Hindawi
Journal of Electrical and Computer Engineering
Volume 2022, Article ID 8612174, 13 pages
https://doi.org/10.1155/2022/8612174

Research Article

Petrochemical Equipment Detection by Improved Yolov5 with Multiscale Deep Feature Fusion and Attention Mechanism

Zhenqiang Wei,[1,2] **Shaohua Dong**,[1] **and Xuchu Wang**[3,4]

[1]*College of Safety and Ocean Engineering, China University of Petroleum (Beijing), Beijing 102249, China*
[2]*CNPC Research Institute of Safety & Environment Technology, Beijing 102206, China*
[3]*Key Laboratory of Optoelectronic Technology and Systems of Ministry of Education, Chongqing University, Chongqing 400040, China*
[4]*College of Optoelectronic Engineering, Chongqing University, Chongqing 400040, China*

Correspondence should be addressed to Zhenqiang Wei; wei-zhenqiang@cnpc.com.cn

Received 4 October 2022; Revised 8 November 2022; Accepted 10 November 2022; Published 2 December 2022

Academic Editor: Yang Li

Petrochemical equipment detection technology plays important role in petrochemical industry security monitoring systems, equipment working status analysis systems, and other applications. In complex scenes, the accuracy and speed of petrochemical equipment detection would be limited because of the missing and false detection of equipment with extreme sizes, due to image quality, equipment scale, light, and other factors. In this paper, a one-stage attention mechanism-enhanced Yolov5 network is proposed to detect typical types of petrochemical equipment in industry scene images. The model considers the advantages of the channel and spatial attention mechanism and incorporates it into the three mainframes. Furthermore, the multiscale deep feature is fused with a bottom-up feature pyramid structure to learn the features of equipment with extreme sizes. Moreover, an adaptive anchor generation algorithm is proposed to handle objects with extreme sizes in a complex background. In addition, the data augmentation strategy is also introduced to handle the relatively small and extremely large sample and to enhance the robustness of the fused model. The proposed model was validated on the self-built petrochemical equipment image data set, and the experimental results show that it achieves a competitive performance in comparison with the related state-of-the-art detectors.

1. Introduction

Object detection is a fundamental topic in the field of computer vision and has played important roles in many industrial applications [1, 2]. Importantly, petrochemical equipment detection that aims to identify and localize the equipment in petrochemical industrial images is a key prerequisite and essential component in many intelligent systems, i.e., petrochemical industry intelligent safety monitoring, automatic equipment localization, robots-assisted inspection, and equipment working status analysis. With the advancement of photography techniques and instruments to shoot industrial images, it has become an active but challenging task towards specific scenarios in the development of these intelligent systems.

Inspired by the appealing performance of deep learning in computer vision tasks, convolutional neural networks (CNN) based detectors for natural images have shown remarkable results in the topic of deep learning-based object detection, which can be divided into two categories: one-stage detectors [3–10] and two-stage detectors [11–14]. In contrast to the two-stage detectors based on the region proposal method, the representative one-stage detector, Yolo [3, 5, 7], uses both classifiers to predict all the categories along with the corresponding confidence and regressors to locate the objects through the predefined anchors, which can speed up the detection greatly yet at the expense of slightly reduced precision. Anchor-free one-stage detectors, such as CenterNet [10], were then proposed to avoid the complicated offset estimation regarding the anchor boxes, which

however were rarely used in multiobjective industrial tasks. Since real-time safety monitoring is highly desirable on the petrochemical working sites, the state-of-the-art Yolov5 appears to be a suitable option thanks to its fast speed.

However, in comparison to the natural images that often capture smaller visual fields and larger object sizes, whereas the petrochemical industry images generally capture information of lower resolution and varying scales of the objects. Petrochemical industry images have a wide covered area and contain lots of tiny and large distributions of equipment with complicated background. Although many object detectors have achieved acceptable performance on natural images, they are not able to obtain desirable detection results on petrochemical industry images.

To solve these issues, an approach is to integrate some general detectors to form a robust and useful detector ensemble towards a specific target. The integrated module usually considers the character of the object to improve the whole model's capability, which has been used in many scenarios since it combines the decision of multiple submodules to upgrade the overall performance. These approaches have been effectively employed for improving accuracy in some object detection tasks [15]. Unfortunately, regarding the complexity and configuration of deep learning-based object detection models, it is not a simple process of incorporating a reasonable submodule to improve the detection performance. On the other hand, one-stage detection with multiscale features instead of multiple detector ensembles has gotten more and more attention in recent years. The feature fusion has been taken into consideration since it has been applied in other applications and obtained desirable performance [16]. The famous RetinaNet [14] introduces the feature pyramid network to build three subnetworks for classification and regression, and also, the state-of-the-art EfficientDet [17] proposes a weighted bidirectional feature pyramid network to make fast multiscale feature fusion. Although the detection algorithms mentioned above had a significant improvement in detection performance, their multiscale features fusion only fused the feature maps directly. In this way, the fused feature layers are restricted by each other, and no other feature related to the interest of region is incorporated, which is not appropriate for petrochemical equipment of huge varying sizes. Incorporating other features to the constraint of feature fusion directly is beneficial to improve the detection performance of multiscale equipment in this scenario. Additionally, when the number of training samples is limited, the data augmentation has been applied for short-term voltage stability assessment of power systems [18], and a similar strategy could be designed in deep learning-based object detection. Therefore, it is necessary to improve the one-stage detector for the specific petrochemical industrial scenario by incorporating the advantage of feature fusion and data augmentation techniques.

Based on these motivations, we propose an object detection algorithm named Yolov5-FA based on the merits of the one-stage detector Yolov5 (you only look once version 5) for petrochemical equipment detection in industrial images. The model designed a network mainframe that incorporated channel and spatial attention mechanisms into the three modules by considering the characteristics of the petrochemical equipment. Furthermore, the multiscale deep feature is fused with a bottom-up feature pyramid structure to learn the features of equipment with extreme sizes. Moreover, an adaptive anchor generation algorithm is proposed to handle objects with extreme sizes in a complex background. In addition, the data augmentation strategy is also introduced to handle the relatively small and extremely large sample and to enhance the robustness of the fused model. The improved model has been validated on the real petrochemical equipment dataset and proved to be reliable and efficient for object detection in this scenario.

To sum up, our work makes the following contributions: (1) we propose an improved network (Yolov5-FA) to integrate the bottom-up feature fusion and attention mechanism in a channel and spatial manner to build the Yolov5-like network for detecting typical equipment in real petrochemical industry images. To the best of our knowledge, it is the first work on petrochemical equipment detection in real petrochemical industry images; (2) we design an effective adaptive anchor generation module to extract the prior petrochemical equipment information; (3) we design a data augmentation strategy to handle the small image size problem in our petrochemical equipment data set.

The rest of the paper is organized as follows: Section 2 introduces the related work about two-stage and one-stage object detection algorithms. Section 3 describes our proposed method in detail. Section 4 introduces the datasets and experimental results and made discussions. Section 5 is a summary of the paper.

2. Related Works

Object detection is a vital technique for realizing the identification and localization of objects in visual images, so it has received much attention in the last two decades. With the rapid development of the deep neural network, the performance of object detection methods has been gradually improved. According to the generation of candidate regions, the state-of-the-art deep learning-based object detection methods can be broadly classified into two categories, namely, two-stage and one-stage methods.

The most representative two-stage detection methods are region-based convolutional neural network (RCNN) series and its variants. RCNN [11–13] is one of the earliest and most effective methods that adopt the deep convolutional neural network (CNN) for object detection, which replaces the traditional hand-crafted feature-extracting process with CNN-based feature learning and improves the accuracy of object detection. In this category, there are usually two separate steps as follows. In the first stage, a series of candidate region proposals that may contain objects is generated. Then, in the second stage, feature maps are extracted by region-of-interest (ROI) pooling from each proposal for classification and localization tasks. In the original RCNN, the selective search technique [19] is adopted to generate almost 2000 region proposals, and this step reduces the detection speed. Fast RCNN [12] generates region proposals

on the feature map rather than the original input images, which improve the detection efficiency to a large extent. Faster RCNN [11] introduces an region proposal network (RPN) to generate region candidates from the convolutional neural network and achieves end-to-end calculation of object recognition. R-FCN [20] employs the full convolution network ResNet to replace the original VGG to improve the effect of feature extraction and classification. Furthermore, Cascade R-CNN [21] proposes multiple repeated networks, and they are connected sequentially, which could increase the amount of samples with high intersection over union (IoU) scores and allow the subsequent module to obtain performance. Lately, BorderDet [22] proposes an efficient border alignment to extract border features from the extreme point of the border to enhance the point feature. A recent You-only-look-one-level feature (YoloF) [23] introduces diluted encoder and uniform matching to optimize detection. Generally speaking, the two-stage methods are easy to generate admirable proposals and control the network depth with parameter adjustment, and they also can achieve admirable accuracy with the cost of speed.

In contrast, the one-stage methods aim to detect objects by directly applying regression and classification analysis strategy, which omits the first stage of generating candidate regions and directly obtains object class and location information. The representative one-stage detectors include Yolo (which is an acronym for you only look once) [3–7], single-shot detector (SSD) [8, 9], CenterNet [10], RetinaNet [14], and RefneDet++ [24]. The methods in this category can perform nearly real-time detection, do not need a proposal generation procedure, and directly conduct object detection in images.

The Yolo family achieves state-of-the-art performance by integrating bounding boxes and subsequent feature resampling in a single stage. The first three versions [3–5] of Yolo received their popularity because of speed and efficiency. Some deep learning-based detection algorithms are unable to detect an object in a single run, but Yolo series, on the other hand, makes the detection in a single forward propagation through a neural network, making it suitable for real-time applications. Intrinsically, Yolov4 [6] and Yolov5 [7] have the sample principle as Yolov3, but with more consideration on different applications and parameter sizes. This method has been applied for detection small objects (such as cars) captured by unmanned aerial vehicle [25]. Besides, in this kind of one-stage methods, SSD [8] is another state-of-the-art real-time object detector. To increase the detection accuracy, SSD predicts category scores and box offsets for a fixed set of default boxes, by using small convolutional filters over multiple scale feature maps. CornerNet [26] is a one-stage method that proposes a model to eliminate anchor boxes, and an object is detected as a pair of the top-left corner and bottom-right corner points of a bounding box. CenterNet [10] is also a kind of one-stage method, in which an object is detected according to one center key point and two key points of a bounding box, which contains the center location and other attributes of an object (e.g., size). This model has been used for aerial object detection by combining with the Yolov5 model [15]. RefineDet++ [24] introduces a module to refine anchor boxes to combine feature fusion and box regression from coarse to fine stages by following the SSD network structure. In general, the two-stage object detection models are more accurate than the one-stage ones, but the one-stage methods are faster and simpler to train the neural network. In our work, the task of petrochemical equipment detection is working under the server by acquiring the images using cameras and videos, so the detecting speed should be more considered, and we focus on the Yolo detection series in our work.

3. Our Method: Yolov5-FA

3.1. Main Structure. The structure of our proposed method is presented in Figure 1, where it follows the flowchart of the backbone mainframe and the multiscale deep feature fusion and attention mechanism modules. The backbone mainframe aims to extract deep features, and the multiscale deep feature fusion is to fuse these features and the sampled ones. After this multiscale feature fusion, four detection modules aim to predict the class and location of petrochemical equipment.

Specifically, we first use the Yolov5 as the basic framework, which combines three modules, i.e., backbone, neck, and prediction. In the backbone, there are a focus layer, convolution layer, multiple CCbam3 layer, and spatial pyramid pooling layer to extract multiscale feature information. In the neck part, we use a channel and spatial attention mechanism module to optimize the feature map in different scales. In the prediction network, we use four detection modules to detect equipment with difference sizes. In addition, we apply the data enhancement strategy to improve the learning ability in occlusion and overlapping situations. In the following subsections, we will present the details of these modules.

3.2. Multiscale Deep Feature Fusion. Considering the background is complex in the petrochemical industry images and the standard mainframe of Yolov5 hardly extracts the features very powerfully, we introduce a convolution kernel group to replace the 3×3 convolution kernel. This group consists of three parallel 3×3, 1×3, and 3×3 convolutional kernels. When applied to the input image, these three kernels mean a weighted 3×3 kernel. Then, the batch normalization (BN) and sigmoid weighted linear unit (SiLU) modules are integrated to form a convolution module in our network. SiLU is a recently appeared activation function for neural networks with the sigmoid function multiplied by its input. If the input value is greater than 0, the SiLU is approximately the same as the ReLU, and if the input value is less than 0, the value of SiLU approaches 0. Compared with the Sigmoid and tanh, the SiLU activation function does not increase monotonously and has a global minimum value of about −0.28. Therefore, an attractive feature of SiLU is self-stability: when the derivative is zero, the global minimum can play the role of

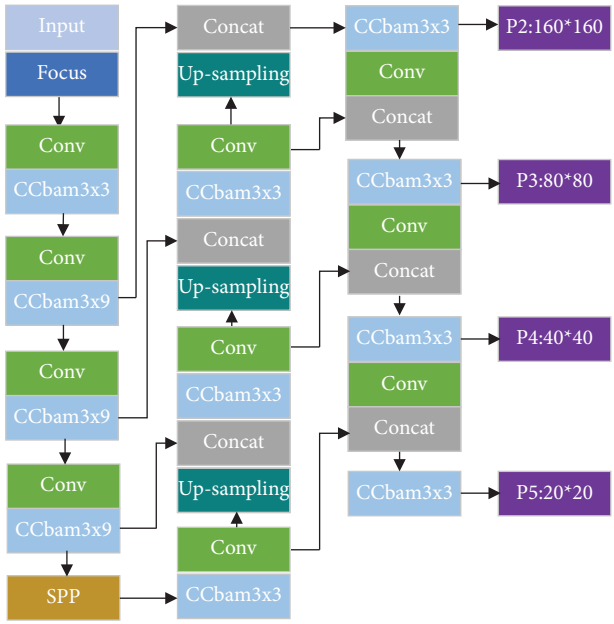

Figure 1: Improved neural network structure in our work.

"soft bottom," which can inhibit the update of large weights by avoiding gradient explosion.

Furthermore, the convolution kernel group is taken as a component of the core module CCbam3 in our network. As shown in Figure 2, the CCbam3 module is a weighting mechanism to emphasize the object information by introducing a channel and spatial attention mechanism. Specifically, the attention mechanism module to focus the potential object integrates the bottleneck in standard Yolov4 [6], and then, the fusion operation is conducted to collect the information from the two channels. In this way, the intrinsic feature of different petrochemical objects could be learned in deep convolution networks.

3.3. Channel and Spatial Attention Module. In the standard Yolov5 framework, the features are extracted only by convolution operations, which lack an effective attention mechanism to drive the network to focus on those more meaningful features. To overcome this problem, we introduce a channel and spatial attention module to optimize the network structure, which includes a channel attention mechanism and spatial attention mechanism [27]. This module can compress and weight the features in channel and spatial dimensions, to improve the focus on important features and suppress the distraction from the background.

The structures of the convolutional block attention module [27] and squeeze-and-excitation module [28] in our work are shown in Figure 3. Specifically, given an input feature map $F_{\text{in}} = R^{C\times H\times W}$, where C denotes the number of channels in the feature map, and H and W denote the height and width of the map, respectively. The average pooling and maximum value pooling are employed to obtain two processed feature maps. These two branches are sent to a multiple perceptron and generate the features with dimension $c/r\times 1\times 1$, where r denotes the compression rate in the hidden layer. These two feature vectors are then added and sent to the nonlinear activation function, and the attention coefficient M_C is obtained as follows:

$$\begin{aligned} M_C(F) &= \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^C_{\text{avg}}\right)\right) + W_1\left(W_0\left(F^C_{max}\right)\right)\right), \end{aligned} \tag{1}$$

where F denotes the input feature map, MaxPool() and AvgPool() denote maximum and average pooling operation, respectively, MLP() denotes perceptron connection layer, and $\sigma(\cdot)$ denotes Sigmoid activation function.

The spatial attention mechanism is improved based on squeeze-and-excitation network [28, 29] but emphasizes the difference between channel and image spatial domains. The attention coefficient is calculated from both dimensions, and it multiplies the feature map by learn the important feature flexibly. Considering this model is light and adjustable, it can be integrated in the main framework of Yolov5; specifically, the expression is as follows:

$$\begin{aligned} M_S(F) &= \sigma\left(f_7\left([\text{AvgPol}(F); \text{MaxPool}(F)]\right)\right) \\ &= \sigma\left(f_7\left(\left[F^S_{\text{avg}}; F^S_{\text{max}}\right]\right)\right), \end{aligned} \tag{2}$$

where f_7 denotes the convolution operation using a filter with 7×7 size, and M_S denotes the spatial attention coefficient.

3.4. Adaptive Anchor Generation. Generally, the accuracy of the detectors could be improved by increasing the resolution of the input image using a deep neural network with strong feature extraction ability and predefined anchor boxes. However, this technique has a negative effect on the detection speed. Since the anchor box technique reduces the detection speed, many systems have been developed to improve the anchor box quality. For instance, several anchor boxes for each category are proposed to improve the detection accuracy of a real-time single-stage object detector in the Yolov2 framework [30].

Intrinsically, the anchor boxes are predefined prior sizes of objects in the training data set. The initial anchor size adopted by Yolov5 is clustered from the object box size in the COCO dataset. Table 1 shows the differences in object size between the petrochemical equipment dataset and COCO dataset, and small objects account for the majority in the chemical equipment dataset. Thus, the original anchor size is not suitable for the petrochemical equipment dataset.

To improve the matching probability of the object box and anchor, we employ the K means++ clustering algorithm [31] to redesign the anchor size. This improvement can reduce the influence of randomly selected initial values on the results. Specifically, we design two steps to obtain the anchor size as follows.

The first step is to determine the initial values of K cluster centers. The distance d is defined as follows:

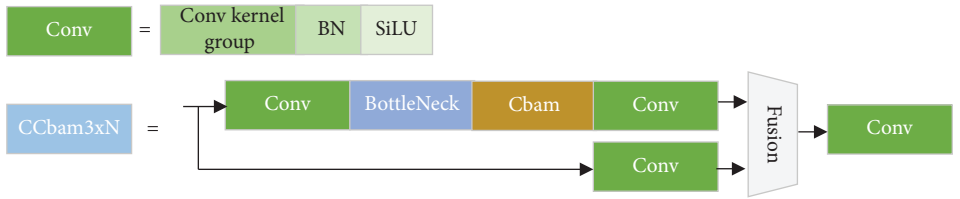

Figure 2: Proposed CCbam3 module in our work.

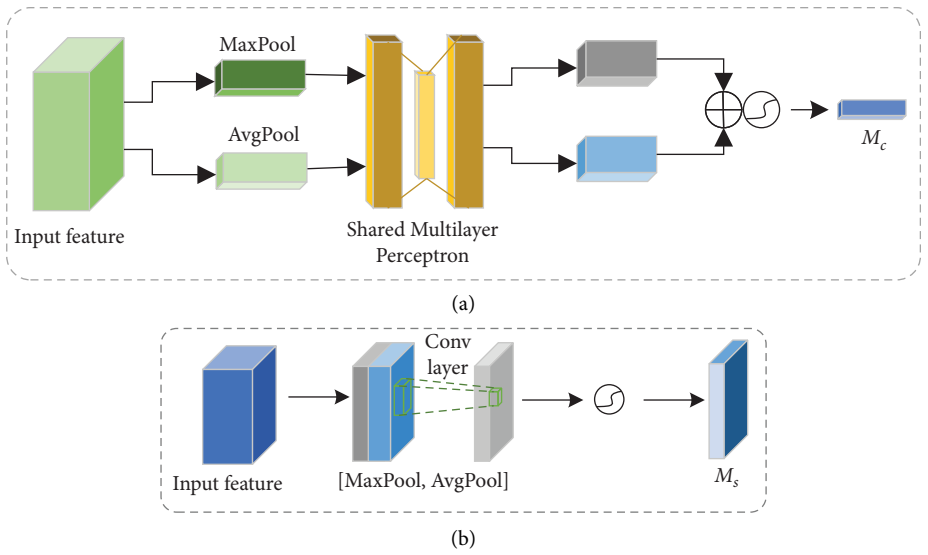

Figure 3: Channel and spatial attention module. (a) Channel attention module. (b) Spatial attention module.

Table 1: Quantitative comparison of object sizes between our chemical equipment data set and COCO.

Size (%)	COCO (%)	Chemical equipment data set (%)
Small (0, 0.3)	46.2	59.3
Medium (0.3, 0.7)	35.7	44.2
Large (0.7, 1)	28.1	6.5

$$d(\text{box}, \text{centroid}) = 1 - \text{MIOU}(\text{box}, \text{centroid}), \quad (3)$$

where "box" is the target box, and "centroid" is the cluster center. The distance from all points to the nearest cluster center is $D(x)$, and $p(x)$ represents the probability of each point to become the next cluster center

$$p(x) = \frac{D(x)}{\sum_{k=0}^{n} D(x)}. \quad (4)$$

According to the probability, the roulette wheel selection (the greater the distance, the greater the probability of being selected as a cluster center) is repeated until K cluster centers are selected.

The second step is to cluster the initial cluster centers selected in the first step: cluster centers are divided into K sets, and each sample is divided into the set to which the nearest cluster center belongs to. The average value of all samples in each set is computed as the new cluster centers, and each sample is subdivided into the set with the shortest distance from the new cluster center; then, the average value of the sample in each set is recalculated. Repeat the previous operation until the change of the average value is less than a certain threshold, and the K cluster centers are the new anchor size. Table 2 reports the redesigned anchor sizes. It is seen that the sizes in our method can handle the varying sizes of different types of petrochemical equipment by extending a middle branch, and the prior sizes fit the actual equipment more adaptively.

3.5. Data Augmentation Strategy. In practical industry images, multiple types of petrochemical equipment are usually heavily overlapped, and some are occluded by complex industry pipelines and other appendages. To overcome this situation, in this paper, we introduce the Mosaic data augmentation method to improve the robustness. This method can adjust four images into one image and then sent to the network for training. In addition, we randomly change

Table 2: Comparison of anchor sizes among different clustering algorithms in our chemical equipment data set.

Size (%)	P5/32	P4/16	P3/8	P2/4
K means in Yolov5	(112, 68)	(221, 176)	(403, 234)	—
	(126, 151)	(405, 96)	(304, 322)	—
	(191, 126)	(225, 268)	(564, 318)	—
K means++ in the proposed method	(89, 69)	(191, 127)	(233, 165)	(304, 321)
	(177, 68)	(160, 218)	(227, 269)	(481, 230)
	(134, 140)	(408, 93)	(344, 222)	(583, 315)

the brightness, contrast, saturation, and angle of the images to reduce the overfitting.

To generate the robustness of the model, we also take a mix virtual sample generation method, specifically, we randomly select two images from the training data set and weight their pixels and labels to obtain the new images and new labels, and these virtual samples are independently sent to the network and refresh the parameters. In this way, the size of the samples can be remarkably enlarged. Figure 4 presents some examples of our data augmentation.

4. Experimental Results and Discussion

In this section, experimental results are presented to demonstrate the performance of the proposed model on a petrochemical equipment image dataset. Ablation studies are performed to evaluate the effectiveness of the proposed method. State-of-the-art detectors, Yolov5 [7], SSD [8], Faster RCNN [11], RetinaNet [14] and Efficient Det [17], are chosen as benchmarks.

4.1. Datasets and Metric

4.1.1. Petrochemical Equipment Image Dataset. The petrochemical equipment automatic detection is essential in computer-assisted equipment examination, repairing, and overhauling of petrochemical factories or oil fields. However, the petrochemical equipment images are difficult to be collected because the safety management rules are usually strict in these places. Thus, to the best of our knowledge, there is no public data set about this scenario. To handle this, we built a petrochemical equipment image dataset which consists of 2644 images. These images were acquisited by digital cameras including camera-equipped explosion-prevented mobile phones and explosion-prevented digital video cameras. Each image scale ranges from 540×960 to 2800×1500 pixels and contains various shapes and scales.

There are 5 types of typical petrochemical equipment in these images, and they were annotated by two petrochemical engineers independently. Their labelling results were exchanged and checked to form a final annotated data set. Specifically, these 5 types of petrochemical equipment include luoganbeng (screw pump), lixingbeng (centrifugal pump), huanreqi (heat interchanger), sphere, and cyclinder. And they are typical key equipment in petrochemical factories or oil fields.

Since the equipment image detection task is still challenging because of class imbalance and object-image size mismatch, in our work, this dataset is utilized for the validation of the proposed method. This dataset in our experiments is randomly divided into three parts as follows: 1696 images for training, 300 for validation, and 648 images for the test.

4.1.2. Evaluation Metrics. The evaluation standard adopted in this paper is the mean average precision (mAP), which is utilized to evaluate the performance of our method relative to other benchmarks. We also computed three different average precision metrics: AP50, AP75, and mAP. For AP50 and AP75, both consider a bounding box prediction as true and overall object categories when the interest over union (IoU) scores between the predicted and the ground-truth bounding box must be larger than 0.5 and 0.75, respectively. The mAP, which takes a value between 0 and 1, is the average of all 10 IoU thresholds from a range of [0.5, 0.95] with a step size of 0.05.

4.1.3. Experimental Platform. All experiments in this paper were conducted on the Ubuntu 18.04LTS system, which has 2.0 GHz Intel CPU and 48 GB RAM. The GPU is NVIDIA RTX 2080Ti. The program environment is Anaconda 5.0.1 (Python 3.7) and PyTorch 1.7. In this paper, the improved network structure of Yolov5s is used for training, with the initial learning rate set at 0.001, the batch size set at 32, and the number of learning epochs set at 150.

4.2. Experimental Results

4.2.1. Performance of Yolov5-FA. We use bottleneck as the backbone for our detection structure, and this model has been pretrained on the ImageNet. Our proposed framework is shown in Figure 2. In the training and testing stage, the input images are resized to 640×640. In the training phase, we trained the model for 150 epochs with one batch size of 6 and a learning rate of 0.001. We have implemented the proposed method on PyTorch 1.7.0 and trained it based on Yolov5. Our proposed model continues to be trained on the Ubuntu 18.04LTS server with an NVIDIA GeForce GTX 2080Ti GPU. In this experiment, we modified the number of Yolov5 outputs, in which only 150 boxes were selected as candidate boxes for each object.

Figure 5(a) reports the three loss curves on the training set and validation set. It is seen that the box loss curve and object loss and classification loss curves are steadily decreased on both training set and validation data set which means that the proposed model and the training parameter sets are effective

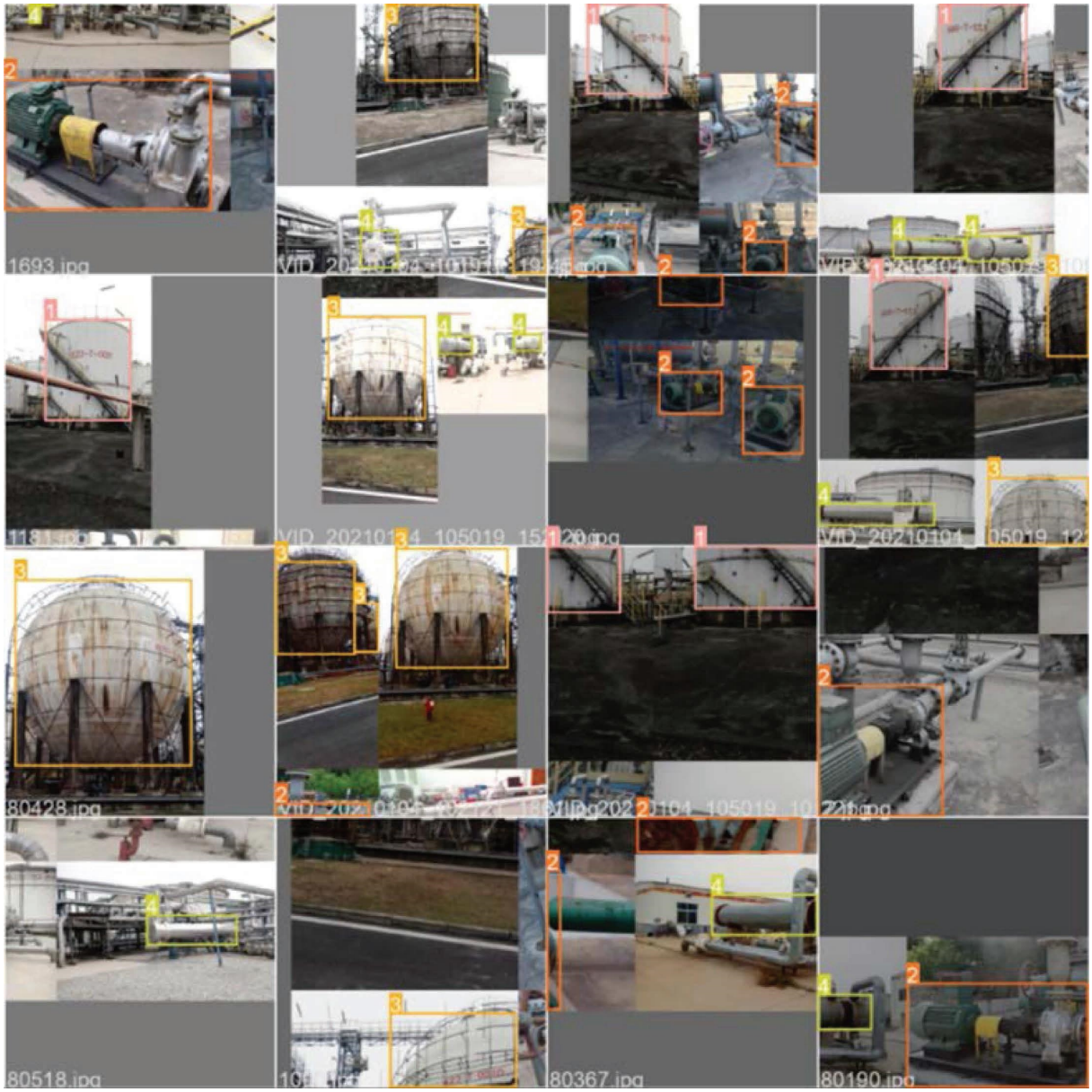

Figure 4: Examples of augmented images in the training set.

to converge the minimum. In fact, the training epochs can be prolonged to 500 if the curves need to be examined in the whole training procedure. However, the precision and recall and mAP have steadily approached to the maximum, so the results are reported only based on 150 epochs.

Figure 5(b) plots the precision-recall curves of different petrochemical equipment localization and identification. It is seen that the average precision is 97.8% for all equipment, and the localization of the cylinder achieves 99.5%, which is the highest value among the five types of equipment. The accuracy of the screw pump (luoganbeng), although in the worst place of the five types of equipment, still approaches to 96.6%. They all can be precisely detected and located by the proposed method. To evaluate the precision and recall in a harmonic mean way, we also plot the $F1$-score of different equipment under different confidence in Figure 5(c). It is seen that the highest $F1$-score can approach to 0.97 and present a desirable result on each equipment.

Considering that our model is improved based on the standard Yolov5 model, we compared our model and two standard Yolov5s and Yolov5x in the same experimental environment. The Yolov5s model is smaller and easier to deploy quickly because it employs the smallest depth and width in the net structure, and on the other side, the Yolov5x has a size of nearly 168 MB but is the most accurate version of its family. Table 3 reports the results of these three models in terms of mAP50, $F1$, Precision, Recall metrics. It is seen that the mAP of the Yolov5x model has increased 2.6% and 0.04 in comparison to the basic Yolo5x in terms of mAP50 and $F1$ scores; however, the model size has increased from 13.70 to 156.2 MB. Our proposed method achieved the competitive results in comparison to the Yolo5x model, but

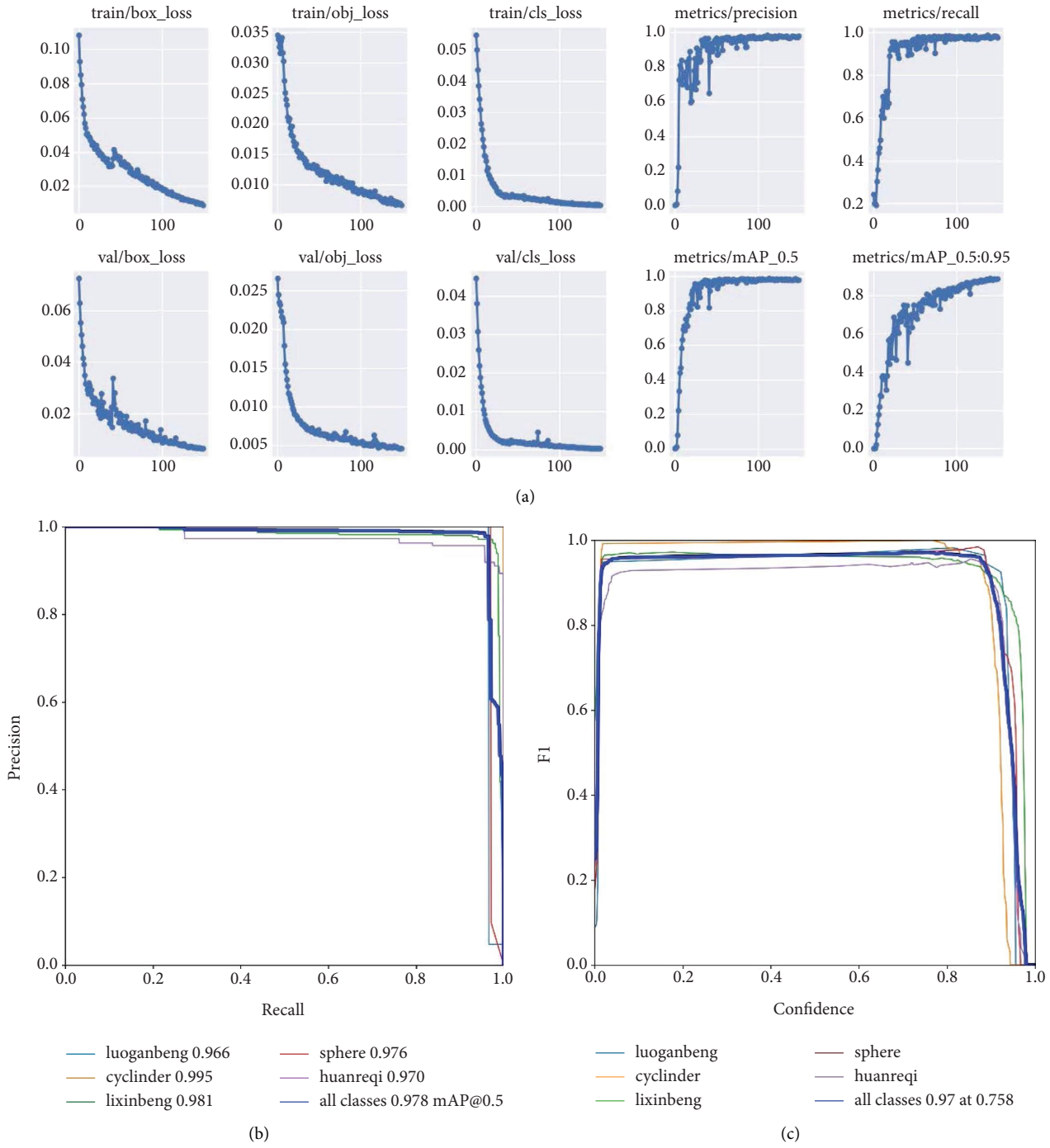

Figure 5: (a) Three loss curves of the training set and validation set and the key metric curves (precision, recall, and mAP). (b) Precision-recall curves of different petrochemical equipment. (c) $F1$ curves of different petrochemical equipment.

Table 3: Performance of the proposed method in different detection precisions and model parameters.

Model	Epochs	mAP50 (%)	$F1$	Precision (%)	Recall (%)	Parameters (MB)
Yolov5s	200	94.6	0.93	93.6	91.9	**13.7**
Yolov5x	200	97.4	**0.97**	**97.3**	97.1	156.2
Proposed	150	**97.8**	**0.97**	**97.3**	**97.2**	**13.7**

Note: The bold values denote the best ones.

it only has 13.7 MB, just 8.8% of the complex Yolo5x model. Our model outperforms Yolo5x in all terms, and it is also better than Yolo5x in terms of mAP50 and Recall which means the added attention mechanism model and adaptive anchor and data augmentation strategies are beneficial to extract the petrochemical equipment feature and perform the inference more accurately.

Figure 6 illustrates the detection results with the proposed model in various petrochemical working scenes, and the detection results on the sixteen images show that the proposed model can successfully detect different targets even when it is tiny or partially appears in the pictures; for instance, the partial lixingbeng equipment in the last column has been accurately detected. In addition, we can find that the proposed model can locate the equipment in different views, just like the huanreqi equipment in the first three columns.

4.2.2. Performance of Comparative Experiments. To further demonstrate the effectiveness and robustness of our model, we compared it with the state-of-the-art detection methods on the testing dataset. The mainstream compared methods are the popular one-stage or two-stage object detection models, such as Faster RCNN, SSD (single-shot detector), RetinaNet and Efficient Det, Yolov5s, and improved Yolov5x. The parameters and experimental setting followed the original model and the pretrained model on COCO data set were also employed.

Table 4 reports the petrochemical equipment detection performance of the proposed model and the related models in terms of the metrices such as mAP50, mAP75, and AP of each equipment. It can be seen that our proposed model achieves 97.8% in mAP50 and 96.3% in mAP75, which are the highest overall mAP50 and mAP75, and the best performance for all 5 categories of objects in comparison to the start-of-the-art models. Specifically, in the mAP50 metric, the proposed model outperforms Faster RCNN, SSD, RetinaNet, and EfficientDet by 2.9%, 2.6%, 2.5%, and 2.5% respectively; in the mAP75 metric, the proposed model outperforms them by 5.1%, 2.5%, 2.8%, and 1.5%, respectively. It is known that the four related models have their characteristics in network structure for handling the object detection; for example, Faster RCNN is a famous two-stage deep convolutional network used for object detection, and it is regarded as an end-to-end and unified network that can accurately predict the locations of different objects, while SSD is known to be a one-stage object detection by introducing small convolutional filters to forecast the object classes and offsets to original boxes. It is usually regarded that the two-stage detector can obtain more accurate results; however, SSD got better performance than Faster RCNN in our petrochemical equipment detection with complex background. Furthermore, RetinaNet is a single-stage object detection model that uses a focal loss function to alleviate the problem of the extreme foreground-background class imbalance during training. EfficientDet is another famous one-stage object detection model that proposed a bidirectional feature pyramid network with fast normalization and feature fusion enhancement. They all achieve better performance than Fast RCNN and SSD. Considering these latest methods incorporate multiscale feature extraction into their backbones, in this viewpoint, it is adorable to adopt multiscale feature fusion when designing a detection network towards a specific task.

Furthermore, compared to the famous Yolov5s and Yolov5x under Yolo network, our proposed model achieves higher mAP50 and mAP75 for five categories of objects and contributes 3.2% and 0.4% to the overall mAP50. Our method also outperformed Yolov5x in terms of mAP50 and mAP75 for 0.4% and 0.7%. Particularly, significant improvement can be found in the detection of four out of five types of petrochemical equipment. Overall speaking, the values of the mAP50 and mAP75 metrics have shown that the proposed model can get more precise petrochemical equipment location information and have great detection performance of equipment with varying sizes compared with other methods.

4.2.3. Ablation and Scalable Experiment. To investigate the effect of the different improved technologies more intuitively on the performance of the proposed model, we conducted an ablation experiment. Specifically, by keeping the structure of Yolov5s unchanged and only improving the extended module, we can observe the impact of the performance. Then, we added the spatial and channel attention mechanism module, and adaptive anchor generation and data augmentation, respectively, to observe the experimental results and analyze their influence. Our ablation experiment also keeps training for 150 epochs. When the training result was stabilized, the training was finished, and the model was tested on the testing data set.

The metric indicators are shown in Table 5. It is found that by introducing the improved spatial and channel attention mechanism module, adaptive anchors, and the added data augmentation, the accuracy indicators of equipment detection have been improved accordingly. When the integration of these three improvements was tested as the final network model, the tested indicators show the best detection accuracy in comparison to the three methods introduced separately. The mAP50, precision, and recall of the proposed model increased by 3.4%, 4.0%, and 5.8%, respectively, in comparison with the base model which means the three modules are meaningful for obtaining the high detection accuracy. It is also seen that the CBAM contributes more than the other two modules for improving the performance. The corresponding mAP50, precision, and recall increased by 2.7%, 4.9%, and 6.3%, respectively. This indicates that the channel and spatial attention mechanism is very useful to capture the key characteristics of petrochemical equipment. The modification on adaptive anchor generation is also helpful for improving the performance.

The proposed model does not have more computational complexity than the standard Yolov5s model. In fact, only CBAM increases the model's complexity because it involves convolution in the backbone, the other components such as adaptive anchors generation and data augmentation can be regarded as the preprocessing that depends on the model

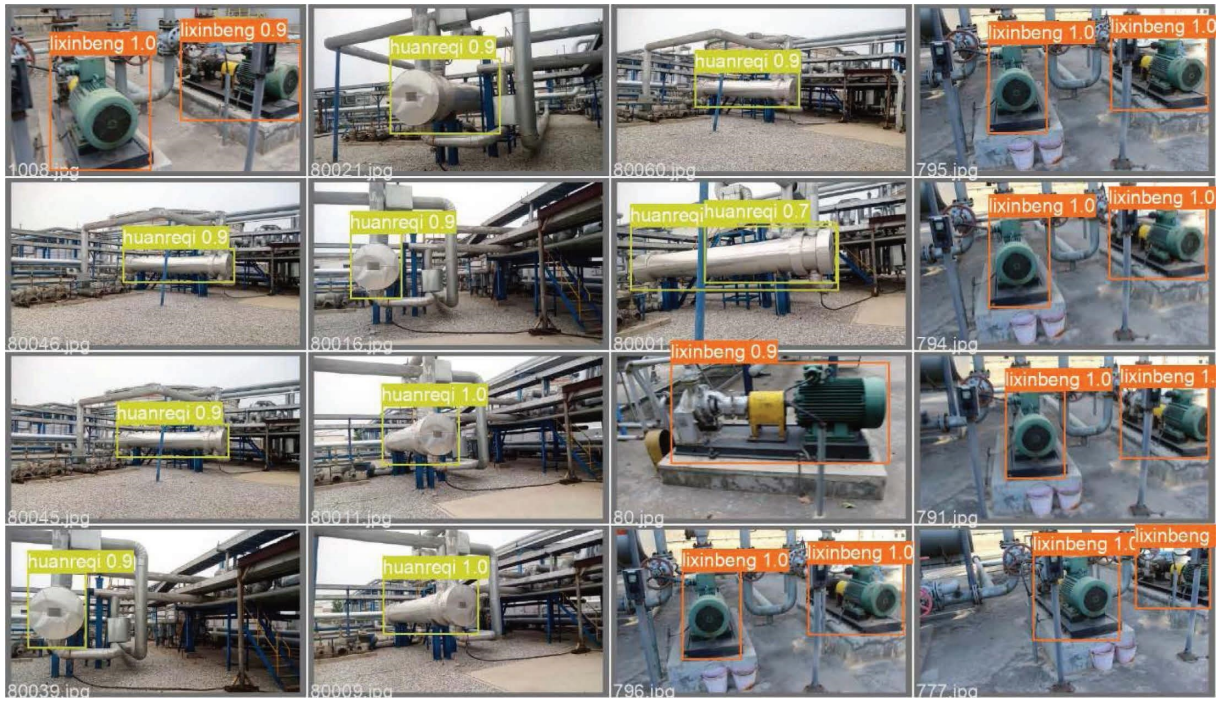

Figure 6: Examples of detected results by the proposed model in the testing set.

Table 4: Quantitative comparison measured by the common metric mAP in percentage with IoU = 0.5 and 0.75 of detection precision on different petrochemical equipment.

Method	mAP50 (all)	mAP75 (all)	Luoganbeng (screw pump)	Cylinder	Lixinbeng (centrifugal pump)	Sphere	Huanreqi (heat interchanger)
Faster RCNN	95.9	91.2	94.3	98.4	97.1	95.4	93.8
SSD	96.2	93.8	95.2	98.1	96.6	95.2	94.9
RetinaNet	96.3	93.5	95.5	99.2	96.1	96.8	94.2
EfficientDet	96.2	94.8	93.8	98.4	96.6	97.2	96.3
Yolov5s	94.6	91.6	95.5	99.0	97.2	97.0	91.5
Yolov5x	97.4	95.6	96.4	**99.5**	98.1	97.5	97.0
Proposed	**97.8**	**96.3**	**96.6**	**99.5**	**98.5**	**97.6**	**97.2**

Note: The bold values denote the best ones.

training and test, and it did not take much computational time for K means++ and virtual sample generation. So, the FLOPs of our model are only larger than 0.1 than that in Yolov5s.

To further investigate the scalability of the proposed model, we extended it to three variants by increasing the depth and width of the network. Specifically, we added two weights (depth coefficient and width coefficient) to make the backbone block and convolutional channel scalable. In our basic model, they were settled as 1/3 and 1/2 for a small model. Then, they are increased to (2/3, 3/4), (1, 1), and (4/3, 5/4) for building the medium, large, and extreme variants. Table 6 reports the experimental results of these variants on the test set of petrochemical equipment data set in terms of the mAP50, precision, recall, FLOPs, and speed where FLOPs denote the floating-point operations per second and can be roughly regarded as a metric for the computational complexity of the model.

It is seen from Table 6 that the detection accuracies consistently increase along with the deeper network and larger convolutional channels. The proposed method with an extremely large deep network and wide channels can achieve 98.9% in mAP50; however, the FLOPs are much larger than that in the proposed method. Considering FLOPs directly represent the parameters capacity, it is suggested to take the proposed model with 1/3 depth coefficient and 1/2 channel coefficient as the major model for similar tasks. On the other hand, the experimental results validate the scalability of the proposed model in different conditions as indicated in the experiment, and the proposed model can be easily extended to the application that puts accuracy in the first place.

4.3. Discussion

4.3.1. Overall Discussion. In this paper, an improved Yolov5-FA model with improved robustness and stability in a complex petrochemical working environment is proposed

Table 5: Ablation study of detection precision on the test set of our petrochemical equipment data set.

Model	CBAM	Adaptive anchors	Data augmentation	mAP50 (%)	Precision (%)	Recall (%)	FLOPs (B)
Yolov5s	—	—	—	94.6	93.6	91.9	**16.3**
M1	√	—	—	96.2	96.1	96.6	16.4
M2	—	√	—	95.8	94.7	95.0	16.3
M3	—	—	√	95.3	95.2	94.4	16.3
Proposed	√	√	√	**97.8**	**97.3**	**97.2**	16.4

Note: The bold values denote the best ones.

Table 6: Scalable study of detection precision on the test set of our petrochemical equipment data set.

Model	Depth coefficient	Channel coefficient	mAP50 (%)	Precision (%)	Recall (%)	FLOPs (B)	Runtime (ms)
Proposed small	**1/3**	½	97.8	97.3	97.2	**16.4**	**6.1**
Proposed medium	2/3	¾	98.3	98.3	97.7	49.2	9.2
Proposed large	1	1	98.7	**98.5**	97.7	108.6	13.8
Proposed extreme	4/3	5/4	**98.9**	**98.5**	**98.2**	212.4	15.4

Note: The bold values denote the best ones.

by using the spatial and channel attention mechanism to focus on important feature information. Besides, we applied the adaptive anchor and data augmentation module to make the model learn more predefined information and improve its accuracy and robustness. Furthermore, a detection layer was embedded to improve the model's detection accuracy for varying equipment. Additionally, using the CIoU as the loss function, it achieves the fastest convergence speed and the best convergence effect function.

The experimental results on the petrochemical equipment image dataset have demonstrated the superiority of Yolov5-FA, and the ablation study of spatial and channel attention mechanism and adaptive anchor modules proves that the combination of attention mechanism and adaptive prior anchor can make a significant improvement on typical petrochemical equipment detection performance. Specifically, the performance of the proposed Yolov5-FA for petrochemical equipment detection is discussed from the following aspects:

(1) The optimization of backbone network.In our work, the bottleneck is added to four, instead of three in basic Yolov5, which increases the effective perceptive field. Additionally, multiple CCbam3 layers and spatial pyramid pooling layer are incorporated into the backbone to extract multiscale meaningful feature information. This channel and spatial attention mechanism module is also added to the neck part for optimizing the feature map in different scales. The experimental results verified this optimization in comparison to the state-of-the-art methods. The proposed method achieved the best performance in the overall mAP and specific AP on each class. The ablation experimental results also reveal the effectiveness of this optimization where the mAP50, precision, and recall increase 1.6%, 2.5%, and 4.7% when only CBAM is added to the standard Yolov5s model. This also indicates there are still many necessities in backbone optimization instead of borrowing standard backbones for equipment detection in specific industry scenarios.

(2) The optimization of anchor generation. Anchors play important roles in our model because it corresponds to the object class in the classification module while the object locates in the regression module. If the IoU-predicted anchors and genuine bounding box are too small, then the performance on both classification and localization will remarkably decrease. Unfortunately, the anchor settled in most general object detection algorithms is not designed for specific tasks. In contrast, the anchors were automatically estimated in our model with finer stages, more numbers, and larger captive sizes, and considering the image sizes of our data set range from 540×960 to 2800×1500 pixels and contain various shapes and scales, this adaptive anchor generation using K means++ can make them cover most equipment in the image set and alleviate the burden of training the network. This can be validated by the experimental results of Yolov5s, Yolov5x, and our proposed model. Our model can achieve competitive results compared to the most powerful Yolov5x in terms of mAP50, $F1$, precision, and recall, but it just needs 13.7 MB parameters while Yolov5x needs 156.2 MB.

(3) The optimization of data augmentation strategy. In petrochemical equipment image data set with limited practical industry images, heavily overlapped and occluded equipment also increase the challenge of applying the deep convolutional networks both in designing and training stage. Furthermore, the image sizes are varying positive images. Traditionally, the data augmentation strategy is usually considered as increasing the variety of the image from an additional source, such as downloading natural images from the web; instead, we introduce the Mosaic data augmentation method to improve the robustness. This method can adjust four images into one image. The brightness, contrast, saturation, and angle of the images are also randomly changed and virtualized to reduce the overfitting. The ablation experimental results verified the effectiveness of this simple

strategy. The corresponding mAP50, precision, and recall increased by 0.7%, 1.6%, and 2.5%, respectively. Although this strategy did not improve much more performance than the optimization of backbone and adaptive anchor generation, it makes use of the powerful nonlinear fitting ability of deep convolutional neural networks without extra acquisition time and cost.

4.3.2. Limitations. There are some limitations of the proposed model should be addressed as follows:

(1) The class imbalance problem was less taken into consideration in the proposed model. Although it optimized the backbone with spatial and channel attention mechanisms and increased the perceptive fields by increasing scales in the bottleneck, the difference between the objects is not explicitly emphasized. If the number of one object in the images is extremely larger or less, the objects with less instances will be diverted by those with larger instances. In our training stage, we built the frequency of every class and then selected images according to their frequencies in each epoch. It could alleviate the class imbalance problem to some extent; however, it was empirical and needed frequency adjustment when there are many subjects in one image. In our future work, we will investigate image data augmentation, image resampling, and weighting in loss functions to solve this limitation.

(2) The hard example mining is not incorporated into our model. Although our model achieved over 96% in the task of petrochemical equipment detection, it still exists numerous objects hard to be detected in many images. For instance, the performance of the proposed model on Luoganbeng (screw pump) and Huanreqi (heat interchanger) is remarkably worse than the other three types of equipment. One reason for this result arises from the huge viewpoint diversity of these equipment in the image data set, and the other reason is that there are much less images that contain these two types of equipment. In our future work, we will explore a new feedback module that can automatically collect the hard examples and feed them to adjust the threshold in the anchor proposal submodule for improving classification and regression accuracy.

(3) The limited size of the image data set is also important for testing the proposed model. As stated above, the specific petrochemical equipment data set is rare, and the size of the training data is relatively small in comparison to the public COCO and ImageNet dataset. This is mainly due to the difficulty of acquisition in high-risk and prohibited scenes. The annotation has also been a challenging and time-consuming subtask in our study. In addition, from the experimental and detection results, we can see that the increase in background complexity and the view of the equipment have little impact on the detection results which indicate that the added attention model and data augmentation strongly support the effectiveness of the design and training of the proposed model. In the future, we will strive on increasing the size of the data set and adding more complex background by taking more images of the working scenes and using the virtual sample generation strategy. On the other hand, the images are mainly taken from the view of workers, and the bird's-eye view could increase the variety of images. In addition, we will make attempts at the automatic or self-automatic annotation of the images to thoroughly evaluate the performance of the proposed model.

5. Conclusion

In this paper, we have presented a one-stage attention mechanism-enhanced Yolov5 network for petrochemical equipment detection in industry images. Considering the advantages of the channel and spatial attention mechanism and adaptive anchor generation to handle objects in complex background, the Yolov5 object detection model is improved by incorporating these two modules. We also made an improvement on the adaptive anchor by applying the K means++ clustering algorithm. In addition, the data augmentation strategy is also introduced to handle the relatively small sample and enhance the robustness of the fused model. The experimental results on the self-built petrochemical equipment image dataset demonstrate the competitive results of our proposed method.

Data Availability

The data set is within the protection of the enterprise privacy and safety policy. The processed data could be accessed upon request to the corresponding author.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was partially supported by the Basic Research and Strategic Reserve Technology Research Program of CNPC (No. 2017D-5008).

References

[1] J. Kang, S. Tariq, H. Oh, and S. S. Woo, "A survey of deep learning-based object detection methods and datasets for overhead imagery," *IEEE Access*, vol. 10, pp. 20118–20134, 2022.

[2] Z. Huang, S. Yang, M. Zhou et al., "Making accurate object detection at the edge: review and new approach," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2245–2274, 2022.

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

[4] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[5] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," *CoRR*, vol. 1, 2020 pages, 2004.

[7] G. Jocher, A. Stoken, and J. Borovec, *Ultralytics/yolov5: V4. 0-Nn.SiLU*, 2021.

[8] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," *Computer Vision – ECCV 2016*, vol. 9905, pp. 21–37, 2016.

[9] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional Single Shot Detector," *CoRR*, vol. 1701, 2017.

[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: keypoint triplets for object detection," in *Proceedings of the 2019 IEEE International Conference on Computer Vision*, Seoul, Korea, November 2019.

[11] S. Ren, K. He, R. Girshick, J. Sun, and R.-C. N. N. Faster, "Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, MIT Press, Cambridge, Massachusetts, United States, 2015.

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

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, June 2014.

[14] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[15] J. Liao, Y. Liu, Y. Piao et al., "GLE-net: a global and local ensemble network for aerial object detection," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, p. 2, 2022.

[16] Y. Li, G. Li, Z. Wang, Z. Han, and X. Bai, "A multifeature fusion approach for power system transient stability assessment using PMU data," *Mathematical Problems in Engineering*, vol. 2015, Article ID 786396, 10 pages, 2015.

[17] M. Tan, R. Pang, and Q. V. Le, "Efficient Det: Scalable and efficient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 2020.

[18] Y. Li, M. Zhang, and C. Chen, "A deep-learning intelligent system incorporating data augmentation for Short-Term voltage stability assessment of power systems," *Applied Energy*, vol. 308, Article ID 118347, 45 pages, 2022.

[19] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] J. Dai, Y. Li, K. He, J. Sun, and R. Fcn, "Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, pp. 379–387, MIT Press, Cambridge, Massachusetts, United States, 2016.

[21] Z. Cai, N. Vasconcelos, and R.-C. N. N. Cascade, "Delving into high quality object detection," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[22] H. Qiu, Y. Ma, Z. Li, S. Liu, and J. Sun, "BorderDet: border feature for dense object detection," in *Proceedings of the 16th European Conference on Computer Vision*, Glasgow, UK, August 2020.

[23] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, PR, USA, June 2021.

[24] S. Zhang, L. Wen, Z. Lei, and S. Z. Li, "RefineDet++: single-shot refinement neural network for object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 674–687, 2021.

[25] W. Zhan, C. Sun, M. Wang et al., "An improved Yolov5 real-time detection method for small objects captured by UAV," *Soft Computing*, vol. 26, no. 1, pp. 361–373, 2022.

[26] H. Law and J. Deng, "CornerNet: detecting objects as paired keypoints," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, September 2018.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, September 2018.

[28] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[30] S. Duan, N. Lu, Z. Lyu, G. Liu, and B. Cao, "An anchor box setting technique based on differences between categories for object detection," *Int. J. Intell. Robot. Appl.*vol. 6, no. 1, pp. 38–51, 2022.

[31] I. Goicovich, P. Olivares, C. Roman et al., "Fiber clustering acceleration with a modified K means algorithm using data parallelism," *Frontiers in Neuroinformatics*, vol. 15, Article ID 727859, 78 pages, 2021.