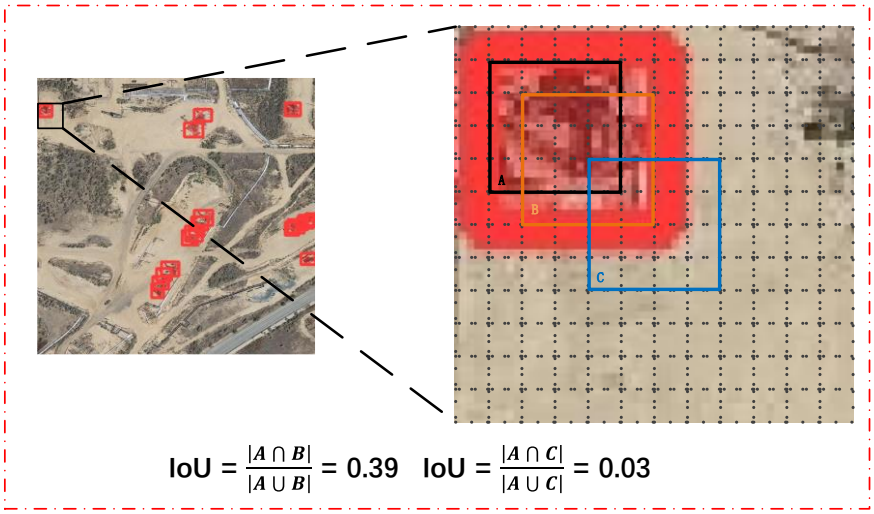


Figure 4. The sensitivity of IoU. (A is the black bounding box, B is the orange bounding box, and C is the blue bounding box).

2.4. Other Modules

The C2f module (Figure 5) in YOLOv8 extracts the features in the image through multiple convolution layers and pooling layers and uses the upper sampling layer to increase the resolution of the feature map. This module is used to replace the C3 structure (Figure 6) in YOLOv5. Due to more residual connections, it has richer gradient flow, thus improving the accuracy and speed of detection.

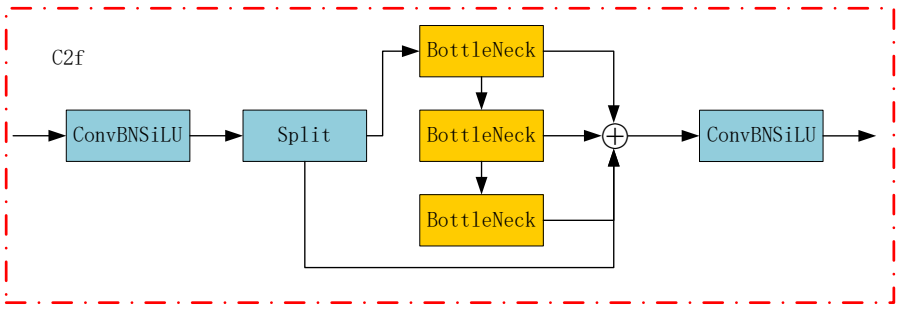


Figure 5. C2f structure.

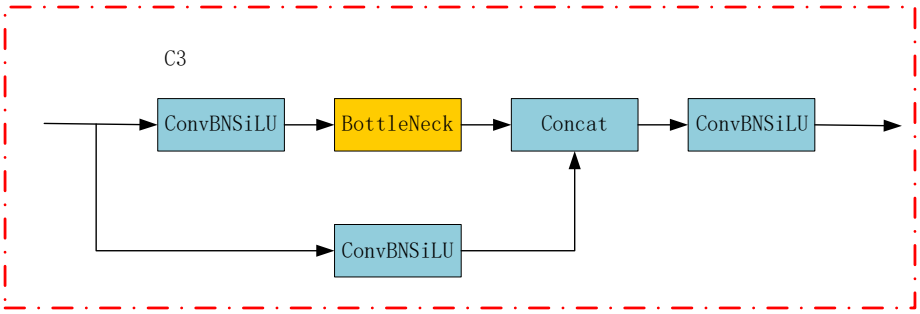


Figure 6. C3 structure.

The ConvNeXt V2 model adds a global response normalization layer (GRN) to enhance the feature competition between channels through global feature aggregation [38], feature normalization, and feature calibration, increasing the contrast and selectivity between channels. Compared with a CNN, a transformer has the advantage of being able to capture context dependencies over long distances with its self-attention mechanism. BiFormer's new universal vision network architecture can focus on a small number of relevant markers in a query-adaptive way without distracting the attention of other irrelevant markers [39]. It has three attention mechanisms: Bi-Level Routing Attention, Attention and AttentionLePE. Swin Transformer not only uses a hierarchical structure (pyramid structure) [40] but also proposes a linear complexity attention calculation, and finally obtains several different sizes of structures. PoolFormer is a structure that fuses information among multiple tokens through average pooling [41]. The RFE module takes full advantage of the receptive field in the feature graph by using extended convolution [9], shared weights across ranches, reducing the number of parameters and the risk of potential overfitting. The EfficientRepGFPN module can use different channel numbers for different scale features [42], flexibly control the expression ability of high-level features and low-level features, and fully exchange high-level semantic information and low-level spatial information.

3. Experimental Results

3.1. Dataset

In this study, we collected our dataset from various oil fields, each of which holds unique significance: Changqing oil field, located in the Ordos Basin, has set the highest annual output of oil and gas in China. The complex terrain such as hills and mountains contributes to the ruggedness of the ground, resulting in changes in the shape, size, direction,

and other features of the oil well. Additionally, this terrain complexity introduces background noise surrounding the oil wells, further complicating the extraction and recognition of the oil well features (see Figure 7a). Huatugou, located within the vast Gobi and the colorful Danxia landform, has the world's highest elevation oil wells, with unique rocks, sand, gullies, and other terrain and large changes in lighting conditions; this may cause some interference to the interpretation and analysis of remote sensing images and reduce the recognition degree of oil well features. In addition, there are many double donkey head oil wells, providing the possibility of diversity of oil wells (see Figure 7b). Daqing oil field, a world-renowned large sandstone oil field, proves that continental strata can generate oil and form large oil fields. The background here is mainly green space and buildings (see Figure 7c). California is a large economic region in the United States and also ranks as one of the nation's most abundant sources of oil reserves. The oil wells here also have their own characteristics in the remote sensing image, not only in shape, but also in direction; many of the oil wells show a distinctive white base (see Figure 7d).

Figure 7. Sample images from our dataset. (**a**) Sample image from Changqing oil field. (**b**) Sample image from Huatugou. (**c**) Sample image from Daqing. (**d**) Sample image from California.

The dataset was derived from Google Earth Imagery [43], which combines a large number of high-definition satellite and aerial images to ensure clarity and lack of cloud cover, but also causes misalignment and distortion due to the variety of image quality. The dataset contains 376 images of Daqing City with 1744 oil well objects, 125 images of Changqing oil field with 598 oil well objects, 91 images of Huatugou with 379 oil well objects, and 130 images of California with 1028 oil well objects, a total of 722 images with

3749 oil well objects. The images in the dataset are 512 × 512 pixels with 0.48 m spatial resolution per pixel.

In this paper, occlusion within 30% was considered as moderate occlusion, and occlusion above 30% was considered as severe occlusion. The dataset format is the Pascal VOC format and YOLO label format. The Pascal VOC format has an image, xml tag, category image tag, and instance image tag (see Figure 8). The YOLO format has images and txt labels. As shown in Figure 9, the remote sensing image containing the oil well object area is divided into 512 × 512-pixel image blocks. The ArcGIS tool was used to mark the image and convert it into a shpfile format file, and the shpfile file was converted into xml format through the GDAL library in python. In xml tags, the size tag stores the size of the image, the object tag stores each object's information, and the bndbox tag stores the object location information. X, Y coordinate information in the lower left and upper right corner of the prior box was used as object attributes, and background and shadow information were used as object attributes. For example, shadow stores whether the object has a shadow. The geo_trans tag stores the six-parameter coordinate conversion model in the GeoTIFF data storage format, including the top-left pixel center X,Y coordinates, pixel resolution in the X,Y direction, and rotation information for the conversion of pixel coordinates (p_x, p_y) and geographic coordinates (geo_x, geo_y). The equations for calculating geographic coordinates are as follows (see Equations (4) and (5)).

$$geo\_x = geo\_trans[0] + geo\_trans[1] * p\_x, \tag{4}$$

$$geo\_y = geo\_trans[3] + geo\_trans[5] * p\_y \tag{5}$$

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

Figure 8. Sample images from our dataset. (**a**) Image; (**b**) classification label; (**c**) instance label.

The oil well is a beam pumping unit, which is composed of a donkey head, beam, and power equipment. In our dataset, we took the donkey head of the oil well as the direction and divided the angle into eight directions, as seen in Figure 10a,b. Occlusion types were divided into six categories, including (1) single well with no occlusion, (2) dense wells with no occlusion, (3) dense wells with occlusion, (4) background occlusion, (5) self-occlusion, and (6) slice occlusion. Single well with no occlusion means that there is no possibility of overlapping prediction boxes. Dense wells with no occlusion mean that ground truth boxes do not overlap but the prediction boxes may be overlap, or the ground truth boxes overlap but the oil wells do not overlap with each other (Figure 10c). Dense wells with occlusion mean that the oil wells overlap with each other (Figure 10d). Background occlusion refers to obstructions caused by the background, while self-occlusion occurs when the oil well's head is partially obscured due to angle-related issues. Self-occlusion becomes particularly evident when prominent features such as the base and floating beam align in a linear fashion (Figure 10e). When slice occlusion is a necessary step for the small object detection of large-scale remote sensing images, the object is segmented (Figure 10f). The object background types were divided into four types, which are lakes, bare land, trees, and

buildings. The degree of occlusion was divided into unoccluded, moderately occluded, and severely occluded scenarios (Figure 10g). For moderately occluded scenarios, the degree of occlusion was less than 30%, and for serious occlusion, it was more than 30%.

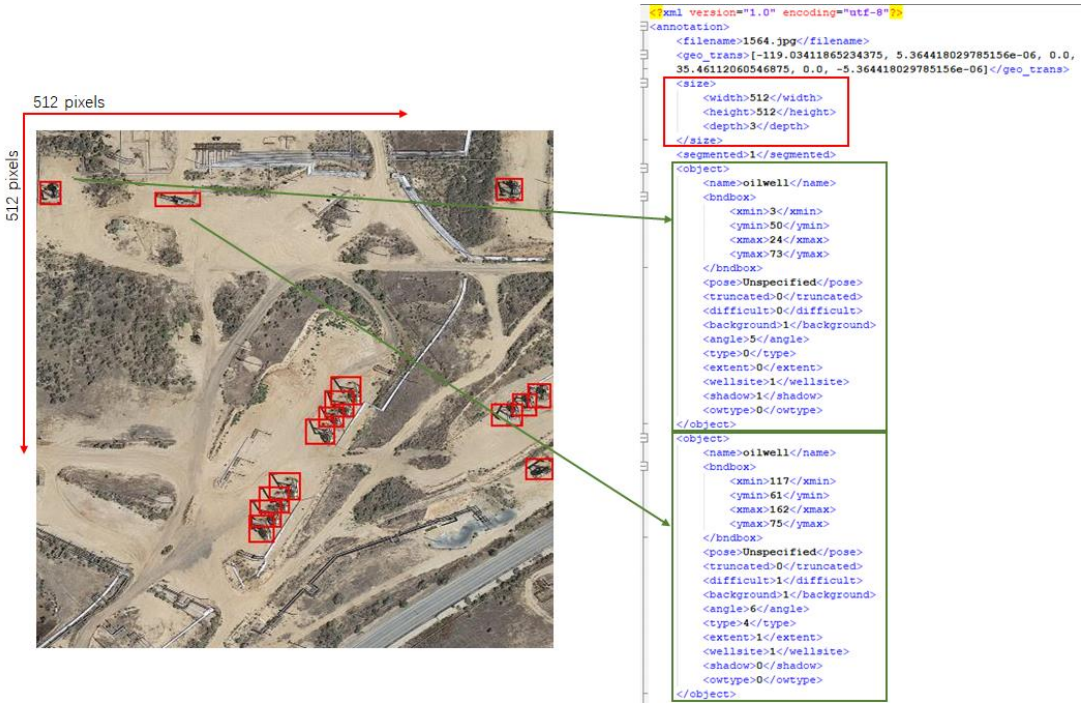


Figure 9. XML annotation details. (the red box is the object detection bounding box annotation).

3.2. Model Training and Improvement

In our experimental evaluation, we utilized the Daqing oil field area, Changqing oil field area, and Huatugou area in China as the training data and validation data, and California was used as the test data. In this paper, the test set for the model consisted of California oil wells situated in a region distinct from the training set. As the landform and obstructed oil well features of this region were not part of the model's training process, it served as a robust validation of our model's capability to detect obstructed oil wells amidst varying geomorphic backgrounds of different complexities. This test set also evaluated the generalization performance of our model across diverse geomorphic backgrounds, offering a preliminary validation for potential extensions of our model to other regions worldwide in the future. As shown in Figure 11, a total of 1637 objects in the domestic area were unoccluded, 394 were moderately occluded, and 690 were severely occluded. In the California dataset, 528 objects were unoccluded, 223 were moderately occluded, and 277 were severely occluded. As shown in Figure 12, in the training and validation set of our dataset, that is, the China region, there were a total of 1042 single wells that were unoccluded, 599 that were dense unoccluded, 474 that were dense occluded, 48 that were background occluded, 130 that were self-occluded, 90 that were slice occluded, and 338 that were multi-class occluded, most of which were dense unoccluded and self-occluded. In the California dataset, 455 single wells were unoccluded, 74 were dense unoccluded, 94 were dense occluded, 46 were background occluded, 241 were self-occluded, 62 slice were occluded, and 56 were multi-class occluded.

We conducted experiments using the YOLOv5 model, YOLOv5 instance segmentation model (YOLOv5s-seg), Faster R-CNN model, Mask R-CNN model, YOLOv7 model, and YOLOv8 model. In YOLOv5, we utilized the pre-trained model's weight YOLOv5s.pt, which was trained on the COCO dataset provided by the official source. For Faster R-CNN, Mask R-CNN, and other similar models, corresponding pre-trained model weights like mask_rcnn_coco.pth are available. The control of the number of layers and channels in the

model was facilitated by utilizing the parameters depth_multiple 0.33 and width_multiple 0.5. The backbone network employs five convolution subsampling processes to extract images of 512 × 512 size. This process results in three distinct feature map sizes: 16 × 16, 32 × 32, and 64 × 64, respectively. YOLOv5 loss is comprised of three components: BCE loss to gauge the prediction accuracy for classes and objectness, and CIoU loss for location. These adjustments aimed to enhance detection performance, especially for occluded and overlapping objects.

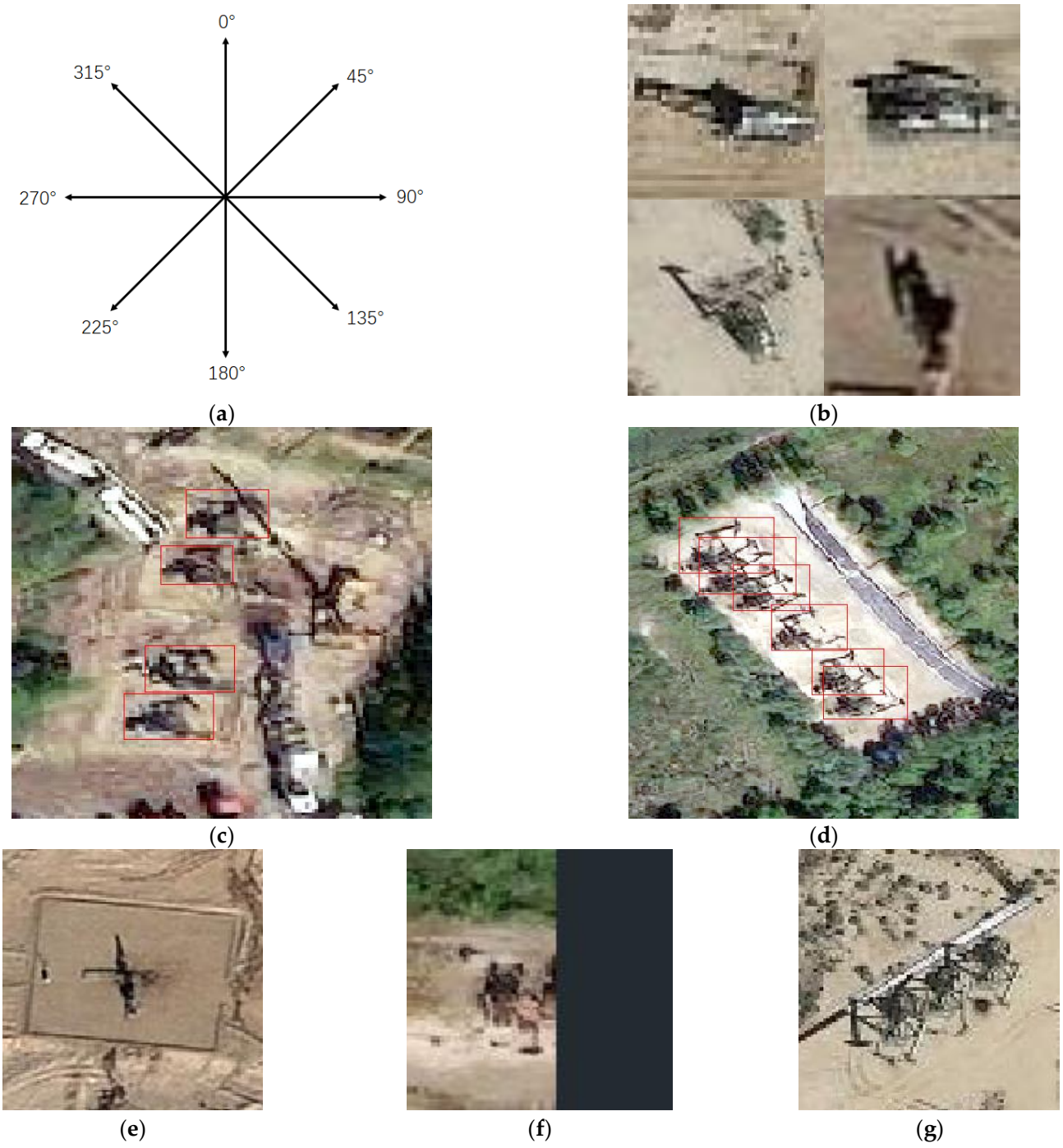


Figure 10. Sample images. (**a**) eight directions. (**b**) sample objects of different directions. (**c**) Dense wells with no occlusion. (**d**) Dense wells with occlusion. (**e**) Self-occlusion. (**f**) slice occlusion. (**g**) different degree of occlusion.

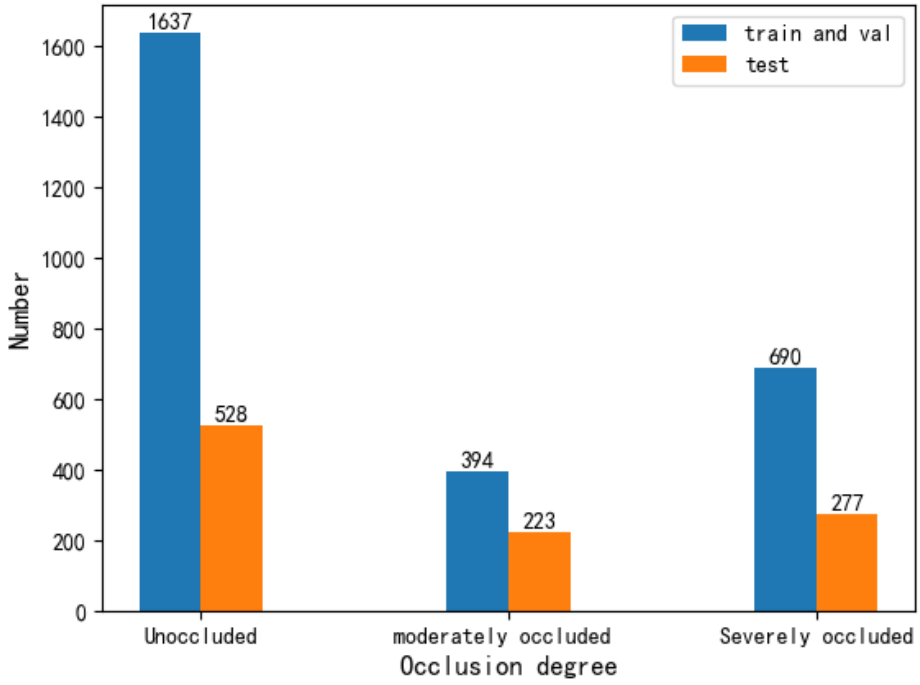


Figure 11. Number of different occlusion degrees.

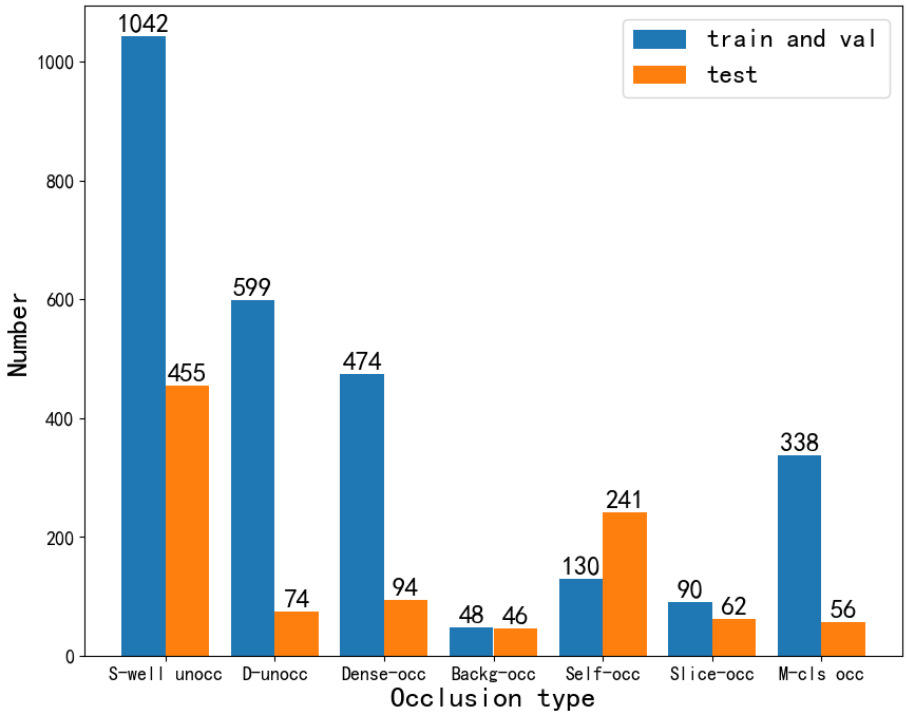


Figure 12. Number of different occlusion types.

In the comparison experiment involving the six general models, YOLOv5s-seg was chosen as the baseline model, and subsequent improvements were implemented. This involved integrating feature modules. The CAM module demonstrated promising performance. Within the YOLOv5s-seg backbone network, a series of five convolutions were applied, and the resulting multi-scale extended convolution features were fused. These fused features were then injected into the feature pyramid network from top to bottom, augmenting context information. Given the heightened sensitivity of occluded oil wells' small objects to position deviation in the Intersection over Union (IOU)-based measurement,

the NWD metric was introduced into the loss function calculation. Specifically, NWD was utilized in the computation of objectness loss and location loss to gauge distance, and an instance segmentation loss was incorporated. The final loss function was obtained through the amalgamation of multiple losses, contributing to the model's improved performance.

3.3. Evaluation Metrics

To evaluate the performance of the model, we selected four metrics in the experiment: precision, recall, F1 score, and average accuracy. These metrics are defined in Equations (6)–(9) below.

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

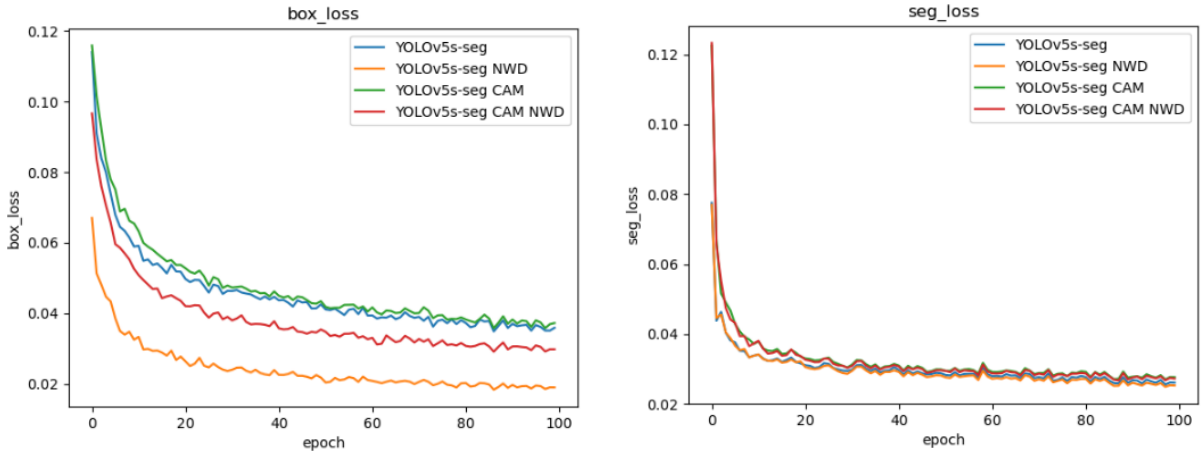
$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

$$AP = \int_0^1 p(r)dr \tag{9}$$

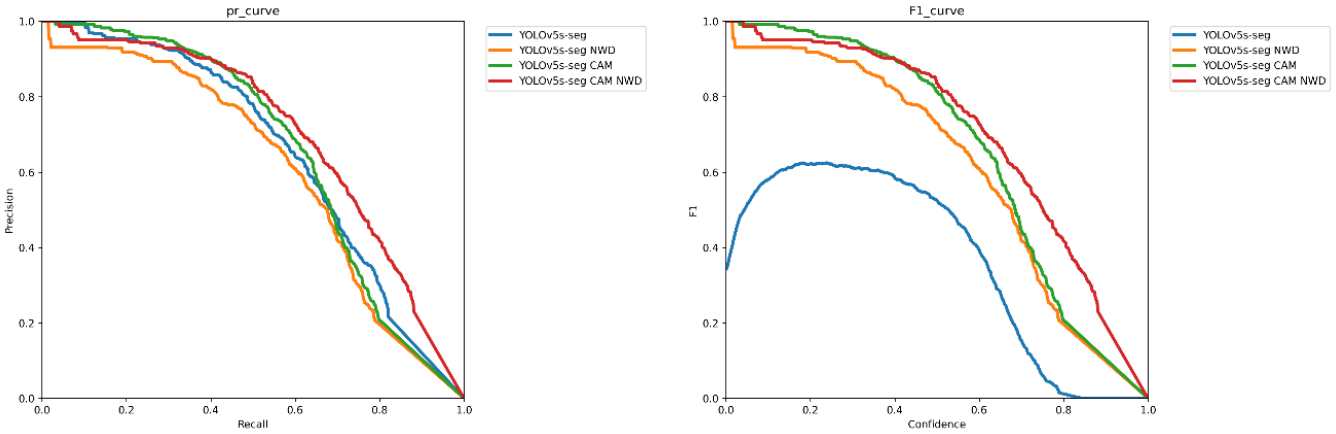
3.4. Experimental Results

The experiment was conducted on a server with an Intel i7-11700 CPU (2.50 GHz) manufactured by Intel Corporation and an NVIDIA GeForce RTX 2060ti GPU (11,264 m) manufactured by NVIDIA Corporation, Both companies are headquartered in Santa Clara, California, USA. Figure 13a,b show the box_loss and seg_loss curves of different models considered in our experimental evaluation. It can be seen that the loss value of the model using NWD metric was the smallest and the closest to the real value. The model was able to detect all oil wells to a large extent; however, because the NWD metric is sensitive to the calculation of similarity between small oil well object boundary boxes under occlusion, the model may produce a high false-alarm rate when dealing with these objects. The CAM module can effectively control the false-alarm rate, which can enhance the model's attention to the object region and suppress the interference in the background region. By integrating the CAM module, the model can better focus on the details of the oil well object, thus reducing the false-alarm rate. The curves in Figure 14a,b intersect, yet when considering the area under the curves, it is evident that the YOLOv5s-seg CAM NWD model exhibited superior performance.



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

Figure 13. Loss curve analysis. (**a**) Box_loss curve. (**b**) Seg_loss curve.



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

Figure 14. Pr_curve and F1_curve analysis. (**a**) Pr_curve. (**b**) F1_curve.

Table 1 shows the precision, recall, and F1 scores achieved by the YOLOv5s-seg, YOLOv5s-seg CAM, YOLOv5s-seg NWD, and YOLOv5s-seg CAM NWD models across the entire dataset, which was randomly divided into an 8:1:1 proportion for testing. Notably, the YOLOv5s-seg CAM NWD model exhibited the best detection accuracy, attaining a remarkable 93.2% F1 score.

Table 1. Evaluation metrics of randomly divided dataset.

Model	P	R	AP50	F1
YOLOv5s-seg	0.907	0.858	0.923	0.882
CAM	0.922	0.894	0.943	0.908
NWD	0.920	0.895	0.949	0.907
CAM NWD	0.933	0.932	0.965	0.932

To further investigate the generalization ability of the model, we performed an additional experiment where the Daqing oil field area, Changqing oil field area, and Huatugou area were used as training sets and verification sets, and California was used as a test set. Table 2 shows the precision, recall, and F1 scores of the baseline models Faster R-CNN, Mask R-CNN, YOLOv5, YOLOv7, YOLOv8, and the YOLOv5 instance segmentation and YOLOv5s-seg. It can be seen that YOLOv5s-seg achieved the best performance. Table 3 shows that the YOLOv5s-seg model, Swin Transformer model, PoolFormer model, convnextv2 model, C2f module, and RFEM model tried to use the backbone. Some attention mechanisms such as BiLevelRoutingAttention, attention, and AttentionLePE were added to the neck, EfficientRepGFPN model, and CAM module, and the NWD metric was used as the comparative experimental result of loss function when calculating loss. Table 4 shows the precision, recall, and F1 scores of the NWD and CAM models with the highest F1 in the training, YOLOv5, YOLOv5s-seg, and YOLOv5s-seg CAM NWD models under different occlusion degrees. Table 5 shows the precision, recall, and F1 scores of each model under different occlusion types.

The results show that the YOLOv5s-seg CAM NWD model can achieve the best F1 score among all the models. The reason why the F1 of the YOLOv5s-seg CAM NWD model is poor in the slice occlusion of different occlusion types is that the number of objects blocked by slice is too small, and when calculating false detection, the falsely detected objects have no occlusion type, so they were added to each occlusion type, resulting in errors.

Table 2. Baseline model evaluation metrics.

Model	P	R	AP50	F1
faster_rcnn	**0.873**	0.201	0.375	0.327
mask_rcnn	0.538	0.457	0.558	0.494
YOLOv5	0.769	0.487	0.584	0.596
YOLOv7	0.568	0.443	0.461	0.498
YOLOv8	0.635	0.513	0.542	0.567
YOLOv5s-seg	0.741	**0.593**	**0.65**	**0.659**

Bold is the optimal model for each column.

Table 3. Improved model evaluation metrics.

Improvement	AP50(Box)	F1	AP50(Mask)	F1
YOLOv5s-seg	0.65	0.659	0.634	0.61
BiLevelRoutingAttention	0.496	0.533	0.484	0.49
Attention	0.558	0.611	0.54	0.546
AttentionLePE	0.593	0.625	0.583	0.574
C2f	0.499	0.575	0.487	0.506
NWD	0.611	0.652	0.62	0.615
Convnextv2	0.503	0.634	0.504	0.541
SwinTransformer	0.447	0.533	0.433	0.465
PoolFormer	0.462	0.55	0.459	0.503
CAM	0.659	0.69	0.626	0.626
EfficientRepGFPN	0.652	0.613	0.621	0.624
RFEM	0.593	0.623	0.577	0.586
CAM NWD	**0.701**	**0.704**	**0.686**	**0.642**

Bold is the optimal model for each column.

Table 4. Occlusion degree evaluation metrics.

	YOLOv5	YOLOv5s-Seg	YOLOv5s-Seg NWD	YOLOv5s-Seg CAM	YOLOv5s-Seg CAM NWD
Unoccluded P	**0.822**	0.774	0.617	0.812	0.691
Unoccluded R	0.612	0.68	0.803	0.693	**0.831**
Unoccluded F1	0.701	0.724	0.698	0.748	**0.755**
Moderately occluded P	0.698	0.66	0.517	**0.769**	0.578
Moderately occluded R	0.395	0.48	**0.7**	0.493	0.668
Moderately occluded F1	0.504	0.556	0.594	0.601	**0.62**
Severely occluded P	0.682	0.731	0.571	**0.788**	0.645
Severely occluded R	0.325	0.52	0.664	0.534	**0.7**
Severely occluded F1	0.44	0.608	0.614	0.639	**0.671**

Bold is the optimal model for each line.

Table 5. Occlusion type evaluation metrics.

	YOLOv5	YOLOv5s-Seg	YOLOv5s-Seg NWD	YOLOv5s-Seg CAM	YOLOv5s-Seg CAM NWD
Unoccluded P	**0.822**	0.783	0.627	0.816	0.703
Unoccluded R	0.637	0.705	0.835	0.712	**0.857**
Unoccluded F1	0.718	0.742	0.716	0.761	**0.772**
Dense unoccluded P	**0.829**	0.704	0.549	0.782	0.617
Dense unoccluded R	0.459	0.514	0.608	0.581	**0.676**
Dense unoccluded F1	0.591	0.594	0.577	**0.667**	0.645
Dense occluded P	0.721	0.69	0.559	**0.724**	0.674
Dense occluded R	0.468	0.521	0.66	0.586	**0.681**
Dense occluded F1	0.568	0.594	0.605	0.647	**0.677**
Background occluded P	**0.75**	0.667	0.525	0.692	0.604
Background occluded R	0.391	0.391	**0.696**	0.391	**0.696**
Background occluded F1	0.514	0.493	0.598	0.5	**0.646**
Self-occluded P	0.64	0.686	0.581	**0.855**	0.637
Self-occluded R	0.237	0.452	0.656	0.465	**0.664**
Self-occluded F1	0.345	0.545	0.616	0.602	**0.65**
Slice occluded P	**0.771**	0.75	0.45	0.725	0.556
Slice occluded R	0.597	0.726	**0.806**	0.597	**0.806**
Slice occluded F1	0.673	**0.738**	0.578	0.655	0.658

Bold is the optimal model for each line.

The TP, FN, and FP of the four network models under the conditions of unoccluded, moderately occluded, and severely occluded were counted and are shown in Figure 15. It can be seen that the accuracy of the YOLOv5s-seg NWD model was improved under different occlusion degrees, but it also caused some false alarms, and YOLOv5s-seg CAM could better suppress false alarms.

The TP, FN, and FP of four network models under conditions of single wells with no occlusion, dense wells with no occlusion, dense wells with occlusion, background occlusion, self-occlusion, and slice occlusion were calculated and are shown in Figure 16. It can be seen that the detection accuracy was improved under different occlusion types, and the false alarm of YOLOv5s-seg NWD was higher.

Below is a sample of the visual test results (see Figure 17). A green box indicates a correct detection, a blue box indicates a false detection, and a red box indicates a missed detection. Figure 18 shows the analysis of different occlusion types and different occlusion degrees under the YOLOv5s-seg CAN NWD model.

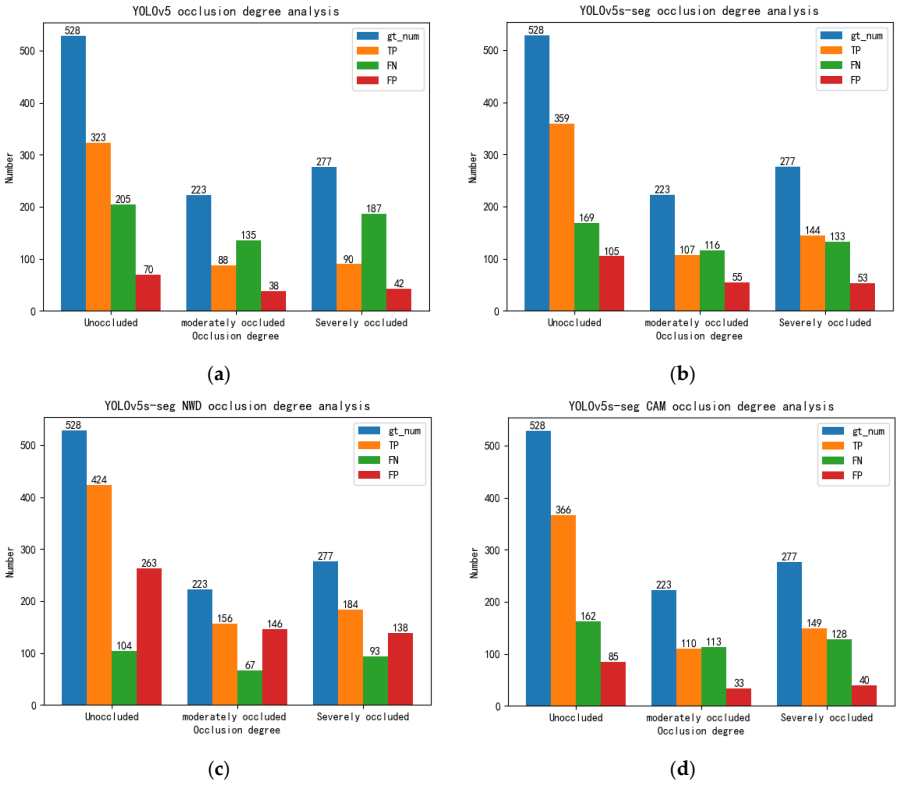


Figure 15. Occlusion degree analysis. (**a**) YOLOv5 occlusion degree analysis. (**b**) YOLOv5s-seg occlusion degree analysis. (**c**) YOLOv5s-seg NWD occlusion degree analysis. (**d**) YOLOv5s-seg CAM occlusion degree analysis.

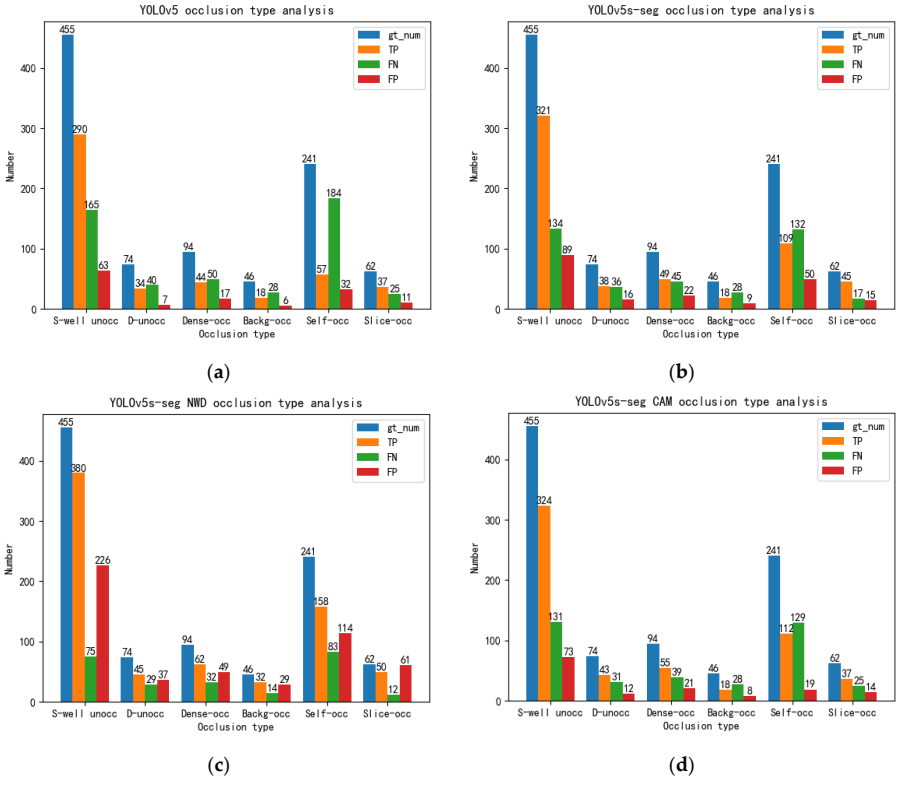


Figure 16. Occlusion type analysis. (**a**) YOLOv5 occlusion type analysis. (**b**) YOLOv5s-seg occlusion type analysis. (**c**) YOLOv5s-seg NWD occlusion type analysis. (**d**) YOLOv5s-seg CAM occlusion type analysis.

(**a**) Raw images.

(**b**) YOLOv5.

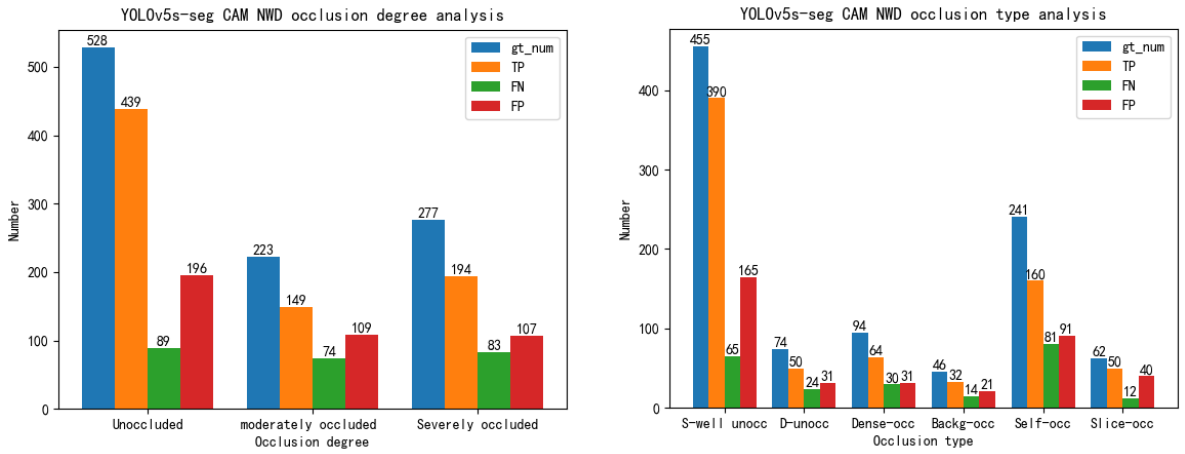
(**c**) YOLOv5s-seg.

(**d**) YOLOv5s-seg NWD.

(**e**) YOLOv5s-seg CAM.

(**f**) YOLOv5s-seg CAM NWD.

Figure 17. Object detection results.



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

Figure 18. Occlusion analysis of YOLOv5s-seg CAM NWD. (**a**) YOLOv5s-seg CAM NWD occlusion degree analysis. (**b**) YOLOv5s-seg CAM NWD occlusion type analysis.

4. Discussions

In this paper, occluded oil well detection in remote sensing images was studied in depth. Firstly, this paper tried to introduce instance segmentation into object detection, which made the model provide pixel-level segmentation results, thus significantly improving the detection accuracy. However, after the introduction of the attention mechanism and some feature enhancement modules (such as the ConvNeXt V2 model), the experimental results show that the detection effect in fact decreased, which may be due to the excessive noise and redundant information learned during the model training process, meaning that not all the extracted features were beneficial for the object detection task. In contrast, the CAM modules showed good performance in enhancing features. Moreover, NWD measurements can better measure the similarity between small oil well object boundary boxes that are obstructed than CIoU. For various occlusion types, NWD showed significant recall improvement in handling both self-occlusion and background occlusion, and the recall rate of self-occlusion achieved 0.419. At the same time, for other types of occlusions, NWD also showed a good recall rate improvement.

To validate the model generalization, the test set used the oil wells in California that were not used to train the model. The F1 score on the California dataset achieved 0.704, which was 0.108 higher than that of the baseline network YOLOv5, which fully verified our model's ability to detect occluded oil wells under geomorphic backgrounds of different complexities and under different target characteristics. This result proves that our model has high robustness and generalization ability.

5. Conclusions

Considering the characteristics of remote sensing images, oil well detection, especially occluded oil well detection, is a great challenge. To address this, we constructed our dataset, considering factors like occlusion degree, occlusion type, background, viewing angle, the presence of shadows, oil well type, and location within a well site.

Our initial comparison involved assessing two-stage detection models like Faster R-CNN against one-stage detection models such as YOLOv5, YOLOv7, and YOLOv8 for object detection. Furthermore, we considered the object detection and instance segmentation models Mask R-CNN and YOLOv5s-seg and found that YOLOv5s-seg exhibited the best performance, with an F1 score of 0.6591. Building upon the YOLOv5s-seg model, we explored the addition of an attention module and replacing the CSP structure with the C2f module. Ultimately, we achieved an F1 score of 0.704 with the YOLOv5s-seg CAM NWD model. The experimental results show that the YOLOv5s-seg CAM NWD based on the optimized YOLOv5 object detection and instance segmentation model can effectively detect

oil wells for both moderately occluded and severely occluded scenarios. In future work, we intend to continue to expand our dataset and continue to improve our methods to achieve better accuracy. In addition, we will study how to combine other data sources (such as topographic maps, etc.) with remote sensing images to improve the accuracy and reliability of oil well detection. We will also explore how the proposed method can be applied to other similar tasks, such as the detection of power lines, pipelines, and other infrastructure.

Author Contributions: Conceptualization, Z.W. and L.B.; methodology, Z.W., L.B. and Y.Z. (Yu Zhang); software, Y.Z. (Yu Zhang); validation, Y.Z. (Yu Zhang); formal analysis, Y.Z. (Yu Zhang); investigation, L.B., Z.W. and Y.Z. (Yu Zhang); resources, L.B., Z.W. and Y.Z. (Yu Zhang); data curation, Y.Z. (Yu Zhang); writing—original draft preparation, Z.W., L.B. and Y.Z. (Yu Zhang); writing—review and editing, L.B., Z.W., Y.Z. (Yu Zhang), M.F., A.J.-L., Y.Z. (Yuqi Zhang), Y.Z. (Ying Zhang), M.Z. and L.C.; visualization, Y.Z. (Yu Zhang); supervision, Z.W. and L.B.; project administration, L.C.; funding acquisition, Z.W. and L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by TUOHAI special project 2020 from Bohai Rim Energy Research Institute of Northeast Petroleum University under Grant HBHZX202002, the project of Excellent and Middle-aged Scientific Research Innovation Team of Northeast Petroleum University under Grant KYCXTD201903, Heilongjiang Province Higher Education Teaching Reform Project under Grant SJGY20200125, and the National Key Research and Development Program of China under Grant 2022YFC330160204.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://drive.google.com/drive/folders/1LijU_JaixSwuTk3Flx7oDVW-mrFw9wWd?usp=drive_link.

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

Abbreviations

CAM	CONTEXT_AUGMENTATION_MODULE
NWD	Normalized Weighted Distance
S-well unocc	Single wells are unoccluded
D-unocc	Dense unoccluded
Dense-occ	Dense occluded
Backg-occ	Background occluded
Self-occ	Self_occluded
Slice-occ	Slice occluded
M-cls occ	Multi-class occluded

References

1. Bp, B. Statistical Review of World Energy 2022. 2023. Available online: https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2022-full-report.pdf (accessed on 11 December 2023).
2. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
3. Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep learning-based object detection techniques for remote sensing images: A survey. *Remote Sens.* **2022**, *14*, 2385. [CrossRef]
4. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
5. Yu, Y.; Yuan, Y.; Guan, H.; Li, D.; Gu, T. Aeroplane detection from high-resolution remotely sensed imagery using bag-of-visual-words based hough forests. *Int. J. Remote Sens.* **2020**, *41*, 114–131. [CrossRef]
6. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
7. Zao, Y.; Shi, Z. Richer U-Net: Learning more details for road detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 3003105. [CrossRef]
8. Han, Q.; Yin, Q.; Zheng, X.; Chen, Z. Remote sensing image building detection method based on Mask R-CNN. *Complex Intell. Syst.* **2021**, *8*, 1847–1855. [CrossRef]

9. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. Yolo-facev2: A scale and occlusion aware face detector. *arXiv* **2022**, arXiv:2208.02019.
10. Du, S.; Zhang, B.; Zhang, P.; Xiang, P.; Xue, H. FA-YOLO: An improved YOLO model for infrared occlusion object detection under confusing background. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1896029. [CrossRef]
11. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
12. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; Yang, J. Feature-attentioned object detection in remote sensing imagery. In Proceedings of the 2019 IEEE international conference on image processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3886–3890.
13. Xiao, J.; Zhao, T.; Yao, Y.; Yu, Q.; Chen, Y. Context Augmentation and Feature Refinement Network for Tiny Object Detection. In Proceedings of the Tenth International Conference on Learning Representations (ICLR), virtual conference, 25 April 2022; pp. 1–11.
14. Sun, X.; Wang, P.; Wang, C.; Liu, Y.; Fu, K. PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 50–65. [CrossRef]
15. Mahmoud, A.; Mohamed, S.; El-Khoribi, R.; AbdelSalam, H. Object detection using adaptive mask RCNN in optical remote sensing images. *Int. J. Intell. Eng. Syst* **2020**, *13*, 65–76. [CrossRef]
16. Tian, D.; Han, Y.; Wang, B.; Guan, T.; Gu, H.; Wei, W. Review of object instance segmentation based on deep learning. *J. Electron. Imaging* **2022**, *31*, 041205. [CrossRef]
17. Liu, Y.; Li, H.; Hu, C.; Luo, S.; Luo, Y.; Chen, C.W. Learning to Aggregate Multi-Scale Context for Instance Segmentation in Remote Sensing Images. *arXiv* **2021**, arXiv:2111.11057.
18. Lin, K.; Zhao, H.; Lv, J.; Li, C.; Liu, X.; Chen, R.; Zhao, R. Face detection and segmentation based on improved mask R-CNN. *Discret. Dyn. Nat. Soc.* **2020**, *2020*, 9242917. [CrossRef]
19. Wang, D.; Wan, J.; Liu, S.; Chen, Y.; Yasir, M.; Xu, M.; Ren, P. BO-DRNet: An improved deep learning model for oil spill detection by polarimetric features from SAR images. *Remote Sens.* **2022**, *14*, 264. [CrossRef]
20. Zhu, M.; Wang, Z.; Bai, L.; Zhang, J.; Tao, J.; Chen, L. Detection of industrial storage tanks at the city-level from optical satellite remote sensing images. In Proceedings of the Image and Signal Processing for Remote Sensing XXVII, Online, 13–17 September 2021; pp. 266–272.
21. Wu, Q.; Zhang, B.; Xu, C.; Zhang, H.; Wang, C. Dense Oil Tank Detection and Classification via YOLOX-TR Network in Large-Scale SAR Images. *Remote Sens.* **2022**, *14*, 3246. [CrossRef]
22. He, H.; Xu, H.; Zhang, Y.; Gao, K.; Li, H.; Ma, L.; Li, J. Mask R-CNN based automated identification and extraction of oil well sites. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102875. [CrossRef]
23. Wang, Z.; Bai, L.; Song, G.; Zhang, J.; Tao, J.; Mulvenna, M.D.; Bond, R.R.; Chen, L. An oil well dataset derived from satellite-based remote sensing. *Remote Sens.* **2021**, *13*, 1132. [CrossRef]
24. Song, G.; Wang, Z.; Bai, L.; Zhang, J.; Chen, L. Detection of oil wells based on faster R-CNN in optical satellite remote sensing images. In Proceedings of the Image and Signal Processing for Remote Sensing XXVI, Online, 20 September 2020; pp. 114–121.
25. Wang, Z.; Bai, L.; Song, G.; Zhang, Y.; Zhu, M.; Zhao, M.; Chen, L.; Wang, M. Optimized faster R-CNN for oil wells detection from high-resolution remote sensing images. *Int. J. Remote Sens.* **2023**, *44*, 6897–6928. [CrossRef]
26. Shi, P.; Jiang, Q.; Shi, C.; Xi, J.; Tao, G.; Zhang, S.; Zhang, Z.; Liu, B.; Gao, X.; Wu, Q. Oil well detection via large-scale and high-resolution remote sensing images based on improved YOLO v4. *Remote Sens.* **2021**, *13*, 3243. [CrossRef]
27. Ribani, R.; Marengoni, M. A survey of transfer learning for convolutional neural networks. In Proceedings of the 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), Rio de Janeiro, Brazil, 28–31 October 2019; pp. 47–57.
28. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3320–3328. [CrossRef]
29. Ruan, D.; Yan, Y.; Chen, S.; Xue, J.-H.; Wang, H. Deep disturbance-disentangled learning for facial expression recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2833–2841.
30. Ma, Y.; Li, H.; Zhang, Z.; Guo, J.; Zhang, S.; Gong, R.; Liu, X. Annealing-Based Label-Transfer Learning for Open World Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 11454–11463.
31. Song, H.; Yang, W. GSCCTL: A general semi-supervised scene classification method for remote sensing images based on clustering and transfer learning. *Int. J. Remote Sens.* **2022**, *43*, 5976–6000. [CrossRef]
32. Alem, A.; Kumar, S. Transfer learning models for land cover and land use classification in remote sensing image. *Appl. Artif. Intell.* **2022**, *36*, 2014192. [CrossRef]
33. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13, 2014. pp. 740–755.
34. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
35. Zhang, J.; Lin, S.; Ding, L.; Bruzzone, L. Multi-scale context aggregation for semantic segmentation of remote sensing images. *Remote Sens.* **2020**, *12*, 701. [CrossRef]

36. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]
37. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.
38. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; pp. 16133–16142.
39. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. BiFormer: Vision Transformer with Bi-Level Routing Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 10323–10333.
40. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
41. Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; Yan, S. Metaformer is actually what you need for vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10819–10829.
42. Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; Sun, X. Damo-yolo: A report on real-time object detection design. *arXiv* **2022**, arXiv:2211.15444.
43. Yu, L.; Gong, P. Google Earth as a virtual globe tool for Earth science applications at the global scale: Progress and perspectives. *Int. J. Remote Sens.* **2012**, *33*, 3966–3986. [CrossRef]

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