Multi-scale information distillation attention network for super-resolution reconstruction of remote sensing images

Bo Huang[1], **Liaoni Wu**[2], **Yiqing Cao**[3], **Mingen Zhong**[4]
[1, 3]School of Electromechanical and Information Engineering, PuTian University, Putian, Fujian, China
[2]School of Aerospace Engineering, Xiamen University, Xiamen, Fujian, China
[4]School of Mechanical and Automotive Engineering, Xiamen University of Technology, Xiamen, Fujian, China
[2]Corresponding author
E-mail: [1]*huangbo@ptu.edu.cn*, [2]*wuliaon@xmu.edu.cn*, [3]*caoyiqing1987@163.com*, [4]*zhongmingen@xmut.edu.cn*

Received 11 July 2024; accepted 11 December 2024; published online 8 January 2025
DOI https://doi.org/10.21595/jme.2024.24351

Check for updates

Abstract. Super-resolution (SR) is an effective and reasonable way to improve the spatial resolution of remote sensing images, which serve as an important information carriers for Earth observations. Compared to natural images, the more complex spatial distributions and more detailed ground information contained within remote sensing data place higher demands on the feature-representation ability of the model. Moreover, considering the deployment of these systems on mobile hardware, the complexity of the model is also an urgent issue. To overcome these problems, this study proposes the multi-size information distillation attention network (MSIDAN) for super-resolution reconstruction of remote sensing images. In the designed residual block, a multi-size information-distillation module is designed to distill and fuse multi-level semantic features step-by-step while reducing the number of model parameters. After this, an enhanced contrast-aware channel attention mechanism is employed to perceive high-frequency information by automatically encoding the weight values of candidate features. A large number of comparative experiments on four typical remote sensing image datasets demonstrate that MSIDAN outperforms other state-of-the-art approaches in both quantitative metrics and visual qualities. Compared to the information multi-distillation network (IMDN), MSIDAN improves the Peak Signal-to-Noise Ratio (PSNR) by 0.03312 dB, 0.06031 dB, 0.05319 dB, and 0.03812 dB on the RSSCN7, WHU-RS19, NWPU VHR-10, and COWC datasets, respectively. Moreover, in comparison to other comparable CNNs-based approaches, MSIDAN achieves a more favorable balance by jointly considering SR performance and model size. This technology provides valuable support for small target measurement and opens new opportunities in the field.

Keywords: super-resolution reconstruction, attention mechanism, information distillation mechanism, remote sensing images.

Nomenclature

SR	Super-resolution
HR	High-resolution
LR	Low-resolution
CNNs	Convolutional neural networks
MSIDAN	Multi-scale information distillation attention network
MSIDAB	Multi-scale information distillation attention block
IMDN	Information multi-distillation network
MSID	Multi-scale information distillation
ECCAM	Enhanced contrast-aware channel attention mechanism
CAM	Channel attention mechanism

CARN	Cascading residual network
IDN	Information distillation network
ECBSR	Edge-oriented convolution block based super-resolution model
CCAM	Contrast-aware channel attention mechanism
GAP	Global average pooling
PAN	Pixel attention network
LRL	Local residual learning
AP	Average pooling
PSNR	Peak signal-to-noise ratio
SSIM	Structural similarity
WDSR	Widely activated super-resolution network
MAFFSRN	Multi-attentive feature fusion super-resolution network
ESRGCNN	Enhanced super-resolution group convolutional neural network

1. Introduction

With the rapid advancement of satellite and measurement technologies, remote sensing images play an increasingly important role in various fields, including intelligent transportation [1], urban planning, geological-resource exploration, and disaster monitoring [2]. However, due to a series of factors such as atmospheric interference, long-distance imaging, and channel transmission capabilities, the clarity of these images often needs to be improved. Generally, designing more accurate remote sensing cameras can improve the resolution of the images obtained, but this inevitably requires high transmission and maintenance costs. Therefore, there is an urgent need to investigate inexpensive and practically applicable image-processing techniques to improve the resolution of acquired remote sensing images. The image super-resolution (SR) reconstruction technique aims to restore high-resolution (HR) images from low-resolution (LR) images. Compared to upgrading hardware to increase image resolution, the SR reconstruction algorithm offers the advantages of lower cost, greater flexibility, and simpler maintenance, making it an efficient approach for enhancing image utilization in the field of remote sensing.

In recent years, computer vision applications based on neural networks have been continuously evolving [3-5]. In the SR problem of general natural images, approaches utilizing convolutional neural networks (CNNs) have proved to be extremely promising [6, 7]. Dong et al. [6] built an SR convolutional neural network that uses a simple three-layer network to achieve better SR performance. Kim et al. [6] constructed a 20-layer CNN; their system introduces the strategy of residual learning, resulting in superior SR results compared to shallow-layer models. Building on the success of this research, Lim et al. [8] and Zhang et al. [9] designed SR models with 69-layer and 400-layer networks, respectively, achieving state-of-the-art performance. Nevertheless, these networks still have some limitations when dealing with the SR reconstruction task of remote sensing images.

Firstly, the wide spatial span of remote sensing images typically encompasses a variety of ground scenes containing complex surface elements, presenting a complicated spatial-structure distribution. This means that the SR network is required to conduct a high-quality analysis of high-frequency details (e.g., contours and edges) in LR images. Moreover, in inherently LR remote sensing images, the feature information available for the SR network as a basis for inference is limited; thus, SR reconstruction of complex remote sensing images is a challenging task.

Secondly, although excellent performance can be achieved by increasing the number of network layers, this leads to a larger number of model parameters, and the consequently greater computing burden restricts practical applications on resource-constrained mobile devices [8, 9]. Moreover, a complex network structure design makes training the model more difficult. Therefore, it is critical to create an efficient and rational architecture that is practical for solving these issues.

To effectively address these challenges, this research proposes a novel remote sensing image

SR reconstruction approach using a multi-scale information distillation attention network (MSIDAN); this introduces the multi-scale information distillation attention block (MSIDAB) as a basic component. Inspired by the information multi-distillation network (IMDN) [10], the MSIDAB employs an information multi-distillation design, and its structure incorporates a multi-scale feature-extraction component and a modified attention mechanism. In comparison to the typical procedure of repeating the convolution layer, information distillation [10] can progressively and efficiently extract plentiful features on the premise of reducing the number of model parameters; this achieves a better balance between performance and model complexity. Specifically, the extracted features are divided into two parts: one is sent for further processing to extract long-path features, and the other is used to store reserved short-path features.

However, the IMDN only uses single-scale convolution to extract image features before information distillation, which limits the learning capability of the network during feature transfer. To address this issue, a multi-scale information distillation (MSID) module is designed in the residual block. This module integrates the information distillation mechanism with the idea of multi-scale feature extraction, reducing the dimensionality of feature channels in the subsequent convolutional layers in the residual block through a channel separation operation, thus progressively refining and fusing multi-scale semantic features with a smaller number of parameters. Compared to previous SR approaches that incorporate multi-scale feature extraction within the information-distillation mechanism [11, 12], the proposed MSID places greater emphasis on the fusion and utilization of distilled features across different receptive fields. The success of the residual channel attention network [9] illustrates that the attention mechanism significantly enhances the SR performance of the network. Thus, an enhanced contrast-aware channel attention mechanism (ECCAM) is constructed, which employs the channel and spatial contrast values of the image to reflect the feature map information and further generates more balanced attention for the perception of high-frequency details.

The contributions are listed as follows: 1) this work presents a unique CNN for the remote sensing image SR task, namely, MSIDAN, to deliver an end-to-end training strategy that is both convenient and efficient; 2) this work proposes the MSID module to gradually distill and extract multi-size features by integrating information multi-distillation with multi-scale feature representation; 3) this work constructs the ECCAM module in the residual block to further strengthen the attention of high-frequency features; 4) Extensive experiments demonstrate that MSIDAN outperforms other comparable CNNs-based approaches, achieving superior results with fewer parameters in both objective evaluation metrics and subjective visual quality.

2. Related works

Recently, deep-learning-based approaches have achieved impressive results in resolving the issue of SISR. Dong et al. [6] were the first to introduce the concept of the neural network to the SR problem, developing a simple three-layer convolution structure to fit the nonlinear mapping between LR images and their HR counterparts. As with the excellent performance of classical ResNet [13] in computer vision tasks, many advanced SR models employ a residual-learning strategy for simplicity of training. Since the skip connection in the residual network allows direct mapping within the residual units, it effectively avoids gradient disappearance and allows faster training. Considering this, Kim et al. [7] constructed a very deep SR network with 20 convolutional layers, demonstrating that a deep network with residual learning can alleviate training difficulties and achieve better SR accuracy. Lim et al. [8] proposed an enhanced deep SR network, expanding the network to 69 layers by incorporating improved residual blocks. Then, Zhang et al. [9] formed a deeper SR network with more than 400 layers, in which a channel attention mechanism (CAM) is used in the residual-in-residual architecture.

Although excellent performance can be achieved by increasing the number of network layers, as noted, the drawback of this is that the large numbers of model parameters and consequently greater computing burden restrict practical applications on resource-constrained mobile devices.

In response to this issue, Ahn et al. [14] presented a cascading residual network (CARN), which achieves a lightweight SR model by introducing a cascading mechanism to incorporate the features from multiple layers. Hui et al. [15] developed a novel information distillation network (IDN), which fuses local short-path information with long-path information by employing a channel-splitting strategy, thus obtaining better reconstruction performance with fewer parameters. Then, Hui et al. [10] extended the IDN to create the lightweight IMDN, which extracts fine-grained image features step-by-step and further streamlines the parameters of the network. Zhang et al. [16] designed an extremely effective block for any SR model: the edge-oriented convolution block (ECB). They further proposed ECB based SR model (ECBSR), which achieves promising real-time SR effects on mobile devices. Gao et al. [17] elegantly integrated CNN and Transformers, achieving a better balance between performance and model size.

Attention mechanisms are a current topic of great interest in computer vision. Just as humans preferentially focus on the most important aspects of an image, the purpose of an attention mechanism in deep learning is to improve efficiency by obtaining important information more quickly and accurately while ignoring irrelevant details. Following the success of work [9], an increasing number of studies have utilized attention-based algorithms in CNNs to achieve superior SR performance. To better capture structure information in low-level vision tasks, Hui et al. [10] presented a contrast-aware channel attention mechanism (CCAM), which replaces the global average pooling (GAP) operation with the summation of the standard deviation and mean from their distillation blocks. Liu et al. [18] developed an enhanced spatial attention block that is lightweight and enables the network to focus on regions of key importance. To take advantage of the large spatial size, Muqeet et al. [19] optimized the ESA block by introducing dilated convolutions. Zhao et al. [20] proposed a pixel attention network (PAN) that generates three-dimensional attention maps instead of a one-dimensional vector or a two-dimensional map, resulting in improved SR outcomes with fewer additional parameters.

3. Design of MSIDAN for remote sensing image SR reconstruction

This section provides a detailed description of MSIDAN. Specifically, the system overview of the network framework and the design structure of MSIDABs will be presented in detail. Additionally, the loss function utilized during the training process is defined to optimize the objective. Here, $I_{LR} \in \mathbb{R}^{H\times W\times C}$ and $I_{SR} \in \mathbb{R}^{sH\times sW\times C}$ are respectively considered the LR input and SR output of MSIDAN, 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. Specifically, C is set as 64, consistent with the work [10].

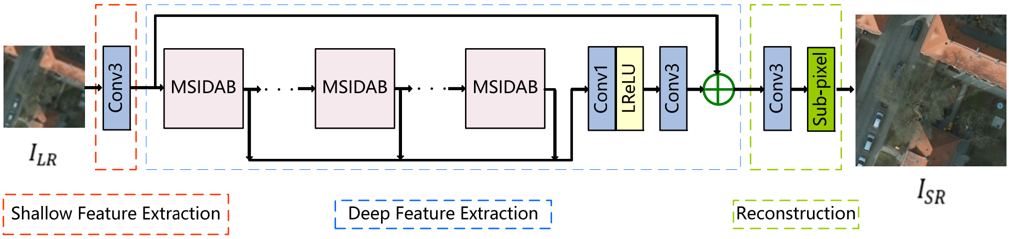


Fig. 1. Overview of the MSIDAN network structure

3.1. System overview

The overall architecture of the proposed MSIDAN model is illustrated in Fig. 1, including three parts: 1) shallow feature extraction, 2) deep feature extraction, and 3) reconstruction. According to a survey of previous reports [9], the output of the shallow feature extraction stage can be calculated as:

$$F_0 = H_{SF}(I_{LR}), \tag{1}$$

where, $H_{SF}(\cdot)$ denotes the 3×3 convolution operation. The resulting F_0 is used in the subsequent deep feature-extraction part, which makes use of stacked MSIDABs. The MSIDABs serve as fundamental components for local residual feature connection. The output of the MSIDAB can be calculated as:

$$F_{b,n} = H_{MSIDAB,n}(F_{b,n-1}), \tag{2}$$

where, $H_{MSIDAB,n}(\cdot)$ denotes the operation of the nth MSIDAB, and $F_{b,n-1}$ and $F_{b,n}$ are the inputs and outputs of the nth MSIDAB, respectively. As depicted in Fig. 2, the MSIDAB was developed using the MSID module, the ECCAM module, and a 1×1 convolution layer, which is employed to reduce the channel dimensionality of features. The MSID module fuses shallow and deep features while reducing the number of training parameters, and the ECCAM module allows the network to adaptively learn the weights of high-frequency detail information by generating more balanced attention information.

After completing the calculations of M MSIDABs, all intermediate features are fused using a concatenation operation:

$$F_{concat} = Concat(F_{b,1}, \dots, F_{b,n}, \dots, F_{b,M}), \tag{3}$$

where, $Concat(\cdot)$ denotes the feature-map concatenation operation. After combining the features of each block, a 1×1 convolution layer is employed to reduce the channel dimensionality of the concatenated features and fuse the spliced channels. Then, an adaptive activation function is adopted to reduce the redundant parameters, and the features are further obtained by a 3×3 convolutional layer. These operators can be expressed as:

$$F_{fused} = H_{fuse}(F_{concat}), \tag{4}$$

where, $H_{fuse}(\cdot)$ denotes a 1×1 convolution operation followed by a LeakyReLU function and a 3×3 convolution operation, and F_{fused} denotes the fused features.

Finally, the fused features F_{fused} are fed to the reconstruction part, which can map LR images to high-dimensional space and generate high-quality SR images. As many CNNs-based SR approaches, the sub-pixel up-sampling operation is adopted to reconstruct HR image; this has been shown to be superior to other up-scaling approaches in terms of the balanced optimization of the SR effect and computational complexity. Furthermore, considering the fusion with shallow features to implement a global residual-learning approach [7], the up-sampling operator can be expressed as:

$$I_{SR} = H_{subpixel}(H_{learnable}(F_{fused} + F_0)), \tag{5}$$

where, $H_{learnable}(\cdot)$ denotes a 3× learnable layer, $H_{subpixel}(\cdot)$ denotes the sub-pixel up-sampling operation, and I_{SR} is the estimated super-resolution image.

3.2. Design of MSIDAB

3.2.1. Design of MSID structure

As noted, deepening the network structure generally improves the SR effect but also increases complexity, potentially causing convergence problems. Considering this, the IDN guarantees recovery results while reducing the number of parameters and increasing test speed by compressing the dimensionality of the feature-map channels in the network. Based on the IDN, the IMDN constructs a progressive refinement module to retrain distilled features, and it further processes other remaining features multiple times. Specifically, the residual blocks in the IMDN divide the 64 feature channels output by the convolutional layer into two parts by 1:3, of which 16 feature channels are directly input into the concatenation operation as distilled features, and the other 48 feature channels serve as input for the next convolutional layer. Using such a channel spliting strategy, the IMDN achieves a better balancing of SR effect against applicability. Nevertheless, there remains room for improvement in the extraction and utilization of intermediate features.

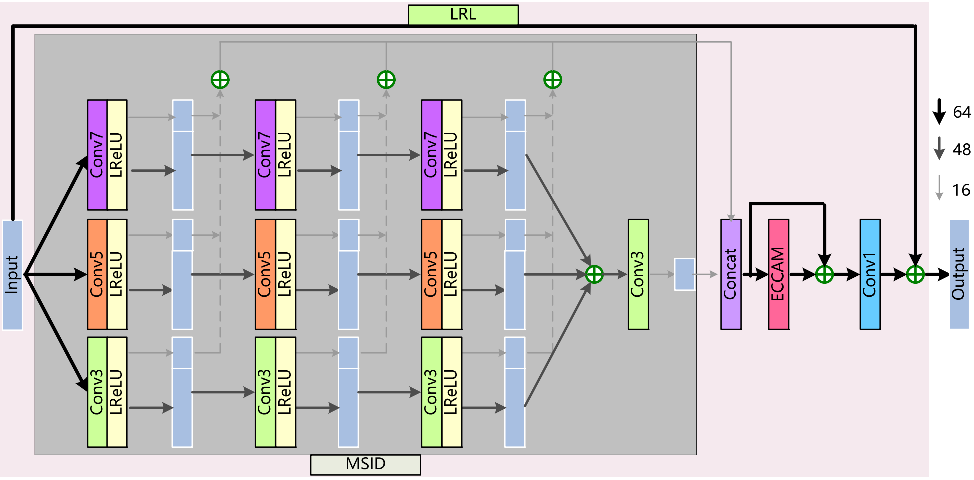


Fig. 2. Detailed structure of MSIDAB; 64, 48, and 16 denote the numbers of output channels of the convolution layers

The wide spatial spans of ground objects in remote sensing images result in a diverse range of scales and shapes, and this requires a network with greater awareness of high-frequency features. However, a traditional CNN usually uses a single-scale convolutional layer to detect the feature information, which often leads to insufficiently accurate information. The success of GoogLeNet [21] was due to the realization that feature extraction could be facilitated by paralleling convolution kernels of different scales. Consequently, a large number of researchers have focused on how to obtain image features on different scales to improve the SR effect. Generally speaking, a large-sized convolution kernel can perceive complex features but tends to lose detailed information, where a small-sized convolution kernel is sensitive to detailed information but lacks the capacity to perceive complex features. Considering this, by fusing the information multi-distillation strategy and the idea of multi-scale feature representations, MSID is leveraged to both reduces the number of model parameters and captures more high-frequency information. Compared to previous SR approaches that fuse multi-scale feature extraction in the information-distillation mechanism [11, 12], the proposed MSID focuses more on the fusion and utilization of distilled features across various receptive fields.

In the MSID module (as marked with the gray background in Fig. 2), the input features are initially processed using a pyramid convolution operation for further multiple successive refining distillation steps. For each step, the preceding extracted features are divided into two parts by adopting a channel-splitting strategy. One part is preserved as the local refined features, and the other part will be processed in the next computing unit as the remaining features. Given the input, this procedure in the MSIDAB can be viewed as:

$$\left[F^{n}_{refined-1-i}, F^{n}_{remaining-1-i}\right] = Split^{n}_{1}\left(Conv^{n}_{1-i}\left(F^{n}_{in}\right)\right), \tag{6}$$

$$\left[F^{n}_{refined-2-i}, F^{n}_{remaining-2-i}\right] = Split^{n}_{2}\left(Conv^{n}_{2-i}\left(F^{n}_{remaining-1-i}\right)\right), \tag{7}$$

$$\left[F_{refined-3-i}^{n}, F_{remaining-3-i}^{n}\right] = Split_{3}^{n}\left(Conv_{3-i}^{n}\left(F_{remaining-2-i}^{n}\right)\right), \tag{8}$$

$$F_{refined-4-3}^{n} = Conv_{4-3}^{n}\left(Sum(F_{remaining-3-i}^{n})\right), \tag{9}$$

where, $Conv_{k-i}^{n}(\cdot)$ and $Split_{k}^{n}(\cdot)$ respectively denote the kth convolution layer (followed by a LeakyReLU activation function) with a kernel size of $i\times i$ and the kth channel separation of the nth MSIDAB; $F_{refined-j-i}^{n}$ and $F_{remaining-j-i}^{n}$ respectively denote the jth refined and remaining features of the branch with a convolutional kernel size of $i\times i$ in the MSIDAB; and $Sum(\cdot)$ denotes the pixel-by-pixel summation operation on the feature map. Such a pyramid structure with parallel convolutions can extract diverse features of different levels, which allows the exploration of detailed information in low-dimension space to reach an optimal solution. Specifically, the pyramidal convolution is implemented internally by grouping convolution to reduce the computational consumption. The values of i in the kth convolutional layer are set to 3, 5, and 7, corresponding to feature-map groups 1, 2, and 4, respectively. Then, the multi-scale extracted features from each pyramid convolutional layer are fused, and followed by a concatenation operation on the fused features at each step to generate the final refined features. These operations can be expressed as:

$$\begin{aligned} &F_{distilled}^{n} \\ &\quad = Concat\left(Sum\left(F_{refined-1-i}^{n}\right), Sum\left(F_{refined-2-i}^{n}\right), Sum\left(F_{refined-3-i}^{n}\right), F_{refined-4-3}^{n}\right), \end{aligned} \tag{10}$$

where, $Sum(\cdot)$ denotes the feature-map summation operation, and $Concat(\cdot)$ denotes the feature-channel concatenation operation.

The concatenated output of the retained feature maps $F_{distilled}^{n}$ is fed into the ECCAM module, then a local-residual-learning (LRL) operation is utilized to preserve the hierarchical features and facilitate the flow of information. The output of the MSIDAB is obtained as:

$$F_{out}^{n} = H_{compress}(ECCAM(F_{distilled}^{n}) + F_{distilled}^{n}) + F_{in}^{n}, \tag{11}$$

where, $ECCAM(\cdot)$ denotes feature extraction using the ECCAM operation, and $H_{compress}(\cdot)$ fuses the obtained features using a 1×1 convolution layer.

3.2.2. Design of ECCAM structure

Classical CNN-based approaches treat all extracted feature channels equally, which leads to a certain loss of high-frequency information during the transformation from LR space to HR space. Inspired by SENet [22], the residual channel attention network deepens the perception of high-frequency information in images by introducing the CAM, thus HR images are recovered. Furthermore, the IMDN presents a CCAM to significantly improve the SR reconstruction performance by adding contrast variables from the original feature map. Based on CAM and CCAM, the ECCAM module is designed by introducing the role of spatial features in the description of global information. As shown in Fig. 3, the input $F = (F^{1}, \dots, F^{k}, \dots, F^{C})$ consists of C channel descriptors with size $H \times W$, along the spatial axis of F, the channel-feature descriptor $D_{Channel}^{k}$ of F^{k} is calculated by a GAP operation:

$$D_{Channel}^{k} = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F^{k}(i,j), \tag{12}$$

where, $F^{k}(i,j)$ denotes the pixel value at location (i,j) of F^{k}. By the GAP operation, the channel-wise statistics $D_{Channel} = (D_{Channel}^{1}, \dots, D_{Channel}^{k}, \dots, D_{Channel}^{C})$ can be obtained, corresponding to $F = (F^{1}, \dots, F^{k}, \dots, F^{C})$, respectively.

Taking into account the variation in the spatial distribution of a remote sensing image, we supplement one spatial-feature descriptor $D_{Spatial}$ with size H×W by applying an average-pooling (AP) operation along the channel axis of F:

$$D_{Spatial}(i,j) = \frac{1}{C}\sum_{k=1}^{C} F^k(i,j), \tag{13}$$

where, $D_{Spatial}(i,j)$ denotes the aggregated spatial descriptor at position (i,j) of F. The enhanced contrast information value is then calculated as:

$$R_{Channel}^k = \sqrt{\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(F^k(i,j) - D_{Channel}^k)^2} + \sqrt{\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(F^k(i,j) - D_{Spatial}(i,j)\right)^2} + D_{Channel}^k, \tag{14}$$

where, $R_{Channel}^k$ denotes the kth channel element output. In this way, compared to the CCAM, the deviation in the proposed method can better reflect the relative spatial importance of each feature map in a whole layer.

After this, similar to SENet, all the channel-wise features are adaptively recalibrated by a multi-layer perceptron structure. Through this operation, valuable information are emphasized among the channels while suppressing useless information. The final output of the ECCAM module is obtained as:

$$F_{ECCAM}^k = f_\sigma\left[W_U\left(ReLU\left(W_D\left(R_{Channel}^k\right) + b_D\right)\right) + b_U\right] \otimes F^k, \tag{15}$$

where, $f_\sigma[\cdot]$ and $ReLU(\cdot)$ denote the Sigmoid gating and ReLU activation functions, respectively; $W_D \in \mathbb{R}^{\frac{C}{r}\times C\times 1\times 1}$ and $b_D \in \mathbb{R}^{\frac{C}{r}}$ denote the weights and bias in the first 1×1 convolution layer, which decreases the channel dimensions of $R_{Channel}$ by the reduction ratio r; $W_U \in \mathbb{R}^{C\times\frac{C}{r}\times 1\times 1}$ and $b_U \in \mathbb{R}^C$ denote the weights and bias in another 1×1 convolution layer, which increases the channel dimensions back to the original number; and $\otimes$ denotes an element-wise multiplication operation. Specifically, the reduction ratio r is defined as 16, which is consistent with the IMDN.

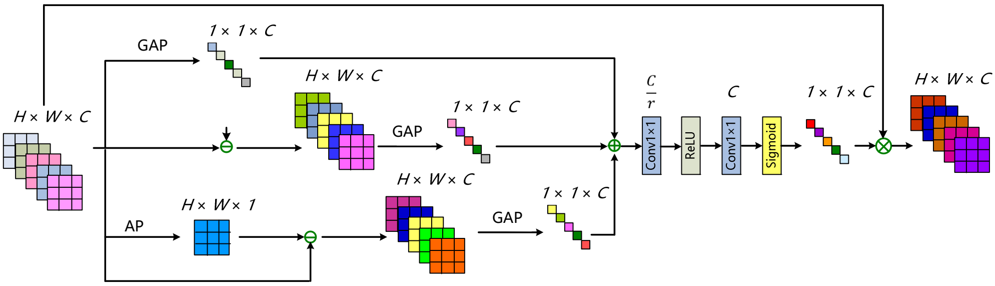


Fig. 3. Detailed structure of ECCAM

3.2.3. Loss function

The loss function in this approach is used to update the gradient parameters by minimizing the distance between the reconstruction result and the ground truth. Following the approach of previous work [10], the L1-norm loss between the ground-truth HR image and the reconstructed HR image is used to measure this distance. Given a training set $\left\{I_{LR}^{i}, I_{HR}^{i}\right\}_{i=1}^{N}$ that contains matching LR-HR pairs, the loss function of MSIDAN can be formulated as:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|H_{\mathrm{MSIDAN}}\left(I_{LR}^{i}\right) - I_{HR}^{i}\|_{1}, \tag{16}$$

where, Θ denotes the training-parameter set of MSIDAN, $\|\cdot\|_1$ denotes the L1 norm, and $H_{\mathrm{MSIDAN}}(\cdot)$ denotes the proposed image SR model.

4. 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.

4.1. Data

Following recent works [23], to facilitate experimental comparisons with other SR networks, the publicly available, high-quality remote sensing dataset, Aerial Image Dataset (AID) [24], was selected as the training dataset, which includes 30 classes of scene images, and each class has approximately 220-420 pieces (a total of 10000 pieces) with 600×600 in the RGB space. The training images are augmented via three strategies: (1) horizontal flipping; (2) vertical flipping; and (3) 90° rotation. After training, the SR models are evaluated using four remote sensing image datasets: RSSCN7 [25], WHU-RS19 [26], NWPU VHR-10 [27], and Cars Overhead With Context (COWC) [28]. In the experiments, 100 samples were randomly selected from the original datasets to create four new datasets for performance evaluation. To generate the image pairs for training and testing, LR samples were created by downsampling the corresponding HR images using a Bicubic interpolation algorithm with scale factors of 2, 3, and 4.

4.2. Evaluation indexes

The average peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [29] are two image-quality evaluation metrics that are commonly employed to objectively evaluate SR models. The PSNR describes the distortion of the reconstructed images caused by random noise; it is expressed as:

$$PSNR(x,y) = 10\log_{10}\left(\frac{I_{max}^{2}}{\frac{1}{W\times H}\sum_{i=1}^{H}\sum_{j=1}^{W}[x(i,j)-y(i,j)]^{2}}\right), \tag{17}$$

where, x and y denote a ground truth and its super-resolved version of size $W\times H$, respectively; and I_{max} denotes the peak pixel value (which is 255 for RGB images). The PSNR is measured in dB, and a higher PSNR score indicates that the super-resolved image has higher pixel fidelity and image quality.

The SSIM index measures the degree of structural similarity between two images by estimating luminance, contrast, and structure. Thus, the evaluation mechanism of SSIM is closer to the human visual system in terms of overall image composition. The SSIM value is computed as:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \tag{18}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \tag{19}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{20}$$

$$SSIM(x,y) = [l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma], \tag{21}$$

where, μ_x, μ_y, σ_x, σ_y, and σ_{xy} denote the average values, variance values, and covariance of the original HR image x and reconstructed SR image y, respectively; c_1, c_2, and c_3 are constants set to avoid calculation instability. In general, the values $\alpha = \beta = \gamma = 1$ are set. The value of SSIM ranges from 0 to 1, a higher SSIM value represents a better-quality SR image. Specifically, as with previous works [8, 9], the SR results are typically evaluated with the PSNR and SSIM values on the luminance (Y) channel in the transformed YCbCr color space.

4.3. Implementation details

To ensure a fair comparison, all models in this study were retrained using the training set mentioned above, without any pre-training and fine-tuning processes. In each training batch, 16 random patches of size 48×48 pixels were cropped from the LR samples as the input for model training, and their HR counterparts of 96×96, 144×144, and 192×192 pixels were generated with scale factors of ×2, ×3, and ×4, respectively. Following the previous work [23], the initial learning rate is 1×10^{-4} and decreases to 10 % every 500 epochs in the process of back-propagation. The Adam algorithm with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ was adopted to optimize the SR model parameters. All SR networks were trained for 1500 epochs in total. The number of MSIDABs was set to 6, in accordance with the IMDN approach. All experiments involved in this paper were conducted on the same device, i.e., a single NVIDIA RTX 3090 GPU and a 3.40 GHz AMD Ryzen 5700X CPU. More setting detail of experiments are listed in Table 1.

Table 1. Setting parameters for proposed MSIDAN

Batch size	48×48
Patch size	16
The numbers of MSIDABs	6
Initial learning rate	1×10^{-4}
Channels	64
Channels-refined (split)	16
Optimizer (Adam)	$\beta_1 = 0.9$, $\beta_2 = 0.999$

4.4. Experimental results

4.4.1. Comparisons with other methods

This section presents a comparison of MSIDAN with other advanced SISR methods: Bicubic interpolation, widely activated super-resolution network (WDSR) [30], multi-attentive feature fusion super-resolution network (MAFFSRN) [19], IMDN [10], edge-oriented convolution block based super-resolution model (ECBSR) [16], and enhanced super-resolution group convolutional neural network (ESRGCNN) [31]. Bicubic interpolation is a representative interpolation algorithm; and others are comparable deep CNN-based approaches.

4.4.1.1. Quantitative results

Table 2 displays the quantitative evaluation results (PSNR/SSIM), with the optimal and second-best values highlighted in bold and underline, respectively. It is clear that MSIDAN achieves the best results for all four datasets. With upscaling factors of ×2 and ×4 on the RSSCN7 dataset, the PSNR gains achieved by MSIDAN in comparison to the second-best algorithm (WDSR) are 0.03989 dB and 0.01394 dB, respectively. For an upscaling factor of ×3 on the RSSCN7 dataset, MSIDAN obtains a gain of 0.00263 dB compared to the second-best algorithm (ESRGCNN). With upscaling factors of ×2 and ×3 on the WHU-RS19 dataset, the PSNR gains achieved by MSIDAN in comparison to the second-best algorithm (MAFFSRN) are 0.03780 dB and 0.01895 dB, respectively. For an upscaling factor of ×4 on the WHU-RS19 dataset, MSIDAN obtains a gain of 0.01874 dB compared to the second-best algorithm (WDSR). With an upscaling factor of ×2 on the NWPU VHR-10 dataset, the PSNR gain achieved by MSIDAN in comparison to the second-best algorithm (ECBSR) is 0.04332 dB. For upscaling factors of ×3 and ×4 on the NWPU VHR-10 dataset, MSIDAN obtains values 0.01962 dB and 0.00675 dB higher than the second-best algorithm (WDSR), respectively. With upscaling factors of ×2 and ×3 on the COWC dataset, the PSNR gains achieved by MSIDAN in comparison to the second-best algorithm (IMDN) are 0.04717 and 0.03364 dB, respectively. For an upscaling factor of ×4 on the COWC dataset, MSIDAN obtains a value 0.00433 dB higher than the second-best algorithm (ESRGCNN). Compared with the IMDN, the MSIDAN improves the PSNR by 0.03312 dB, 0.06031 dB, 0.05319 dB, and 0.03812 dB on the RSSCN7, WHU-RS19, NWPU VHR-10, and COWC datasets, respectively, and the SSIM by 0.0011, 0.0011, 0.0012, and 0.0010 respectively.

Table 2. Comparison of SR used different methods. Bold and underlining represent optimal and second-best performance, respectively

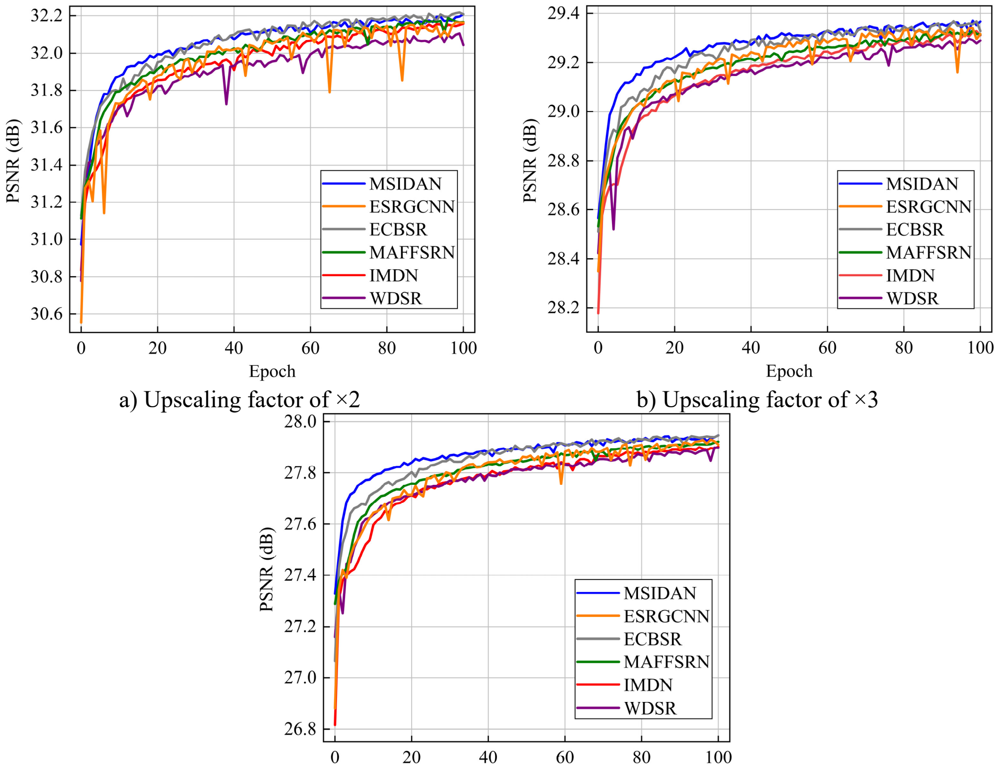
Scale	Method	RSSCN7 [25]	WHU-RS19 [26]	NWPU VHR-10 [27]	COWC [28]
×2	Bicubic	30.82776/0.8460	34.68401/0.9232	32.79005/0.9094	32.21539/0.8772
	WDSR [30]	32.33108/0.8811	36.48032/0.9444	35.22087/0.9322	34.25776/0.8976
	MAFFSRN [19]	32.31966/0.8812	36.51930/0.9448	35.21966/0.9323	34.23883/0.8978
	IMDN [10]	32.32418/0.8811	36.48802/0.9444	35.21106/0.9320	34.29290/0.8982
	ECBSR [16]	32.32222/0.8813	36.49464/0.9447	35.22308/0.9325	34.22104/0.8977
	ESRGCNN [31]	32.32399/0.8813	36.49642/0.9446	35.22092/0.9324	34.25008/0.8980
	MSIDAN	32.36407/0.8821	36.55710/0.9452	35.26640/0.9329	34.34007/0.8991
×3	Bicubic	28.42559/0.7313	30.82741/0.8294	29.62905/0.8271	30.03914/0.8064
	WDSR [30]	29.48540/0.7752	32.53147/0.8744	31.93973/0.8712	31.51455/0.8335
	MAFFSRN [19]	29.47462/0.7753	32.53849/0.8744	31.91697/0.8706	31.49830/0.8334
	IMDN [10]	29.47146/0.7741	32.50749/0.8736	31.91500/0.8705	31.55127/0.8338
	ECBSR [16]	29.46520/0.7746	32.50125/0.8738	31.89514/0.8704	31.49211/0.8331
	ESRGCNN [31]	29.48778/0.7754	32.53231/0.8744	31.92378/0.8713	31.54567/0.8342
	MSIDAN	29.49041/0.7755	32.55744/0.8747	31.95935/0.8715	31.58491/0.8345
×4	Bicubic	27.21346/0.6543	28.76653/0.7477	27.92820/0.7634	28.76707/0.7546
	WDSR [30]	28.04068/0.6975	30.28405/0.8071	29.88932/0.8172	30.13465/0.7881
	MAFFSRN [19]	28.03602/0.6975	30.28001/0.8067	29.87874/0.8165	30.10831/0.7874
	IMDN [10]	28.01410/0.6963	30.24089/0.8056	29.83619/0.8156	30.10782/0.7872
	ECBSR [16]	28.01418/0.6964	30.24093/0.8058	29.82262/0.8155	30.11199/0.7874
	ESRGCNN [31]	28.03523/0.6972	30.27773/0.8068	29.85653/0.8167	30.13846/0.7880
	MSIDAN	28.05462/0.6977	30.30279/0.8073	29.89607/0.8174	30.14279/0.7882

Fig. 4 shows a comparison of the PSNR values between the MSIDAN and the other networks using the RSSCN7 dataset in the epoch range of 0 to 100. As depicted in Fig. 4, MSIDAN (blue lines) shows notably superior SR accuracy and faster convergence than ESRGCNN (orange line), WDSR (purple line), IMDN (red line), and MAFFSRN (green line). This further demonstrates the improvements provided by MSIDAN in remote sensing image reconstruction.

4.4.1.2. Qualitative comparisons

Fig. 5 shows some super-resolved examples of the different methods. Specifically, for clearer presentation and easier comparison, some regions are denoted by a red rectangle in the original HR samples, and the corresponding super-resolved results are enlarged. From Fig. 5, MSIDAN obtains the best PSNR and SSIM, which is closer to original HR images than other models.

Fig. 5(a) lists the qualitative comparison of various methods for "Industry" scene ×4 SR reconstruction. The building reconstructed by the Bicubic is quite blurry, and the results do not effectively present the original spatial structure. Compared to WDSR, MAFFSRN, IMDN, ECBSR, and ESRGCNN, MSIDAN achieves the highest structural integrity while exhibiting minimal distortion. Fig. 5(b) lists the qualitative comparison of various methods for "Airplane" scene ×4 SR reconstruction. The lines recovered by Bicubic have a mosaic effect and become serrated. The MAFFSRN has the second worst result, only slightly higher than the Bicubic method. More recent methods (WDSR, IMDN, ECBSR, and ESRGCNN) can obtain global contrast information, but with significant unreal artifacts. In comparison to other approaches, MSIDAN achieves clearer edges and produces a sharper image. Due to the low quality of LR images, most models are unable to generate more spatial details of remote sensing images, resulting in blurred boundaries of reconstructed objects. In contrast, MSIDAN represents a significant improvement in the restoration of LR input images, and this is consistent with quantitative comparisons.



a) Upscaling factor of ×2

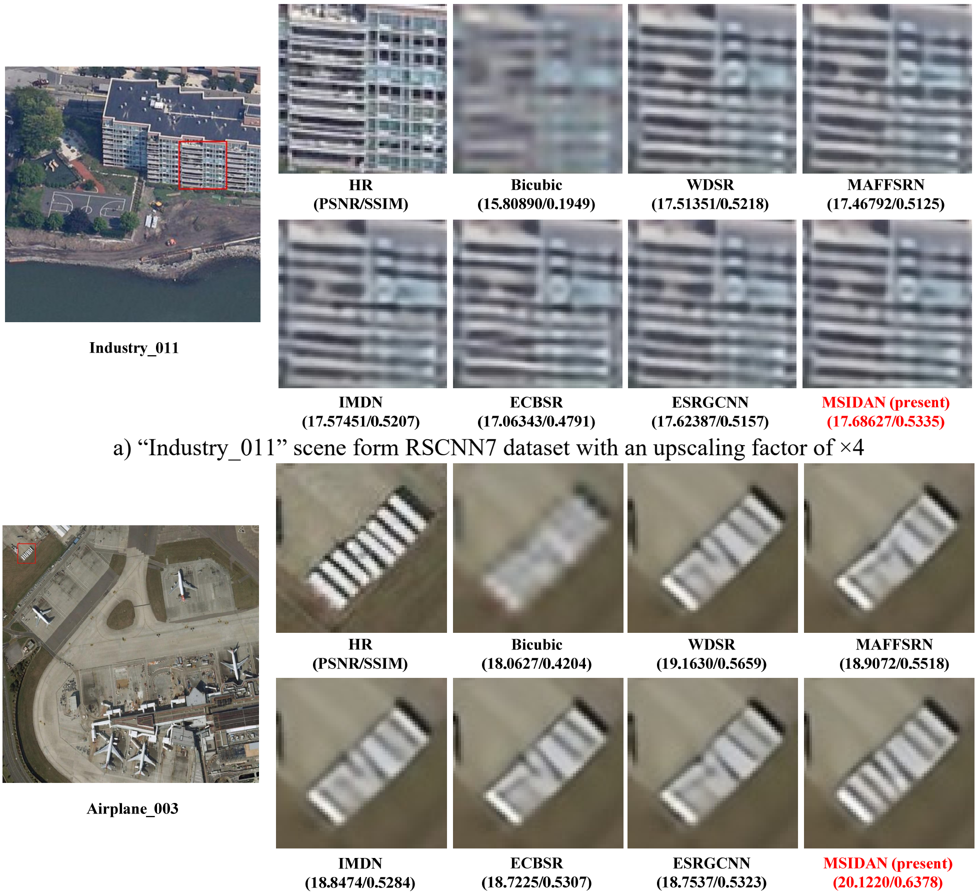
b) Upscaling factor of ×3

c) Upscaling factor of ×4

Fig. 4. Curves for different approaches representing the PSNR with the RSSCN7 dataset in the epoch range 0 to 100

4.4.1.3. Comparison of trade-offs

The model size is an important parameter to consider when evaluating a lightweight SR model. To better represent the parametric efficiency of the algorithms, other commonly used lightweight models such as VDSR [7], PAN [20], IDN [15], CARN [14], RFDN [18], and LBNet [17] are also statistically used for comparative analysis. Fig. 6(a) shows the trade-off of PSNR vs. number of parameters on the WHU-RS19 dataset with an upscaling factor of ×2. The x-axis represents the SR model size, and the y-axis represents the average PSNR value. On the one hand, the proposed MSIDAN achieves an optimal PSNR indicator with a model size that is roughly half that of ECBSR. On the other hand, although VDSR and PAN have smaller numbers of parameters, the SR results of these two models are significantly worse than those from MSIDAN.



a) “Industry_011” scene form RSCNN7 dataset with an upscaling factor of ×4

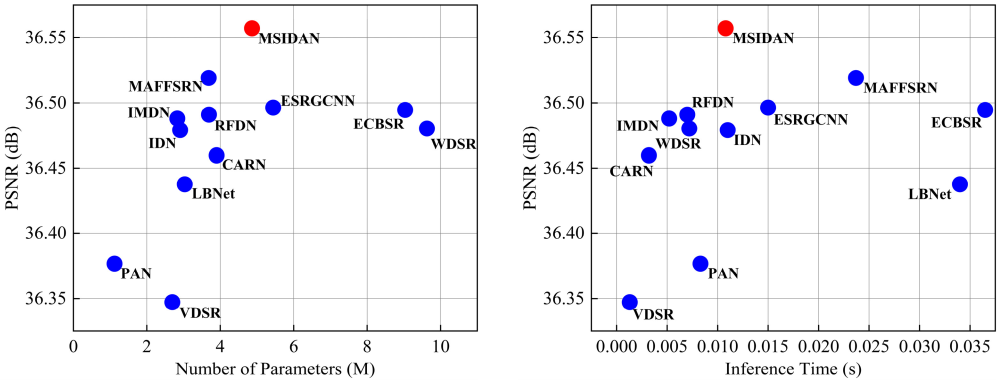
b) “Airplane_003” scene form NWPU VHR-10 dataset with an upscaling factor of ×4

Fig. 5. Qualitative comparisons of super-resolved results

Fig. 6(b) shows the trade-off of PSNR vs. inference time on the WHU-RS19 dataset with an upscaling factor of ×2. Obviously, the MSIDAN achieves the best PSNR results under the premise of comparable execution time. From these findings, it can be concluded that MSIDAN achieves better parameter efficiency, making it more feasible to deploy it on mobile devices.

4.5. Analysis of MSIDAB

MSIDAB consists of two main modules: MSID and ECCAM. To validate the necessity of these components, several MSIDAN ablation experiments were performed with an upscaling factor of ×2. The baseline structure for comparison is set as follows: The MSID was replaced with a single-scale convolution (3×3) information-distillation operation and the ECCAM was removed. Table 3 lists the ablation results.



a) Model size

b) Inference time

Fig. 6. Comparison of trade-offs of different SR networks on the WHU-RS19 dataset with an upscaling factor of ×2

Table 3. Ablation comparison of MSIDAN with ×2 upscale. Bold indicates the optimal performance

MSID	ECCAM	RSSCN7 [25]	WHU-RS19 [26]	NWPU VHR-10 [27]	COWC [28]
×	×	32.30724/0.8808	36.45777/0.9443	35.19072/0.9320	34.20195/0.8972
✓	×	32.35897/0.8819	36.54685/0.9450	35.26435/0.9328	34.33416/0.8991
×	✓	32.33146/0.8812	36.49090/0.9444	35.21422/0.9320	34.29329/0.8982
✓	✓	32.36407/0.8821	36.55710/0.9452	35.26640/0.9329	34.34007/0.8991

To perform a comprehensive analysis of MSID, a series of model variants with various convolutional kernel settings under different branches was designed, and a quantitative comparison of the results is presented in Table 4. It is clear from these results that the models with multiple branches (when the number of branches is 2 or 3) achieve better performance. In other words, the network achieves better feature-information acquisition through the structural design of multiple branches with different convolutional-kernel scales.

Table 4. Comparison using various convolutional kernel settings in the MISD module with ×2 upscale. Bold indicates the optimal performance

3×3	5×5	7×7	RSSCN7 [25]	WHU-RS19 [26]	NWPU VHR-10 [27]	COWC [28]
✓			32.30724/0.8808	36.45777/0.9443	35.19072/0.9320	34.20195/0.8972
	✓		32.27799/0.8804	36.42902/0.9440	35.14493/0.9315	34.18073/0.8971
		✓	32.24320/0.8798	36.40860/0.9438	35.10586/0.9311	34.14091/0.8966
✓	✓		32.34245/0.8816	36.52401/0.9449	35.24245/0.9326	34.27895/0.8983
✓		✓	32.33847/0.8814	36.51078/0.9447	35.23264/0.9324	34.28733/0.8982
	✓	✓	32.31441/0.8810	36.47043/0.9444	35.20010/0.9321	34.25179/0.8977
✓	✓	✓	32.35897/0.8819	36.54685/0.9450	35.26435/0.9328	34.33416/0.8991

Table 5. Comparison using different attention mechanisms with an upscaling factor of ×2. Bold indicates the optimal performance

Mechanism	RSSCN7 [25]	WHU-RS19 [26]	NWPU VHR-10 [27]	COWC [28]
/	32.30724/0.8808	36.45777/0.9443	35.19072/0.9320	34.20195/0.8972
CAM	32.32394/0.8811	36.48303/0.9442	35.21061/0.9320	34.28494/0.8980
CCAM	32.32418/0.8811	36.48802/0.9444	35.21106/0.9320	34.29290/0.8982
ECCAM	32.33146/0.8812	36.49090/0.9444	35.21422/0.9321	34.29329/0.8982

To further demonstrate the improvements brought about by the ECCAM module, the impact of the different attention mechanisms on model performance was also verified. Specifically, the multi-scale convolutions operation was removed for quick verification. It can be seen from Table 5 that the ECCAM module outperforms the CAM and CCAM modules on all datasets. The CAM module uses a GAP operation to generate channel-wise statistics. Although the GAP operation indeed leads to performance improvements in terms of PSNR values (e.g., 0.01670 dB on RSSCN7, 0.02526 dB on WHU-RS19, 0.01989 dB on NWPU VHR-10, and 0.08299 dB on COWC), it lacks the information sensitivity regarding high-frequency details (e.g., structures, edges, and textures) that is essential for SR recovery of remote sensing images. Considering this, the CCAM module replaces the GAP operation with the summation of the standard deviation and mean of each feature map to correct the attention value. In this way, the CCAM module achieves better performance gains in terms of PSNR values (e.g., 0.01694 dB on RSSCN7, 0.03025 dB on WHU-RS19, 0.02034 dB on NWPU VHR-10, and 0.09095 dB on COWC). Compared with the CCAM module, the deviation in the ECCAM module further reflects the relative spatial importance of each feature map in a whole layer by introducing the spatial-feature descriptor, thus generating more balanced attention for the perception of high-frequency details. In summary, these comparisons consistently demonstrate the superiority of our proposed MSID and ECCAM.

In general, deeper networks achieve better SR performance. MSIDAB is used as a deep feature-extraction component to enhance the perception of high-frequency information in LR spaces. Therefore, to verify the influence of using different numbers of MSIDABs in MSIDAN, we trained models with different depths ($M=$ 6, 9, and 12), and a corresponding quantitative comparison of the results is presented in Table 6. These results show that a deeper network achieves a better SR reconstruction effect. The PSNR gains of MSIDAN-M12 over MSIDAN-M6 are 0.00752, 0.12675, 0.11851, and 0.12904 dB with an upscaling factor of ×2 on RSSCN7, WHU-RS19, NWPU VHR-10, and COWC, respectively. From this finding, we can observe that the MSIDAB can serve as a commendable feature-extraction component in LR space for training deep SR models, and it allows the depth of the network to be flexibly adjusted to the capabilities of specific hardware devices.

Table 6. Comparison using different network depths with an upscaling factor of ×2. Bold indicates the optimal performance

Network depths	RSSCN7 [25]	WHU-RS19 [26]	NWPU VHR-10 [27]	COWC [28]
MSIDAN-M6	32.32394/0.8811	36.48303/0.9442	35.21061/0.9320	34.28494/0.8980
MSIDAN-M9	32.39592/0.8826	36.59280/0.9455	35.31486/0.9335	34.38315/0.8996
MSIDAN-M12	32.40296/0.8827	36.60978/0.9456	35.32912/0.9336	34.41398/0.9000

5. Conclusions

The technique of super-resolution reconstruction is crucial for measurement engineering applications, as high-resolution images can provide more accurate boundary and spatial information of land features, which helps to perform accurate quantitative analysis and interpretation of surface features. This paper presents a MSADAN for remote sensing imagery super-resolution, aiming to overcome some critical limitations inherent in the existing CNN-based SR methods. The MSIDAN framework effectively stacks several MSIDABs together, in which the hierarchical features are progressively refined step-by-step while the number of model parameters is decreased by a channel-splitting operation. The core modules of MSIDAB are MSID and ECCAM. In this system, inspired by the strategies of information distillation and multi-size feature learning, the MSID module is employed to distill and fuse multi-level semantic features step-by-step while reducing the number of model parameters. The ECCAM module is then capable of perceiving high-frequency information in remote sensing images by automatically encoding the weight values of candidate features. The experimental results on RSSCN7, WHU-RS19, NWPU VHR-10, and COWC datasets indicate that: (1) MSIDAN achieves significant improvements in both the quantitative evaluation indicators and qualitative visualization results compared to other comparable approaches. Compared with IMDN, MSIDAN improves the PSNR by 0.03312 dB, 0.06031 dB, 0.05319 dB, and 0.03812 dB on the RSSCN7, WHU-RS19, NWPU VHR-10, and COWC datasets, respectively. (2) In comparison with other advanced SR methods, MSIDAN achieves a more appropriate balance by jointly considering SR performance and model size. A statistical comparison of PSNR vs. number of parameters on the WHU-RS19 dataset with an upscaling factor of ×2 shows that MSIDAN achieves an optimal PSNR indicator with a model size that is roughly half that of ECBSR. (3) MSIDAB is structured in a rational and efficient way, and it can be used as a building block for deep SR reconstruction networks. Future work will focus on achieving end-to-end joint optimization of different networks for measurement applications with other tasks. Furthermore, the ground object segmentation based on supervised learning will be studied to expand the application scenarios of the proposed MSIDAN.

Nevertheless, the MISDAN does have some shortcomings. Supervised training requires a large number of paired datasets. However, difficult to obtain high quality images corresponding to degraded images. Therefore, the use of semi-supervised or unsupervised methods is an important research direction for subsequent attention in this work.

Acknowledgements

The research is supported by the National Natural Science Foundation of China (Grant No. 62205168), the National Natural Science Foundation of China (Grant No. 51276151), the Natural Science Foundation of Fujian Province (Grant No. 2023J011439), 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, funding acquisition. Liaoni Wu: resources, supervision, validation, funding acquisition. Yiqing Cao: data Curation, supervision, software, Validation, funding acquisition. Mingen Zhong: methodology, writing – review, investigation, funding acquisition.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] 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

[2] H. Luo, J. Liao, and G. Shen, "Combining remote sensing and social media data for flood mapping: a case study in Linhai, Zhejiang Province, China," *Journal of Applied Remote Sensing*, Vol. 17, No. 2, p. 02450, Apr. 2023, https://doi.org/10.1117/1.jrs.17.024507

[3] C. Pany, U. K. Tripathy, and L. Misra, "Application of artificial neural network and autoregressive model in stream flow forecasting," *Journal of Indian Water Works Association*, Vol. 33, No. 1, pp. 61–68, 2001.

[4] H. Ma and Y. Han, "Logo recognition of vehicles based on deep convolutional generative adversarial networks," *Journal of Measurements in Engineering*, Vol. 12, No. 2, pp. 353–365, Jun. 2024, https://doi.org/10.21595/jme.2024.23849

[5] 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

[6] 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

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

[8] 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

[9] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2018, pp. 294–310, https://doi.org/10.1007/978-3-030-01234-2_18

[10] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *MM '19: The 27th ACM International Conference on Multimedia*, pp. 2024–2032, Oct. 2019, https://doi.org/10.1145/3343031.3351084

[11] J. Wu, L. Cheng, M. Chen, T. Wang, Z. Wang, and H. Wu, "Super-resolution infrared imaging via multi-receptive field information distillation network," *Optics and Lasers in Engineering*, Vol. 145, p. 106681, Oct. 2021, https://doi.org/10.1016/j.optlaseng.2021.106681

[12] H. Zang, Y. Zhao, C. Niu, H. Zhang, and S. Zhan, "Attention network with information distillation for super-resolution," *Entropy*, Vol. 24, No. 9, p. 1226, Sep. 2022, https://doi.org/10.3390/e24091226

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Jun. 2016, https://doi.org/10.1109/cvpr.2016.90

[14] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Lecture Notes in Computer Science*, pp. 256–272, Oct. 2018, https://doi.org/10.1007/978-3-030-01249-6_16

[15] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 723–731, Jun. 2018, https://doi.org/10.1109/cvpr.2018.00082

[16] X. Zhang, H. Zeng, and L. Zhang, "Edge-oriented convolution block for real-time super resolution on mobile devices," in *MM '21: ACM Multimedia Conference*, pp. 4034–4043, Oct. 2021, https://doi.org/10.1145/3474085.3475291

[17] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, and T. Zeng, "Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer," *arXiv:2204.13286*, Jan. 2022, https://doi.org/10.48550/arxiv.2204.13286

[18] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in *Lecture Notes in Computer Science*, Vol. 12537, Cham: Springer International Publishing, 2021, pp. 41–55, https://doi.org/10.1007/978-3-030-67070-2_2

[19] A. Muqeet, J. Hwang, S. Yang, J. Kang, Y. Kim, and S.-H. Bae, "Multi-attention based ultra lightweight image super-resolution," in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2021, pp. 103–118, https://doi.org/10.1007/978-3-030-67070-2_6

[20] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "efficient image super-resolution using pixel attention," in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2021, pp. 56–72, https://doi.org/10.1007/978-3-030-67070-2_3

[21] C. Szegedy et al., "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, https://doi.org/10.1109/cvpr.2015.7298594

[22] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 8, pp. 2011–2023, Aug. 2020, https://doi.org/10.1109/tpami.2019.2913372

[23] 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

[24] G.-S. Xia et al., "AID: a benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55, No. 7, pp. 3965–3981, Jul. 2017, https://doi.org/10.1109/tgrs.2017.2685945

[25] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, Vol. 12, No. 11, pp. 2321–2325, Nov. 2015, https://doi.org/10.1109/lgrs.2015.2475299

[26] D. Dai and W. Yang, "satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and Remote Sensing Letters*, Vol. 8, No. 1, pp. 173–176, Jan. 2011, https://doi.org/10.1109/lgrs.2010.2055033

[27] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 54, No. 12, pp. 7405–7415, Dec. 2016, https://doi.org/10.1109/tgrs.2016.2601622

[28] 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," in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2016, pp. 785–800, https://doi.org/10.1007/978-3-319-46487-9_48

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, Apr. 2004, https://doi.org/10.1109/tip.2003.819861

[30] J. Yu et al., "Wide activation for efficient and accurate image super-resolution," *arXiv:1808.08718*, Jan. 2018, https://doi.org/10.48550/arxiv.1808.08718

[31] C. Tian, Y. Yuan, S. Zhang, C.-W. Lin, W. Zuo, and D. Zhang, "Image super-resolution with an enhanced group convolutional neural network," *Neural Networks*, Vol. 153, pp. 373–385, Sep. 2022, https://doi.org/10.1016/j.neunet.2022.06.009

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

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 with 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 method of optical system.

Mingen Zhong received the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 2008. He is currently an Professor with the School of Mechanical and Automobile Engineering, Xiamen University of Technology. His research interests include Machine Vision and Artificial Intelligence.