Small targets detection in low-resolution remote sensing images based on super-resolution joint optimization

Bo Huang[1]**, Jian Lin**[2]**, Enqi Huang**[3]**, Liaoni Wu**[4]**, Yiqing Cao**[5]
[1, 2, 5]School of Electromechanical and Information Engineering, Putian University, Putian, Fujian, China
[3]School of Electronic Engineering, Xidian University, Xi'an, Shanxi, China
[4]School of Aerospace Engineering, Xiamen University, Xiamen, Fujian, China
[1]Corresponding author
E-mail: [1]*huangbo@ptu.edu.cn*, [2]*17880349219@163.com*, [3]*huangenqi@stu.xidian.edu.cn*, [4]*wuliaon@xmu.edu.cn*, [5]*caoyiqing1987@163.com*

Received 18 March 2025; accepted 22 May 2025; published online 16 July 2025
DOI https://doi.org/10.21595/jme.2025.24904

Abstract. While convolutional neural networks have driven remarkable progress in remote sensing object detection, persistent challenges remain in detecting small targets within low-resolution imagery due to their limited pixel representation and feature degradation during hierarchical downsampling. To address this, this study proposed the joint super-resolution and detection network (JSRDN), which synergistically optimizes SR reconstruction through task-specific detection feedback, significantly enhancing small target recognition in LR remote sensing imagery. Firstly, generator in generative adversarial network incorporates improved residual blocks, enabling enhanced perception of complex deep-level features in the SR reconstruction process. Then, a perceptual loss function is introduced into the adversarial training process, which captures perceptual discrepancies in high-level features between reconstructed images and original HR references. After that, an edge-enhancement network is designed to dynamically detect edges in intermediate features restored by the generator, prioritizing edge influence across network layers to generate discriminative features for target recognition. Furthermore, the JSRDN implements detection-driven feedback by backpropagating object recognition loss through the generator, enforcing the super-resolution process to prioritize detection-salient feature recovery. Evaluated on 64×64 low-resolution COWC datasets, JSRDN achieves 0.1819 dB peak signal-to-noise ratio (PSNR) and 7.18 % average precision (AP) improvements over the deep residual dual-attention network (DRDAN), with ablation studies and visualizations confirming its balanced optimization of reconstruction fidelity and detection-oriented feature learning. This technology can provides valuable support for small target measurement and opens new opportunities in the field.

Keywords: super-resolution reconstruction, small object detection, generative adversarial networks, remote sensing images.

Nomenclature

CNNs	Convolutional neural networks
HR	High-resolution
LR	Low-resolution
SR	Super-resolution
CNNs	Convolutional neural networks
GAN	Generative adversarial network
JSRDN	Joint super-resolution and detection network
EEN	Edge-enhancement network
SRGAN	Super-resolution generative adversarial network
GT	Ground truth
ESRGAN	Enhanced super-resolution generative adversarial network
RRDBs	Residual-in-residual dense blocks

RaD	Relativistic average discriminator
HR-MSRN	High-resolution representation and multi-stage region-based network
HRFPN	High-resolution feature pyramid network
RPN	Region proposal network
COWC	Cars overhead with context
PSNR	Peak signal-to-noise ratio
AP	Average precision
SRCNN	Super-resolution convolutional neural network
VDSR	Very deep convolutional networks SR
LGCNet	Local-global combined network
EDSR	Enhanced deep super-resolution network
DRDAN	Deep residual dual-attention network
I_{LR}	Low-resolution images
I_{HR}	High-resolution images
I_{Gen}	Generated intermediate images
G_{θ_G}	Network parameters of the generator
σ	Sigmoid function
E	Expectation operation over the mini-batch of data
D_{Ra}	Relative average discriminant function
L_D^{Ra}	Adversarial loss for discriminator
L_G^{Ra}	Adversarial loss for generator
L_{percep}	Perceptual loss function
$L_{content}$	Content loss function
L_G	Total loss function for generator
$E_{Laplacian}$	Laplacian edge detection operation
$I_{Laplacian}$	Laplacian edge maps
C_{DFESN}	Operation of deep feature extraction sub-network
C_M	Mask branch
DS	Downsampling operation
US	Upsampling operation
I_{Edge}	Enhanced edge maps
I_{SR}	Final super-resolution images
ρ	Charbonnier penalty function
L_{img_cst}	Consistency loss function for images
L_{edge_cst}	Consistency loss function for edges
L_{EEN}	Total loss function for EEN
Det_{cls}	Classifier for objector
Det_{reg}	Regressor for objector
L_{cls}	Classification loss function for detector
L_{reg}	Regression loss function for detector
L_{det}	Total loss function for detector

1. Introduction

The advent of cutting-edge satellite platforms and sensor measurement technologies has catalyzed exponential growth in earth observation capabilities, precipitating an urgent need for advanced analytical methodologies in remote sensing image interpretation. Modern geospatial intelligence frameworks demonstrate transformative potential across heterogeneous domains, including but not limited to: autonomous transportation optimization [1], intelligent agriculture, resource and environment detection [2], and disaster response systems [3].

With the development of deep learning in recent years, remote sensing image object detection based on deep convolutional neural networks (CNNs) shows powerful advantages [4], [5]. Cheng et al. [6] improved the feature pyramid network by integrating atrous convolution and multi-scale feature fusion, along with an attention mechanism to emphasize critical object features, thereby significantly boosting detection accuracy in remote sensing imagery. Huang et al. [7] introduced a high-resolution (HR) feature pyramid network as the backbone network within the detection framework, enabling hierarchical feature aggregation while preserving spatial fidelity across multi-scale representations. Zheng et al. [8] developed a plug-and-play feature enhancement module that integrates multi-scale contextual features, effectively bolstering the detection performance for small objects. Sun et al. [9] developed a multi-dimensional interaction mechanism within the Transformer backbone network to enhance global contextual awareness while harmonizing local and global feature representations, resulting in significant improvements in small object detection accuracy.

The abovementioned methods can achieve good results for object detection in remote sensing images. However, practical implementations frequently encounter low-resolution (LR) imagery compromised by atmospheric disturbances, long-range imaging constraints, and transmission channel artifacts. The low accuracy of object detection in LR images of remote sensing scenarios is primarily attributed to the following reasons: 1) a large number of small objects (less than 32×32 pixels) are presented in the scene, exhibiting insufficient discriminative features susceptible to noise interference, and 2) progressive information erosion through successive downsampling operations in conventional CNN architectures, where critical features of small objects are attenuated in deeper network layers. These challenges fundamentally constrain the capacity of deep neural architectures to extract discriminative features from small targets in LR remote sensing imagery.

To solve the above problems, the prevailing strategy integrates super-resolution (SR) reconstruction with object detection to enhance feature representation in LR imagery. SR techniques specifically aim to recover HR representations from degraded LR inputs, thereby amplifying discriminative feature signatures of small targets. Cao et al. [10] combined sparse coding-based SR with SSD detectors [11] to improve vehicle detection in satellite imagery. Li et al. [12] leveraged intermediate-layer feature maps and saliency-guided SR to amplify target-specific features, experimentally validating enhanced focus on critical object attributes. Wang et al. [13] developed a SR framework based on generative adversarial network (GAN) [14] to enhance LR remote sensing imagery, coupled with an optimized YOLOv4 architecture [15] to achieve enhanced multi-target recognition accuracy. Ji et al. [16] designed an unsupervised hybrid framework based on SRGAN [17] for vehicle detection in remote sensing imagery, eliminating dependence on paired LR/HR training data while enhancing small-object discriminability through adversarial SR. The aforementioned approaches, which integrate image SR reconstruction networks with object detectors, have achieved certain results in LR image analysis. However, these frameworks still insufficiently consider and utilize the specificity of the two visual tasks. Current SR algorithms primarily prioritize human perceptual quality through visual fidelity enhancement. For detecting small objects in LR scenarios, the reconstructed results should not only consider the human-centered visual effects, but also ensure sufficient clarity of the target edge details. This is crucial to ensure structural features (e.g., shape, contour, texture) remain discernible in shallow layers of detection networks, enabling robust feature extraction critical for small-object analysis.

Considering the above issues, this study proposes an end-to-end trained joint super-resolution and detection network (JSRDN). The proposed JSRDN does not simply combine a SR network with an object detection network. Instead, it integrates the feature-aware guidance module and the task-specific optimization module by leveraging the downstream detection task, and trains the entire network end-to-end. This framework enables the SR network to generate reconstructed images that are more conducive to detection, thereby achieving superior object detection performance. The contributions are listed as follows: 1) the JSRDN is proposed to jointly optimize dual remote sensing vision tasks through a unified framework, enhancing LR image resolution to achieve higher detection accuracy; 2) the GAN-based method is used in this work, which allows the model to perceive more complex deep features by optimizing the residual structure in the generator; 3) this work introduces a perceptual loss function in GAN that provide strong supervision for brightness consistency and texture recovery; and 4) this work designs an edge-enhancement network (EEN) that enhances contour delineation in intermediate feature maps through noise suppression while preserving edge sharpness and structural details.

2. Design of JSRDN

This section provides a detailed description of JSRDN, a multi-stream framework designed to address the critical challenge of small object detection in LR remote sensing imagery through resolution-aware feature learning. Towards this goal, JSRDN implements a synergistic optimization comprising two modules: a GAN-based SR network and a detector network. As illustrated in Fig. 1, the whole network establishes a training paradigm by integrating the optimization of HR and LR image pairs. Architecturally, the SR module comprises three functionally specialized components: 1) generator network, 2) discriminator network, and 3) EEN. The generator network initially generates intermediate SR images (I_{Gen}), which are subsequently refined by an EEN to produce final SR outputs (I_{SR}). The discriminator network implements adversarial learning to differentiate ground truth (GT) images from generator-derived intermediate SR images. The detector network aims to recognize objects from the final SR images, giving the probabilistic localization distributions and categorical confidence scores. By designing such a cross-task optimization mechanism, the detection loss gradients are backpropagated through the SR network architecture, further improving the application of the SR technique to the object detection task.

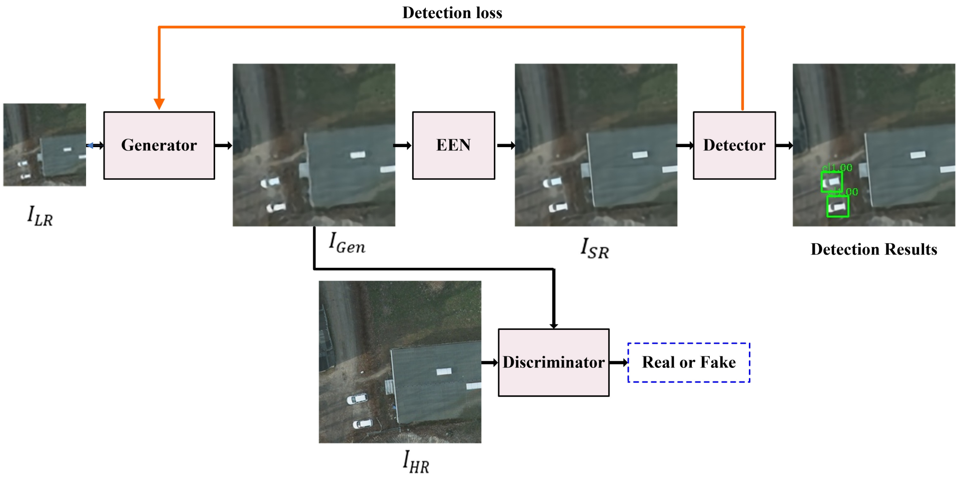


Fig. 1. Overview of the JSRDN framework

2.1. Design of generator and discriminator

2.1.1. Design of generator

The GAN used in this study is derived from enhanced super-resolution generative adversarial network (ESRGAN) [18] and some processing details have been modified. The generator structure is shown in Fig. 2. As illustrated in Fig. 2, the generator network employs residual-in-residual dense blocks (RRDBs) as its fundamental operational units. Each RRDB module integrates three hierarchically interconnected dense blocks through cascaded residual pathways, and the batch normalization layer is removed from the dense block to reduce computational complexity. To mitigate training instability arising from network depth escalation, this study also scales down the residuals by multiplying a scaling factor (empirical constant of 0.2) before adding them to the main path. Each dense block strategically leverages densely connected convolutional layers to extract rich local feature representations, as illustrated in Fig. 2. This architectural optimization facilitates systematic integration of hierarchical features across network layers, thereby enhancing model capacity through multi-scale feature utilization while simultaneously elevating image reconstruction fidelity through improved feature preservation and propagation mechanisms.

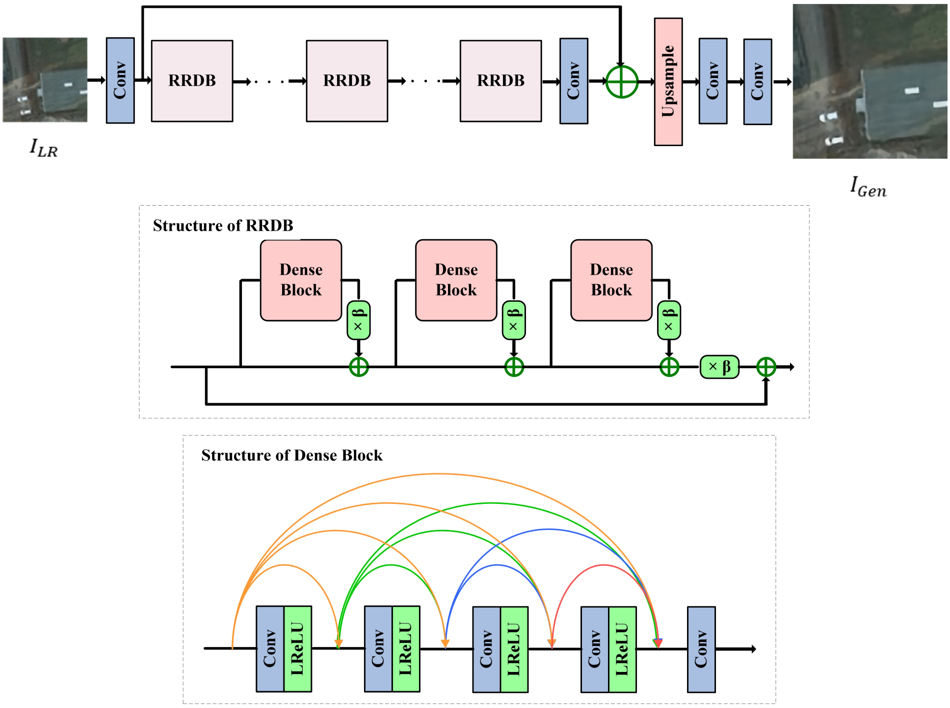


Fig. 2. Structure of generator network

2.1.2. Design of discriminator

As shown in Fig. 3, the discriminator employs VGG-19 as backbone network, augmented with relativistic adversarial learning through a relativistic average discriminator (RaD) [19]. In RaD, the task of discriminator is perform relative true and fake discrimination between two images. Here, $I_{LR} \in \mathbb{R}^{H\times W\times C}$ is considered the LR input of generator, $I_{HR} \in \mathbb{R}^{sH\times sW\times C}$ and $I_{Gen} \in \mathbb{R}^{sH\times sW\times C}$ are respectively considered the HR input and generated intermediate input of discriminator, where: H and W denote the height and width of the input LR image, respectively; s denotes the upscaling factor; and C denotes the number of channels. The mathematical relationship between I_{LR} and I_{Gen} is as follows:

$$I_{Gen} = G_{\theta_G}(I_{LR}), \tag{1}$$

where, G_{θ_G} denotes the network parameters of the generator. The mathematical expression of the RaD is expressed as follows:

$$D_{Ra}(I_{HR}, I_{Gen}) = \sigma(C(I_{HR}) - \mathbb{E}_{I_{Gen}}[C(I_{Gen})]) \rightarrow 1 \text{ More Realistic Than Fake Data?}, \tag{2}$$

$$D_{Ra}(I_{Gen}, I_{HR}) = \sigma(C(I_{Gen}) - \mathbb{E}_{I_{HR}}[C(I_{HR})]) \rightarrow 0 \text{ More realistic than real data?}, \tag{3}$$

where, $\sigma(\cdot)$ and $C(\cdot)$ denote the Sigmoid function and the non-transformed discriminator output, respectively; $\mathbb{E}[\cdot]$ denotes the expectation operation over the mini-batch of data, e.g., $\mathbb{E}_{I_{Gen}}[\cdot]$ and $\mathbb{E}_{I_{HR}}[\cdot]$ denote the operation of calculating mean for all generated intermediate images and original HR images in a mini-batch, respectively; D_{Ra} denotes a relative average discriminant function. When the real image exhibits superior photorealism and perceptual fidelity compared to the generated image, the result of $D_{Ra}(I_{HR}, I_{Gen})$ tends to be 1 (Eq. (2)); If the generated image demonstrates degraded quality relative to the real image, the result of $D_{Ra}(I_{Gen}, I_{HR})$ tends to be 1 (Eq. (3)).

Building upon the theoretical framework of RaD, the adversarial loss for discriminator can be defined as follows:

$$L_D^{Ra} = -\mathbb{E}_{I_{HR}}[\log(D_{Ra}(I_{HR}, I_{Gen}))] - \mathbb{E}_{I_{Gen}}[\log(1 - D_{Ra}(I_{Gen}, I_{HR}))]. \tag{4}$$

Then, the adversarial loss for generator is expressed as follows:

$$L_G^{Ra} = -\mathbb{E}_{I_{HR}}[\log(1 - D_{Ra}(I_{HR}, I_{Gen}))] - \mathbb{E}_{I_{Gen}}[\log(D_{Ra}(I_{Gen}, I_{HR}))]. \tag{5}$$

Therefore, the generator in RaD benefits from the gradients from both generated data and real data in adversarial training, thus generating rich edge textures.

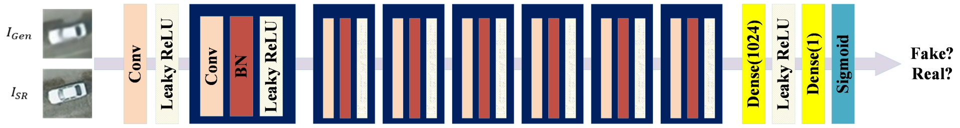


Fig. 3. The structure of discriminator network

2.1.3. Design of perceptual loss

Grounding in human visual perception characteristics that emphasize hierarchical feature processing [20], the seminal work [17] established perceptual loss through feature representation from deep layers of pre-trained networks. While SRGAN extended this paradigm by computing loss on post-activation features, subsequently analyses reveal critical limitations: 1) Activation-induced feature sparsity in deep networks diminishes gradient signal integrity for supervision; 2) Nonlinear activation distortions introduce luminance inconsistency between reconstructed and HR references, destabilizing loss landscape optimization.

To overcome these constraints, this study uses the features before the activation layers. Specifically, given the strong generalization capability of the pre-trained VGG-19 features, this study compute the perceptual loss using the features extracted from VGG-19 network before the activation function layers to leverage pre-activation convolutional responses. This approach preserves richer gradient information and structural primitives-particularly enhancing edge acuity and texture fidelity through undistorted low-level feature preservation. The perceptual loss function is defined as follows:

$$L_{percep} = \mathbb{E}_{I_{LR}} \|\varphi(I_{Gen}) - \varphi(I_{HR})\|_1, \tag{6}$$

where, $\varphi(\cdot)$ denotes the calculation process before the activation layers of a fine-tuned VGG-19 network, $\|\cdot\|_1$ denotes the L1 norm, and $\mathbb{E}_{I_{LR}}[\cdot]$ denote the operation of calculating mean for all input LR images in a mini-batch.

In addition to the perceptual loss, this study introduces the content loss for measuring the 1-paradigm distance between I_{Gen} and I_{HR}. The content loss function is defined as follows:

$$L_{content} = \mathbb{E}_{I_{LR}} \|I_{Gen} - I_{HR}\|_1. \tag{7}$$

The total loss function for the generator is:

$$L_G = L_{percep} + \lambda L_G^{Ra} + \eta L_{content}, \tag{8}$$

where, λ and η are denote the weight parameters, which are taken as 0.01 and 0.001, respectively.

2.2. Design of EEN

Although adversarial learning-based SR reconstruction methods demonstrate enhanced capacity for synthesizing realistic texture, their synthesized high-frequency features (e.g., edge structures) frequently exhibit distributional discrepancies relative to authentic image statistics, resulting in hallucinatory artifacts. Notably, primitive GAN-based methods exhibits heightened susceptibility to noise amplification, often introducing non-semantic high-frequency artifacts that lack correspondence to the underlying image content. These limitations critically degrade the utility of reconstructed images in downstream vision applications, particularly in precision-sensitive tasks such as small object detection. To address these challenges, this study introduces the EEN, a dedicated architecture for refining edge-aware feature representations to improve discriminative detail preservation. The EEN mainly integrates five functionally specialized modules: 1) Laplacian edge extraction, 2) downsampling, 3) deep feature extraction sub-network (DFESN), 4) denoising branch, and 5) upsampling. The structure of the EEN is illustrated in Fig. 4. As shown in Fig. 4, the process begins by extracting edges from the input image using the Laplacian operator. Then, these extracted Laplacian edges are sequentially processed through a downsampling operation and the DFESN to estimate and enhance edge information. Moreover, a dedicated denoising branch is integrated subsequent to the downsampling operation, specifically designed to remove edge noise while preserving critical structural features. Finally, the upsampling operation with subpixel convolution to project refined edge maps into HR spaces.

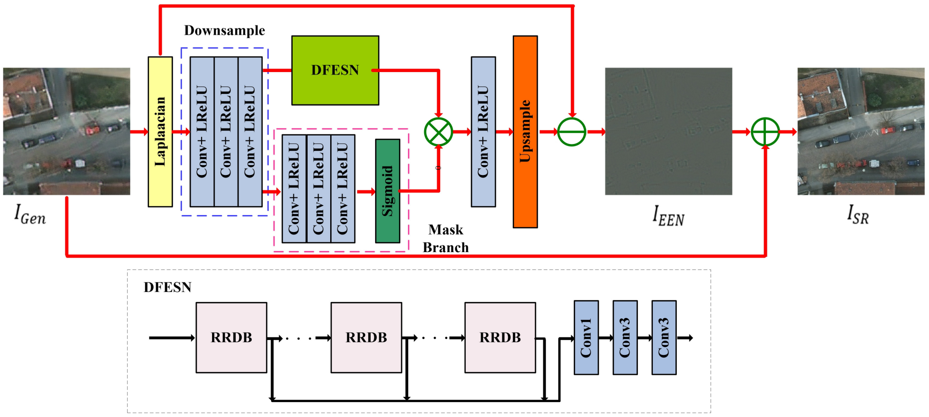


Fig. 4. The structure of the EEN

2.2.1. Design of EEN structure

For the images reconstructed by the generator, the Laplacian operator [21] is initially utilized for detecting and extracting edges. In particular, this study employs the discrete convolution mask of the Laplacian defined as ([−1, −1, −1], [−1, 8, −1], [−1, −1, −1]). Using the convolution kernel, Laplacian procedure is outlined as follows:

$$I_{Laplacian} = E_{Laplacian}(I_{Gen}), \tag{9}$$

where, $E_{Laplacian}(\cdot)$ and $I_{Laplacian}$ denote the Laplacian edge detection operation and the extracted edge maps, respectively. After the edge information is extracted, this study firstly utilizes the strided convolution to extract edge maps while simultaneously transforming them into LR space. The downsampling operation consists of three strided convolutional layers (followed by a LeakyReLU activation function), each with a kernel size of 3×3. Specifically, the number of feature maps for each convolutional layer, in order, is 128, 256, and 64, with stride sizes of 2, 2, and 1, respectively. Compared to operation in HR space, the aforementioned strategy reduces computational complexity. Then, this study employs a DFESN containing multiple RRDB blocks to estimate and enhance edge information in EEN. The structural diagram of the DFESN is shown in Fig. 4.

To address the inherent edge noise in $I_{Laplacian}$, a dedicated denoising branch is constructed to dynamically learn the image mask for detecting and eliminating isolated noise artifacts introduced during the Laplacian edge extraction process. As shown in Fig. 4, this branch employs a hierarchical architecture comprising three stacked convolutional layers with a kernel size of 3×3 followed by a Sigmoid activation function to learn a spatially variant weight matrix. This operation produces a probabilistic edge map with normalized intensity values in the range (0, 1). This biologically-inspired mechanism mimics the selective attention mechanism in human visual perception, prioritizing salient edge features while suppressing spurious noise and erroneous edge responses. The refined edge maps are subsequently upsampled to HR space via a learnable upsampling module. The process of converting $I_{Laplacian}$ to enhanced edge maps I_{Edge} in EEN can be formulated as follows:

$$I_{Edge} = US(C_{DFESN}(DS(I_{Laplacian})) \otimes C_M(DS(I_{Laplacian}))), \tag{10}$$

where, $C_{DFESN}(\cdot)$ denotes the operation of the DFESN to achieve hierarchical feature extraction and cross-layer feature fusion, $C_M(\cdot)$ denotes the mask branch to learn the image mask for suppressing the noises and the false edges, $DS(\cdot)$ denotes the downsampling operation with strided convolution to project Laplacian edge maps into LR spaces, and $US(\cdot)$ denotes the upsampling operation with subpixel convolution to project refined edge maps into HR spaces.

In the terminal phase of EEN, a residual learning mechanism is employed to amplify discriminative feature variations between I_{Edge} and $I_{Laplacian}$. Specifically, the pixel-wise discrepancy between I_{Edge} and $I_{Laplacian}$ is computed, generating a residual map that encapsulates high-frequency edge details. This residual signal is then fused with I_{Gen} through channel-wise summation, yielding the final SR reconstruction results I_{SR}:

$$I_{SR} = I_{Edge} - I_{Laplacian} + I_{Gen}. \tag{11}$$

2.2.2. Design of loss functions in EEN

To optimize the perceptual and structural fidelity of reconstructed imagery while mitigating artifacts, this study introduces a pixel-based Charbonnier penalty loss [22] to enhance the consistency of image contents between I_{SR} and I_{HR}. The consistency loss function is formulated as follows:

$$L_{img\_cst} = \mathbb{E}_{I_{SR}}[\rho(I_{HR} - I_{SR})]\,, \tag{12}$$

where, $\rho(\cdot)$ and $\mathbb{E}_{I_{SR}}[\cdot]$ denote Charbonnier penalty function and the operation of calculating mean for all final SR images in a mini-batch, respectively. Considering the challenges of geometric distortion and artifactual noise propagation in edge representations, this study adopts a consistency loss calculation strategy that explicitly enforces geometric and photometric consistency between reconstructed and GT edge maps. The consistency loss function for the image edges is formulated as follows:

$$L_{edge\_cst} = \mathbb{E}_{I_{Edge}}[\rho(I_{Edge\_HR} - I_{Edge})], \tag{13}$$

where, $I_{Edge\_HR} = E_{Laplacian}(I_{HR})$, denotes the extracted edge maps of the original HR images; and $\mathbb{E}_{I_{Edge}}[\cdot]$ denotes the operation of calculating mean for all Laplacian edge maps in a mini-batch. Finally, the supervised learning of EEN is performed with the total loss L_{EEN} that aggregates task-specific objectives for both holistic image reconstruction and edge structure preservation:

$$L_{EEN} = L_{img\_cst} + L_{edge\_cst}. \tag{14}$$

2.3. Design of target detector

The target detector aims to recognize the pixel locations and the corresponding categories of objects from construction images. Follow the previous work [7], this study uses the high-resolution representation and multi-stage region-based network (HR-MSRN) for detector network. The working mechanism of HR-MSRN in this study is as follows: firstly, the high-resolution feature pyramid network (HRFPN) is introduced as a backbone network in the whole framework to extract and fuse multi-level image feature maps and maintain high-resolution feature representation; secondly, target candidate boxes are generated by defining many anchors through Region Proposal Network (RPN); thirdly, high quality object detection is achieved with a combination of cascaded bounding box regression. The structure of the EEN is illustrated in Fig. 5.

It should be emphasized that the HR-MSRN framework exclusively addresses classification and object localization tasks on images reconstructed via the SR network. Consequently, the detector's loss function is formulated as a composite of classification loss L_{cls} and regression loss L_{reg}, where the classification term is mathematically expressed as:

$$L_{cls} = \mathbb{E}_{I_{LR}}[-\log(Det_{cls}(I_{SR}))], \tag{15}$$

where, $Det_{cls}(\cdot)$ denotes the classifier for the HR-MSRN. The regression loss is defined as:

$$L_{reg} = \mathbb{E}_{I_{LR}}[Smooth_{L1}(Det_{reg}(I_{SR}), t^*)], \tag{16}$$

where, $Det_{reg}(\cdot)$ denotes the regressor for the HR-MSRN, t^* denotes the GT bounding box coordinates, and $Smooth_{L1}$ denotes the Smooth L1 loss function. The total detection loss L_{det} is defined as:

$$L_{det} = L_{cls} + L_{reg}. \tag{17}$$

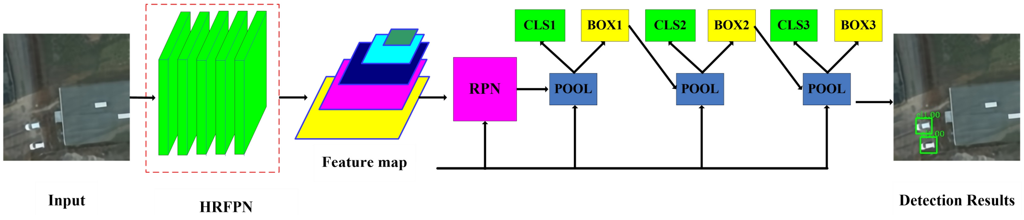


Fig. 5. The structure of the HR-MSRN

3. Experiments

This section will firstly describe the experimental settings, including the datasets, image-quality evaluation indexes, and implementation details of the model. Then, quantitative and qualitative experimental comparisons and a detailed analysis of the results will be reported.

3.1. Data

Cars overhead with context (COWC) dataset [23] comprises satellite imagery with a ground sampling distance of 15 cm per pixel, collected across six geographically distinct regions: Toronto (Canada), Selwyn (New Zealand), Potsdam (Germany), Vaihingen (Germany), Columbus (USA), and Utah (USA). This dataset encompasses 32716 vehicle instances. Out of these six regions, this study used the parts of the dataset from Toronto, Selwyn, and Potsdam; hence, all subsequent references to the COWC dataset in this study specifically pertain to these three regions.

The experiments considered the dataset having only one class, vehicle, with all images titles standardized to dimensions of 256×256 pixels and constrained to contain at least one vehicle instance. Vehicle dimensions were quantified as follows: average length ranged between 24-48 pixels, and width varied from 10-20 pixels. Consequently, the projected vehicle are spanned 240-960 pixels, categorizing these objects as small-size targets typically encountered in remote sensing imagery. For LR image generation from the COWC dataset, this study employed Bicubic interpolation with a ×4 downscaling factor. This resampling process yielded 64×64 pixel LR images where vehicle instances became particularly diminutive, substantially increasing detection complexity. Each image tile was accompanied by textual annotation files containing bounding box coordinates for all vehicle instances.

In this study, a total of 1600 images from COWC are utilized for training the SR model to enhance image quality. Furthermore, 5120 new images from COWC are used for the training of the object detection model and another 900 and 900 new images are used for the validation and testing, respectively. To optimize model robustness, the training images are augmented via three principal transformations: 1) horizontal flipping; 2) vertical flipping; and 3) 90° rotation.

3.2. Evaluation indexes

The average peak signal-to-noise ratio (PSNR) [24] serves as a principal quantitative image quality metric for SR model evaluation, with its mean value extensively adopted in comparative analyses. The PSNR describes the distortion of the reconstructed images caused by random noise; it is expressed as:

$$PSNR(x,y) = 10\log_{10}\left(\frac{I_{max}^2}{\frac{1}{W \times H}\sum_{i=1}^{H}\sum_{j=1}^{W}[x(i,j) - y(i,j)]^2}\right), \tag{18}$$

where, $x \in \mathbb{R}^{H\times W}$ denote the GT image and $y \in \mathbb{R}^{H\times W}$ denote its corresponding super-resolved counterpart, where H and W denote the height and width of the image, respectively; and I_{max} denotes the peak intensity value in the pixel domain (which is 255 for RGB images). The PSNR is quantified in dB, and a higher PSNR value means superior perceptual quality in the super-resolved image.

The average precision (AP) metric is employed to quantify the efficacy of object detection model. To compute the AP, the procedure necessitates the prior computation of precision and recall, defined formally as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{19}$$

$$Recall = \frac{TP}{TP + FN}, \tag{20}$$

where, TP denotes the count of correctly identified positive instances, FP denotes the count of erroneously classified negative instances, and FN denotes the count of undetected positive instances. Then, the AP can be given as:

$$AP = \int_0^1 P(R)dR, \tag{21}$$

where, P and R denote the precision and recall values at different confidence levels.

3.3. Implementation details

The architecture can be trained in separate steps or jointly in an end-to-end way. In separate training, the SR network was first optimized to convergence, followed by detector network training using the enhanced SR outputs. In the end-to-end training framework, the entire architecture undergoes integrated optimization where detector loss gradients propagate backward through the super-resolution network. This configuration enables the generator module to simultaneously assimilate gradient signals originating from both the detection and discrimination components.

To ensure a fair comparison, all models in this study are retrained using the training set mentioned above, without any pre-training and fine-tuning processes. In each training batch, two LR images of size 64×64 pixels as the input for model training, and their SR counterparts of 256×256 pixels are generated with scale factor of ×4. The initial learning rate is 1×10^{-4} and decreases to 50 % every 30000 iterations in the process of back-propagation. The Adam algorithm with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ is adopted to optimize the architecture parameters until convergence. According to the previous work [25], the loss curve gradually stabilizes when the model is trained on the COWC dataset up to the 150-th epoch. Therefore, to ensure the fairness of the comparison, all networks are trained for 150 epochs in total. In the generator and the EEN, the depths of RRDB are set to 23 and 3, respectively. The HR-MSRN uses HRFPN-W18 for feature extraction. All detector configurations and hyperparameters strictly adhered to the default settings in the MMDetection open-source library, unless otherwise explicitly specified. The operating system is Ubuntu 20.04. All experiments adopts the deep-learning framework Pytorch, and four Nvidia GTX-2080Ti GPUs are used to train all models.

3.4. Experimental results and analysis

To objectively analyze the performance of the proposed methodology, ablation experiments were conducted on different modules of the framework using the COWC dataset, thereby enabling a more precise assessment of the impact of each module on the overall algorithm. In addition, this study introduced some mainstream representative algorithms, for comparative experiments to verify the effectiveness of the proposed JSRDN.

3.4.1. Ablation study

To ensure comparative fairness, the experimental protocol employed original HR images and SR outputs from isolated SR models (SR1: adversarial framework [generator + discriminator]; SR2: SR1 + perceptual loss; SR3: SR2 + EEN) to train dedicated detectors, subsequently evaluating them across SR-restored test sets. As shown in Table 1, quantitative analysis reveals fundamental resolution-detection correlations: HR→LR detection achieved 40.50 % AP versus HR→HR upper bound of 78.10 % AP (+37.60 %), indicating the large impact of the resolution to the object detection quality and establishing theoretical performance limits for SR-based approaches. The adversarial framework was employed to preprocess LR images, with the detector subsequently trained on pristine HR data. This configuration yielded an 11.58 % AP improvement over direct LR detection, confirming the GAN's capacity to recover spatially discriminative features for small vehicle recognition. Progressive architectural enhancements revealed systematic gains: introducing perceptual loss supervision elevated both reconstruction fidelity (+0.0672 dB PSNR) and detection precision (+1.24 % AP), empirically validating its role in aligning high-level semantics between SR and HR domains. Further integration of an EEN amplified these benefits (+0.0672 dB PSNR, +2.58 % AP), attributable to its edge-texture refinement critical for detection-oriented SR. Additionally, as evident from Table 1, detectors trained on outputs from any SR variant generalized robustly to the SR testing set, still achieving a better recognition result.

The SR networks were also jointly optimized with diverse architectural configurations and HR-MSRN in an end-to-end training paradigm to evaluate the efficacy of the proposed methodology. The detector's loss gradient propagated backward through the SR network, thereby enabling synergistic optimization that enhanced the quality of LR imagery. The SR module was trained using LR-HR image pairs during the training phase. The generated SR outputs were subsequently fed into the detection network. During the testing phase, only LR inputs were processed through the integrated network architecture, which sequentially performed SR reconstruction followed by object detection. As quantified in Table 2, the end-to-end framework demonstrated statistically significant improvements across both SR reconstruction and detection tasks. Compared to separate training paradigm utilizing different SR networks, the progressive incorporation of adversarial learning ("generator + discriminator"), perceptual loss ("generator + discriminator + perceptual loss"), and EEN ("generator + discriminator + perceptual loss + EEN") yields incremental quantitative enhancements, achieving PSNR improvements of 0.0679 dB, 0.1105 dB, and 0.1294 dB alongside corresponding AP metric gains of 1.25 %, 1.89 %, and 2.43 %, respectively. Fig. 6 visually demonstrates the enhancement of AP values achieved through end-to-end joint training of super-resolution and object detection networks across diverse architectural configurations. Furthermore, the complete JSRDN architecture (integrating generator, discriminator, perceptual loss, and EEN) demonstrates the optimal performance in both SR reconstruction fidelity and object detection accuracy for remote sensing imagery, confirming the synergistic benefits of joint optimization strategies.

Table 1. Performance results on different images with separately trained SR network

Training set	Testing set	PSNR / dB	AP / %
HR	HR	/	78.10
LR	LR	/	40.50
HR	SR1	30.4495	52.08
SR1			52.72
HR	SR2	30.5167	53.32
SR2			54.55
HR	SR3	30.6600	55.90
SR3			58.73

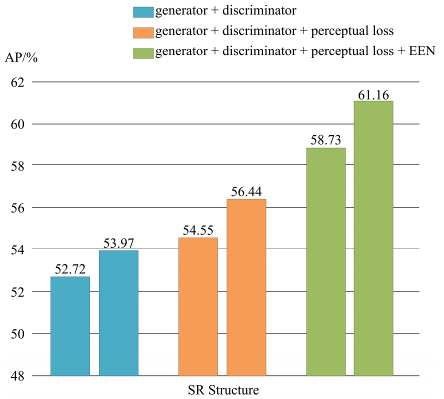
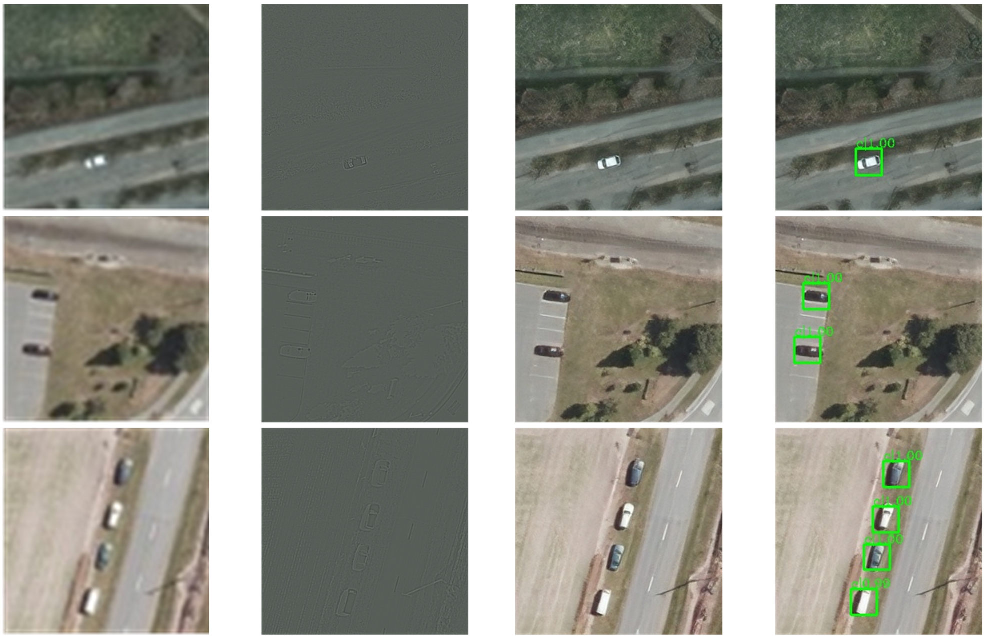


Fig. 6. Comparison of the effectiveness of end-to-end training; The left and right charts represent separate training and end-to-end training, respectively

Table 2. Performance results of SR-Detection end-to-end training

SR Structure	PSNR / dB	AP / %
generator + discriminator	30.5174	53.97
generator + discriminator + perceptual loss	30.6272	56.44
generator + discriminator + perceptual loss + EEN	30.7894	61.16

Moreover, a visual analysis is performed to evaluate the effectiveness of EEN. As illustrated in Fig. 7, qualitative visual representations demonstrate that the EEN can effectively perceives high-frequency detail information within the input LR images. Subsequent super-resolved images display enhanced edge and textual granularity in subjective visual assessment, achieving sufficient discriminative clarity for vehicular object identification.



a) Input LR images b) Enhanced edges c) SR results d) Detection results

Fig. 7. Visual effects of EEN

3.4.2. Comparison with other SR approaches

Furthermore, this study conducted a comprehensive comparative analysis encompassing seven typical SR approaches, including Bicubic, super-resolution convolutional neural network (SRCNN) [26], very deep convolutional networks SR (VDSR) [27], local-global combined network (LGCNet) [28], SRGAN, enhanced deep super-resolution network (EDSR) [29], and deep residual dual-attention network (DRDAN) [30], to perform ×4 upscaling operations on LR input data. The SR outputs were evaluated using the detector based on the HR-MSRN. To ensure a fair comparison, the object detector was trained exclusively on original HR images, while the testing phase utilized SR images generated by each reconstruction model. The experimental outcomes are mainly compared from two aspects, one is the comparative analysis of the objective quantitative indicators of image reconstruction and object detection, the other is the subjective assessment of visual performance focusing on detection localization precision and structural preservation in reconstructed outputs.

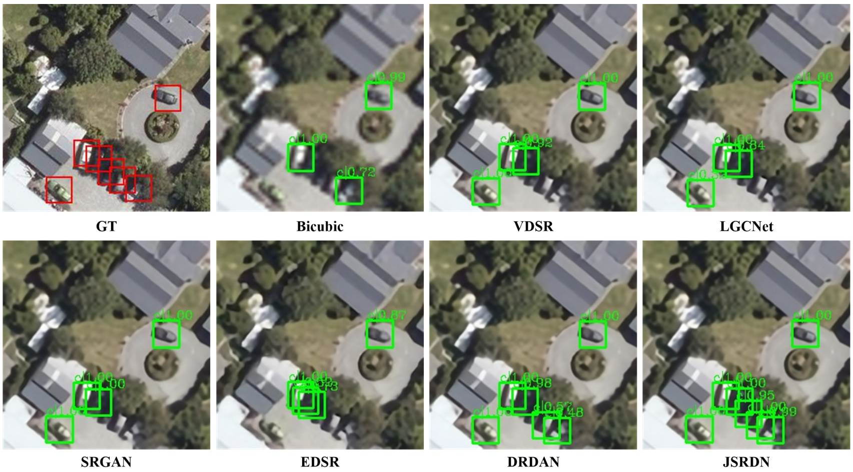
Table 3 quantifies the PSNR and AP metrics for SR images generated by different algorithms on the COWC dataset. The data reveals a direct positive correlation between SR algorithm performance and vehicle recognition accuracy, indicating that enhanced spatial resolution from advanced SR methodologies directly improves object detection efficacy. Among CNN-based approaches (SRCNN, VDSR, LGCNet, SRGAN, EDSR, and DRDAN), all exhibit substantial AP improvements over the Bicubic algorithm, with gains of 2.66 %, 3.34 %, 3.06 %, 5.45 %, 5.59 %, and 6.74 % respectively. Notably, the proposed JSRDN framework achieves superior performance, surpassing DRDAN by 0.1819 dB in PSNR and 7.18 % in AP. The quantitative comparisons presented above demonstrate the effectiveness and superiority of the proposed method.

Table 3. Performance comparisons of JSRDN with other approaches

SR method	PSNR / dB	AP / %
Bicubic	28.6981	46.21
SRCNN [23]	29.9163	48.87
VDSR [24]	30.4242	49.55
LGCNet [25]	30.0989	49.27
SRGAN [14]	30.4307	51.66
EDSR [26]	30.5048	51.80
DRDAN [27]	30.6075	52.95
JSRDN	30.7894	60.13

Fig. 8 visualizes the detection performance of SR images using different approaches. Other approaches struggle with leakage or misdetection due to challenges like blurring, small object sizes, complex lighting/shadow conditions, and inter-object interference. In contrast, by integrating GAN and EEN structures, JSRDN effectively preserves high-frequency details in remote sensing imagery. Furthermore, the joint optimization of a unified SR and object detection architecture substantially enhances detection robustness, particularly improving recognition reliability for small targets in complex environments. Significantly, the proposed JSRDN method generates detection results that exhibit enhanced spatial alignment with GT annotations, demonstrating superior consistency in localization precision.

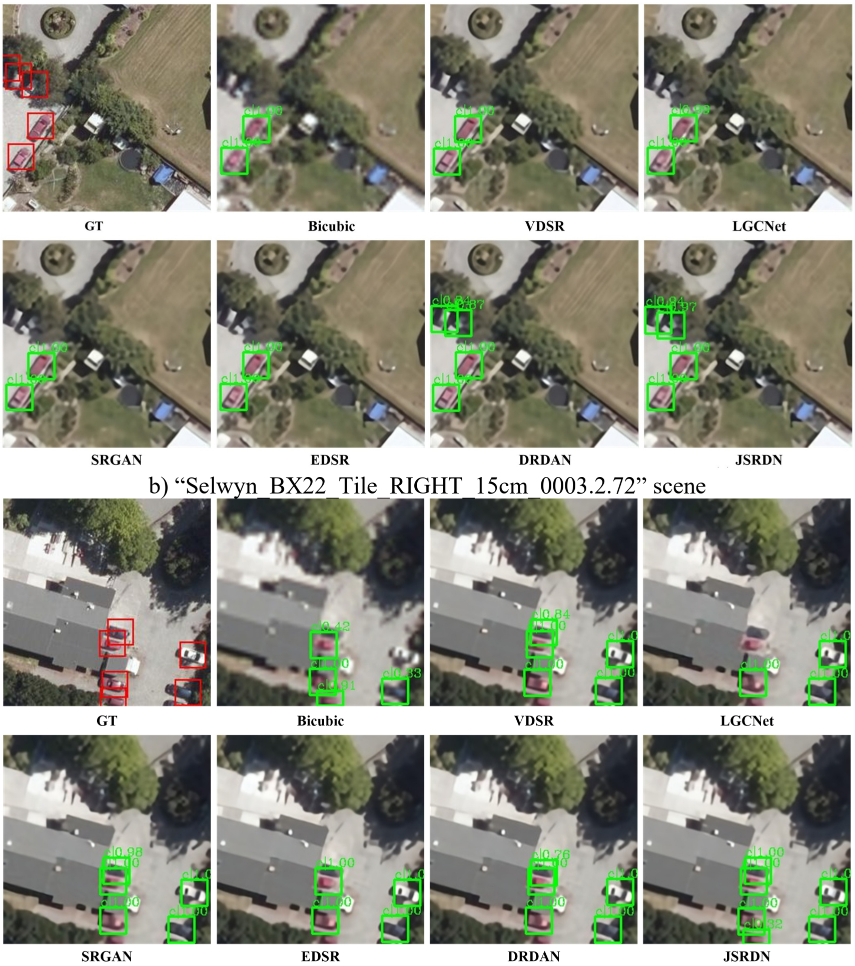
Fig. 9 provides a magnified visual comparison of vehicle targets generated by various SR models. In comparison to other approaches, the proposed JSRDN excels in restoring the color, texture, and edge information of the vehicle targets, while achieving enhanced structural congruence with GT annotations. This advancement contributes to measurable improvements in detection efficacy of small objects with cluttered environments, particularly through optimized feature representation in complex scenarios.



GT Bicubic VDSR LGCNet

SRGAN EDSR DRDAN JSRDN

a) “Selwyn_BX22_Tile_RIGHT_15cm_0003.1.71” scene



b) “Selwyn_BX22_Tile_RIGHT_15cm_0003.2.72” scene

c) “Selwyn_BX22_Tile_RIGHT_15cm_0003.8.53” scene

Fig. 8. Visual comparisons of the detection effects of different SR models

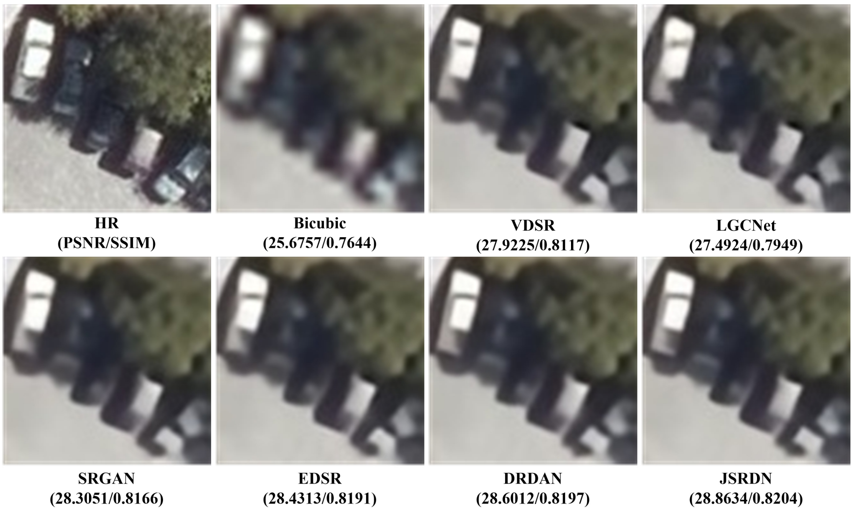


Fig. 9. Visual comparisons of reconstruction effects of different SR models

In summary, the proposed JSRDN achieves advanced performance through the synergistic integration of images SR reconstruction and small object detection tasks within a unified learning paradigm. The methodology demonstrate dual-domain superiority, excelling not only in reconstructing high-fidelity remote sensing imagery with preserved structural integrity but also in addressing the critical challenge of small object detection in LR scenarios under complex environmental conditions. The co-optimized architecture establishes JSRDN as a holistic framework capable of providing a comprehensive solution for challenging remote sensing image analysis tasks.

4. Conclusions

To address the challenge of small target detection in LR remote sensing imagery, this study proposes an end-to-end trained JSRDN that synergistically optimizes image reconstruction and object recognition tasks through joint supervision, where SR reconstruction compensates for missing details in LR images, thereby improving subsequent target detection precision. The JSRDN integrates a GAN-supervised SR reconstruction network and an object detection network. Firstly, the generator incorporates improved residual blocks, enabling enhanced perception of complex deep-level features in the SR reconstruction process. Then, a perceptual loss function is introduced into the adversarial training process, which captures perceptual discrepancies in high-level features between reconstructed images and original HR references, thereby enforcing structural and semantic similarity constraints. After that, to address noise sensitivity and artifact-prone edge reconstruction in GAN-based image restoration, an EEN is designed to dynamically detect edges in intermediate features, prioritizing edge influence across network layers to generate discriminative features for target recognition. Furthermore, the framework backpropagates detection loss to the generator, steering the SR reconstruction toward producing detection-optimized images, thereby improving the application of the SR technique to the remote sensing target detection. The proposed method is validated on 64×64 pixel LR images from the COWC dataset, with comprehensive quantitative metric evaluations and qualitative visual comparisons demonstrating its effectiveness. Compared to the advanced DRDAN, the proposed JSRDN achieves improvements of 0.1819 dB in PSNR and 7.18 % in AP. Future work will focus on achieving end-to-end joint optimization of different networks for measurement applications with other tasks.

However, the JSRDN framework remains constrained by certain limitations. Its supervised learning paradigm necessitates extensive paired training data, while the acquisition of precisely aligned high-quality reference images corresponding to degraded inputs remains practically challenging. Considering that, semi-supervised or unsupervised learning paradigms are critically required, as they enable the extraction of discriminative features and latent knowledge from unlabeled or sparsely annotated datasets, effectively addressing the limitations of fully supervised frameworks in annotation-scarce scenarios. For instance, pre-trained models from other domains or tasks, such as object detection models trained on natural scenes in general-view imagery, can be leveraged to initialize or fine-tune remote sensing object detection models, thereby harnessing knowledge from heterogeneous domains or tasks to enhance detection performance.

Acknowledgements

The research is supported by the National Natural Science Foundation of China (Grant No. 51276151), the Guiding science project of Fujian province (Grant No. 2024H0022), the Natural Science Foundation of Fujian Province of China (Grant No. 2024J01887), the Startup Fund for Advanced Talents of Putian University (Grant No. 2023132), and the China Scholarship Council (CSC) (Grant No. 202106310094).

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Bo Huang: conceptualization, methodology, validation, investigation, writing, supervision, funding acquisition. Jian Lin: validation, formal analysis, investigation, writing-review. Enqi Huang: methodology, validation, software. Liaoni Wu: resources, supervision, validation, funding acquisition. Yiqing Cao: data curation, supervision, software, funding acquisition.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] S. Zhao, F. Kang, and J. Li, “Intelligent segmentation method for blurred cracks and 3D mapping of width nephograms in concrete dams using UAV photogrammetry,” *Automation in Construction*, Vol. 157, p. 105145, Jan. 2024, https://doi.org/10.1016/j.autcon.2023.105145

[2] M. T. Amare, S. T. Demissie, S. A. Beza, and S. H. Erena, “Land cover change detection and prediction in the Fafan Catchment of Ethiopia,” *Journal of Geovisualization and Spatial Analysis*, Vol. 7, No. 2, p. 19, Jul. 2023, https://doi.org/10.1007/s41651-023-00148-y

[3] Y. Li, Y. Lei, B. Chen, and J. Chen, “Evaluation of geological hazard susceptibility based on the multi-kernel density information method,” *Scientific Reports*, Vol. 15, No. 1, p. 7982, Mar. 2025, https://doi.org/10.1038/s41598-025-91713-6

[4] Z. Wu, Y. Tang, B. Hong, B. Liang, and Y. Liu, “Enhanced precision in dam crack width measurement: leveraging advanced lightweight network identification for pixel-level accuracy,” *International Journal of Intelligent Systems*, Vol. 2023, No. 1, Sep. 2023, https://doi.org/10.1155/2023/9940881

[5] K. Hu, Z. Chen, H. Kang, and Y. Tang, “3D vision technologies for a self-developed structural external crack damage recognition robot,” *Automation in Construction*, Vol. 159, p. 105262, Mar. 2024, https://doi.org/10.1016/j.autcon.2023.105262

[6] Y. Cheng et al., “A multi-feature fusion and attention network for multi-scale object detection in remote sensing images,” *Remote Sensing*, Vol. 15, No. 8, p. 2096, Apr. 2023, https://doi.org/10.3390/rs15082096

[7] B. Huang, B. He, L. Wu, and Z. Guo, “High-resolution representations and multistage region-based network for ship detection and segmentation from optical remote sensing images,” *Journal of Applied Remote Sensing*, Vol. 16, No. 1, p. 01200, Aug. 2021, https://doi.org/10.1117/1.jrs.16.012003

[8] X. Zheng, Y. Qiu, G. Zhang, T. Lei, and P. Jiang, “ESL-YOLO: small object detection with effective feature enhancement and spatial-context-guided fusion network for remote sensing,” *Remote Sensing*, Vol. 16, No. 23, p. 4374, Nov. 2024, https://doi.org/10.3390/rs16234374

[9] H. Sun, G. Yao, S. Zhu, L. Zhang, H. Xu, and J. Kong, “SOD-YOLOv10: small object detection in remote sensing images based on YOLOv10,” *IEEE Geoscience and Remote Sensing Letters*, Vol. 22, pp. 1–5, Jan. 2025, https://doi.org/10.1109/lgrs.2025.3534786

[10] L. Cao, C. Wang, and J. Li, “Vehicle detection from highway satellite images via transfer learning,” *Information Sciences*, Vol. 366, pp. 177–187, Oct. 2016, https://doi.org/10.1016/j.ins.2016.01.004

[11] W. Liu et al., “SSD: Single Shot MultiBox Detector,” in *Lecture Notes in Computer Science*, pp. 21–37, Sep. 2016, https://doi.org/10.1007/978-3-319-46448-0_2

[12] J. Li, Z. Zhang, Y. Tian, Y. Xu, Y. Wen, and S. Wang, “Target-guided feature super-resolution for vehicle detection in remote sensing images,” *IEEE Geoscience and Remote Sensing Letters*, Vol. 19, pp. 1–5, Jan. 2022, https://doi.org/10.1109/lgrs.2021.3112172

[13] Z. Wang, C. Wang, Y. Chen, and J. Li, “Target detection algorithm based on super – resolution color remote sensing image reconstruction,” *Journal of Measurements in Engineering*, Vol. 12, No. 1, pp. 83–98, Mar. 2024, https://doi.org/10.21595/jme.2023.23510

[14] I. Goodfellow et al., “Generative adversarial networks,” *Communications of the ACM*, Vol. 63, No. 11, pp. 139–144, Oct. 2020, https://doi.org/10.1145/3422622

[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: optimal speed and accuracy of object detection,” *arXiv: 2004.10934*, Jan. 2020, https://doi.org/10.48550/arxiv.2004.10934

[16] H. Ji, Z. Gao, T. Mei, and B. Ramesh, “Vehicle detection in remote sensing images leveraging on simultaneous super-resolution,” *IEEE Geoscience and Remote Sensing Letters*, Vol. 17, No. 4, pp. 676–680, Apr. 2020, https://doi.org/10.1109/lgrs.2019.2930308

[17] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, Jul. 2017, https://doi.org/10.1109/cvpr.2017.19

[18] X. Wang et al., “ESRGAN: enhanced super-resolution generative adversarial networks,” *Lecture Notes in Computer Science*, Vol. 11133, pp. 63–79, Jan. 2019, https://doi.org/10.1007/978-3-030-11021-5_5

[19] A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard GAN,” *arXiv: 1807.00734*, Jan. 2018, https://doi.org/10.48550/arxiv.1807.00734

[20] J. Bruna, P. Sprechmann, and Y. Lecun, “Super-Resolution with Deep Convolutional Sufficient Statistics,” arXiv: 1511.05666, Jan. 2015, https://doi.org/10.48550/arxiv.1511.05666

[21] B. Kamgar-Parsi, B. Kamgar-Parsi, and A. Rosenfeld, “Optimally isotropic Laplacian operator,” *IEEE Transactions on Image Processing*, Vol. 8, No. 10, pp. 1467–1472, Oct. 1999, https://doi.org/10.1109/83.791975

[22] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Fast and accurate image super-resolution with deep laplacian pyramid networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 11, pp. 2599–2613, Nov. 2019, https://doi.org/10.1109/tpami.2018.2865304

[23] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, “A large contextual dataset for classification, detection and counting of cars with deep learning,” *Lecture Notes in Computer Science*, pp. 785–800, Sep. 2016, https://doi.org/10.1007/978-3-319-46487-9_48

[24] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in *20th International Conference on Pattern Recognition (ICPR)*, pp. 2366–2369, Aug. 2010, https://doi.org/10.1109/icpr.2010.579

[25] B. Huang, Z. Guo, L. Wu, B. He, X. Li, and Y. Lin, “Pyramid information distillation attention network for super-resolution reconstruction of remote sensing images,” *Remote Sensing*, Vol. 13, No. 24, p. 5143, Dec. 2021, https://doi.org/10.3390/rs13245143

[26] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 2, pp. 295–307, Feb. 2016, https://doi.org/10.1109/tpami.2015.2439281

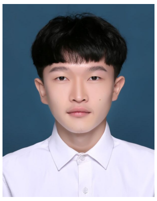
[27] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654, Jun. 2016, https://doi.org/10.1109/cvpr.2016.182

[28] S. Lei, Z. Shi, and Z. Zou, “Super-resolution for remote sensing images via local-global combined network,” *IEEE Geoscience and Remote Sensing Letters*, Vol. 14, No. 8, pp. 1243–1247, Aug. 2017, https://doi.org/10.1109/lgrs.2017.2704122

[29] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1132–1140, Jul. 2017, https://doi.org/10.1109/cvprw.2017.151

[30] B. Huang, B. He, L. Wu, and Z. Guo, “Deep residual dual-attention network for super-resolution reconstruction of remote sensing images,” *Remote Sensing*, Vol. 13, No. 14, p. 2784, Jul. 2021, https://doi.org/10.3390/rs13142784

Bo Huang received the Ph.D. degree in Xiamen University, Xiamen, China, in 2023. He is currently a Lecturer with the School of Electromechanical and Information Engineering, Putian University. His research interests include remote sensing image super resolution and deep learning.



Jian Lin received the B.S. degree from Putian University, Putian, China, in 2024. He is currently pursuing his M.S. degree at Putian University. His research interests include artificial intelligence and deep learning.

Enqi Huang is currently pursuing his B.S. degree at Xidian University. His research interests focus on multi-stage information collaborative mining of remote sensing big data, including multidimensional data fusion, spatiotemporal feature extraction, and intelligent analysis of earth observation data.

Liaoni Wu received the M.S. degree and Ph.D. degree from the Nanjing University Of Aeronautics and Astronautics, Nanjing, China, in 2005 and 2009, respectively. He is currently an Associate Professor at the School of Aerospace Engineering, Xiamen University. His research interests include unmanned aerial vehicle (UAV) technology and application.

Yiqing Cao received the Ph.D. degree from the Shanghai University, Shanghai, China, in 2018. He is currently an Associate Professor with the School of Electromechanical and Information Engineering, Putian University. His research interests include imaging analysis and design methods of optical system.