2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Arbitrary-Scale Super-Resolution for Remote Sensing Images with Multi-Branch Feature Enhancement and Scale-Specific Dictionary Attention

Xiaoxuan Ren, Qianqian Wang, Xin Jin(✉), and Qian Jiang

[1] School of Software, Yunnan University, Kunming, 650000, China
renxiaoxuan@stu.ynu.edu.cn; wangqianqian@stu.ynu.edu.cn; xinxin_jin@163.com; jiangqian_1221@163.com

Abstract. For remote sensing image processing, high quality images are particularly essential because of their more detailed texture information and clearer edges. Image super-resolution (SR) could reconstruct the high-resolution (HR) image from its low-resolution (LR) counterpart, overcoming the limitations of devices and environmental conditions. The shortcoming of most SR methods is that they could only be applied to fixed-scale SR task, which requires more training and deployment cost. Therefore, arbitrary-scale SR approaches are proposed to restore HR images of different scales with a single model. However, most approaches only use simple MLPs or the local attention mechanism in the decoding phase, which limits the representative power of the model. In this work, we propose an arbitrary-scale super-resolution method for remote sensing images with Multi-branch Feature Enhancement and Scale-specific Dictionary Attention (MFESDA). We use a Multi-branch Feature Enhancement (MFE) module which combines global information and scale-aware attention to capture more informative features. Moreover, we design a Scale-specific Multi-level Dictionary Attention Modulation (SMDAM) module in the decoding process which makes use of scale-specific priors to improve the performance. The experimental results have shown that the proposed model performs better than other arbitrary-scale SR approaches and our visual quality is higher than other approaches.

Keywords: Remote sensing, super resolution, implicit neural representation, attention mechanism.

1 Introduction

Single Image Super-resolution (SISR) is a classical computer vision task with the goal of restoring the high-resolution (HR) image from the low-resolution (LR) counterpart. In the practical applications of remote sensing image processing, high quality images are advantageous to many tasks such as object detection, semantic segmentation and image classification. However, due to the limitations of devices and the effects of environmental conditions, it is difficult to obtain high quality images at times. Therefore, image SR has highly attracted attention on account of its great performance of restoring images with clearer texture and structure.

In recent years, many deep learning-based methods [2,4,5,7,10-12,14,15,18,30,34-36] have been proposed to tackle the image SR problem. However, most SR methods could only be applied to images with a fixed scale factor, which means it is required to train and save models for each different scale factor, costing more time for training and occupying more space to save parameters of each scale factor. In order to address this problem, there have been some arbitrary-scale image SR methods [1,3,8,13] based on the implicit neural representation (INR). These methods expect to train a single model for all SR tasks with different scales and usually use the autoencoder architecture consisting of an alternative encoder used for feature extraction and an implicit decoder used for upsampling. However, in the decoding process, these methods usually use simple MLPs or the local attention mechanism without well-designed modules for capturing rich features, which limits the performance of the model.

In this work, we propose an arbitrary-scale SR method for remote sensing images with Multi-branch Feature Enhancement and Scale-specific Dictionary Attention (MFESDA) to tackle this problem. We propose a Multi-branch Feature Enhancement (MFE) module where we combine both global statistics and scale-aware attention mechanism to enhance features extracted by the encoder. Furthermore, considering the arbitrary-scale SR model is responsible for different scale factors, we design a Scale-specific Multi-level Dictionary Attention Modulation (SMDAM) module in which we use scale-specific priors to generate the modulation for the implicit decoder. We have conducted experiments to demonstrate that the proposed model has better performance on objective metrics and the visual quality of images reconstructed by the proposed model is higher than other approaches.

In Summary, the contributions of this work are as follows:

- We propose a Multi-branch Feature Enhancement (MFE) module to combine global statistics and scale-aware features.
- We design a Scale-specific Multi-level Dictionary Attention Modulation (SMDAM) module to exploit scale-specific information in the decoding phase.
- We have conducted experiments to show the performance of the proposed model is superior to other arbitrary-scale SR approaches and our visual quality is higher than other approaches.

2 Related Work

2.1 Single Image Super-resolution

With the rapid development of deep learning techniques, there have been many significant SR methods based on deep learning. SRCNN [5] firstly used a Convolutional Neural Network (CNN) to map the LR image to the corresponding HR image. DRCN [10] proposed a recursive network with recursive-supervision to utilize large context without increasing parameters. ESPCN [18] proposed a sub-pixel convolution operation for efficient upsampling. LapSRN [11] used the laplacian pyramid structure to progressively reconstruct sub-band residuals of HR images in a coarse-to-fine fashion. EDSR [15] improved the structure of residual blocks and generalized the single-scale model to multi-scale architecture. RDN [35] used the dense architecture to exploit hierarchical features. DBPN [7] combined the error feedback mechanism with the iterative upsampling and downsampling process. To alleviate the smooth problem caused by MSE loss, SRGAN [12] used a Generative Adversarial Network (GAN) [6] for image SR to generate more realistic images with higher perceptual quality. After that, ESRGAN [30] enhanced the performance of SRGAN with the Relativistic average GAN (RaGAN) [9] and improved the network structure and the perceptual loss. In order to model dependencies of features across channels or regions, some methods introduced attention mechanism to the network design. RCAN [34] used the channel attention mechanism to model channel-wise dependencies. SAN [4] used second-order feature statistics to learn more discriminative representations. PAN [36] proposed an efficient pixel attention mechanism to decrease the amount of parameters. Considering the great performance of Transformer [28] in vision tasks, some SR methods based on Transformer have been proposed recently. SwinIR [14] introduced Swin Transformer [16] to image SR and used the local attention mechanism and the cross-window interaction to improve the performance. HAT [2] combined channel attention with the window-based self-attention mechanism and proposed an overlapping cross-attention module to improve the window-based self-attention. However, these SR models require to be trained individually and we need to save parameters for each scale factor, leading to more training time and more deployment cost. Moreover, these methods are only used for integer scale factors and are not generalized to non-integer scale factors.

2.2 Arbitrary-scale Super-resolution

The arbitrary-scale image super-resolution task is put forward with the aim of reconstructing HR images of arbitrary scale factors with a single model. Inspired by implicit neural representation (INR), LIIF [3] learnt a continuous representation for images and used the coordinate information and local latent codes around the coordinate to predict the pixel value. LTE [13] proposed a dominant-frequency estimator to facilitate the learning of fine details. CiaoSR [1] proposed an attention-in-attention network where the attention mechanism was applied in order to adaptively predict the ensemble weights and aggregate local features. LMF [8] proposed a computational optimal paradigm which decoupled the HR High-Dimensional (HD) decoding process into decoding in LR HD latent space and HR Low-Dimensional (LD) rendering space. However, these methods tend to use simple MLPs or the local attention mechanism for the implicit decoder, which limits the improvement of the performance. In this work, we will use the Multi-branch Feature Enhancement (MFE) module and Scale-specific Multi-level Dictionary Attention Modulation (SMDAM) module to capture more informative features.

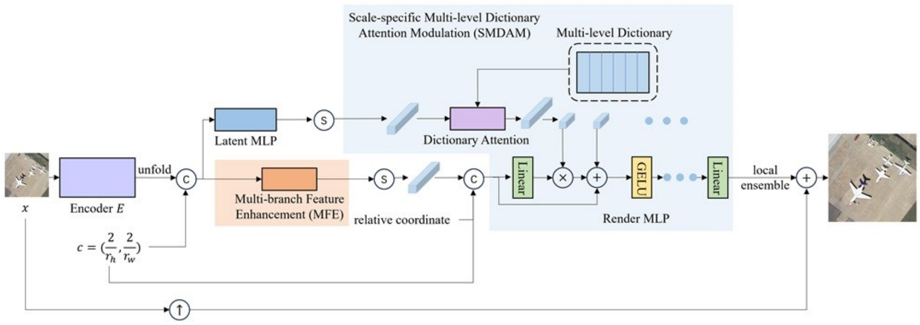

Fig. 1. The overall framework of the proposed MFESDA.

3 Proposed Method

In this section, we will provide a detailed description of the proposed model. As shown in Fig. 1, the proposed model is composed of an encoder and a decoder. The main branch uses an encoder to extract latent features which are then input to two branches. The one branch is connected to the Multi-branch Feature Enhancement (MFE) module to capture rich features, while the other branch uses a latent MLP to generate the initial modulation content of the render MLP. Then for each HR coordinate, the sampled modulation is input to the Scale-specific Multi-level Dictionary Attention Modulation (SMDA) module to generate the final modulation of the render MLP, while the sampled features are input to the render MLP to predict the pixel value. We will then introduce the motivation of the proposed method and then explain the structure design.

3.1 Preliminary and Motivation

Given an input image $I^{LR} \in \mathbb{R}^{H \times W \times 3}$ and arbitrary scale factors r_h and r_w, the aim of arbitrary-scale SR task is to reconstruct the HR image $I^{HR} \in \mathbb{R}^{r_h H \times r_w W \times 3}$. In most arbitrary-scale SR methods, the LR image is firstly input to the encoder network to obtain latent features which are then unfolded to aggregate local features. Then a decoding function is used to predict the pixel value of a 2D coordinate $a \in \mathcal{A}$ in the continuous image domain with the latent code z nearest to the coordinate, which is defined as followed:

$$s(a, I^{LR}) = f_\theta(z, a), \tag{1}$$

where $f_\theta(z, \cdot)$ represents an implicit image function that maps HR coordinates to pixel values. In practice, we usually use the relative coordinate a^* and cell decoding technique [3] to predict the pixel value, which is defined as followed:

$$s(a, I^{LR}) = f_\theta(z, [a^*, c]), \tag{2}$$

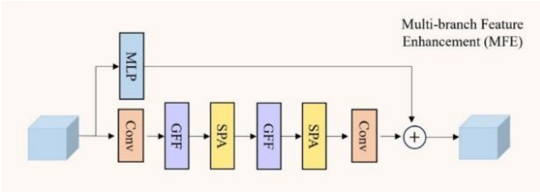

Fig. 2. The structures of the Multi-branch Feature Enhancement (MFE) module.

where $c = (2/r_h, 2/r_w)$. Recently, LMF [8] proposed to decouple the decoding process into latent decoding shared by adjacent coordinates and independent rendering for each HR coordinate, which is defined as followed:

$$\begin{aligned} [m, z_{in}] &= f_{\theta_l}(z, c), \\ s(a, I^{LR}) &= f_{\theta_r}(z_{in}, [a^*, c], m), \end{aligned} \tag{3}$$

where f_{θ_l} denotes the latent MLP with parameters θ_l, f_{θ_r} denotes the render MLP with parameters θ_r, m denotes the modulation of the render MLP, and z_{in} denotes the latent code that is input to the render MLP. In this work, we adopt the similar paradigm as LMF [8] where we use two branches in the decoding process. In order to further improve the performance of the network, we propose to use a Multi-branch Feature Enhancement (MFE) Module to replace the simple MLP which is used to extract features before predicting the pixel value. Besides, considering the representations of input images with different scale factors are also various [29], we design a Scale-specific Multi-level Dictionary Attention Modulation (SMDAM) module to exploit the scale-specific information.

3.2 Multi-branch Feature Enhancement Module

In this section, we will introduce the design of the Multi-branch Feature Enhancement (MFE) module. We will firstly introduce the overall structure of MFE module, then we will explain detailed structures of the Scale-aware Pixel Attention (SPA) Module and the Global Feature Fusion (GFF) module.

According to previous works [13,24], a simple MLP is biased towards learning low-frequency components. Therefore, we design a MFE module which serves as a complement to the decoder. The attention mechanism has been proved to be beneficial to many computer vision tasks [19] because of the capacity to adaptively assign weights to features based on the global statistics. In this work, we combine different attention mechanisms to utilize global information with the proposed MFE module. As shown in Fig. 2, the Multi-branch Feature Enhancement (MFE) module is composed of two branches. The first branch uses a simple MLP to extract low frequency features, and the other branch combines global information and scale-aware features to learn more informative features. We use the convolution operation before and after the modules, and the outputs of branches are aggregated with the summation operation. The process of MFE module is denoted as:

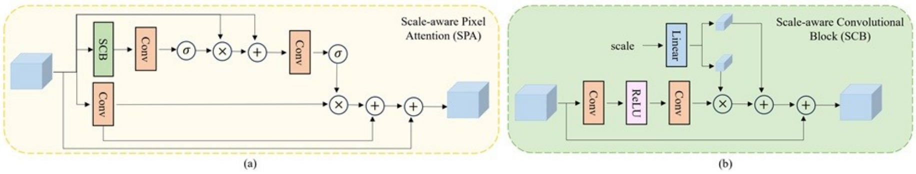

Fig. 3. The structures of the proposed modules. (a) The Scale-aware Pixel Attention (SPA). (b) The Scale-aware Convolutional Block (SCB).

$$\begin{aligned} F_{b_1} &= MLP(z), \\ F_{b_2} &= Conv\left(SPA_n\left(GFF_n\left(\dots SPA_1\left(GFF_1\left(Conv(z)\right)\right)\dots\right)\right)\right), \\ F_{MFE} &= F_{b_1} + F_{b_2}, \end{aligned} \tag{4}$$

where z is the unfolded latent features after the encoder, F_{b_i} represents the output of the i-th branch, F_{MFE} represents the output of MFE module and $n = 2$ represents the number of SPA or GFF modules.

Scale-aware Pixel Attention Module. According to the previous work [29], the learned representations of images with different scales are also diverse and the scale information could be used to learn discriminative features to improve the performance. Inspired by this, we propose a Scale-aware Pixel Attention (SPA) Module to incorporate scale information into the model. There have been some works [20,21,29] which use the external variable to interact with the network. In this work, we also use the scale factor as a variable to adaptively modulate the network. As shown in Fig. 3, we perform the modulation within each Scale-aware Convolutional Block (SCB) where we use the scale factor to generate the scale modulation and the shift modulation for the output of the convolution operation, and the process of SCB is denoted as followed:

$$\begin{aligned} F_{SCB_i} &= Conv\left(ReLU\left(Conv\left(F_{GFF_{i-1}}\right)\right)\right), \\ \left[M_{i_1}, M_{i_2}\right] &= Linear(s), \\ F_{SCB_i} &= F_{SCB_i} \odot M_{i_1} + M_{i_2}, \\ F_{SCB_i} &= F_{SCB_i} + F_{GFF_{i-1}}, \end{aligned} \tag{5}$$

where $F_{GFF_{i-1}} \in \mathbb{R}^{C\times H\times W}$ represents the input features of SCB, s represents the scale factor, $M_{i_1}, M_{i_2} \in \mathbb{R}^{C\times 1\times 1}$ represents the modulations generated by a linear layer, $\odot$ represents the element-wise multiplication.

Then we use SCB to construct SPA module. The pixel attention [36] is a simple attention mechanism which could improve the representation capacity. Here we integrate SCB into cascaded pixel attention blocks with residual connections. To be specific, the output of SCB is input to a convolution layer followed by the Sigmoid function, then we combine the scale-aware features with the input features to generate pixel-wise attention weights which are used to enhance large-scale features, and the process is denoted as followed:

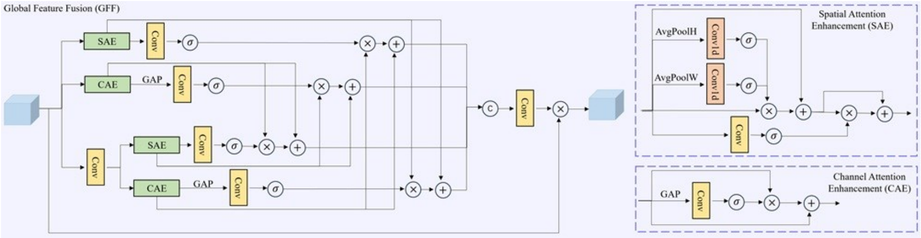

Fig. 4. The structure of the Global Feature Fusion (GFF) module.

$$
\begin{aligned}
F_{A_i} &= \sigma\left(Conv\left(F_{SCB_i}\right)\right) \odot F_{GFF_{i-1}} + F_{GFF_{i-1}}, \\
F_{A_i} &= \sigma\left(Conv\left(F_{A_i}\right)\right), \\
F_{SPA_i} &= Conv\left(F_{GFF_{i-1}}\right), \\
F^*_{SPA_i} &= Reshape\left(F_{SPA_i} \odot F_{A_i} + F_{SPA_i}\right) + F^*_{GFF_{i-1}},
\end{aligned} \tag{6}
$$

where σ represents the Sigmoid function, $F_{GFF_{i-1}} \in \mathbb{R}^{C\times H\times W}$ represents the grouped features of $F^*_{GFF_{i-1}} \in \mathbb{R}^{C^*\times H\times W}$ where $C = C^*/N_{group}$.

Global Feature Fusion Module. In order to exploit the global statistics, we combine multi-scale channel-wise information and spatial-wise information and perform interactions between different branches. The channel attention mechanism [22,25] usually uses the global average pooling to generate channel-wise statistics which are input to fully connected layers or convolutional layers to model channel-wise dependencies, then the features are scaled by the attention weights. This mechanism could also be combined with spatial attention for better performance [23,26]. Therefore, we combine this two mechanisms to capture global information. As shown in Fig. 4, we design a Spatial Attention Enhancement (SAE) module and a Channel Attention Enhancement (CAE). In the SAE module, we apply average pooling and 1-dimensional convolution along the width and height directions to extract the cross-spatial information of the two directions which is used to aggregate input features, then the output is combined with the $1 \times H \times W$ spatial-wise statistics extracted along the channel dimension, which is denoted as followed:

$$
\begin{aligned}
F_{SAE_{H_i}} &= \sigma\left(Conv1d\left(AvgPoolH\left(z_{SAE_i}\right)\right)\right), \\
F_{SAE_{W_i}} &= \sigma\left(Conv1d\left(AvgPoolW\left(z_{SAE_i}\right)\right)\right), \\
F_{SAE_{i_a}} &= z_{SAE_i} \odot F_{SAE_{H_i}} \odot F_{SAE_{W_i}} + z_{SAE_i}, \\
F_{SAE_{i_b}} &= \sigma\left(Conv\left(z_{SAE_i}\right)\right), \\
F_{SAE_i} &= F_{SAE_{i_a}} \odot F_{SAE_{i_b}} + F_{SAE_{i_a}},
\end{aligned} \tag{7}
$$

where $z_{SAE_i} \in \mathbb{R}^{C\times H\times W}$ represents the input features of the SAE module. In the CAE module, we apply global average pooling followed by the 1×1 convolution operation

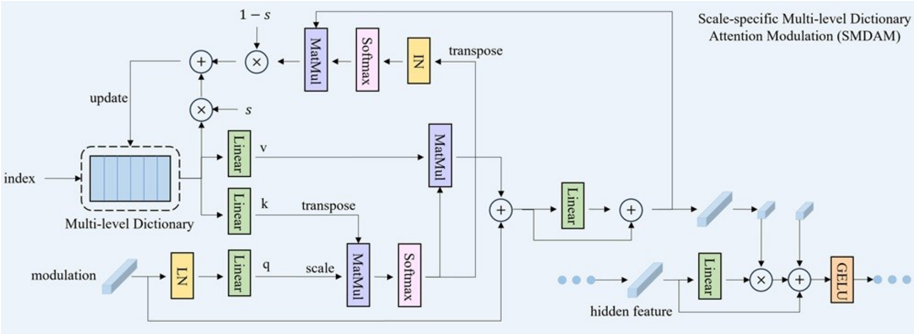

Fig. 5. The structure of the Scale-specific Multi-level Dictionary Attention Modulation (SMDAM) Module.

and the Sigmoid function to obtain channel attention weights which are used to combine input features, and the process is denoted as followed:

$$\begin{aligned} F_{CAE_i} &= \sigma\left(Conv\left(GAP\left(z_{CAE_i}\right)\right)\right), \\ F_{CAE_i} &= F_{CAE_i} \odot z_{CAE_i} + z_{CAE_i}, \end{aligned} \tag{8}$$

where $z_{CAE_i} \in \mathbb{R}^{C\times H\times W}$ represents the input features of the CAE module. In the first two branches, we use the SAE and CAE module to obtain $F_{SAE_{i_1}}$ and $F_{CAE_{i_1}}$, while in the other two branches, we use the SAE and CAE module after the convolution operation to capture larger-scale features $F_{SAE_{i_2}}$ and $F_{CAE_{i_2}}$. Then we perform interactions between multi-scale features, which is denoted as followed:

$$\begin{aligned} F_{SAE_{i_1}} &= F_{SAE_{i_1}} \odot \sigma\left(Conv\left(GAP\left(F_{CAE_{i_2}}\right)\right)\right) + F_{SAE_{i_1}}, \\ F_{CAE_{i_1}} &= F_{CAE_{i_1}} \odot \sigma\left(Conv\left(F_{SAE_{i_2}}\right)\right) + F_{CAE_{i_1}}, \\ F_{SAE_{i_2}} &= F_{SAE_{i_2}} \odot \sigma\left(Conv\left(GAP\left(F_{CAE_{i_1}}\right)\right)\right) + F_{SAE_{i_2}}, \\ F_{CAE_{i_2}} &= F_{CAE_{i_2}} \odot \sigma\left(Conv\left(F_{SAE_{i_1}}\right)\right) + F_{CAE_{i_2}}, \end{aligned} \tag{9}$$

We concatenate the outputs of these branches followed by the convolution operation, and the obtained global information is then incorporated into the input features, which is denoted as followed:

$$\begin{aligned} F_{GFF_i} &= Conv\left[F_{SAE_{i_1}}; F_{CAE_{i_1}}; F_{SAE_{i_2}}; F_{CAE_{i_2}}\right] \odot F_{GFF_{i-1}}, \\ F_{GFF_i} &= Reshape\left(F_{GFF_i}\right) + F^*_{GFF_{i-1}}, \end{aligned} \tag{10}$$

where $[\cdot;\cdot]$ represents the concatenating operation, $F_{GFF_{i-1}} \in \mathbb{R}^{C\times H\times W}$ is the grouped features of $F^*_{GFF_{i-1}} \in \mathbb{R}^{C^*\times H\times W}$.

3.3 Scale-specific Multi-level Dictionary Attention Modulation Module

The token dictionary [32] could be used to integrate external priors with cross-attention mechanism. In this work, we use the token dictionary-based cross-attention to generate scale-specific modulation for the decoder MLP. To be specific, we propose a Scale-specific Multi-level Dictionary Attention Modulation (SMDAM) Module where we learn a modulation dictionary for each scale factor in $\{1.1, 1.2, \ldots, 4.0\}$ to utilize scale-specific priors. As shown in Fig. 5, the modulation M_l of each layer l generated by the latent MLP is input to the linear layer after the Layer Normalization operation to generate query Q_l, and then we use the scale factor as the index $Index_l$ to fetch the corresponding dictionary $D[Index_l]$ and apply the linear projection to generate key K_l and value V_l. Next, we calculate the attention map with Q_l and K_l, and use the attention map to aggregate value V_l, which is then followed by a linear layer with weights $W_{M,l}$. The process of the dictionary attention is denoted as followed:

$$\begin{aligned} Q_l &= LN(M_l)W_{Q,l}, \\ K_l &= D[Index_l]W_{K,l}, \\ V_l &= D[Index_l]W_{V,l}, \\ M_l &= Softmax\left(\frac{1}{\sqrt{d}} Q_l K_l^T\right) V_l + M_l, \\ M_l &= M_l W_{M,l} + M_l, \end{aligned} \tag{11}$$

where $M_l \in \mathbb{R}^{HW \times C_{mod}}$ is the modulation of the i-th layer, LN represents the Layer Normalization operation, $W_{Q,l}, W_{K,l}, W_{V,l} \in \mathbb{R}^{C_{mod} \times C_{mod}}$ represent the weights of the query, key and value of the i-th layer, $Q_l \in \mathbb{R}^{HW \times C_{mod}}$ represents the query of the i-th layer, $D \in \mathbb{R}^{N_D \times N_M \times C_{mod}}$ is a multi-level dictionary with number of levels N_D and number of modulation codes N_M which is accesses by the index $Index_l$ of the i-th layer, $D[Index_l] \in \mathbb{R}^{N_M \times C_{mod}}$ represents the gathered dictionary values, $K_l, V_l \in \mathbb{R}^{N_M \times C_{mod}}$ represent the key and value of the i-th layer, $d = C_{mod}$ is used to scale the dot products, $W_{M,l}$ represents the weights of the modulation of the i-th layer. Then the generated modulation M_l is split into scale modulation $M_{\gamma,l}$ and shift modulation $M_{\beta,l}$ for the render MLP which consists of FiLM layers [8,17], and the hidden features h^l is modulated by $M_{\gamma,l}$ and $M_{\beta,l}$. The process is denoted as followed:

$$\begin{aligned} h^l &= W^l h^{l-1} + b^l, \\ h^l &= GELU\left(h^l \odot M_{\gamma,l} + M_{\beta,l} + h^{l-1}\right), \end{aligned} \tag{12}$$

where h^l denotes the l-th hidden features, $M_{\gamma,l}$ is the scale modulation of the l-th hidden layer, $M_{\beta,l}$ is the shift modulation of the l-th hidden layer, W^l and b^l are the weights and bias of the l-th hidden layer respectively. The final pixel value of the decoder is the local ensembled outputs within the local 2×2 region added by the bilinear upsampled version of the input sample. Inspired by the adaptive refining strategy of ATD [32], we then use the attention map to adaptively update the dictionary. We perform Instance Normalization and the Softmax operation on the transposed attention map A_l^T, then we conduct matrix multiplication with the output of the attention module

Table 1. Quantitative results (PSNR/SSIM) on UCMerced dataset. Bold indicates the best performance.

Method \ Scale	×1.5	×2	×2.5	×3	×3.5	×4
EDSR-baseline [15]	-	36.01/0.9411	-	31.68/**0.8692**	-	29.25/0.7968
LIIF [3]	39.66/0.9712	35.84/0.9395	33.23/0.9003	31.49/0.8637	30.21/0.8286	29.08/0.7890
LTE [13]	39.92/0.9721	35.99/0.9407	33.43/0.9046	31.65/0.8680	30.34/0.8326	29.20/0.7935
CiaoSR [1]	39.88/0.9718	35.98/0.9400	33.40/0.9031	31.46/0.8622	30.32/0.8299	29.21/0.7922
LMF [8]	39.60/0.9709	35.77/0.9386	33.18/0.8996	31.44/0.8632	30.16/0.8264	29.05/0.7867
MFESDA (ours)	**40.00/0.9723**	**36.09/0.9415**	**33.51/0.9048**	**31.75**/0.8680	**30.47/0.8357**	**29.33/0.7976**

M_l to get the updated dictionary codes. We then use a learnable parameter $s = \sigma(s^*)$ to weight the original codes and the updated codes, which is denoted as followed:

$$D[Index_l] = sD[Index_l] + (1 - s)Softmax\left(IN(A_l^T)\right)M_l, \tag{13}$$

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. In this experiment, we adopt UCMerced [31] dataset and RSSCN7 [27] dataset for the arbitrary-scale SR task. The UCMerced dataset consists of 2100 images of 21 classes in 256×256 resolution, and the categories contain agricultural, airplane, baseballdiamond, beach, buildings, chaparral, denseresidential, forest, freeway, golfcourse, harbor, intersection, mediumresidential, mobilehomepark, overpass, parkinglot, river, runway, sparseresidential, storagetanks and tenniscourt. We use 1260 images as training set, 420 images as validation set and 420 images as testing set. The RSSCN7 dataset consists of 2800 images of 7 classes in 400×400 resolution, and the categories contain grass, field, industry, riverLake, forest, resident and parking. We use 1680 images as training set, 560 images as validation set and 560 images as testing set. The Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [33] are chosen as evaluation metrics.

Implementation Details. In this work, the MLP used for the SMDAM module is a 7-layer MLP with 128 dimension, the number of levels of the scale-specific dictionary $N_D = 30$, and the size of each level $N_M = 64$. The EDSR [15] network without the upsampling module is selected as the alternative encoder for all the arbitrary-scale SR methods. The LR patches of size 48×48 are bicubic downsampled from the HR patches with scale factors uniformly sampled from a set of $\{1.1, 1.2, \ldots, 4.0\}$, and LR patches are randomly flipped for data augmentation. The batch size is 16 and the model

Table 2. Quantitative results (PSNR/SSIM) on RSSCN7 dataset. Bold indicates the best performance.

Method \ Scale	×1.5	×2	×2.5	×3	×3.5	×4
EDSR-baseline [15]	-	33.83/0.9067	-	30.71/0.8152	-	29.14/0.7445
LIIF [3]	37.08/0.9542	33.74/0.9042	31.89/0.8563	30.65/0.8134	29.78/0.7750	29.09/0.7417
LTE [13]	37.22/0.9552	33.82/0.9056	31.95/0.8579	30.71/0.8152	29.83/0.7770	29.14/0.7436
CiaoSR [1]	37.13/0.9545	33.78/0.9048	31.94/0.8574	30.61/0.8116	29.81/0.7761	29.12/0.7429
LMF [8]	37.06/0.9540	33.73/0.9042	31.87/0.8562	30.63/0.8135	29.77/0.7745	29.08/0.7411
MFESDA (ours)	**37.53/0.9580**	**33.88/0.9071**	**31.99/0.8588**	**30.75/0.8162**	**29.87/0.7786**	**29.18/0.7455**

is trained for 1000 epochs with Adam optimizer. The learning rate is initialized to 1e-4 and decays by factor 0.5 every 200 epochs.

4.2 Evaluation

Quantitative Results. The results of the quantitative performance on the UCMerced dataset are shown in Table 1. We use the fixed-scale method EDSR [15] and arbitrary-scale methods LIIF [3], LTE [13], CiaoSR [1] and LMF [8] as comparison methods, among which LMF uses the LM-LIIF structure. We evaluate the performance on both integer scale factors $\times 2, \times 3, \times 4$ and non-integer scale factors $\times 1.5, \times 2.5, \times 3.5$. As shown in Table 1, the proposed method has higher PSNR and SSIM than other methods except scale factor $\times 3$ in both integer scale factors and non-integer scale factors. Besides, we also evaluate the quantitative performance on the RSSCN7 dataset in Table 2. It could be seen that the proposed method has the highest performance on PSNR and SSIM.

Qualitative Results. The results of the qualitative comparison on the UCMerced dataset are shown in Fig. 6. We display the visual results of the UCMerced dataset with scale factor $\times 3$. It could be observed that the proposed model could reconstruct images with clearer edges and more detailed texture information. For example, in the first row of Fig. 6, our model could reconstruct a sharper edge of the road. In the second row, the proposed method could generate a clearer edge for the contour of the building. In the third row, the right side of the building reconstructed by the proposed method is clearer. In the fourth row, the car and lane lines generated by the proposed model are more visible than those of other methods. The results of the qualitative comparison on the RSSCN7 dataset are shown in Fig. 7. As shown in the first row of the figure, the bottom edge of the land restored by the proposed method is clearer. In the second row, the proposed method could generate clearer contour of the land. In the third row, the field generated by the proposed method has clearer edges. In the fourth row, the pavement marking and cars in the parking lot generated by the proposed model are also more visible.

Fig. 6. Qualitative results on UCMerced dataset of scale factor × 3. The small red box on the image denotes the highlighted region, and the large red box at the bottom denotes the enlarged region. The first group of images belong to ``airplane" class, the second group of images belong to ``buildings" class, the third group of images belong to ``denseresidential" class, and the fourth group of images belong to ``freeway" class.

Table 3. Results of the ablation study (PSNR/SSIM) of the proposed modules on UCMerced dataset.

SMDAM	MFE	×1.5	×2	×2.5	×3	×3.5	×4
		39.93/0.9721	36.02/0.9406	33.41/0.9030	31.65/0.8672	30.37/0.8331	29.24/0.7950
√		39.95/0.9721	36.03/0.9408	33.47/**0.9049**	31.71/**0.8683**	30.42/0.8341	29.28/0.7954
√	√	**40.00/0.9723**	**36.09/0.9415**	**33.51**/0.9048	**31.75**/0.8680	**30.47/0.8357**	**29.33/0.7976**

Ablation Studies. In this section, we conduct ablation experiments to validate the effectiveness of the proposed modules. We replace the MFE module with a 2-layer MLP with 128 dimension and replace the SMDAM module with a 7-layer render MLP [8] with 128 dimension in the first model. In the second model, we add the SMDAM module to the network. The last row is the proposed model with both SMDAM and MFE modules. As shown in Table 3, the SMDAM module could improve the PSNR and SSIM especially in larger scales which indicates that it could assist with the image reconstruction using the scale-specific dictionary attention, and the MFE

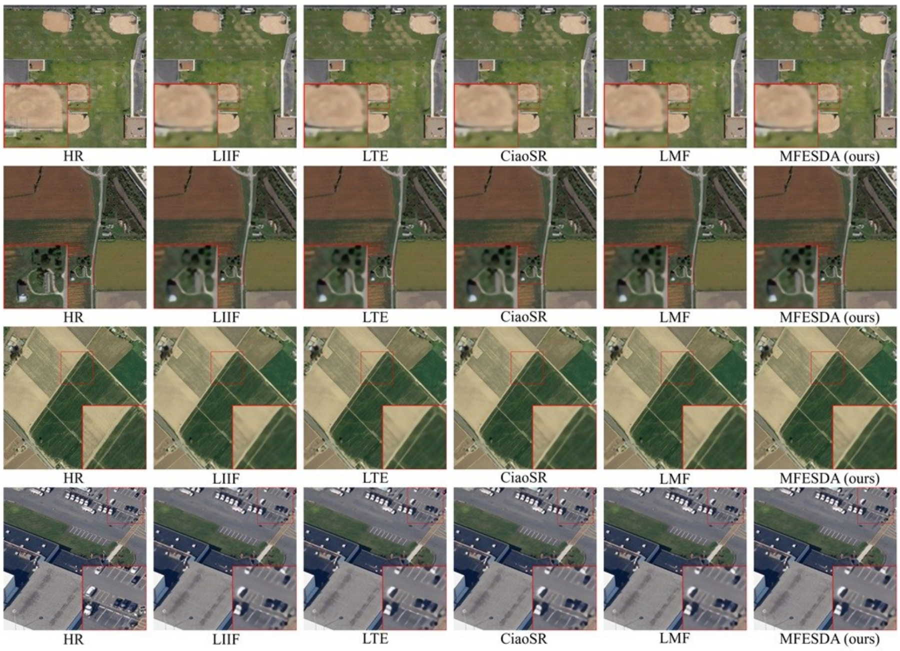

Fig. 7. Qualitative results on RSSCN7 dataset of scale factor × 4. The small red box on the image denotes the highlighted region, and the large red box at the bottom denotes the enlarged region. The first group of images belong to ``Grass" class, the second group of images belong to ``Field" class, the third group of images belong to ``Field" class, and the fourth group of images belong to ``Industry" class.

Table 4. Results of the comparison of the evaluation metrics, the number of parameters and the inference time on the UCMerced dataset.

Method	PSNR↑	SSIM↑	LPIPS↓	Params(M)	Time(ms)
LIIF [3]	35.84	0.9395	0.0957	1.6	33.67
LTE [13]	35.99	0.9407	0.0969	1.7	32.24
CiaoSR [1]	35.98	0.9400	0.0978	2.6	132.91
LMF [8]	35.77	0.9386	0.0977	1.4	13.27
MFESDA (ours)	36.09	0.9415	0.0942	9	170.23

module could further improve the PSNR of the model although the SSIM in × 2.5 and × 3 will slightly decrease.

Model Efficiency and Perceptual Quality. In this section, we analyze the number of parameters and the computational efficiency of models, and we also provide the results of Learned Perceptual Image Patch Similarity (LPIPS) [33] for a more comprehensive comparison. The experiment is conducted on UCMerced dataset with scale factor $\times 2$ on NVIDIA RTX 3090. As shown in Table 4, the LPIPS of the proposed method is lower than other methods which indicates better perceptual quality. However, the disadvantage of the proposed method is that it will lead to larger quantity of parameters and more inference time, which could restrict the application of the proposed method.

Limitations. Although the proposed method performs well in in-distribution scales $\times 1 - \times 4$, the dictionary used in the SMDAM module only includes scale-specific information for $\times 1 - \times 4$ scales but not generalizes to out-of-distribution scales that are unseen in the training process, which is required to be improved in the future.

5 Conclusion

In this work, we propose an arbitrary-scale super-resolution method for remote sensing images with Multi-branch Feature Enhancement and Scale-specific Dictionary Attention (MFESDA). We use a Multi-branch Feature Enhancement (MFE) module which consists of Scale-aware Pixel Attention (SPA) modules and Global Feature Fusion (GFF) modules to capture rich features. Besides, we propose a Scale-specific Multi-level Dictionary Attention Modulation (SMDAM) module to improve the performance of decoding process. The experiments have demonstrated that the proposed model has better performance and visual quality than other methods.

Acknowledgments. This study is supported by the National Natural Science Foundation of China (Nos. 62261060), Yunnan Fundamental Research Projects (Nos. 202301AW070007, 202201AU070033, 202201AT070112, 202301AU070210, 202401AT070470, 202005AC160007), Yunnan Province Major Science and Technology Project (No. 202202AD080002), and Yunnan Province Expert Workstations (202305AF150078), and Xingdian Talent Project in Yunnan Province.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, J., Wang, Q., Xian, Y., Li, Y., Ni, B., Pi, Z., Zhang, K., Zhang, Y., Timofte, R., Van Gool, L.: CiaoSR: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1796–1807 (Jun 2023)
2. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 22367–22377 (Jun 2023)
3. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 8624–8634 (Jun 2021)

4. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11057–11066 (Jun 2019)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proc. Eur. Conf. Comput. Vis. (ECCV). pp. 184–199 (Sep 2014)
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. Adv. Neural Inf. Process. Syst. (NIPS). pp. 2672–2680 (Dec 2014)
7. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for superresolution. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1664–1673 (Jun 2018)
8. He, Z., Jin, Z.: Latent modulated function for computational optimal continuous image representation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 26026–26035 (June 2024)
9. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. In: Int. Conf. Learn. Represent. (ICLR) (May 2019)
10. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1637–1645 (Jun 2016)
11. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5835–5843 (Jul 2017)
12. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image superresolution using a generative adversarial network. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 105–114 (Jul 2017)
13. Lee, J., Jin, K.H.: Local texture estimator for implicit representation function. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1928–1937 (Jun 2022)
14. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW). pp. 1833–1844 (Oct 2021)
15. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). pp. 1132–1140 (Jul 2017)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 9992–10002 (Oct 2021)
17. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proc. AAAI Conf. Artif. Intell. pp. 3942–3951 (Feb 2018)
18. Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1874–1883 (Jun 2016)
19. Guo, Meng-Hao and Xu, Tian-Xing and Liu, Jiang-Jiang and Liu, Zheng-Ning and Jiang, Peng-Tao and Mu, Tai-Jiang and Zhang, Song-Hai and Martin, Ralph R. and Cheng, Ming-Ming and Hu, Shi-Min: Attention mechanisms in computer vision: A survey. Comput. Vis. Media 8(3), 331–368 (Sep 2022)

20. He, Jingwen and Dong, Chao and Liu, Yihao and Qiao, Yu: Interactive MultiDimension Modulation for Image Restoration. IEEE Trans. Pattern Anal. Mach. Intell. 44(12), 9363–9379 (2022)
21. He, Jingwen and Dong, Chao and Qiao, Yu: Modulating Image Restoration with Continual Levels via Adaptive Feature Modification Layers. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11048–11056 (Jun 2019)
22. Hu, Jie and Shen, Li and Sun, Gang: Squeeze-and-Excitation Networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 7132–7141 (2018)
23. Jongchan Park and Sanghyun Woo and Joon-Young Lee and In So Kweon: BAM: Bottleneck Attention Module (2018)
24. Rahaman, Nasim and Baratin, Aristide and Arpit, Devansh and Draxler, Felix and Lin, Min and Hamprecht, Fred A. and Bengio, Yoshua and Courville, Aaron: On the Spectral Bias of Neural Networks. In: Proc. Int. Conf. Mach. Learn. (ICML). pp. 5301–5310 (Jun 2019)
25. Wang, Qilong and Wu, Banggu and Zhu, Pengfei and Li, Peihua and Zuo, Wang meng and Hu, Qinghua: ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11531–11539 (2020)
26. Woo, Sanghyun and Park, Jongchan and Lee, Joon-Young and Kweon, In So: CBAM: Convolutional Block Attention Module. In: Proc. Eur. Conf. Comput. Vis. (ECCV). pp. 3–19 (Sep 2018)
27. Zou, Qin and Ni, Lihao and Zhang, Tong and Wang, Qian: Deep Learning Based Feature Selection for Remote Sensing Scene Classification. IEEE Geosci. Remote Sens. Lett. 12(11), 2321–2325 (Nov 2015)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proc. Adv. Neural Inf. Process. Syst. (NIPS). pp. 5998–6008 (Dec 2017)
29. Wang, L., Wang, Y., Lin, Z., Yang, J., An, W., Guo, Y.: Learning a single network for scale-arbitrary super-resolution. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 4781–4790 (2021)
30. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: ESRGAN: Enhanced super-resolution generative adversarial networks. In: Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW). pp. 63–79 (Sep 2019)
31. Yang, Y., Newsam, S.D.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS). pp. 270–279 (Nov 2010)
32. Zhang, L., Li, Y., Zhou, X., Zhao, X., Gu, S.: Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2856-2865 (2024)
33. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 586–595 (Jun 2018)
34. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proc. Eur. Conf. Comput. Vis. (ECCV). pp. 294–310 (Sep 2018)
35. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2472–2481 (Jun 2018)
36. Zhao, H., Kong, X., He, J., Qiao, Y., Dong, C.: Efficient image super-resolution using pixel attention. In: Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW). pp. 56–72 (Aug 2020)