Ganster R-CNN: Occluded Object Detection Network Based on Generative Adversarial Nets and Faster R-CNN

KELEI SUN, QIUFEN WEN, AND HUAPING ZHOU
School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China
Corresponding author: Huaping Zhou (13805549155@163.com)

ABSTRACT Object detection has shown noticeably rapid improvement, despite most existing methods still scrabbling in occluded object detection. In response to this problem, this paper proposes a method for occluded object detection called Ganster R-CNN, which is on a basis of improved Generative Adversarial Nets (IGAN) and Faster R-CNN to enhance the detection ability of occluded objects. IGAN combines the generator of Generative Adversarial Nets and the detector of Faster R-CNN. By considering the lack of diversity of information in the feature maps, we first integrated feature maps from the shallow layer to the deep layer using Feature Pyramid Network. Next, the generator can generate occluded fake samples, and the scale of the training samples and the proportion of occluded objects in the dataset are expanded. Thus, the precision rate of occluded objects can be improved. Thus, the adversarial learning strategy can improve the detection ability of Faster R-CNN detector. Experiments show that compared with Faster R-CNN, this method achieves an improvement of +10.3 AP on the MS COCO dataset, and the mean average precision of this method is improved by 4.31% on the VOC2007 dataset and 3.92% on the VOC2012 dataset. Compared with classically existing models on PASCAL VOC datasets and some Transformer based models on MS COCO dataset, this method improves the average precision value and the mean average precision value of occluded objects.

INDEX TERMS Adversarial learning strategy, feature maps, feature pyramid network, generative adversarial nets.

I. INTRODUCTION Object detection is a challenging task in computer vision, and occluded targets are easily missed. Objects occluded by obstacles or other objects may lose important information during the detection. In recent years, several methods have been proposed for generating occluded objects. Keserwani and Roy [1] proposed a method to accurately erase specified words in images to occlude some private information in a dataset. Dervies [2] discarded parts of a feature before inputting it into the network to enhance robustness, which is called Cutout. Based on their example, we propose the generation of occluded samples by Generative Adversarial Nets [3](GAN) and sending real and fake samples into the detector of Faster R-CNN.

Responding to the problem of unsuitable feature maps for occlusion, this paper proposes a module that combines Region Proposal Network [4](RPN) and Feature Pyramid Network as an RFPN. With regard to research on object detection, an existing work called TridentNet [5] developed an object detection network using the concept of a scale-aware Trident. TridentNet proposed a method of three weights shared branches in training and used only one branch in testing to reduce the calculation. This model effectively
detected objects of different size, but failed to enhance the detection performance of occluded targets. In contrast to this method, for the sake of fusing more detailed information, such as semantic information, location information, spatial information, and other detailed information, into feature maps, Ganster R-CNN combines the RPN and FPN as the RFPN module. We then take such feature maps as input into the generator to generate occluded realistic samples. Thus, we can enhance the detection ability of occluded objects.

To obtain more occluded samples for training and testing, we used a generator of improved Generative Adversarial Nets (IGAN) to improve the proportion of occluded objects. Then, we take samples as input into the detector of Faster R-CNN. In the first step, the ResNet-101 network was chosen as the backbone network to obtain the feature maps. In recent years, residual networks [6](ResNet), which are affected by deep learning, have been universally used in feature extraction. R-FCN uses ResNet-101 as the backbone network, and Atrous Convolution replaces a part of the original convolution layers in ResNet-101. Owing to Atrous Convolution, R-FCN can output feature maps including a larger scale of images than the original feature maps. In classification, while objects keep moving in the images, ResNet-101 can always classify them accurately. However, ResNet-based detection models have limited abilities in accurately detecting occluded objects, and ResNet-101 cannot accurately detect occluded objects. In summary, improving the precision rate of occluded objects is crucial for enhancing detection capability.

Considering this problem, this paper proposes a new method for occluded objects based on IGAN and Faster R-CNN, so we name the method Ganster R-CNN. The name is a combination of IGAN and Faster R-CNN. This method aims to enhance the precision rate of occluded objects that are easily missed during the detection. In contrast to RPN in Faster R-CNN, Ganster R-CNN first combines RPN and FPN as RFPN module. The advantage of this displacement is that we can integrate feature maps from low to high levels and output richer feature maps with semantic and spatial information. Furthermore, Ganster R-CNN contains the IGAN module, which is combined with the generator of the GAN and the detector of Faster R-CNN. This process increases the size of dataset and improves the precision rate of the occluded objects via adversarial learning. Fig. 1 illustrates the construction of the Ganster R-CNN.

Our study mainly consists of the following three aspects.

1) We added FPN into RPN module of Faster R-CNN to integrate the feature maps of various scales extracted by convolutional neural networks. Subsequently, high semantic information from high-level feature maps and high-resolution information from low-level feature maps can be integrated into new multi-integration feature maps. These multi-integration feature maps are then input into the generator.

2) We constructed a new module with two branches, called IGAN, to generate fake samples with occlusions and enhance the detection ability of the model. IGAN applies the concept of an adversarial learning strategy between a generator in GAN and a detector in Faster R-CNN. The generator adds a mask to the feature maps, synthesizes them with the original images, and inputs them together with the original feature maps into the detector.

3) We combined RFPN and IGAN as Ganster R-CNN to enhance the detection ability of occluded objects. We experimented on three public datasets to verify the validity of Ganster R-CNN.

The Ganster R-CNN can obtain feature maps that include more semantic, spatial, and other detailed information, as well as generate fake samples with occlusion. Subsequently, we can enhance the precision rate of detection, particularly the success rate of objects with occlusions. Compared with the feature maps from Faster R-CNN, we obtain richer and high-resolution feature maps using RFPN. These feature maps can be better detected and help the generator generate more realistic samples. The construction of the IGAN enables the generator and detector to learn from each other, thereby improving the detection ability of the Ganster R-CNN. With regard to the size of the dataset, the method presented in this paper has witnessed advantages in enlarging the dataset during the period of generation and reducing the risk of overfitting.

II. RELATED WORKS

In recent years, feature fusion and context information-based algorithms have been applied to fully utilize multiscale feature maps to reduce overfitting and enhance the detection abilities of algorithms.

A. FEATURE BLENDING

There is a large body of literature on feature blending methods. The FTT [7] module is designed to obtain more detailed information from the extended Feature Pyramid Network. Guided-Anchorings [8] is aimed at guiding the generation of anchors through feature maps. Widely used feature fusion and feature enhancement methods include the integration of feature maps extracted by different convolution layers (e.g., D-DETR [9], NAS-FPN [10], and Qu et al.[11]) and rich semantic information of feature maps through semantic segmentation branching and global activation module(DES [12] and FCOS [13]). CE-FPN [14], D-DETR [9], and AugFPN [15] are aimed at resolving the low detection accuracy of multi-scale objects. CBNet [16] integrates high-resolution and low-resolution features from different backbone networks. Tian et al. [13] combine FCOS, which contains FPN and the centerness branch, to obtain more semantic information and reduce the number of parameters. Many scholars have used the FPN [17] as the core to build object detection models, including Ren et al. [18], to combine the FPN and SSD. Inspired by the subpixel FCOS [13] to obtain more semantic information, AugFPN [15] makes the most use of multi-scale features, NAS-FPN [10] fuses
features from different regions, and CE-FPN [14] outputs information across scales.

B. CONTEXT INFORMATION INTEGRATION
Several papers have proposed methods for using scene context information and relationships between objects in an image to enlarge the detection area (SIN [19] and ThunderNet [20]). DSSD [21] incorporates additional contextual information into the calculation to improve the detection ability. Parallel multi-branch architectures (DetectoRS [22] and Trident [5]) have received increasing attention for efficient and accurate visual recognition. Each branch in Trident has the same conversion parameters. Pica et al. [23] construct pixel-to-pixel contextual attention. Chen et al. [24] fuse detailed information from a low level, semantic information from a high level, and contextual information from a global level to improve detection completeness.

Inspired by these methods, we combined FPN and Faster R-CNN to obtain multi-scale feature maps that contain more detailed information. We then input these feature maps into a generator to obtain more truthful images with occlusions.

III. GANSTER R-CNN
In this section, the proposed method, named Ganster R-CNN, consists of two modules: RFPN and IGAN. To enhance the classification ability of Faster R-CNN, we combine RPN and FPN as an RFPN module, and use an adversarial learning strategy and the method of feature mapping to construct the IGAN module. For anchors, we used three scales with box areas of $128^2$, $256^2$, and $512^2$ pixels and three aspect ratios of 1:1, 1:2, and 2:1. Hence, we can enhance the average precision of detection, particularly the success rate of objects with occlusions.

A. GANSTER R-CNN
Our method for object detection, named Ganster R-CNN, mainly consists of two modules. The first component is RFPN, which deals with the unbalanced distribution information problem of feature maps by connecting feature maps of different scales. The second component is IGAN, integrated by the generator of the GAN and the detector of the Faster R–CNN to generate occluded samples and train the detection ability toward occluded objects. Therefore, the name Ganster R-CNN is combined with IGAN and Faster R-CNN.

In the first step, we input the feature maps of the original images into the RFPN to obtain the real samples. Real samples contain more semantic information from high-level feature maps and more location information from low-level feature maps. In the second step, real samples are input into the generator. After convolution, pooling, and residual operations, a new feature sequence is generated. The fake sample is composed of a new feature sequence and an original feature map. In the third step, authentic and imitative samples are placed into the detector. The detector verifies whether the sample is real or fake and outputs the verifying results with the object box marked. The network structure of the Faster R-CNN that integrates RFPN and IGAN(Ganster R-CNN) is shown in Fig. 1.

Generator and detector iteratively learn from each other to improve the authenticity of imitative images and the detection capability of the detector to enhance the detection capability of objects with occlusion.

B. RFPN
The first step of Ganster R-CNN is feature extraction. The detection ability of Ganster R-CNN depends on the quality of the feature maps. Therefore, we combine RPN and FPN as RFPN module to integrate feature maps from various scales. Then we can improve the quality of the feature maps and fuse more detailed information. Feature maps from a low level include the location and detailed information. High-level feature maps contain rich semantic information. The feature fusion structure of the RFPN module is illustrated in Fig. 2.

We name the output of the last residual blocks $\{C1, C2, C3, C4, C5\}$ for conv1, conv2, conv3, conv4, and conv5 outputs. The top-down pathway hallucinates higher resolution features by upsampling spatially coarser features, and contains more semantic information. These features are combined with those from the bottom-up pathway by lateral connections. We iterated the process of merging the upsampled
maps and the corresponding bottom-up maps by element-wise addition. We add a $3 \times 3$ convolution to each merged map to obtain the final feature map, called $P_1, P_2, P_3, P_4, P_5$. $P_5$ is downsampled to form $P_6$ that is only for RPN. We retain the $P_1$ feature map that contains more detailed information and is more suitable for detecting smaller objects. We form $\{P_1, P_2, P_3, P_4\}$ as in (1), $P_5$ as in (2), and $P_6$ as in (3).

\[
P_x = R(H(P_{x+1}) + T(C_x)) \quad (1)
\]
\[
P_5 = R(T(C_5)) \quad (2)
\]
\[
P_6 = Y(P_5) \quad (3)
\]

In (1), $x$ represents level1 to level4.

$Y(x)$ means we downsample the spatial resolution by a factor of 0.5.

$H(x)$ means we upsmaple the spatial resolution by a factor of 2 (using nearest neighbor upsampling for simplicity).

$T(x)$ means we use a $1 \times 1$ convolutional layer to reduce channel dimensions.

$R(x)$ means we append a $3 \times 3$ convolution on the merged map to generate the final feature map, which reduces the aliasing effect of upsampling.

RFPN supports bottom-up, top-down, and lateral connections. Therefore, after inputing a single-scale image, we can output feature maps at multiple levels. RFPN integrates location information from high-resolution lower-level feature maps and semantic information from higher-level feature maps. After improving feature extraction and integrating different-scale feature maps, we can enhance the ability and efficiency of the feature extraction of the Ganster R-CNN.

C. IGAN

Occluded objects hold a small proportion of the datasets and are easily missed during object detection. Enlarging the proportion of occluded objects is the first step to enhance the detection ability of Ganster R-CNN for occluded objects. Therefore, we constructed a module named IGAN composed of a GAN generator and Faster R-CNN detector. The IGAN generates fake occluded images to increase the number and proportion of occluded objects. Subsequently, the generator and detector in the IGAN can learn from each other to enhance the detection capability of the Ganster R-CNN.

The generator is an important component of Generative Adversarial Nets (GAN). It samples noise variables from a random uniform distribution and outputs fake samples. The discriminator outputs the probability rate value of the input sample as authentic or imitative. After training, the imitative samples were identical to the authentic samples, which misled the discriminator.

IGAN combines a generator in GAN and a detector in Faster R-CNN. The generator design provides a method for increasing the size of the occluded targets in the dataset.

We input noise variables sampled from a random uniform distribution and the feature map of real sample $x$ into the generator. Then, the generator outputs a fake sample $z$, and we define the distributions $P_{\text{data}}$ over $x$ and $P_z$ over $z$. The value function of IGAN is expressed as follows:

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim P_z(z)}[\log(1 - D(G(z)))] \quad (4)
\]

In (4), the generator is represented by $G$. $G$ learns how to map the distribution of fake samples $P_z(z)$ to $P_{\text{data}}(x)$. The detector is represented by $D$. $D$ calculates the rate value from the distribution of data $x$. Maximizing the probability of the detector error is the ultimate goal of the generator training process. Using this adversarial learning strategy, we can generate fake occluded samples and improve the detection ability of the Ganster R-CNN toward occluded objects.

Mathematically, it is assumed that $X$ is an object candidate generated by the RFPN, and the object detection network is $F(X)$. The Faster R-CNN detector outputs the category, width, height, and coordinates of the bounding box of the objects. The category is represented by $F_c$, and the location of the bounding box is represented by $F_l$. We assume that $C$ is the correct category of $X$ and $L$ is its spatial location. The loss function of the detector is defined as:

\[
L_F = L_{\text{softmax}}(F_c(X), C) + |C \neq \text{bg}| L_{\text{bbox}}(F_l(X), L) \quad (5)
\]

In (5), $L_{\text{softmax}}(F_c(X), C)$ represents the SoftMax loss of the classification. $|C \neq \text{bg}| L_{\text{bbox}}(F_l(X), L)$ is the loss of predicting the location of the bounding box and accurately annotating box location for only foreground classes.

Assume that $A(X)$ represents the generator of IGAN. $A(X)$ outputs a feature $X$ according to the input image. We input feature $X$ into the generator to obtain a new fake sample $A(X)$.

The new $A(X)$ is placed in the detector for training. In addition, IGAN must learn to predict unsuccessful features. IGAN then strengthens the abilities of the generator and detector. The loss function of the training IGAN is expressed as:

\[
L_A = 1 - L_{\text{softmax}}(F_c(X), C) \quad (6)
\]

Without enlarging the training dataset, the generator outputs fake samples by learning the distribution of occluded authentic samples, hence, after training, the imitative samples are difficult to identify by the detector. An adversarial learning strategy improves the detection capability of the detector in Ganster R-CNN.

We input real samples into the generator, including high-resolution and low-resolution feature maps, between which the generator learns different eigenvalues. Then the generator generates new eigenvalues that are similar to the original eigenvalues. We form a high-resolution feature map by adding new eigenvalues to a low-resolution feature map. After training, the distribution of fake samples is closer to that of real samples. A flowchart of the feature mapping of IGAN is shown in Fig. 3.

To address the problem of a low proportion of occluded objects, we add a mask to form the fake-occluded samples. After the RoI pooling layer, the generator generates
mask from the convolutional features and learns uncommon occluded invariance. The mask can be automatically adjusted according to the loss function to create occlusion. The mask determines the part of the value in the discarded feature map. The structure of IGAN is shown in Fig. 4.

During training, real and fake samples are placed in the detector. Then the detector determines whether the samples are real or fake. The detector outputs 1 for authentic samples and 0 for imitative samples, as well as the categories and locations of the objects. The generator and the detector can learn from each other. Network training can then learn uncommonly occluded objects and enhance the classification and detection ability of the detector.

IV. EXPERIMENTS

We propose a Ganster R-CNN based on RFPN and IGAN, and evaluate it on PASCAL VOC 2007, 2012 datasets and MS COCO 2017 dataset. We primarily evaluated the detection rate of average precision (AP) and mean average precision (mAP), because these are the actual measurements in object detection.

A. TEST GANSTER R-CNN AND FASTER R-CNN ON VOC2007 AND MS COCO DATASETS

The AP value is the average precision of the detection category in a training session. The higher the AP value, the better is the performance of the algorithm. After training for 10 epochs on PASCAL VOC 2007 and MS COCO datasets, the AP values for 20 categories by Faster R-CNN and Ganster R-CNN are shown in the radar map in Fig. 5. Each concentric circle in the radar map in Fig. 5 is an AP value from 0 to 100 from inside to outside.

Fig. 5 shows that the orange circle surrounded by Ganster R-CNN is larger than the blue circle surrounded by Faster R-CNN. Ganster R-CNN combines multiple feature maps and adopts adversarial learning strategy to generate occluded images and improve the classification capability of the detector. After training for 10 epochs, the AP values of 14 categories of objects increased. Both Ganster R-CNN and Faster R-CNN have better feature extraction and detection classification ability for large objects than tiny objects in the detection field. For example, in object detection of animals, the detection accuracy of large animals such as dog, horse, sheep, and cow are 15% to 25% higher than that of small animals such as birds and cats. Generally, the detection performance of Ganster R-CNN on the VOC2007 dataset is better than Faster R-CNN.

And the AP values for 80 categories in MS COCO datasets by Faster R-CNN and Ganster R-CNN are shown in Fig. 6.

Fig. 6 shows that the orange line represented by Ganster R-CNN is higher than the blue line represented by Faster R-CNN. Ganster R-CNN learns from real images and produces fake occluded images to enlarge training datasets. Thus, Ganster R-CNN gets more sufficient training for detection. Ganster R-CNN has better feature fusion and detection capability for large objects than tiny objects in the detection field. For example, in target detection of animals, vehicles, tableware and fruits, the detection accuracy of large animals and vehicles such as elephant, giraffe, train, truck and airplane are 30% to 45% higher than that of tableware and fruits such as fork, knife spoon and apple. However, after training for 10 epochs, Ganster R-CNN still badly performs in detecting small targets such as fork, knife and spoon. Generally, the detection performance of Ganster R-CNN on the MS COCO dataset is better than Faster R-CNN.

B. COMPARISON AND ANALYSIS OF MAP VALUE OF FOUR ALGORITHMS ON VOC2007 DATASET

The mAP value is another indicator for evaluating the capability of algorithm. The mAP is the mean average precision of all detection categories during training. The model was trained for 10 epochs. The higher the mAP value, the better is the performance of the algorithm. The mAP values of Faster R-CNN, RFPN(combined with FPN and Faster R-CNN), IGAN(combined with the generator of GAN and detector of Faster R-CNN), and Ganster R-CNN(combined with RFPN and IGAN) on the dataset VOC 2007 are shown in Fig. 7.
TABLE 1. Comparison of mAP values on the VOC2007 test set of detection algorithms.

| Method   | Backbone | Training data      | Testing data   | mAP(%)  |
|----------|----------|--------------------|----------------|---------|
| Faster R-CNN | ResNet-101 | VOC2007+2012       | VOC2007        | 76.40   |
| R-FCN     | ResNet-101 | VOC2007+2012       | VOC2007        | 80.50   |
| SSD513    | ResNet-101 | VOC2007+2012       | VOC2007        | 80.60   |
| RFPN      | ResNet-101 | VOC2007+2012       | VOC2007        | 80.61   |
| IGAN      | ResNet-101 | VOC2007+2012       | VOC2007        | 80.64   |
| Ganster R-CNN | ResNet-101 | VOC2007+2012       | VOC2007        | 80.71   |

Fig. 7 demonstrates that the red broken line of Ganster R-CNN is generally higher than those of the other three algorithms. Among the other three algorithms, the higher the accuracy of the broken line of the Ganster R-CNN, the better is the capability of the Ganster R-CNN classification. RFPN is combined with FPN and Faster R-CNN to integrate feature maps from various scales. Thus, we can improve the quality of feature maps and provide more detailed information. Trained for 10 epochs, RFPN obtains a higher mAP value than Faster R-CNN.

IGAN is combined with the generator of GAN and the detector of Faster R-CNN. We use an adversarial learning strategy to improve the detection capability of the detector of Faster R-CNN. IGAN generates fake occluded images by learning the distribution of occluded authentic samples to increase the number and proportion of occluded objects. Thus, we obtain more occluded objects for training. Trained for 10 epochs, IGAN obtains a higher mAP value than Faster R-CNN and RFPN.

The greater the mean average precision rate, the more accurate is the model prediction. Ganster R-CNN has the advantages of RFPN and IGAN. After 10 epochs of training, the other three methods reach the fitting state, whereas the Ganster R-CNN has not yet reached the fitting state. It has been proven that Ganster R-CNN has better detection performance than the other three algorithms. It integrates the features of each layer using bottom-up, top-down, and lateral connections. In summary, we are able to improve the classification ability of the detector. The mAP value of Ganster R-CNN exceeds that of the other three algorithms.

C. TEST OF RESNET-101 BASED NETWORK ON PASCAL VOC AND MS COCO DATASETS

To verify that the Ganster R-CNN has good detection performance, we train and test other algorithms in the same environment. Tables 1 and 2 illustrate the comparison results of the mAP values of Ganster R-CNN, Faster R-CNN [4], SSD513 [25], and R-FCN [26] on the PASCAL VOC 2007 and 2012 datasets.

Table 1 shows that on the VOC2007 dataset, compared with Faster R-CNN, RFPN has an mAP of 80.61%, which is an increase of 4.21%. IGAN has an mAP of 80.64%, which is an increase of 4.24%. Ganster R-CNN has an mAP of 80.71%, which is an increase of 4.31%.

Table 2 shows that, on the VOC2012 dataset, compared with the Faster R-CNN, RFPN has an mAP of 76.23%, which is increased by 2.43%. IGAN has an mAP of 77.06%,
TABLE 2. Comparison of mAP values on the VOC2012 test set of detection algorithms.

| Method     | Backbone | Training data | Testing data | mAP(%) |
|------------|----------|---------------|--------------|--------|
| Faster R-CNN | ResNet-101 | VOC2007+2012  | VOC2012      | 73.80  |
| R-FCN      | ResNet-101 | VOC2007+2012  | VOC2012      | 77.60  |
| SSD513     | ResNet-101 | VOC2007+2012  | VOC2012      | 75.40  |
| RFPN       | ResNet-101 | VOC2007+2012  | VOC2012      | 76.23  |
| IGAN       | ResNet-101 | VOC2007+2012  | VOC2012      | 77.06  |
| Ganster R-CNN | ResNet-101 | VOC2007+2012  | VOC2012      | 77.72  |

which is increased by 3.26%. Ganster R-CNN has an mAP of 77.72%, which is increased by 3.92%.

Ganster R-CNN fuses the features of each layer through bottom-up, top-down, and lateral connections. We adopt adversarial learning to generate occluded images and enhance the detection capability of the detectors.

The reasons for the poor results from Faster R-CNN, SSD, and R-FCN are as follows. First, the single feature map extracted by the original Faster R-CNN cannot cover all the features. Using classification scores and NMS to filter overlapping candidate boxes may filter out parts of the multiple boxes of overlapping objects, resulting in missed detection. Second, SSD uses a large number of anchors and outputs of multiple convolution layers for classification and regression to improve recognition accuracy. The SSD weighs the speed against the accuracy. Third, compared with the Ganster R-CNN, the R-FCN has a deeper shared convolution layer, which can obtain more abstract features. Consequently, a fully convolutional network was constructed. The R-FCN weighs the speed against the accuracy.

To verify that the Ganster R-CNN shows well detection performance on MS COCO 2017 dataset, we test some Transformer based algorithms in the same environment. Table 3 illustrates the comparison results of the AP values of Ganster R-CNN, Faster R-CNN [4], FCOS [13], DETR [27], D-DETR [9] and DINO [28] on MS COCO dataset.

Table 3 shows that on MSCOCO dataset, compared with Faster R-CNN, RFPN yields an improvement of +1.8 AP, IGAN yields an improvement of +8.0 AP, and Ganster R-CNN achieves an improvement of +10.3 AP. As shown in Table 3, Ganster R-CNN achieves the best results of 50.5 AP on MSCOCO val2017 dataset, which demonstrates its strong object detection capabilities. In MSCOCO 2017 dataset, Faster R-CNN holds the speed of 21 FPS, and Ganster R-CNN holds the speed of 18.6 FPS. In the RFPN module, we retain the P1 feature map that contains more detailed information and is more suitable for detecting smaller objects. So we weigh the performance against speed and space.

We reconstruct the Faster R-CNN by adding FPN to fuse feature maps from different levels and the generator of GAN to generate fake occluded objects. Therefore, we can get more training data and get more sufficient training for detector of Ganster R-CNN. After that, we can enhance the detection ability of occluded objects of Ganster R-CNN.

Transformer based models require a larger training dataset and longer training time to achieve better results. Transformer based object detection algorithms contain attention modules that fuse context information globally. Simultaneously, the similarity between same-level-features increases, which decreases the capabilities of Transformer based algorithms.

TABLE 3. Comparison of AP values on the MSCOCO 2017 dataset of detection algorithms.

| Method     | Dataset   | AP   | AP_{50} | AP_{75} | AP_{S} | AP_{M} | AP_{L} |
|------------|-----------|------|---------|---------|--------|--------|--------|
| Faster R-CNN | MSCOCO 2017 | 40.2 | 61.0    | 43.8    | 24.2   | 43.5   | 52.0   |
| FCOS       | MSCOCO 2017 | 41.0 | 59.8    | 44.1    | 26.2   | 44.6   | 52.2   |
| DETR       | MSCOCO 2017 | 42.0 | 62.4    | 44.2    | 20.5   | 45.8   | 61.1   |
| D-DETR     | MSCOCO 2017 | 43.8 | 62.6    | 47.7    | 26.4   | 47.1   | 58.0   |
| DINO       | MSCOCO 2017 | 50.4 | 68.3    | 54.8    | 33.3   | 53.7   | 64.8   |
| RFPN       | MSCOCO 2017 | 42.0 | 62.5    | 45.9    | 25.2   | 45.6   | 54.6   |
| IGAN       | MSCOCO 2017 | 48.2 | 65.4    | 53.1    | 30.9   | 52.8   | 63.4   |
| Ganster R-CNN | MSCOCO 2017 | 50.5 | 68.7    | 55.2    | 33.0   | 54.2   | 65.4   |
Therefore, Ganster R-CNN outperforms the DETR and D-DETR.

In our experiments, we found that the AP and mAP of Ganster R-CNN are higher than those of the other algorithms. The results in Tables 1, 2, and 3 indicate that the Ganster R-CNN performs better on PASCAL VOC datasets and MSCOCO dataset. Some of the test results of the Ganster R-CNN on PASCAL VOC and MS COCO datasets are shown in Fig. 8 and Fig. 9.

V. CONCLUSION

To detect occluded objects more accurately, feature maps must contain more detailed semantic information. Moreover, the detector must be trained on a dataset that contains more occluded objects to enhance its ability to detect objects with occlusion. According to the current concept of the Feature Pyramid Network, the Ganster R-CNN builds an adversarial learning network around the Faster R-CNN. Ganster R-CNN consists of two modules. The first component is RFPN, which deals with the unbalanced distribution information problem using feature fusion by connecting feature maps of different levels. The second component is IGAN, which is integrated by the generator of the GAN and the detector of the Faster R-CNN. Without enlarging the training dataset, IGAN improves the detection capability of the Ganster R-CNN in detecting occluded objects. After RoI pooling, low-resolution feature maps were used to form new high-resolution feature maps to improve the detection capability of the Ganster R-CNN. Experiments show that adversarial learning and feature mapping can enhance the classification ability and solve the unbalanced distribution information problem of feature maps at different levels. The feature work includes further analysis of the mask generated by the generator. For example, we can study the relationship between mask rate and detection ability.

REFERENCES

[1] P. Keserwani and P. P. Roy, “Text region conditional generative adversarial network for text concealment in the wild,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 5, pp. 3152–3163, May 2022, doi: 10.1109/TCSTVT.2021.3103922.
[2] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” 2017, arXiv:1708.04552.
[3] I. Goodfellow, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–6.
[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[5] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, “Scale-aware trident networks for object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6064–6063.
[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[7] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, “Extended feature pyramid network for small object detection,” IEEE Trans. Multimed., vol. 24, pp. 1968–1979, 2022.
[8] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, “Region proposal by guided anchoring,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2965–2974.
[9] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” 2020, arXiv:2010.04159.
[10] G. Ghiasi, T.-Y. Lin, and Q. V. Le, “NAS-FPN: Learning scalable feature pyramid architecture for object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7036–7045.
[11] J. Qu, C. Su, Z. Zhang, and A. Razi, “Dilated convolution and feature fusion SSD network for small object detection in remote sensing images,” IEEE Access, vol. 8, pp. 82832–82843, 2020, doi: 10.1109/ACCESS.2020.2991439.
[12] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, “Single-shot object detection with enriched semantics,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 5813–5821.
[13] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9627–9636.
[14] Y. Luo, X. Cao, J. Zhang, J. Guo, H. Shen, T. Wang, and Q. Feng, “CE-FPN: Enhancing channel information for object detection,” Multimed. Tool. Appl., vol. 81, pp. 30685–30704, Apr. 2022, doi: 10.1007/s11042-022-11940-1.
[15] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, “AugFPN: Improving multi-scale feature learning for object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12595–12604.
[16] Y. Liu, “CBNet: A novel composite backbone network architecture for object detection,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, 2020, pp. 11653–11660, doi: 10.1609/aaai.v34i07.6834.
[17] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2117–2125.
[18] K. Ren, L. Huang, C. Fan, H. Han, and H. Deng, “Real-time traffic sign detection network using DS-DetNet and lite fusion FPN,” J. Real-Time Image Process., vol. 18, no. 6, pp. 2181–2191, Dec. 2021, doi: 10.1007/s11554-021-01102-1.
[19] Y. Liu, R. Wang, S. Shan, and X. Chen, “Structure inference net: Object detection using scene-level context and instance-level relationships,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6985–6994.
[20] Z. Qin, Z. Li, Z. Zhang, Y. Bao, G. Yu, Y. Peng, and J. Sun, “ThunderNet: Towards real-time generic object detection on mobile devices,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6718–6727.
[21] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DSSD: Deconvolutional single shot detector,” 2017, arXiv:1701.06659.
[22] S. Qiao, L.-C. Chen, and A. Yuille, “DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 10213–10224.
[23] N. Liu, J. Han, and M.-H. Yang, “PiCANet: Learning pixel-wise contextual attention for saliency detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[24] Z. Chen, Q. Xu, and R. Cong, “Global context-aware progressive aggregation network for salient object detection,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 10599–10606, doi: 10.1609/aaai.v34i07.6633.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[26] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 379–387.

[27] N. Carion, “End-to-end object detection with transformers,” in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 213–229.

[28] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, “DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection,” 2022, arXiv:2203.03605.

**KELEI SUN** was born in Huainan, Anhui, in 1980. He is a Professor and a Doctor with the School of Computer Science and Engineering, Anhui University of Science and Technology. His main research interests include target detection and recommendation systems.

**QIUFEN WEN** was born in Huzhou, Zhejiang, in 1999. She is currently pursuing the master’s degree with the School of Computer Science and Engineering, Anhui University of Science and Technology. Her main research interests include object detection and multiple object tracking.

**HUAPING ZHOU** was born in Tanghe, Henan, in 1979. She is a Professor and a Doctor with the School of Computer Science and Engineering, Anhui University of Science and Technology. Her main research interests include machine learning and target detection.