IECE
Chinese Journal of Information Fusion
http://dx.doi.org/10.62762/CJIF.2025.845143

RESEARCH ARTICLE

EFSOD: Enhanced Feature based Small Object Detection Network in Remote Sensing Images

Jiawei Yi[1,2], **Ying Liu**[1,2], **Yanshan Li**[1,2,*] **and Weixin Xie**[1,2]

[1] Institute of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China
[2] Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China

Abstract

Due to the poor imaging quality of remote sensing images and the small size of targets, remote sensing small target detection has become a current research difficulty and hotspot. Recent years have seen many new algorithms. Remote sensing small target detection methods based on image super-resolution reconstruction have attracted many researchers due to their excellent performance. However, these algorithms still have problems such as weak feature extraction capability and insufficient feature fusion. Then, we propose Enhanced Feature based Small Target Detection Network in Remote Sensing Images (EFSOD), which includes a Edge Enhancement Super-Resolution Reconstruction Module (EESRM) and a Cross-Model Feature Fusion Module (CMFFM). EESRM enhances the recognizability of small target contours by fusing extracted edge features with the original features through residual connections, alleviating the constraints of feature blurring on detection performance. CMFFM achieves deep integration of the detailed features extracted by the EESRM network with the semantic features extracted by the target detection network, improving the model's sensitivity and accuracy in recognizing small targets in complex backgrounds. Additionally, considering the effects of blurring, noise, illumination changes, and atmospheric scattering on remote sensing images, a remote sensing image degradation simulation algorithm is proposed. This algorithm realistically simulates the generation process of low-resolution remote sensing images under natural conditions, providing more realistic training and testing data. The experimental results show that the proposed EFSOD significantly enhances the performance of small object detection in remote sensing.

Keywords: remote sensing images, super-resolution, small object detection, cross-model feature fusion.

1 Introduction

In recent years, with the development of deep neural networks, methods based on deep learning have achieved remarkable performance in various visual tasks, and object detection is one of the

Academic Editor:
Tiancheng Li

Submitted: 12 March 2025
Accepted: 21 April 2025
Published: 27 April 2025

Vol. 2, **No.** 2, 2025.
10.62762/CJIF.2025.845143

***Corresponding author:**
✉ Yanshan Li
lys@szu.edu.cn

Citation
Yi, J., Liu, Y., Li, Y., & Xie, W. (2025). EFSOD: Enhanced Feature based Small Object Detection Network in Remote Sensing Images. *Chinese Journal of Information Fusion*, 2(2), 127–143.

key problems in computer vision. Due to the limitations of the environment and remote sensing image acquisition technology, remote sensing images often have characteristics such as low resolution and high noise, which are detrimental to the detection of small targets in remote sensing. Among the current mainstream algorithms, small target detection technology based on image super-resolution reconstruction has attracted the attention of many researchers.

Current methods for detecting small targets in low-resolution remote sensing images mainly follow the "reconstruct first, then detect" technical paradigm. Typical methods include attention mechanism models based on single-image reconstruction [1–4], multimodal spectral fusion frameworks, and cross-scale feature migration networks [5–8]. However, there are three main limitations in existing methods: First, decoupling the reconstruction process from the detection task leads to a mismatch in feature representation. Super-resolution networks focus on optimizing PSNR metrics while neglecting high-frequency edge features that are crucial for detection. Second, multi-stage architectures cause feature redundancy, as separate reconstruction and detection modules require repeated feature extraction, leading to redundant consumption of system storage space and computational resources. Third, there is insufficient degradation modeling; existing methods often use simple bicubic downsampling to simulate image degradation, failing to fully consider actual physical constraints such as illumination, atmospheric scattering, and sensor noise.

In this paper, We present an EFSOD to address the above issues. Its innovation is reflected in two aspects: First, an edge enhancement module is designed to effectively alleviate the problem of detail loss by introducing abundant high-frequency edge features, thereby improving the accuracy and precision of target detection. Second, a systematic analysis of the intrinsic correlation between super-resolution and target detection models in terms of feature representation is conducted, and a multi-layer feature fusion mechanism is proposed. This mechanism achieves complementary and synergistic optimization of both types of features through cross-modal feature interaction, significantly enhancing detection performance.

Our main contributions can be summarized into three folds:

- We introduce the EFSOD. The network effectively enhances the accuracy of small target detection without increasing complexity.
- We propose a collaborative architecture of edge enhancement super-resolution reconstruction module and cross-model feature interaction mechanism. The former enhances the edge features of targets through an edge-enhanced dense residual network. The latter addresses the issues of detail loss and feature mismatch by integrating the complementarity and consistency of super-resolution features and target detection features.
- We propose a remote sensing image degradation method. This method simulates the impact of lighting and atmospheric conditions on the remote sensing imaging process through Gaussian blurring, adding noise, downsampling, and simulating illumination and atmospheric scattering. This significantly enhances the model's robustness in complex real-world scenarios.

2 Related Work

Small object detection is an important task in computer vision. It focuses on identifying target objects that are smaller in size and have a lower pixel proportion in images or videos. Current methods proposed for small object detection mainly include multi-scale representation [9–12], contextual information [13–16] , region proposals [17–20], and image super-resolution methods [21–23].

In object detection networks based on image super-resolution, the method addresses the issue of small objects covering few pixels by utilizing generative adversarial networks (GAN) [24] to transform low-resolution original images into higher-resolution versions, thereby implementing object detection on these high-definition images. Li et al. [25] proposed a target detection method based on two stages, which realized the framework of automatic detection and search of potential target regions. Krishna et al. [26] proposed a task-driven super-resolution method that combines low-level image processing with high-level visual objectives. Perceptual GAN [27] enhances the representation of small objects to super-resolution representations, providing more substantial discernment capabilities. To obtain more features, Bai et al. [28] introduced image-level super-resolution on candidate boxes for small objects in their SOD-MTGAN. EESRGAN

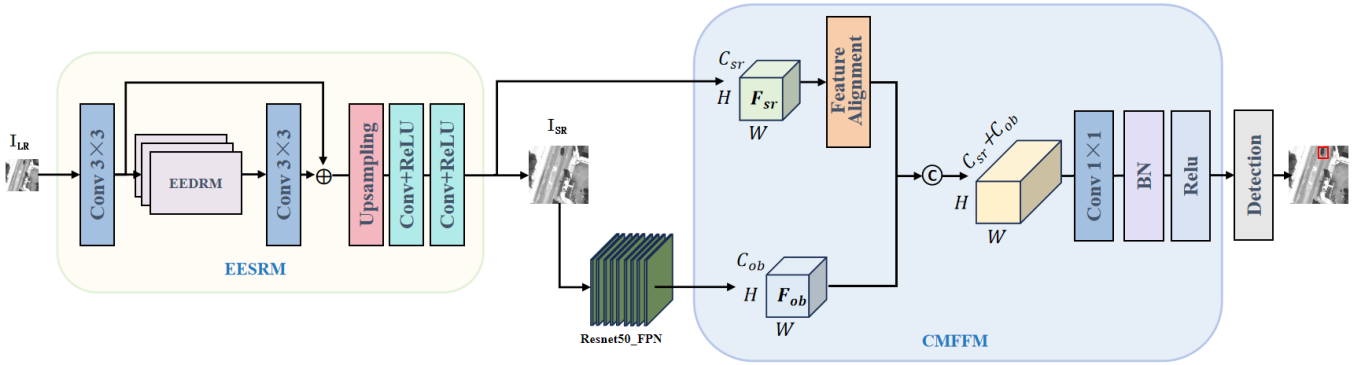

Figure 1. The structural diagram of the proposed EFSOD.

employs ESRGAN [29] in the generator to obtain intermediate super-resolution images (ISR) and uses EEN to enhance edge features, resulting in the final SR images that are inputted into the detection network for detection.

Existing small object detection methods based on super-resolution reconstruction have achieved considerable advancements. However, these methods still have some shortcomings. 1) Most approaches rely on complex super-resolution generators or adversarial networks, leading to high computational costs and difficulty meeting real-time detection requirements. 2) The target features of super-resolution models are easily disrupted in noisy and complex images, which may affect detection accuracy. 3) Most current models employ separate super-resolution and detection networks, failing to utilize the complementary relationship between their features fully. Therefore, this paper considers enhancing edge features and cross-model feature fusion for low-quality remote sensing images to improve the accuracy of small object detection while maintaining the same computational complexity.

3 Methodology

3.1 Model Architecture

This paper proposes the EFSOD to address the shortcomings of existing small object detection algorithms based on super-resolution reconstruction. The EESRM and the CMFFM improve image quality and enhance target edge features, increasing detection accuracy. EFSOD shares information at the feature level, allowing super-resolution reconstruction results to optimize the detection task more directly. The overall structure of the EFSOD network is shown in Figure 1.

In EFSOD, the generator mainly consists of the EESRM, which generates high-quality super-resolved images I^{SR} from low-resolution remote sensing images I^{LR}. To better capture and preserve edge information in the images, this paper proposes the Edge Enhanced Dense Residual Module (EEDRM), which extracts edge information from the feature maps of each layer within the residual blocks and connects these edge details through residual connections.

In the generator's super-resolution reconstruction network, the low-resolution remote sensing image I^{LR} is first processed through an initial convolutional layer $(Conv)$ to extract the basic features F_0. Then, the extracted features are applied to three EEDRM to learn deeper features within the feature maps, using residual connections to facilitate the flow of information and obtaining the feature map F_d after deep feature learning. Subsequently, the feature map is further processed through a convolutional layer $(Conv)$ to extract the fused deep features F_{deep}. Finally, the learned feature map is upsampled using pixel shuffle technology to increase the image to the target resolution. Additional convolutional layers further enhance and refine the features, and a final convolutional layer converts the feature map into a high-resolution image I^{SR}. The I^{SR} undergoes feature extraction using the convolutional neural network ResNet50 [30] to obtain the feature map.

The object detection network is responsible for detecting and identifying objects in the generated images. We propose the CMFFM module to extend the object detection component. Firstly, to ensure dimensional consistency between the feature map $F_{sr} \in \mathbb{R}^{H_s \times W_s \times C_s}$ extracted from the super-resolution reconstruction network and the feature map $F_{ob} \in \mathbb{R}^{H_d \times W_d \times C_d}$ in the object detection network, we use a Feature Alignment function $FA(\cdot)$ to adjust F_{sr} to match the size of F_{ob} as a feature map:

$$\tilde{F}_{sr} = FA(F_{sr}) \in \mathbb{R}^{H_d \times W_d \times C_d} \quad (1)$$

Then, $\tilde{F}_{sr}$ and F_{ob} are concatenated at the channel level, denoted as:

$$F_{cat} = Concat(\tilde{F}_{sr}, F_{ob}) \in \mathbb{R}^{H_d \times W_d \times (C_s + C_d)} \quad (2)$$

where H_d , W_d and C_s+C_d respectively denote the height, width, and channels of the feature maps. $\tilde{F}_{sr}$ is the super-resolution feature map after feature alignment. F_{cat} represents the feature map obtained by channel-wise concatenation and fusion of the aligned super-resolution feature map $\tilde{F}_{sr}$ with the object detection feature map Fob.

This step effectively combines the features of the two networks, where the super-resolution feature map adds more detailed information and the object detection feature map incorporates richer high-level semantic information. The number of channels in the concatenated feature map will increase, and to meet the input requirements of the subsequent layers of the object detection network, it is necessary to adjust the number of channels in the feature map using a convolution operation $Conv(\cdot)$, convert F_{cat} into a feature map with $C^{'}$ target channels.

$$F_{conv} = Conv(F_{cat}) \in \mathbb{R}^{H_d \times W_d \times C^{'}} \quad (3)$$

where $C^{'}$ respectively denotes the adjusted target number of channels to match the input requirements of the object detection network.

This convolutional layer adjusts the number of channels and helps further integrate features from two different networks. It is then followed by a batch normalization layer (BN) to standardize the features, adjusting and scaling the incoming features to promote rapid convergence of the model.

$$F_{bn} = BN(F_{conv}) \quad (4)$$

where F_{bn} respectively denotes the feature map after batch normalization, $BN(\cdot)$ denotes the batch normalization operation, standardizing the input features so that their mean is close to 0 and variance is close to 1.

Subsequently, the ReLU activation function is applied to maintain non-linear characteristics.

$$F_{relu} = RELU(F_{bn}) \quad (5)$$

where F_{relu} is the feature map after RELU activation, the ReLU activation function is applied to enhance the features' nonlinear expressive capability.

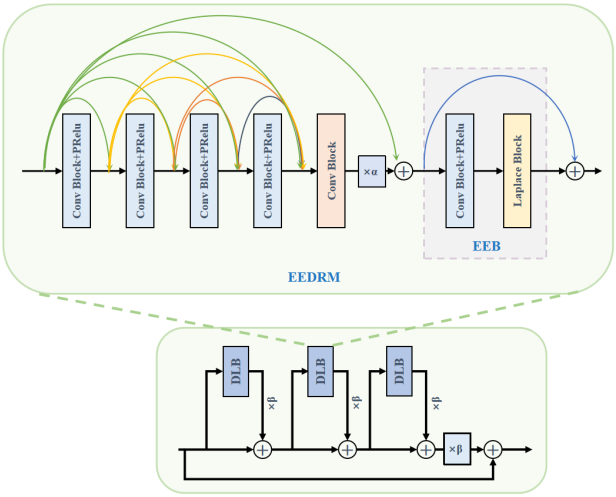

Figure 2. Diagram of the EEDRM structure.

Finally, the fused features output by the CMFFM module are:

$$F_{fusion} = F_{relu} \in \mathbb{R}^{H_d \times W_d \times C^{'}} \quad (6)$$

where F_{fusion} denotes the fused feature map output by the CMFFM.

This feature effectively integrates the detailed edge and texture information provided by the super-resolution network with the high-level semantic information from the object detection network, thereby enhancing the detection accuracy of small targets in remote sensing images.

3.2 Edge Enhancement Super-Resolution Reconstruction Module

This paper proposes the EESRM to better capture and preserve edge information in images. The EESRM primarily comprises the Edge-Enhanced Dense Residual Module (EEDRM), whose network structure is shown in Figure 2. Based on the dense residual structure, EEDRM incorporates an Edge Enhance Block (EEB) consisting of a convolution layer, an activation layer, and a Laplacian operator.

The EEB extracts high-frequency components from the feature layers extracted by the residuals using a Laplacian operator [31] and connects them with the original feature layers through residual connections. This process can be expressed as:

$$F_e^i = \alpha F^i + F_{le}^i \quad (7)$$

where F^i represents the feature map extracted by the ith dense residual module, F_{le}^i represents the edge information extracted by the edge enhancement

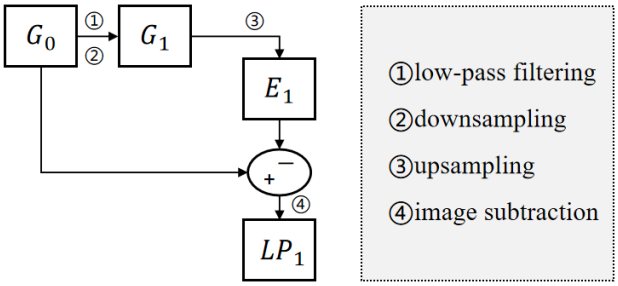

Figure 3. Structure of the Laplace Block.

module of F^i, α denotes the adjustment factor, and F_e^i represents the edge-enhanced feature layer output through residual connection.

Motivated by the Laplacian operator, The structure is illustrated in Figure 3, the original image undergoes low-pass filtering and downsampling, followed by upsampling to restore the dimensions. Finally, the original image and the upsampled result are subtracted, a process that can be expressed as:

$$L_i = G_i - PyrUp(PyrDown(G_i)) \tag{8}$$

where $PyrUp$ is the upsampling, $PyrDown$ is the downsampling.we employ the Dense Residual Laplacian Block (DLB) as the basic unit to extract features from the original image. This approach preserves more high-frequency edge information without incurring significant computational overhead, thereby improving the accuracy of small object detection. The ith DLB, the feature map F^i is obtained through dense residual feature extraction. Due to repeated convolution operations, the edge information of small objects becomes weakened, resulting in ghosting artifacts in the generated images. Therefore, we embed an additional computation for edge information extraction within each dense residual block to enhance the high-frequency details of the image better, thereby achieving a more precise and refined super-resolution reconstruction.

For the feature extraction network of the super-resolution model, a 3×3 convolutional layer is first used to extract the shallow feature map F^S from the image. This operation reduces the spatial dimensions of the feature maps, increases the receptive field of the network layers—thus enabling the capture of a broader range of features—and simultaneously decreases the volume of data subsequent layers need to process. This process can be expressed as:

$$F^S = Conv_{3\times3}(I_{LR}) \tag{9}$$

where $Conv_{3\times3}$ denotes the operation of a 3×3 convolutional layer.I_{LR} x denotes the low-resolution image.

Next, the shallow feature map F^S is passed through a series of EEDRM to extract deep features containing edge information. In this process, the number of EEDRM blocks is set to three. This module employs dense residual connections through a multi-scale residual network and an edge feature extraction module to extract rich semantic information, thereby enhancing the feature representation capability of the target feature extraction module.Simultaneously, the output of each convolutional layer is connected via residual connections with all preceding convolutional layers to prevent the loss of feature information. Furthermore, the EEDRM employs the Laplacian operator to extract edge information from the image, and this extracted edge information is then integrated through residual connections with the previously extracted features, significantly enhancing the utilization of edge features. After the series of EEDRM blocks, an additional convolutional layer is applied, and its output is added to the shallow feature map to obtain the final deep feature F^d. This process can be expressed as:

$$F^d = Conv_{3\times3}\left(EEDRM(F^i)\right) + F^S, i = 1, ..., 3 \tag{10}$$

Then, the extracted deep features undergo a feature upsampling operation to increase their resolution to the desired size. A convolutional layer extracts features while preserving spatial resolution, followed by an activation function to introduce non-linearity for learning complex features.

Finally, a second convolution and activation operation is performed to refine the target features further, enhancing detail restoration and deep feature extraction. The ultimately extracted target features F^{SR} can be expressed as:

$$F^{SR} = \sigma(Conv_{3\times3}(\sigma\left(Conv_{3\times3}\left(Upsample(F^d)\right)\right))) \tag{11}$$

3.3 Cross-Model Feature Fusion Module

To integrate super-resolution features with target detection features and obtain superior detection performance, this paper proposes the CMFFM. The core objective of this module is to effectively fuse the target features extracted by the super-resolution

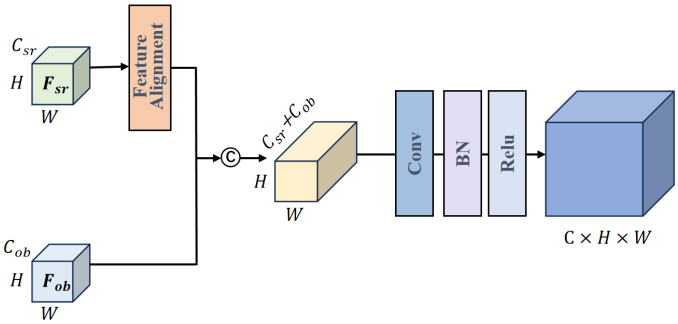

Figure 4. Cross-model feature fusion module.

reconstruction network into the feature maps extracted by the target detection network. By leveraging the detail enhancement capability of the super-resolution network and the recognition and localization expertise of the target detection network, the proposed method jointly improves detection performance. This approach introduces the super-resolution network's sensitivity to fine details without compromising the target detection network's spatial perception capability, thereby achieving effective feature fusion between the two networks and enhancing overall target detection performance.

The network structure of CMFFM is shown in Figure 4. C_{sr} represents the number of channels in the super-resolution feature map, and C_{ob} represents the number of channels in the target detection feature map. H and W denote the height and width of the feature maps, respectively. $Conv-1$ refers to a 1×1 convolutional layer, BN represents Batch Normalization, and Relu denotes the nonlinear activation function.

During the cross-model feature fusion process, the given inputs include the target features $F_{sr}\epsilon F^{C_{sr}\times H\times W}$ extracted from the super-resolution reconstruction network and the features $F_{ob}\epsilon F^{C_{ob}\times H\times W}$ extracted from the backbone of the target detection network.First, it is necessary to ensure that the target features generated by the super-resolution model match the spatial dimensions of the final output layer of the target detection backbone network. To achieve this, the size of the additional features is adjusted using bilinear interpolation to align with the dimensions of the backbone features. The proposed feature fusion network employs a series of feature fusion operations using concatenation $Concat$ to merge cross-model features along the channel dimension. This process can be expressed as:

$$F'_{multi} = Concat(F_{sr}, F_{\mathrm{ob}}) \tag{12}$$

where $Concat$ is the feature concatenation operation along the channel dimension.

Then, the merged features are transformed into the target number of channels through the 1×1 convolutional layers in CMFFM. Subsequently, batch normalization and the ReLU activation function are applied to enhance feature representation and the network's nonlinear capability. The final fused feature F_{multi} can be expressed as:

$$F_{multi} = \delta(BN\left(Conv_{1\times1}(F'_{multi})\right)) \tag{13}$$

where $Conv_{1\times1}$ is the 1×1 convolution operation, BN denotes batch normalization processing, and δ represents the application of Relu as the nonlinear activation function.

After the CMFFM processing, the fused features effectively combine multi-level information, incorporating rich high-level semantic representations. This mechanism successfully integrates critical attributes such as target categories and properties into the final feature encoding, enabling object detection models to leverage these advanced semantic cues for more precise localization and classification. Notably, the feature fusion module preserves sufficient spatial resolution throughout the fusion process, ensuring the retention of spatial fidelity at the feature level without compromising detection accuracy. The resultant fused features, enriched with comprehensive detail and semantic depth, significantly enhance the model's capability to capture subtle characteristics and small targets. By synergistically integrating heterogeneous feature sources, the framework improves the model's contextual understanding of inter-object relationships in complex scenes. It demonstrates outstanding performance advantages in detecting small targets within challenging environmental contexts.

3.4 Loss function

The network involved in this paper consists of a super-resolution network based on Generative Adversarial Networks (GANs) and a target detection network. During the training process, an end-to-end training approach is used, where the loss from the target detection is backpropagated to the generation network, guiding the generation network to reconstruct images that are more conducive to target detection.

We introduce a new loss in addition to the traditional adversarial loss, perceptual loss L_{per}, and L1 loss to prevent the edge enhancement module from over-enhancing the edges. Specifically, we use

Charbonnier loss, i.e., the consistency loss L_{char}, between the super-resolved image (SR) and the high-resolution image (HR).

$$L_{per} = E_{xf}||vgg_{fea}(x_f) - vgg(x_r)||_1 \tag{14}$$

$$L_1 = E_{xf}||x_f - x_r||_1 \tag{15}$$

$$L_{char} = \rho(I_{HR} - I_{SR}) \tag{16}$$

where E_{xf} is the average value of all generated images in a batch. At the same time, $vgg_{fea}(x_f)$ and $vgg(x_r)$ denote the feature representations extracted by the convolutional neural network when the actual image x_r and the super-resolved image (x_f are fed into the network, respectively. $\rho(\cdot)$ is Charbonnier function.

Finally, we obtain the total loss of the edge-enhanced super-resolution generation network by adding the loss of the edge enhancement module to the original generative network's loss, with the empirical value set to $\gamma_1 = 1$, $\gamma_2 = 0.001$, $\gamma_3 = 0.01$, $\gamma_4 = 5$

$$L_G = \gamma_1 L_{per} + \gamma_2 L_G^{Ra} + \gamma_3 L_1 + \gamma_4 L_{char} \tag{17}$$

where L_G^{Ra} is the adversarial loss of the generator.

In the target detection network, Faster R-CNN, regression loss, and localization loss exist for the detected objects. Both losses are computed using the smoothed L1 loss. The classification loss L_{cls}, regression loss L_{reg}, and total detection loss L_{det} can be expressed as:

$$L_{cls} = E_{I_{LR}}[-log(Det_{cls}(G_G(I_{LR})))]_1 \tag{18}$$

$$L_{reg} = E_{I_{LR}}[smooth_{L_1}(Det_{reg}(G_G(I_{LR}), T_*))]_1 \tag{19}$$

$$L_{det} = L_{cls} + \alpha L_{reg} \tag{20}$$

where T_* is the ground truth target coordinates, ($G_G(\cdot)$ denotes the super-resolution reconstruction network, Det_{cls} is the classification loss in the target detection network, Det_{reg} denotes the regression loss in target detection, $smooth_{L_1}(\cdot)$ refers to the smoothed L_1 loss, and α is the balancing parameter. In the proposed network,

The total loss of the entire discriminator L_{D_det} can be expressed as:

$$L_{D\_det} = L_D^{Ra} + \eta L_{det} \tag{21}$$

where L_D^{Ra} is the adversarial loss, and η is the balancing parameter for the discriminator loss, which measures the contribution of the target detection network to the discriminator. Based on empirical experience, η is set to 1.

Finally, the overall loss of the entire network architecture, $L_{overall}$, can be expressed as:

$$L_{overall} = L_G + L_{D\_det} \tag{22}$$

4 Experiments

This section designs an image degradation method to simulate actual remote sensing image formation to validate the effectiveness of the proposed EFSOD. This ensures that the super-resolution target detection model has better generalization ability and adapts more effectively to real-world scenarios. First, we conduct comparative experiments between EFSOD and baseline networks. Then, we compare the EFSOD algorithm with other super-resolution target detection methods. Finally, an ablation analysis is performed on the proposed modules.

4.1 Dataset

Currently, the two mainstream datasets in the field of super-resolution target detection are the COWC [32] and RSOD [33] datasets. The COWC (Cars Overhead With Context) dataset consists of satellite images collected from six different geographic locations, with an image size of 256×256 pixels. The average target length ranges from 24 to 48 pixels, while the width ranges from 10 to 20 pixels. This dataset focuses solely on small targets of the "car" category and includes 3954 images for training and testing. 3164 images were randomly selected as the training set, while 790 were used as the test set. The RSOD dataset includes four types of objects: airplanes, sports fields, overpasses, and oil tanks. These objects exhibit diverse characteristics and have targets of varying sizes. The experiments divided this dataset into training and test sets in an 8:2 ratio.

4.2 Degradation processing of remote sensing image

Most methods for small target detection in super-resolution remote sensing rely on downsampling to create low-resolution datasets without considering the imaging characteristics of remote sensing images in practical scenarios. This leads to models that perform well on specific datasets but fail to generalize well on others, undermining their robustness and generalization ability. Figure 5 shows remote sensing image degradation methods, including blurring, downsampling, noise, and image compression.

Before model training, to obtain paired high/low-resolution images, it is necessary to

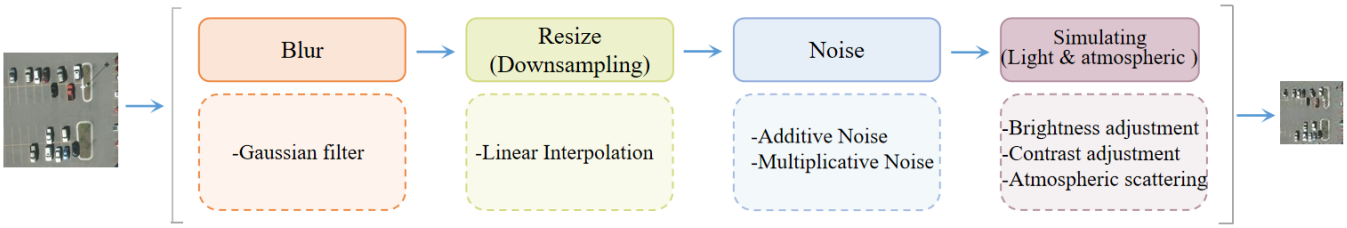

Figure 5. remote sensing image degradation module.

preprocess the images in the dataset. A crucial part of this process is degrading the high-resolution remote sensing images to simulate the real remote sensing image generation environment. Let $I^{HR} \in \mathcal{R}^{C_{in}\times H\times W}$ and $I^{LR} \in \mathcal{R}^{C_{in}\times H\times W}$ represent the initial high-resolution image and its degraded low-resolution version, respectively. C_{in}, H, and W correspond to the number of channels, height, and width of the input image.

First, Gaussian blurring is applied to the input high-resolution image I^{HR}. The blur kernel is based on a Gaussian probability density function with zero mean to achieve $N(0,\Sigma)$, resulting in $I^{HR}_{blur} \in \mathcal{R}^{C_{in}\times H\times W}$. This process can be represented as:

$$I^{HR}_{blur} = G_{blur}(I^{HR}) \tag{23}$$

where $G_{blur}(\cdot)$ is the Gaussian blur function.

Secondly, considering that remote sensing images are inevitably affected by noise during the generation process, this paper focuses on two types of noise: additive noise in optical imaging systems and multiplicative features in remote sensing image imaging systems. Additive and multiplicative features are sequentially added to the blurred image. This process can be expressed as:

$$I^{HR}_{noise} = (I^{HR}_{blur} + N_1) \times N_2 \tag{24}$$

where N_1 is additive noise, which follows a normal distribution $X\sim N(0,\sigma^2)$. N_2 represents multiplicative features, where each image pixel is multiplied by a gain factor N_2. This gain factor follows an exponential distribution with unit mean $N_2\sim Exp(1)$.

The resolution of the image is reduced through downsampling to simulate the situation of a low-resolution remote sensing image, resulting in the low-resolution remote sensing image $I^{LR}_I \in R^{C_{in}\times H\times W}$. This process can be expressed as:

$$I^{LR}_I = D(I^{HR}_{noise}) \tag{25}$$

where $D(\cdot)$ represents the downsampling operation applied to the image.

Finally, the image parameters are adjusted to simulate remote sensing images $I^{LR}_B \in \mathcal{R}^{C_{in}\times H\times W}$ under different sunlight conditions. This process can be expressed as:

$$I^{LR}_B = \alpha(I^{LR}_I + \Delta I - \bar{I}) + \bar{I} \tag{26}$$

where ΔI is the brightness adjustment factor, α is the contrast adjustment factor, and $\bar{I}$ is the average brightness of the image.

In addition, to account for light attenuation caused by atmospheric scattering and the increase in ambient light due to scattering, the image after atmospheric scattering processing is the final low-resolution remote sensing image I^{LR}. The effect of atmospheric scattering on the image can be expressed as:

$$I^{LR}(x) = I^{LR}_B(x)\cdot e^{-\beta d(x)} + A\cdot(1-e^{-\beta d(x)}) \tag{27}$$

where $I^{LR}(x)$ is the image brightness at position x after the influence of atmospheric scattering, $I^{LR}_B(x)$ is the brightness of the original image at position x, β is the atmospheric scattering coefficient, $d(x)$ is the distance at position x in the scene, and A is the ambient light intensity.

Through the above degradation processing, this paper simulates more realistic low-resolution remote sensing images in the data processing, providing a solid foundation for the subsequent training and testing of the network.

4.3 Experimental Setup

For the images in the dataset, this paper first processes the original images using the proposed degradation method to obtain low-resolution images, forming high/low-resolution image pairs. During training, the low-resolution and high-resolution images are fed into the network for training. The parameter settings are as follows: the batch_size is set to 5, the learning rate for the super-resolution reconstruction network is initialized to 500, and the Adam optimizer parameters are set to $\delta_1 = 0.9$ and $\delta_2 = 0.99$. For the object detection network, the initial learning rate

is set to 0.005, and the momentum parameter of the SGD optimizer is set to 0.9. The network is trained in an end-to-end manner. During the testing phase, only the low-resolution images must be input for super-resolution object detection without requiring high-resolution images.

This paper uses the MS COCO [22] evaluation matrix to evaluate the object detection results. It selects AP^{50}, AP^S, AP^M, AP^L, AR, AR^S, AR^M and AR^L as the verification criteria for the effectiveness of the object detection experiments to assess the performance of the degradation model.

4.4 Performance Evaluation

This paper evaluates the proposed EFSOD's object detection performance on remote sensing images from both subjective and objective perspectives. Considering the specific characteristics of remote sensing images, the original images are processed with blurring, downsampling, noise, simulated lighting changes, and atmospheric scattering when generating low-resolution images. In the comparative experiments with the baseline networks, ESRGAN+FRCNN and EESRGAN networks are compared with the proposed network, evaluating the networks from both object detection metrics and parameter/computation complexity aspects. EESRGAN adopts an end-to-end structure similar to our method. Additionally, this paper compares current mainstream object detection algorithms such as Faster RCNN++ [34], RetinaNet [35], CenterNet [36], and PP-YOLOv2 [37], as well as remote sensing small target detection algorithms like AVDNet [38], the target feature super-resolution vehicle detection algorithm TGFSR-VD [39], and ASAHI [40]. All of the above algorithms are based on the official code provided, and experiments are conducted with consistent experimental parameters as described in the papers, trained, and tested on the COWC and RSOD datasets.

4.4.1 Validity experiment of remote sensing image degradation method

In order to verify the effectiveness of the degradation model, this section is based on the COWC data set, and the models proposed in this chapter are trained on the image data set after bicubic downsampling and the image data set after degradation method processing, respectively, then the test is performed uniformly on the low-quality degraded images that simulate the actual application scenarios. The experimental results are shown in Table 1.

As can be seen from the experimental results in Table 1, the model trained with the degradation data set, the Peak Signal-to-Noise Ratio (PSNR) is improved by 3.02, and the target detection accuracy is only selected as the evaluation parameter in this section. It can be seen that the target detection-related indicators have been greatly improved, including 29.9%, 32%, and 34%. The experimental results show that compared with the simple bicubic downsampling method, the degradation method adopted in this chapter can better simulate factors such as illumination and atmospheric scattering in the practical application of remote sensing images.This method can more accurately simulate the performance of images under different lighting conditions, including light intensity, shadow change, etc., which makes the model more robust and generalization ability in training. At the same time, considering the influence of atmospheric scattering and other phenomena on the image, the degradation method can also simulate the influence of these factors on image details and contrast, and provide more prosperous and more realistic training samples for the model. Through this detailed degradation simulation, the generated low-resolution image is closer to the actual remote sensing image, which makes the trained model more robust and reliable in the real scene. This method provides important support for improving the robustness and stability of the model and helps the model achieve better performance in practical applications.

Therefore, in the subsequent experiments, this paper uniformly adopts the degraded image data set as the low-resolution data set, and by using these low-resolution data that are closer to the real situation, we can improve the accuracy of image processing, models can be better trained to adapt to changes and challenges in real-world environments. Such training data helps to improve the generalization ability and robustness of the model, making it more robust and reliable in actual scenarios.

4.4.2 Comparative Experiments of EFSOD with Baseline

To verify the small target detection performance of EFSOD, this section conducts comparative experiments between the baseline networks ESRGAN+FRCNN, EESRGAN, and EFSOD on the COWC and RSOD datasets, respectively. The experimental results are shown in Tables 2 and 3.

As shown in Table 2, on the COWC dataset, EFSOD significantly improves all object detection metrics compared to ESRGAN+FRCNN. Specifically, in terms

Table 1. Experimental results of the degradation method.

Method	Train/test	PSNR	AP^{50}	AP^S	AP^M	AP^L
EFSOD	Bicucic/Degradation	28.18	57.10	0.409	0.443	-
EFSOD	Degradation/Degradation	31.20	87.00	0.731	0.783	-

Table 2. Comparative experimental results of EFSOD, ESRGAN+FRCNN, and EESRGAN on the COWC dataset.

Method	AP^{50}	AP^S	AP^M	AP^L	AR	AR^S	AR^M	AR^L
ESRGAN+FRCNN	0.843	0.709	0.744	-	0.715	0.729	0.786	-
EESRGAN	**0.893**	0.744	0.779	-	0.768	0.771	0.815	-
EFSOD	0.892	**0.773**	**0.841**	-	**0.799**	**0.801**	**0.847**	-

Table 3. Comparative experimental results of EFSOD, ESRGAN+FRCNN, and EESRGAN on the RSOD dataset.

Method	AP^{50}	AP^S	AP^M	AP^L	AR	AR^S	AR^M	AR^L
ESRGAN+FRCNN	0.859	0.720	0.753	0.799	0.722	0.734	0.781	0.825
EESRGAN	0.902	0.771	0.806	0.839	0.763	0.781	0.822	0.858
EFSOD	**0.913**	**0.779**	**0.854**	**0.865**	**0.808**	**0.820**	**0.861**	**0.877**

Table 4. Comparison of experimental results between parameter quantity and computational quantity.

Method	Param	FLOPs
ESRGAN+FRCNN	**53.51**	120.32
EESRGAN	106.69	217.41
EFSOD	60.60	**106.86**

of accuracy, the most critical metric, AP^{50}, increases by 5.8%, while the accuracy for small object detection, AP^S, improves by 9.0%. The average accuracy for medium-sized objects, AP^M, sees the highest increase of 13.0%. Since the COWC dataset does not contain large objects, the corresponding experimental results are denoted as "-". In terms of recall, EFSOD also demonstrates notable improvements. The overall recall rate, AR, increases by 11.7%, while the average recall for small objects, AR^S, improves by 9.9%, and the average recall for medium-sized objects, AR^M, increases by 7.8%.

Compared to the EESRGAN network, although the overall object detection accuracy decreases by 0.1%, EFSOD performs well in other metrics. Specifically, small object detection AP^S accuracy increases by 3.9%, while the average accuracy of medium-sized objects AP^M achieves the highest improvement of 7.9%. In terms of recall, EFSOD also shows significant enhancements. The overall recall rate AR increases by 4.0%, while the average recall for small objects AR^S improves by 3.89%, and the average recall for medium-sized objects AR^M increases by 3.9%. As shown in Table 3, compared to ESRGAN+FRCNN, EFSOD achieves the highest values across all metrics. Specifically, AP^{50} reaches 91.3%, the average accuracy for small objects AP^S reaches 77.9%, the average accuracy for medium-sized objects AP^M reaches 85.4%, and the average accuracy for large objects AP^L reaches 86.5%. Similarly, recall performance is significantly improved. The overall recall rate AR reaches 80.8%, the average recall for small objects AR^S reaches 82.0%, the average recall for medium-sized objects AR^M reaches 86.1%, and the average recall for large objects AR^L reaches 87.7%. EFSOD also achieves promising results on the RSOD datase.

The proposed EFSOD effectively integrates features from the super-resolution reconstruction network into the object detection task, enhancing the utilization of super-resolution features. This allows the network structure of the super-resolution reconstruction-based object detection algorithm to be more compact, achieving efficient use of super-resolution features in object detection. Consequently, it has a significant positive impact on the accuracy of small object detection.

Considering the existence of resource constraints in practical applications, the number of model parameters (Params) and Floating Point Operations (FLOPs) are added as evaluation indicators in this paper, where the number of parameters reflect the size and complexity of the model, and the FLOPs reflects the complexity of the model, the higher the number of parameters is, the stronger the expression ability

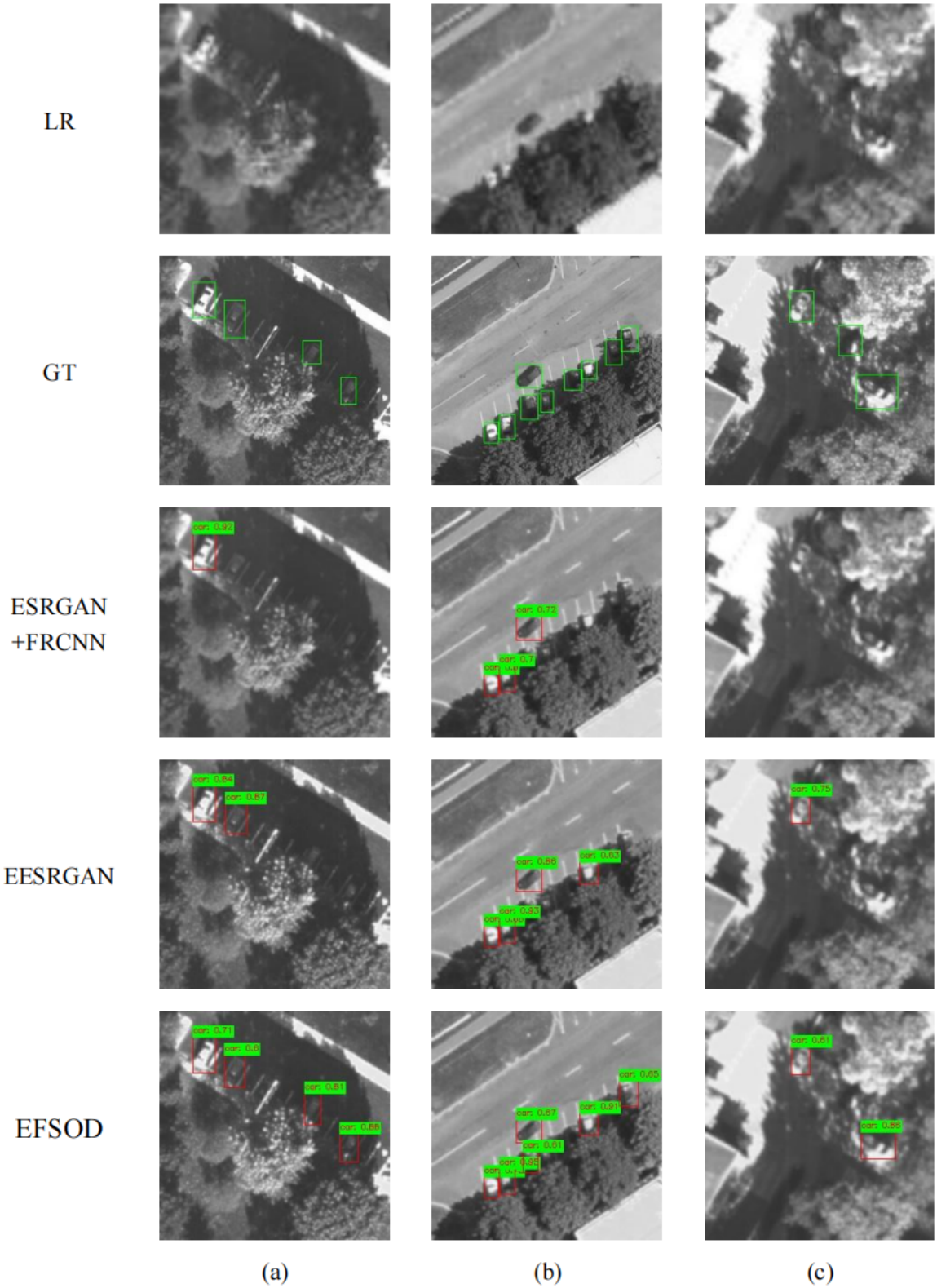

Figure 6. Representative test results based on the COWC dataset.

of the model is, and the more computing resources are required, which affects the storage requirements and model complexity. The computational amount measures the computational resources required for a forward propagation of the model, that is, the number of floating-point operations required for a single inference. Flops are usually used to measure the computational efficiency and speed of the model. The experimental results are shown in Table 4, because the EESRGAN increases a parameter

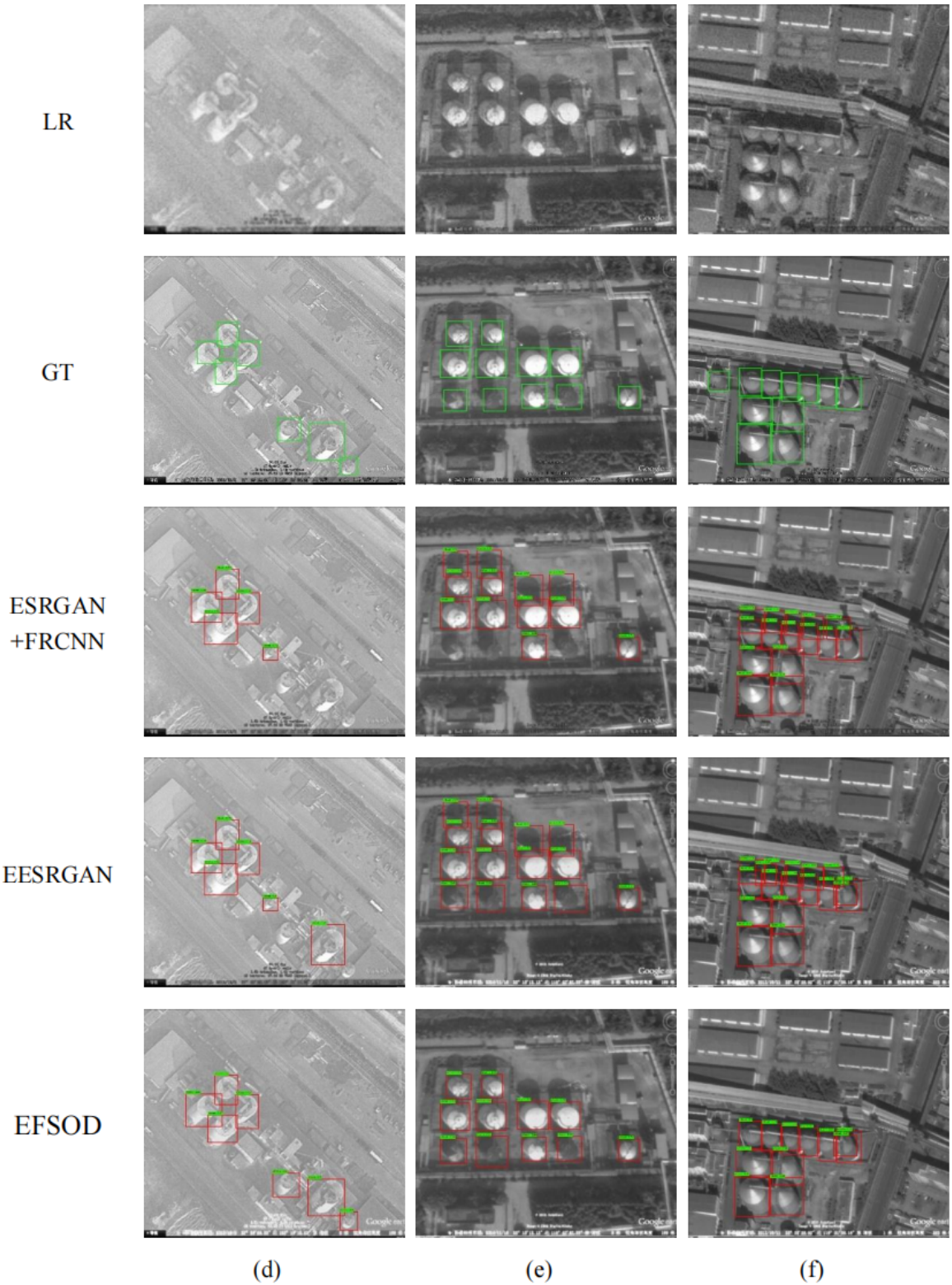

Figure 7. Representative test results based on the RSOD dataset.

amount and computational complexity relative to the ESRGAN+FRCNN network, although the parameter amount of the model is reduced relative to EFSOD. The model is more efficient than the ESRGAN+FRCNN network, however, the required network parameters and computing resource consumption increase significantly. In real applications, this increase may hinder the effective deployment and operation of models in resource-constrained environments due to the limitations of storage capacity and computing speed. Compared with EESRGAN, the EFSOD network reduces the number of parameters by 46.09

m and the amount of calculation by 110.55 GFLOPS, which can effectively extract the edge information of the image without increasing the number of parameters and the amount of calculation, to improve network performance.

In addition, EFSOD compared to the benchmark network ESRGAN+FRCNN, the number of parameters of the model increased by 7.09 m due to the addition of the edge-enhanced dense residual block EEDRM, which shows that EEDRM does not increase too many parameters and computational burden, and it can be used to improve the accuracy of the model, is a relatively lightweight edge enhancement module. In addition, the floating-point operation required by the algorithm per inference is reduced by 13.46 GFLOPS, which is because ESRGAN+FRCNN uses the FasterRCNN algorithm that comes with torchvision and uses a pre-trained model, which can be used to improve the performance of the algorithm, as a result, the amount of parameters and calculation of the model in the target detection network FasterRCNN is relatively large, while EFSOD uses the improved FasterRCNN and does not use the pre-training model, which reduces the calculation of the network.

This section conducts an effectiveness analysis of EFSOD on three representative images selected from the COWC and RSOD datasets, as shown in Figures 6 and 7, to further verify and analyze its effectiveness. In the COWC dataset, the selected images contain partially occluded objects. In contrast, in the RSOD dataset, the chosen images feature objects that are difficult to distinguish from the background due to contrast factors. A comparative analysis is performed on these two types of images, and the detection visualization results are presented. Figures 6 and 7 consist of three columns from left to right: (a), (b), and (c) in one figure, and (d), (e), and (f) in the other, each representing one of the three selected representative images. Vertically, there are six rows corresponding to: the low-resolution images obtained after degradation processing, ground-truth, ESRGAN+FRCNN, EESRGAN, EESR-SOD and EFSOD. The confidence score of the detected objects is displayed in the upper right corner of the bounding boxes, with values ranging from 0 to 1.

Figures 6 and 7 demonstrate that ESRGAN+FRCNN fails to detect small objects with low contrast against the background or partial occlusion due to its lack of edge information extraction and utilization of other super-resolution features, as shown in Figures 6(a) and 7(e). EESRGAN, benefiting from the presence of an edge enhancement network, is more sensitive to object edges and can detect some low-contrast objects, as illustrated in Figures 6(b), 6(c), 7(e), and 7(f). However, due to the absence of other super-resolution feature applications, it fails to detect corresponding objects in images with complex lighting and significant noise, such as 6(c).In contrast, the proposed EFSOD integrates high-frequency edge information, which is crucial for object detection, and incorporates super-resolution object features extracted from the super-resolution network into the object detection feature map. As a result, EFSOD achieves better performance in scenarios with low contrast and high noise.

In summary, EFSOD significantly improves feature extraction and feature enhancement capabilities. While maintaining the parameter count, it achieves enhancements across various object detection metrics, with particularly noticeable improvements in small object detection. Specifically, EFSOD exhibits more substantial recognition capabilities, enabling better identification of small objects in low-contrast and high-noise scenarios.

4.4.3 *Comparison of EFSOD and typical target detection algorithms*

To validate the model's performance, this section presents a comparative analysis of EFSOD against other state-of-the-art general object detection algorithms and specialized small object detection algorithms for the remote sensing domain. The detailed results are shown in Tables 4 and 6. The comparison includes the following methods: Faster RCNN++ [34], RetinaNet [35], CenterNet[36], PP-YOLOv2[37], AVDNet[38], TGFSR-VD[39], and ASAHI [40].

Tables 5 and 6 show that EFSOD demonstrates outstanding performance on the COWC and RSOD datasets. The highest values among all compared methods are highlighted in bold, while the second-best values are underlined. On the COWC dataset, EFSOD achieves the highest scores across all metrics, with AP^{50} reaching 89.2%, AP^{S} at 77.3%, and AP^{M} at 84.1%, reflecting its accuracy in object detection tasks. In terms of recall, AR reaches 79.9%, exceeding the second-best value by 5%, while AR^{S} reaches 80.1%, highlighting its significant advantage in small object recall tasks. Additionally, AR^{M} reaches 84.7%, emphasizing its superior performance in object detection. Since the COWC dataset does not contain

Table 5. Comparison of EFSOD and typical target detection models in COWC dataset.

Method	AP^{50}	AP^S	AP^M	AP^L	AR	AR^S	AR^M	AR^L
FasterRCNN	0.652	0.476	0.526	-	0.542	0.535	0.632	-
RetinaNet	0.610	0.489	0.588	-	0.502	0.511	0.600	-
CenterNet	0.693	0.575	0.622	-	0.587	0.605	0.669	-
PP-YOLOv2	0.751	0.595	0.731	-	0.691	0.711	0.755	-
AVDNet	0.719	0.598	0.604	-	0.652	0.616	0698	-
TGFSR-VD	0.851	0.714	0.739	-	0.714	0.727	0.779	-
ASAHI	0.866	0.763	0.759	-	0.744	0.701	0.773	-
EFSOD	**0.892**	**0.773**	**0.841**	-	**0.799**	**0.801**	**0.847**	-

Table 6. Comparison of EFSOD and typical target detection models in RSOD dataset.

Method	AP^{50}	AP^S	AP^M	AP^L	AR	AR^S	AR^M	AR^L
FasterRCNN	0.655	0.547	0.625	0.700	0.622	0.555	0.632	0.659
RetinaNet	0.702	0.586	0.645	0.693	0.663	0.566	0.676	0.690
CenterNet	0.711	0.591	0.680	0.720	0.612	0.695	0.709	0.719
PP-YOLOv2	0.817	0.750	0.769	0.798	0.749	0.758	0.790	0.822
AVDNet	0.762	0.602	0.709	0.752	0.630	0.709	0.721	0.759
TGFSR-VD	0.876	0.745	0.774	0.805	0.756	0.759	0.806	0.818
ASAHI	0.882	0.751	**0.785**	0.811	0.769	0.770	0.808	0.811
EFSOD	**0.913**	**0.758**	0.781	**0.865**	**0.808**	**0.820**	**0.861**	**0.877**

large objects, the corresponding results are marked as "-". On the RSOD dataset, EFSOD continues to perform exceptionally well, achieving AP^{50} of 91.3%, AP^S of 75.8%, AP^M of 78.1%, and AP^L of 86.5%, further confirming its outstanding performance in object detection tasks. In terms of recall, AR reaches 80.8%, AR^S reaches 82.0%, AR^M reaches 86.1%, and AR^L reaches 87.7%, demonstrating EFSOD's overall superior capability. The EFSOD network proposed in this paper achieves the best overall experimental results in small target detection.

The main reasons are as follows: First, the proposed Edge-Enhanced Dense Residual Module effectively extracts edge information of small objects, facilitating their distinction from the background and reducing the interference of lighting and color variations in detection. Second, the Edge-Enhanced Super-Resolution Reconstruction Module possesses intense feature extraction and enhancement capabilities, effectively extracting features beneficial to object detection tasks. This module also identifies and leverages the correlation between super-resolution and object features. Finally, the Cross-Model Feature Fusion Module in the object detection network effectively integrates the extracted super-resolution object features with object detection features, significantly improving detection performance. By deeply integrating the super-resolution reconstruction network with the object detection network, the EFSOD framework fully utilizes and efficiently fuses features beneficial to object detection from the super-resolution network. This approach offers a novel perspective for small object recognition based on super-resolution reconstruction, significantly improving detection accuracy and recall rates.

4.5 Ablation analysis

To verify the proposed CMFFM and EEDRM contribution to target detection accuracy, this section designs four comparative experiments on the baseline ESRGAN+FRCNN network using two different datasets, COWC and RSOD. Tables 7 and 8 show the performance comparison of each module on target detection.

From the performance shown in Tables 7 and 8, in the COWC dataset, both the CMFFM and EEDRM modules contribute significantly to improving the target detection accuracy of the baseline network ESRGAN+FRCNN. Regarding various metrics, the contribution of CMFFM is more significant than that of EEDRM. Specifically, in terms of accuracy, the most critical metric, AP^{50} improves by 5.8%. In contrast, the small target detection accuracy, an important metric in this paper, AP^S, improves by 9.0%, and AP^M improves by 13.0%. There is also a substantial

Table 7. Ablation experiment of the CMFFM module on the COWC dataset.

Method	CMFFM	EEDRM	AP^{50}	AP^S	AP^M	AP^L	AR	AR^S	AR^M	AR^L
ESRGAN+FRCNN	×	×	0.843	0.709	0.744	-	0.715	0.729	0.786	-
	✓	×	0.875	0.744	0.778	-	0.754	0.761	0.81	-
	×	✓	0.87	0.731	0.783	-	0.756	0.759	0.806	-
EFSOD	✓	✓	**0.892**	**0.773**	**0.841**	-	**0.799**	**0.801**	**0.847**	-

Table 8. Ablation experiment of the CMFFM module on the RSOD dataset.

Method	CMFFM	EEDRM	AP^{50}	AP^S	AP^M	AP^L	AR	AR^S	AR^M	AR^L
ESRGAN+FRCNN	×	×	0.859	0.720	0.753	0.799	0.722	0.734	0.781	0.825
	✓	×	0.907	0.750	0.788	0.834	0.753	0.772	0.820	0.865
	×	✓	0.895	0.758	0.781	0.826	0.745	0.761	0.815	0.851
EFSOD	✓	✓	**0.913**	**0.758**	**0.781**	**0.865**	**0.808**	**0.820**	**0.861**	**0.877**

improvement in recall, with AR increasing by 11.7%, AR^S by 9.8%, and AR^M by 7.7%. In the RSOD dataset, similar to the experimental results in the COWC dataset, both modules, when used individually, effectively enhance the target detection performance compared to the baseline network ESRGAN+FRCNN. However, fully integrating the edge-enhanced dense residual block and the cross-model feature fusion module leads to better performance. Regarding accuracy, when $\sigma = 0.5$, AP^{50} increases by 6.3%, AP^S by 5.3%, AP^M by 3.7%, and AP^LL by 8.3%. In terms of recall, AR increases by 11.9%, AR^S by 11.7%, AR^M by 10.2%, and AR^LL by 6.3%.

Introducing the CMFFM module in the target detection network dramatically enhances the feature extraction and fusion capabilities, significantly improving overall performance compared to the baseline network. The success of this improvement can be attributed to the effective combination of two key factors: First, the edge-enhanced super-resolution reconstruction network effectively restores high-frequency edge details of the image and fully exploits the relevant features beneficial for target detection within the super-resolution reconstruction network, significantly improving the quality of super-resolution target features. Second, the CMFFM module precisely and effectively integrates these excellent super-resolution target features with target detection-related features, further enhancing the performance of the target detection feature extraction network while maintaining the superior features of the super-resolution reconstructed image. This fusion strategy not only helps improve the accuracy of small target detection but also enhances the robustness and generalization ability of the network under challenging conditions such as complex backgrounds and lighting variations.

5 Conclusion

To address the issue of insufficient feature utilization in super-resolution object detection networks, we propose the EFSOD. The core advantage of this network lies in its ability to fully exploit the consistent features between the super-resolution reconstruction network and object detection, providing intense feature extraction and enhancement capabilities. The EESRM enhances the edge information of the reconstructed image, solving the problem of unclear, small target edges that lead to detection difficulties, thereby improving the detection accuracy of small targets. The CMFFM also enables the effective fusion of the edge-enhanced super-resolution reconstruction network with the object detection network. Experimental results show that the EFSOD network has strong feature extraction capabilities. It performs well in complex scenes and scenarios with dense, small target distributions.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62076165 and under Grant 62471317; in part by the Innovation Team Project of Department of Education of Guangdong Province under Grant 2020KCXTD004; in part by the Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Fu, C. Y., Liu, W., Ranga, A., Tyagi, A., & Berg, A. C. (2017). Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

[2] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing. [CrossRef]

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[4] Singla, K., Pandey, R., & Ghanekar, U. (2022). A review on Single Image Super Resolution techniques using generative adversarial network. *Optik, 266*, 169607. [CrossRef]

[5] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

[6] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems, 28*.

[7] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems, 29*.

[8] Li, W., Li, W., Yang, F., & Wang, P. (2019, July). Multi-scale object detection in satellite imagery based on YOLT. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (pp. 162-165). IEEE. [CrossRef]

[9] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

[10] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 764-773).

[11] Wei, X., Li, Z., & Wang, Y. (2025). SED-YOLO based multi-scale attention for small object detection in remote sensing. *Scientific Reports, 15*(1), 3125. [CrossRef]

[12] Chen, Y., Yuan, X., Wang, J., Wu, R., Li, X., Hou, Q., & Cheng, M. M. (2025). YOLO-MS: rethinking multi-scale representation learning for real-time object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [CrossRef]

[13] Tang, X., Du, D. K., He, Z., & Liu, J. (2018). Pyramidbox: A context-assisted single shot face detector. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 797-813).

[14] Shen, W., Qin, P., & Zeng, J. (2019). An indoor crowd detection network framework based on feature aggregation module and hybrid attention selection module. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 0-0).

[15] Zhao, Z., Du, J., Li, C., Fang, X., Xiao, Y., & Tang, J. (2024). Dense tiny object detection: A scene context guided approach and a unified benchmark. *IEEE Transactions on Geoscience and Remote Sensing, 62*, 1-13. [CrossRef]

[16] Wei, W., Cheng, Y., He, J., & Zhu, X. (2024). A review of small object detection based on deep learning. *Neural Computing and Applications, 36*(12), 6283-6303. [CrossRef]

[17] Wang, G., Guo, J., Chen, Y., Li, Y., & Xu, Q. (2019). A PSO and BFO-based learning strategy applied to faster R-CNN for object detection in autonomous driving. *IEEE Access, 7*, 18840-18859. [CrossRef]

[18] Haris, M., Shakhnarovich, G., & Ukita, N. (2021). Task-driven super resolution: Object detection in low-resolution images. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28* (pp. 387-395). Springer International Publishing. [CrossRef]

[19] Yadav, S. P., Jindal, M., Rani, P., de Albuquerque, V. H. C., dos Santos Nascimento, C., & Kumar, M. (2024). An improved deep learning-based optimal object detection system from images. *Multimedia Tools and Applications, 83*(10), 30045-30072. [CrossRef]

[20] Gui, S., Song, S., Qin, R., & Tang, Y. (2024). Remote sensing object detection in the deep learning era—a review. *Remote Sensing, 16*(2), 327. [CrossRef]

[21] Bai, Y., Zhang, Y., Ding, M., & Ghanem, B. (2018). Finding tiny faces in the wild with generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 21-30).

[22] Ji, H., Gao, Z., Mei, T., & Ramesh, B. (2019). Vehicle detection in remote sensing images leveraging on simultaneous super-resolution. *IEEE Geoscience and Remote Sensing Letters, 17*(4), 676-680. [CrossRef]

[23] Jiang, T., Yu, Q., Zhong, Y., & Shao, M. (2024). PlantSR: Super-Resolution Improves Object Detection in Plant Images. *Journal of Imaging, 10*(6), 137. [CrossRef]

[24] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of*

the ACM, 63(11), 139-144. [CrossRef]

[25] Li, Y., Xu, J., Xia, R., Wang, X., & Xie, W. (2019). A two-stage framework of target detection in high-resolution hyperspectral images. *Signal, Image and Video Processing, 13*, 1339-1346. [CrossRef]

[26] Krishna, H., & Jawahar, C. V. (2017, November). Improving small object detection. In *2017 4th IAPR Asian conference on pattern recognition (ACPR)* (pp. 340-345). IEEE. [CrossRef]

[27] Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., & Yan, S. (2017). Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1222-1230).

[28] Bai, Y., Zhang, Y., Ding, M., & Ghanem, B. (2018). Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 206-221).

[29] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops* (pp. 0-0).

[30] Theckedath, D., & Sedamkar, R. R. (2020). Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. *SN Computer Science, 1*(2), 79. [CrossRef]

[31] Paris, S., Hasinoff, S. W., & Kautz, J. (2015). Local Laplacian filters: edge-aware image processing with a Laplacian pyramid. *Communications of the ACM, 58*(3), 81-91. [CrossRef]

[32] Mundhenk, T. N., Konjevod, G., Sakla, W. A., & Boakye, K. (2016). A large contextual dataset for classification, detection and counting of cars with deep learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14* (pp. 785-800). Springer International Publishing. [CrossRef]

[33] Long, Y., Gong, Y., Xiao, Z., & Liu, Q. (2017). Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing, 55*(5), 2486-2498. [CrossRef]

[34] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems, 28*.

[35] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

[36] Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.

[37] Huang, X., Wang, X., Lv, W., Bai, X., Long, X., Deng, K., ... & Yoshie, O. (2021). PP-YOLOv2: A practical object detector. *arXiv preprint arXiv:2104.10419*.

[38] Mandal, M., Shah, M., Meena, P., Devi, S., & Vipparthi, S. K. (2019). AVDNet: A small-sized vehicle detection network for aerial visual data. *IEEE Geoscience and Remote Sensing Letters, 17*(3), 494-498. [CrossRef]

[39] Li, J., Zhang, Z., Tian, Y., Xu, Y., Wen, Y., & Wang, S. (2021). Target-guided feature super-resolution for vehicle detection in remote sensing images. *IEEE geoscience and remote sensing letters, 19*, 1-5. [CrossRef]

[40] Zhang, H., Hao, C., Song, W., Jiang, B., & Li, B. (2023). Adaptive slicing-aided hyper inference for small object detection in high-resolution remote sensing images. *Remote Sensing, 15*(5), 1249. [CrossRef]

Jiawei Yi received the master degree from Shenzhen University.

Her research interests include image super-resolution algorithms, object detection, and automatic annotation systems. (Email: 15007962908@163.com)

Ying Liu received the master degree from University of Chinese Academy of Sciences (UCAS) and is currently pursuing a Ph.D degree at Shenzhen University.

Her research interests include object detection and large language model. (Email: 2453043003@mails.szu.edu.cn)

Yanshan Li received the Ph.D. degree in the South China University of Technology. He is currently a Researcher and Doctoral Supervisor with the Institute of Intelligent Information Processing and Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University.

His research interests include computer vision, machine learning, and image analysis. (Email: lys@szu.edu.cn)

Weixin Xie received the degree from Xidian University, Xi'an. He was a Faculty Member with Xidian University in 1965. From 1981 to 1983, he was a Visiting Scholar at the University of Pennsylvania, USA. In 1989, he was a Visiting Professor with the University of Pennsylvania. He is currently with the School of Information Engineering, Shenzhen University, China.

His research interests include intelligent information processing, fuzzy information processing, image processing, and pattern recognition. (Email:wxxie@szu.edu.cn)