Large-factor Super-resolution of Remote Sensing Images with Spectra-guided Generative Adversarial Networks

Yapeng Meng, Wenyuan Li, Sen Lei, Zhengxia Zou, and Zhenwei Shi*, *Member IEEE*

***Abstract*—Large-factor image super-resolution is a challenging task due to the high uncertainty and incompleteness of the missing details to be recovered. In remote sensing images, the sub-pixel spectral mixing and semantic ambiguity of ground objects make this task even more challenging. In this paper, we propose a novel method for large-factor super-resolution of remote sensing images named "Spectra-guided Generative Adversarial Networks (SpecGAN)". In response to the above problems, we explore whether introducing additional hyperspectral images to GAN as conditional input can be the key to solving the problems. Different from previous approaches that mainly focus on improving the feature representation of a single source input, we propose a dual branch network architecture to effectively fuse low-resolution RGB images and corresponding hyperspectral images, which fully exploit the rich hyperspectral information as conditional semantic guidance. Due to the spectral specificity of ground objects, the semantic accuracy of the generated images is guaranteed. To further improve the visual fidelity of the generated output, we also introduce the Latent Code Bank with rich visual priors under a generative adversarial training framework so that high-resolution, detailed, and realistic images can be progressively generated. Extensive experiments show the superiority of our method over the state-of-art image super-resolution methods in terms of both quantitative evaluation metrics and visual quality. Ablation experiments also suggest the necessity of adding spectral information and the effectiveness of our designed fusion module. To our best knowledge, we are the first to achieve up to 32x super-resolution of remote sensing images with high visual fidelity under the premise of accurate ground object semantics. Our code can be publicly available at https://github.com/YapengMeng/SpecGAN.**

***Index Terms*—Super-resolution, remote sensing image, hyperspectral image, deep convolutional neural networks, generative adversarial networks**

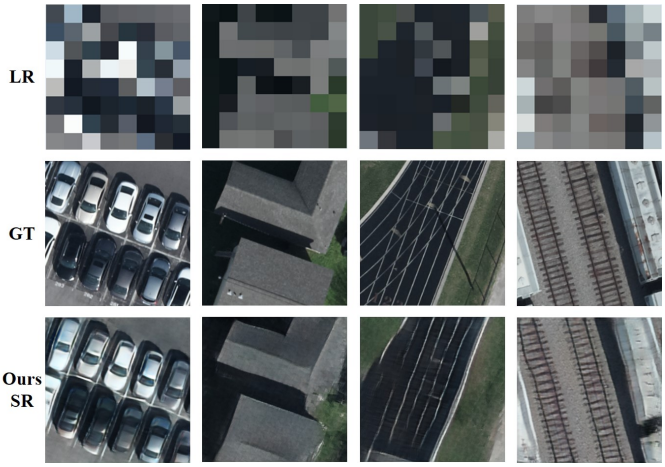

Fig. 1. We propose a novel method for large-factor super-resolution of remote sensing images, which achieves up to 32x super-resolution with high visual fidelity under the premise of accurate ground object semantics. The 1st - 3rd rows show the input low-resolution images, ground truth, and our super-resolution output, respectively.

I. Introduction

Remote sensing technology expands the way humans understand the earth and improves the timeliness and accuracy of earth observation. Remote sensing images play an important role in many application fields, such as land classification [1, 2], agricultural monitoring [3, 4], urban planning [5, 6], disaster disposal [7, 8], and etc. However, due to the limitations of imaging conditions and costs, the resolution of remote sensing images varies greatly, ranging from centimeters to hundreds of meters per pixel. Low-resolution remote sensing images have the advantage of large-scale ground observation, however, the missing details of ground objects bring difficulties to downstream remote sensing image applications.

In order to improve the resolution of remote sensing images without increasing the imaging hardware cost, image super-resolution (SR) technology has brought great attention to the remote sensing field. Early super-resolution methods are mainly based on neighborhood embedding [9], sparse representation learning [10, 11], local linear regression [12–14] and wavelet transform [15]. With the development of deep learning in recent years, the nonlinear mapping between low-resolution images and high-resolution images can be directly learned with deep neural networks in an end-to-end fashion [16, 17], which greatly improves the reconstruction accuracy. Particularly, the emergence of the Generative Adversarial Network (GAN) [18] further improves the visual perceptual quality of reconstruction results [19, 20]. In the field of remote sensing image super-resolution, many deep learning and GAN-based approaches

The work was supported by the National Key Research and Development Program of China (Titled "Brain-inspired General Vision Models and Applications"), the National Natural Science Foundation of China under Grant 62125102, and the Fundamental Research Funds for the Central Universities. (Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))

Yapeng Meng, Wenyuan Li, and Zhenwei Shi are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and with State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory.

Sen Lei is with AVIC Chengdu Aircraft Industrial (Group) Company Ltd., Chengdu, China.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory.

have been proposed recently. For example, Lei et al. proposed HSENet [21] based on the self-similarity of remote sensing images. Qin et al. [22] designed a gradient-aware loss to preserve important gradient information in remote sensing images.

Despite the recent progress, most of the current research on the super-resolution of remote sensing images only focuses on low zoom-in magnitudes such as $\times 2$, $\times 4$, and there are few studies on the problem of larger factor super-resolution. More importantly, in remote sensing images, it is often necessary to solve the problem of converting unrecognized objects (e.g., cars) into clear ones from a considerable low-resolution input image.

In Fig. 1, we give an example where the first row is the low-resolution remote sensing images that might be encountered in practice, while the second row is the ground truth we expect to get. As we can see, the main difficulties of this problem can be summarized as follows: 1) Subpixel blending and semantic ambiguity. Some small-scale ground objects are completely integrated into a single-pixel grid, and the semantics of ground objects in pixels cannot even be clarified by relying only on the information of RGB bands because the RGB band radiation of different ground objects may be the same. It is even more difficult to reconstruct the ground objects mixed with each other only with the help of RGB information. 2) The details of the ground objects are seriously missing, mainly reflected in the texture and outline structure. Because the downsampling factor is very high, the contours of ground objects in low-resolution images are no longer continuous, and texture details are completely lost. It is necessary to restore image details with the help of prior information.

In response to the above problems, we propose a novel method named "Spectra-guided Generative Adversarial Networks (SpecGAN)" for large-factor remote sensing image super-resolution, and explore whether introducing additional hyperspectral images (also has low spatial resolution) as conditional input can be the key to solving the problems. Different from previous approaches that mainly focus on improving the feature representation of a single source input, we construct a dual-branch spectral information fusion and extraction module to achieve the fusion of RGB images and hyperspectral images, thereby ensuring that the targets in the super-resolution results have accurate semantic information. We show that with the proposed method, low-resolution, unrecognized ground objects can be accurately converted into clear ones with vivid visual appearance and accurate semantics. Inspired by the GLEAN [20] method, we introduce the Latent Code Bank into the network to add rich details to SR results, which helps to improve image quality and avoid checkerboard effects. Experimental results on the dataset provided in the IEEE data fusion contest 2019 [23] show that our method can effectively achieve large-factor super-resolution. According to perceptual evaluation metrics LPIPS [24], our method outperforms state-of-art remote sensing image super-resolution method HSENet by 58.3% and improves the performance by 33.6% compared with the GLEAN dedicated to large-factor super-resolution. Our network can effectively reconstruct targets on the premise of semantic accuracy and has high visual quality texture and edge details.

The main contributions of this paper include:

- We propose a large-factor super-resolution network suitable for remote sensing images, which achieves 32x super-resolution with high visual fidelity under the premise of accurate ground object semantics.
- We explore a novel idea for remote sensing image super-resolution by using hyperspectral images as conditional input. With the help of the spectral specificity of different ground objects, using the high spectral resolution to make up for the lack of spatial resolution. The introduction of hyperspectral information ensures the semantic accuracy of the super-resolution results.
- We propose a dual-branch network that effectively fuses hyperspectral information and RGB information, which can improve the possible alignment bias between RGB images and hyperspectral images, and provide effective high-level semantic information and necessary low-level features for the Latent Code Bank and decoder of the network.

II. Related Works

A. Image super-resolution based on deep learning

Deep learning has made great advantages in nearly all branches of computer vision. For image super-resolution, Dong et al. [16, 17] proposed SRCNN, which is the first CNN network designed for super-resolution tasks, established end-to-end mapping between LR images and SR results.

The subsequent improvements mainly focus on improving network structure to increase network nonlinearity, reuse low-level feature maps, and increase receptive fields. VDSR [25] and EDSR [26] adapt deeper networks to reuse low-level features by introducing residual connections. Kim et al. [27] use a recursive module to construct the DRCN, increase the network receptive field, and reuse the network parameters, which achieves a better reconstruction effect with fewer parameters and reduces the difficulty of model training. Inspired by DenseNet [28], Zhang et al. [19] propose the residual dense block and build a super-resolution network based on it. This network structure has also been used by the most advanced super-resolution models [20, 29] in recent years.

After the emergence of the Generative Adversarial Network (GAN) [18], researchers use the adversarial loss to make generated results have a distribution similar to ground truth, so as to generate more realistic image texture, improve the excessive smoothing problem [30–33] caused by only optimizing pixel-level loss such as MSE, and improve the visual quality of output. Ledig et al. [31] first apply GAN in the super-resolution field and proposed SRGAN. Menon et al. [34] propose PULSE, which iteratively optimizes the latent space of StyleGAN [35] based on GAN inversion to achieve face super-resolution.

The diffusion probabilistic model [36, 37] has also been successfully adapted to the image super-resolution task. Li et al. [38] propose Srdiff, aimed at tackling the over-smoothing, mode collapse, and large footprint problems in previous

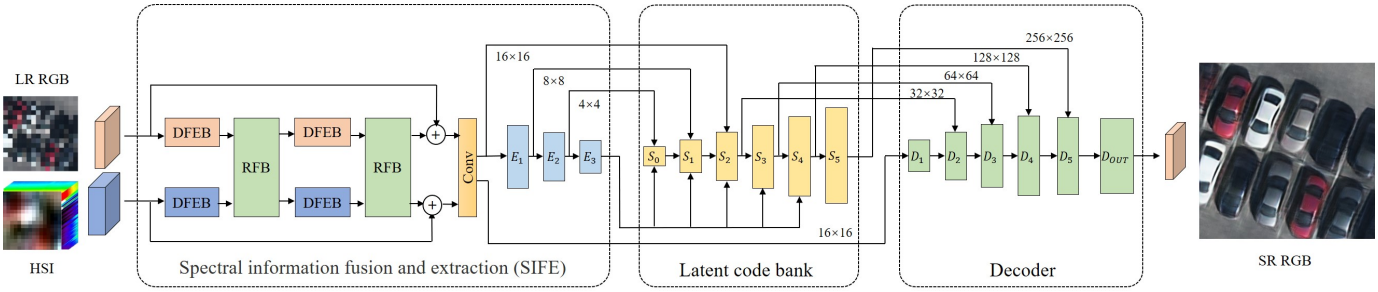

Fig. 2. An overview of the proposed SpecGAN. The proposed network takes in both the low-resolution RGB image and hyperspectral image and produces up to 32x super-resolution output. The network consists of a dual-branch Spectral Information Fusion and Extraction (SIFE) module, a Latent Code Bank for collecting priors, and an up-sampling decoder.

PSNR-oriented, GAN-driven models. Saharia et al. [39] propose SR3 and get photo-realistic outputs outperforming GAN-based methods. Rombach et al. [40] propose the Latent Diffusion Model and evaluate it on super-resolution tasks. They apply the diffusion model in the latent space rather than pixel space, which reduces the computational complexity and improves the image quality.

In recent years, more research has challenged the natural image large-factor super-resolution task. RFB-ESRGAN [41] is based on ESRGAN [42] and uses multi-scale receptive field blocks for $\times 16$ super-resolution. Dong et al. propose the DSSR network, which uses the dense sampling mechanism and introduces a wide feature block (WAB) to enhance the representation ability of the neural network. The RDN-LIIF network proposed by Chen et al. uses the implicit neural representation to generate arbitrary resolution images, achieving $\times 30$ or even higher image super-resolution. GLEAN [20] proposed by Chan et al. has a good effect on natural scenes super-resolution ($\times 16$ to $\times 64$). GLEAN use pre-trained StyleGAN as Latent Code Bank which contains a large amount of texture prior to bring vivid and realistic reconstruction effect of missing texture details.

In large-factor super-resolution tasks, an important factor affecting the quality of the generated image is the checkerboard effect. Odena et al. [43] first study the chessboard effect caused by deconvolution in the image generation process based on the neural network, and point out that the nearest neighbor or bilinear up-sampling method can be used to replace deconvolution to avoid checkerboard effects. In large-factor super-resolution tasks, due to a large number of up-sampling processes, it is very easy to cause the chessboard effect. Sugawara et al. [44] also study ways to avoid the checkerboard effect in super-resolution tasks.

B. Super-resolution for remote sensing image

Early studies in remote sensing image super-resolution are mostly based on sparse representation [10, 11] and discrete wavelet transform [15].

The emergence of deep learning has greatly improved the quality of remote sensing image SR. Lei et al. [45] first use the deep learning method in the remote sensing image super-resolution task and proposed LGCNet. New neural network architecture has also been applied to remote sensing image SR tasks. Haut et al. [46] and Pan [47] apply dense connection, residual connection, jump connection to remote sensing image SR. Yang et al. [48] use transformer [49] network and attention mechanism to migrate texture from the high-resolution reference image to the original image. Lei et al. [50] propose a transformer-based enhancement network to exploit image features at different levels.

More recently, many researchers develop SR methods with the help of characteristics in remote sensing images. HSENet [21] explores the multi-scale self-similarity commonly contained in remote sensing images and make full use of the spatial attention mechanism. Some researchers [51, 52] transform remote sensing images into wavelet domain and learn SR methods in the transform domain. Qin et al. [22] design a gradient-aware-loss to retain important gradient information in remote sensing images. People also find the special effects of using GAN network on remote sensing image SR. Lei [53] find discrimination ambiguity problem in process of GAN-based remote sensing image SR and propose coupled-discriminated GAN.

Despite the above progress, in the remote sensing field, few studies focus on large-factor super-resolution tasks (e.g. $\times 16$ or more). Such tasks require not only restoring object details but also reconstructing small-scale objects that are heavily mixed in a single-pixel grid. The idea of using hyperspectral information to improve the super-resolution of RGB images is also rarely studied.

III. Methodology

The overall network structure of the proposed SpecGAN is shown in Fig. 2. The network is designed based on the idea of GAN-based image super-resolution, and the low-resolution hyperspectral images are added as auxiliary information in the generation stage. To make full use of the semantic information in the hyperspectral image and provide more refined features for the decoder, we design a fusion network for RGB images and hyperspectral images. Considering that there may be alignment deviation between images from different sources, we design a strategy to avoid deviation confusion, that is, further registration between two branches based on the attention mechanism, which effectively improves the performance of large-factor SR. To provide more sufficient priors for the network to reconstruct detailed texture and avoid the

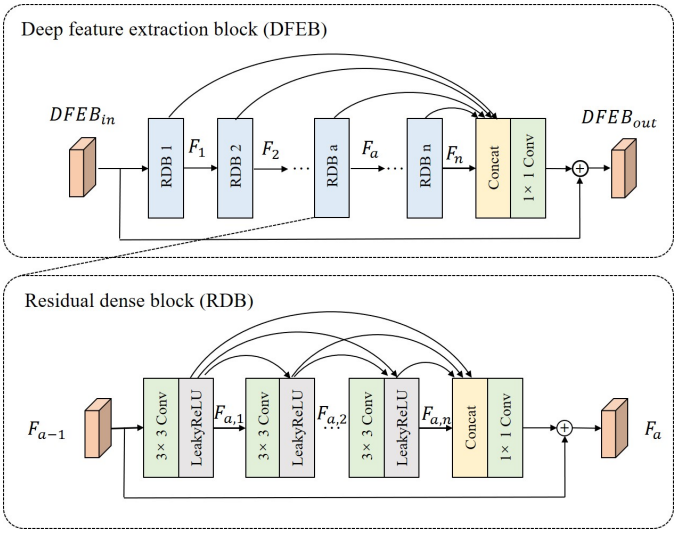

Fig. 3. Detailed structure of the proposed Deep Feature Extraction Block (DFEB). DFEB is composed of several Residual Dense Blocks (RDB) in series.

checkerboard effect in the up-sampling process, we add a pre-trained GAN to the network as the Latent Code Bank. In order to improve the perceptual quality of generated image, we use the discriminator structure in StyleGAN as adversarial loss and introduce the perceptual loss.

A. Spectral information fusion and extraction module (SIFE)

In order to effectively integrate the semantic information from hyperspectral images into the super-resolution process, we design a dual-branch spectral information fusion and extraction module. We design independent input branches for the hyperspectral image and RGB image respectively. The module consists of three parts: Deep Feature Extraction Block (DFEB) in each branch, Registration and Fusion Block (RFB) between two branches, and down-sampling to obtain high-level semantic information.

Deep Feature Extraction

We design Deep Feature Extraction Block (DFEB) in each branch. DFEB is composed of several Residual Dense Blocks (RDB) in series. The overall structure is shown in Fig. 3. The module uses the dense connection, local residual connection, and global residual connection to ensure that each branch can take into account low-level details and high-level semantic information of the image.

The operation of each convolution layer in RDB is described as follows:

$$F_{a,b} = \sigma(W_{a,b}[F_{a-1}; F_{a,1}; F_{a,2}; ...; F_{a,b-1}]), \tag{1}$$

where $F_{a,b}$ is the $b-th$ feature map in the $a-th$ RDB, $W_{a,b}$ is the kernel weight of this convolution layer, σ is the LeakyReLU [54] activation function, $[;]$ means that feature maps are stacked in feature dimension. In this design, the input information of each convolution layer includes the output information of the previous RDB and the operation results of each previous convolution layer in the current RDB.

The output result of each RDB is linearly superimposed by each convolution result in the current block. Previous RDB outputs are added through local residual connection:

$$F_a = F_{a-1} + W_a([F_{a,1}; F_{a,2}; ...; F_{a,n}]), \tag{2}$$

where F_a is the $a-th$ RDB output, W_a is a convolution layer with 1×1 kernel, which is equivalent to linear combine all feature maps in current block.

The output result of the whole DFEB combines all the RDB output and further adds a global residual connection between the input and output:

$$DFEB_{out} = DFEB_{in} + W_{out}([F_1; F_2; ...; F_n]), \tag{3}$$

where W_{out} is combined by a convolution layer with 1×1 kernel and a convolution layer with 3×3 kernel to linearly stack and transform every RDB output feature map.

Registration and Fusion

The introduction of the dual branch network clearly separates the low-resolution RGB image from the hyperspectral image and helps to adjust incomplete correspondence between two images on pixels during data acquisition. In order to better realize multi-source data fusion and enhance the semantic guidance of hyperspectral images, we design Registration and Fusion Block (RFB) between two branches. The structure is shown in Fig. 4. The whole process is divided into Branch stacking, linear transformation, and registration based on similarity.

1. Branch stacking can be expressed as:

$$F_{fused} = H_F([F_A; F_B; M_{A/B}]), \tag{4}$$

where M_A represents a mask layer, which is used to identify whether the fusion information is used to transform channel A or channel B. H_F is composed of several convolution layers, which transform the features after stacking.

2. Linear transform of each branch is expressed as: (Using branch A as an example)

$$F_A^{1'} = (F_A \odot X + Y), \tag{5}$$

where $\odot$ represents pixel by pixel multiplication, X and Y are obtained by F_{fused} through several convolution layers, as shown in the upper right of Fig. 4.

3. Registration based on the similarity between two branches

We first calculate the similarity matrix F_s between any two positions:

$$F_s(i,j) = e^{H_\theta^T(F_{fused}(i))H_\phi(F_{fused}(j))}, \tag{6}$$

where $F_s(i,j)$ is similarity between the $i-th$ position and $j-th$ position.

The similarity matrix is used to register and align the information in each branch:

$$F_A^{2'}(i) = \frac{\sum_j F_s(i,j)H_g(F_{Aj})}{\sum_j F_s(i,j)}, \tag{7}$$

where H_θ, H_ϕ, H_g is composed of convolution layers with 1×1 kernel, as shown in the bottom right of Fig. 4.

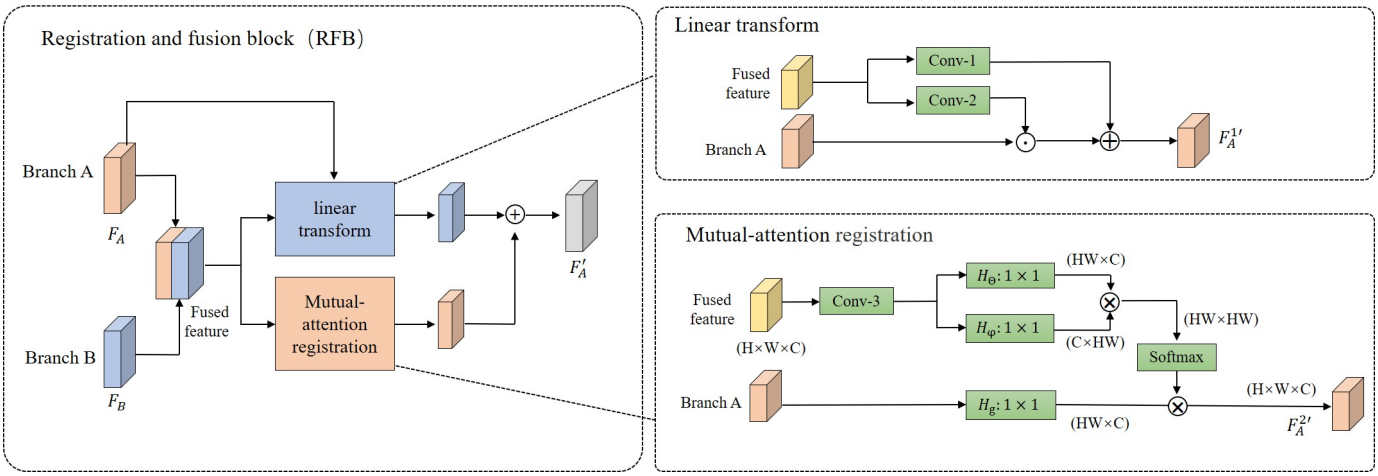

Fig. 4. Detailed structure of the proposed Registration and Fusion Block (RFB). This module is composed of three parts: Branch stacking, linear transform, and registration based on similarity.

4. Total transformation in RFB is defined as:

$$F_A^{'} = F_A + F_A^{1'} + F_A^{2'}. \tag{8}$$

Down-sampling

As shown in Fig. 2, the last step in the spectral information fusion and extraction stage is to down-sample the feature map layer by layer, so that the network can extract high-level semantic information from the input image. The down-sampling is used as constraint information for subsequent small-scale object disambiguation and reconstruction.

$$e_i = E_i(e_{i-1}), \quad i \in \{1, ..., N\}, \tag{9}$$

where e_i is the $i-th$ feature map, E_i is the $i-th$ convolution layer. Each group of convolution layers is composed of a stride-2 convolution layer and a stride-1 convolution layer. e_i will then be used to constrain progressive texture generation in the Latent Code Bank.

At the end of the down-sampling stage, a set of fully connected layers are used to generate Latent Vector c_i, providing high-level semantic information for the Latent Code Bank by adjusting the mean and variance of feature maps in the Latent Code Bank.

B. Loss functions

Our network aims to generate SR results with accurate object position and high visual quality. In order to achieve this goal, we apply the following loss function:

Pixel Loss: The basic requirement of image super-resolution is content accuracy. Introducing this loss function makes SR results as close to the ground truth as possible at the pixel level. In addition, early studies [16, 42] show that introducing the mean square error between generated image and high-resolution ground truth in GAN-based super-resolution task can help to avoid image distortion such as artifacts in the generated image.

$$\mathcal{L}_{pxl}(G) = \mathbb{E}_{x,y\sim p_{data}}\{\|G(x) - y\|_2^2\}. \tag{10}$$

Adversarial Loss: We use the adversarial loss to encourage the network to generate realistic results with texture details. Although only optimizing the network with pixel level loss can get higher peak signal-to-noise ratios(PSNR), relevant studies [30–33] show that this will get an excessively smooth output, and adding adversarial loss [18] is more conducive to generating image texture and edge. The loss function is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv}(G, D) = &\mathbb{E}_{y\sim p_y}\{\log D(y)\} \\ &+ \mathbb{E}_{x\sim p_x}\{\log(1 - D(G(x)))\}.\end{aligned} \tag{11}$$

Perceptual Loss: To make the super-resolution results have a better visual effect, we introduce Perceptual Loss [55], which also helps to speed up the network training:

$$\mathcal{L}_{pcp}(G) = \mathbb{E}_{x,y\sim p_{data}}\{\|\Phi^L_{G(x)} - \Phi^L_y\|_2^2\}, \tag{12}$$

where Φ^L represents the feature map obtained from the $L-th$ layer of a pre-trained deep convolution network. Here we use the feature map of the 21st layer in the pre-trained VGG16 network [56].

Objective Function: Combining all the above losses, the overall objective function we aim to optimize is:

$$\begin{aligned}\min_G \max_D \mathcal{L}(G, D) = &\lambda_{pxl}\mathcal{L}_{pxl}(G) + \lambda_{pcp}\mathcal{L}_{pcp}(G) \\ &+ \lambda_{adv}\mathcal{L}_{adv}(G, D),\end{aligned} \tag{13}$$

where λ_{pxl}, λ_{adv} and λ_{pcp} are predefined positive factors for balancing the weights between different loss terms.

C. Avoiding Chessboard Effect

In large-factor super-resolution tasks, due to a large number of up-sampling operations, it is easy to produce chessboard artifacts [43]. In order to eliminate the checkerboard effect in generated images, we add the Latent Code Bank and use the progressive bilinear interpolation up-sampling method.

the Latent Code Bank We add the Latent Code Bank proposed in GLEAN [20] to learn rich and diverse image priors such as object texture and edge in the dataset. The detailed

TABLE I
THE DETAILED CONFIGURATION OF OUR NETWORK

	Layer	Kernel	stride	Input	Output	Activation function
DFEB	Conv layer x in each RDB ($x \in [1,8]$)	3×3	1	$(128+64(x-1))\times H\times W$	$64\times H\times W$	LeakyReLU
	Output layer in each RDB	1×1	1	$(128+64*8)\times H\times W$	$128\times H\times W$	None
	Output layer in each DFEB	1×1	1	$(128*8)\times H\times W$	$128\times H\times W$	None
		3×3	1	$128\times H\times W$	$128\times H\times W$	None
RFB	H_F in branch stacking	3×3	1	$(128*2+2)\times H\times W$	$128\times H\times W$	LeakyReLU
		3×3	1	$128\times H\times W$	$128\times H\times W$	LeakyReLU
		3×3	1	$128\times H\times W$	$128\times H\times W$	None
	Conv-1 in linear transform	3×3	1	$128\times H\times W$	$128\times H\times W$	None
	Conv-2 in linear transform	3×3	1	$128\times H\times W$	$128\times H\times W$	None
	Conv-3 in registration	3×3	1	$128\times H\times W$	$128\times H\times W$	None
	H_θ in registration	1×1	1	$128\times H\times W$	$64\times H\times W$	None
	H_ϕ in registration	1×1	1	$128\times H\times W$	$64\times H\times W$	None
	H_g in registration	1×1	1	$128\times H\times W$	$64\times H\times W$	None
Decoder	D_1	3×3	1	$C_{H,W}\times H\times W$	$C_{H,W}\times H\times W$	LeakyReLU
		Bilinear Interpolate		$C_{H,W}\times H\times W$	$C_{H,W}\times 2H\times 2W$	None
		3×3	1	$C_{H,W}\times 2H\times 2W$	$C_{H,W}\times 2H\times 2W$	LeakyReLU
		3×3	1	$C_{H,W}\times 2H\times 2W$	$C_{H,W}\times 2H\times 2W$	LeakyReLU
	$D_x(x\in[2,n-1])$	3×3	1	$(2*C_{H,W})\times 2^{x-1}H\times 2^{x-1}W$	$C_{H,W}\times 2^{x-1}H\times 2^{x-1}W$	LeakyReLU
		Bilinear Interpolate		$C_{H,W}\times 2^{x-1}H\times 2^{x-1}W$	$C_{H,W}\times 2^{x}H\times 2^{x}W$	None
		3×3	1	$C_{H,W}\times 2^{x}H\times 2^{x}W$	$C_{H,W}\times 2^{x}H\times 2^{x}W$	LeakyReLU
		3×3	1	$C_{H,W}\times 2^{x}H\times 2^{x}W$	$C_{H,W}\times 2^{x}H\times 2^{x}W$	LeakyReLU
	D_n	3×3	1	$(2*C_{H,W})\times 2^{n-1}H\times 2^{n-1}W$	$64\times 2^{n-1}H\times 2^{n-1}W$	LeakyReLU
		3×3	1	$64\times 2^{n-1}H\times 2^{n-1}W$	$3\times 2^{n-1}H\times 2^{n-1}W$	LeakyReLU

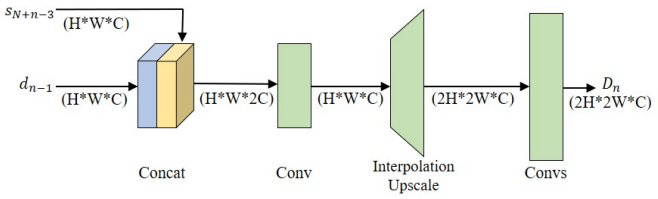

Fig. 5. A single layer-block in the Decoder

structure is also similar to the Latent Code Bank in GLEAN, which borrows the structure of the StyleGAN [35] generator.

$$s_i = \begin{cases} S_i(c_i, e_{N-i-1}) & i=0, \\ S_i(c_i, e_{N-i-1}, s_{i-1}) & i>0, \end{cases} \tag{14}$$

where S_i represents $i-th$ convolution layer in the Latent Code Bank. Each layer first connects e_i generated by SIFE in the feature dimension, so as to better constrain the semantics of the generated image. The Latent Vector c_i is used to control the mean and variance of the feature map in the Latent Code Bank.

Progressive bilinear interpolation up-sampling We design progressive bilinear interpolation up-sampling, which combines feature maps from the Latent Code Bank and the previous layer in the decoder as input to expand the feature map resolution, as shown in Fig. 5. The transformation of each layer is defined as:

$$d_n = \begin{cases} D_{n,2}(U_2(D_{n,1}(d_{n-1}))) & n=1, \\ D_{n,2}(U_2(D_{n,1}([d_{n-1}; s_{N+n-3}]))) & n>1, \end{cases} \tag{15}$$

where $D_{n,i}$ is $i-th$ convolution layer with 3×3 kernel in $n-th$ decoder, U_2 is a $\times2$ up-sampling layer. We use bilinear interpolation to realize feature map up-sampling, avoiding possible chessboard effects in process of deconvolution and pixel-shuffle.

D. Implementation Details

We show the detailed structure of out network in Tab. I. The number of feature channels $C_{H,W}$ in the Decoder is related to the height and width of feature map. $C_{H,W}$ is 512,256,128,64 when the size of feature map is equal of lower than 64, 128, 256, 512 respectively.

In our experiment, we mainly focus on $\times16$, $\times32$ super-resolution. In the training stage, we randomly cut out 256×256 high-resolution image patches from the training set as ground truth. The LR RGB and HSI are 16×16 pixels (for $\times16$ SR) or 8×8 pixels (for $\times32$ SR) down-sampled image patches. We use a random combination of horizontal and vertical flipping as our data augmentation strategy. We set hyper-parameters $\lambda_{pxl}=1.0$, $\lambda_{adv}=0.1$ and $\lambda_{pcp}=0.001$. We use Adam [57] optimizer to train network parameters, and set $\beta_1=0.9, \beta_2=0.99$. The initial learning rate is set to 10^{-4}, mimi-batch [58] size is set to 4. During the training of 1,000,000 iters, the learning rate decreases to 10^{-8} according to the cosine law. The method we proposed is implemented using the Pytorch [59] framework, with the help of code in the mmediting [60] image editing toolbox. The network is trained, verified, and tested on an NVIDIA GeForce GTX 1080Ti graphics card.

IV. Experimental Results and Analyses

In this section, we will give a detailed description of our experimental dataset, evaluation metrics, ablation analysis, and experimental results.

Fig. 6. Visual comparison of different methods. From 1st to 8th column: low resolution input (32x), bilinear interpolation, bicubic interpolation, HSENet [21], GLEAN [20], RDN-liif [29], our result, and ground truth. It can be seen that for lawns, roofs, playgrounds, parking lots, roads, rails, and other types of ground objects, our output results are superior to similar networks in terms of semantic accuracy, edge, texture, and other details. For small-scale targets that are mixed in a single-pixel grid, our super-resolution idea can effectively realize sub-pixel level unmixing, and clearly define the meaning and geographical location of objects. On this basis, our network can effectively generate image texture with the help of rich and diverse priors. For example, the unique texture of the parking lot line, playground runway line, roof, and lawn texture.

A. Experimental Dataset

We use the dataset [23, 61] from the "IEEE data fusion contest 2019" to prove the effectiveness of our proposed method. The dataset contains multi-view and multi-band remote sensing images sampled from the city, including residential areas, green spaces, roads, schools, factories, railways, and other common urban landscapes. It contains 0.05m/pixel RGB images and 1m/pixel hyperspectral images (HSI) with 48 spectra. In the experiment, we use the bilinear interpolation method to down-sample high-resolution RGB images to get low-resolution RGB images and use the bilinear interpolation method to resize the HSI images to the same size as the low-resolution RGB images. The target size is 256×256 pixels after super-resolution, and the input low resolution image is 8×8 pixels or 16×16 pixels. We divide the whole dataset into three parts: training set, valid set, and test set. The best-performing model in the valid set is applied to the test set for image quality evaluation.

B. Comparison with the state of the art methods

On the same data set, we compare our method with the state-of-art large-factor super-resolution methods, including

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS

	Method	PSNR↑	SSIM↑	NIQE↓	LPIPS↓
×16	Bilinear	23.31	0.6233	21.81	0.6457
	Bicubic	22.88	0.6140	16.25	0.6671
	HSENet [21]	24.51	0.6580	13.40	0.4533
	RDN-liif [29]	**24.76**	**0.6661**	16.15	0.4130
	GLEAN [20]	22.80	0.5312	6.64	0.3259
	Ours	23.08	0.5577	**5.54**	**0.2345**
×32	Bilinear	21.01	0.6294	24.34	0.6296
	Bicubic	20.56	0.6211	20.13	0.6500
	HSENet [21]	22.04	0.6444	15.58	0.5804
	RDN-liif [29]	**22.29**	**0.6529**	17.41	0.4994
	GLEAN [20]	20.33	0.4809	**4.90**	0.3825
	Ours	20.80	0.5376	5.56	**0.2863**

GLEAN [20], which propose the Latent Code Bank to make full use of prior information in the dataset, and RDN-liif [29], which use implicit neural representation to infinitely extrapolate the resolution of images. We also compare with the latest remote sensing image super-resolution method HSENet [21], which uses characteristics of remote sensing images and design network based on mixed scale attention. The bilinear and bicubic interpolation are used as the baseline of our comparison methods.

For a fair comparison, for all the comparison methods, our experiments stack the low-resolution RGB image and the corresponding HSI image in the channel dimension as input. The number of input channels for the first convolution layer of the comparison super-resolution networks is adjusted accordingly.

Our visual comparison results are shown in Fig. 6. The quantitative comparison results are shown in Tab. II. We use three groups of evaluation metrics, 1) the traditional image evaluation metrics including PSNR and SSIM [62], 2) the unsupervised image evaluation metric NIQE [63], and 3) the latest image perceptual metric LPIPS [24], which is the closest to human's visual effect. From Tab. II, we can see that our method achieved better scores and better visual effects than the state-of-art methods above.

From Fig. 6, we can see that our network effectively reconstructs small-scale features such as cars, paths, and railway tracks. This is due to our network effectively exploiting the rich and diverse spectral information provided by hyperspectral images. We achieve sub-pixel level unmixing, clarify the semantics and location of these small-scale objects in the image. At the same time, our network also has a good effect on reconstructing internal detailed texture features (e.g., parking lines in parking lots and texture of lawns and roofs), which should be attributed to the prior captured by the network during the training process. Because even in the hyperspectral image, these fine textures (e.g., twigs and roof textures) cannot be accurately located, only by relying on priors can we produce texture realistic images.

It is worth mentioning that from Tab. II, we can see that in the large-factor super-resolution task, the PSNR values of state-of-art methods and our method are low. Especially in the $\times 32$ super-resolution experiment, the PSNR values obtained by GLEAN [20] and our method are even lower than using bilinear interpolation. However, previous studies [30–33] show that the PSNR score is not consistent with the human visual perception effect, and this metric is not objective in image super-resolution. From the output results of the RDN-liif network with the highest PSNR value, it can be found that a higher PSNR value may correspond to an excessively smooth visual effect but does not reflect the unique texture features of various ground objects. Therefore, we also test the latest image perceptual metric LPIPS. In the $\times 32$ SR experiment, our proposed method is significantly ahead of the latest super-resolution network GLEAN 33.6%, and ahead of the latest remote sensing image super-resolution network HSENet 58.3%, which proves that our super-resolution results have a better visual perception score. From the visual effect of the output image, our method is also significantly better than other SR networks.

TABLE III
ABLATION STUDY OF FOUR DIFFERENT COMPONENTS OF OUR METHOD: 1)HSI INPUT, 2) DUAL BRANCH SIFE, 3) THE LATENT CODE BANKS, AND 4) RGB INPUT.

	PSNR↑	SSIM↑	NIQE↓	LPIPS↓
our full implementation	20.80	**0.5376**	5.56	**0.2863**
w/o HSI input	20.90	0.5210	5.93	0.2932
w/o dual branch SIFE	**21.55**	0.5285	**5.55**	0.2898
w/o the Latent Code Bank	21.15	0.5153	6.00	0.2987
w/o the RGB input	17.50	0.4482	5.55	0.3635

C. Ablation Studies

In this part, we carry out some ablation studies, mainly to verify 1) the necessity of introducing hyperspectral information, 2) the effectiveness of our dual branch spectral information fusion and extraction module, 3) the role of the Latent Code Bank, and 4) the necessity of introducing RGB information. Our evaluation value is shown in Tab. III, and visual results are shown in Fig. 7.

To verify the necessity of hyperspectral image (HSI), we only input RGB images into both branches of SIFE, which ensures that the total number of computing units and computing structure of the network remains unchanged. To verify the effectiveness of the dual branch SIFE structure design, we replace the dual branch structure with a single branch network with the same number of DFEB blocks and stack hyperspectral image and RGB image directly in spectral dimension as input. To verify the effectiveness of the Latent Code Bank, we remove the Latent Code Bank from the network but keep the appropriate stacking connection. The necessity of introducing hyperspectral information as a semantic constraint and the network structure we designed can be proved in both the evaluation value and visual effect, which effectively improves the quality of the output image.

According to the comparison between the SR effect of each ablation experiment and the original network, it can be concluded that the accurate interpretation of the meaning of the ground object mainly depends on hyperspectral information, while the dual branch structure and the Latent Code Bank

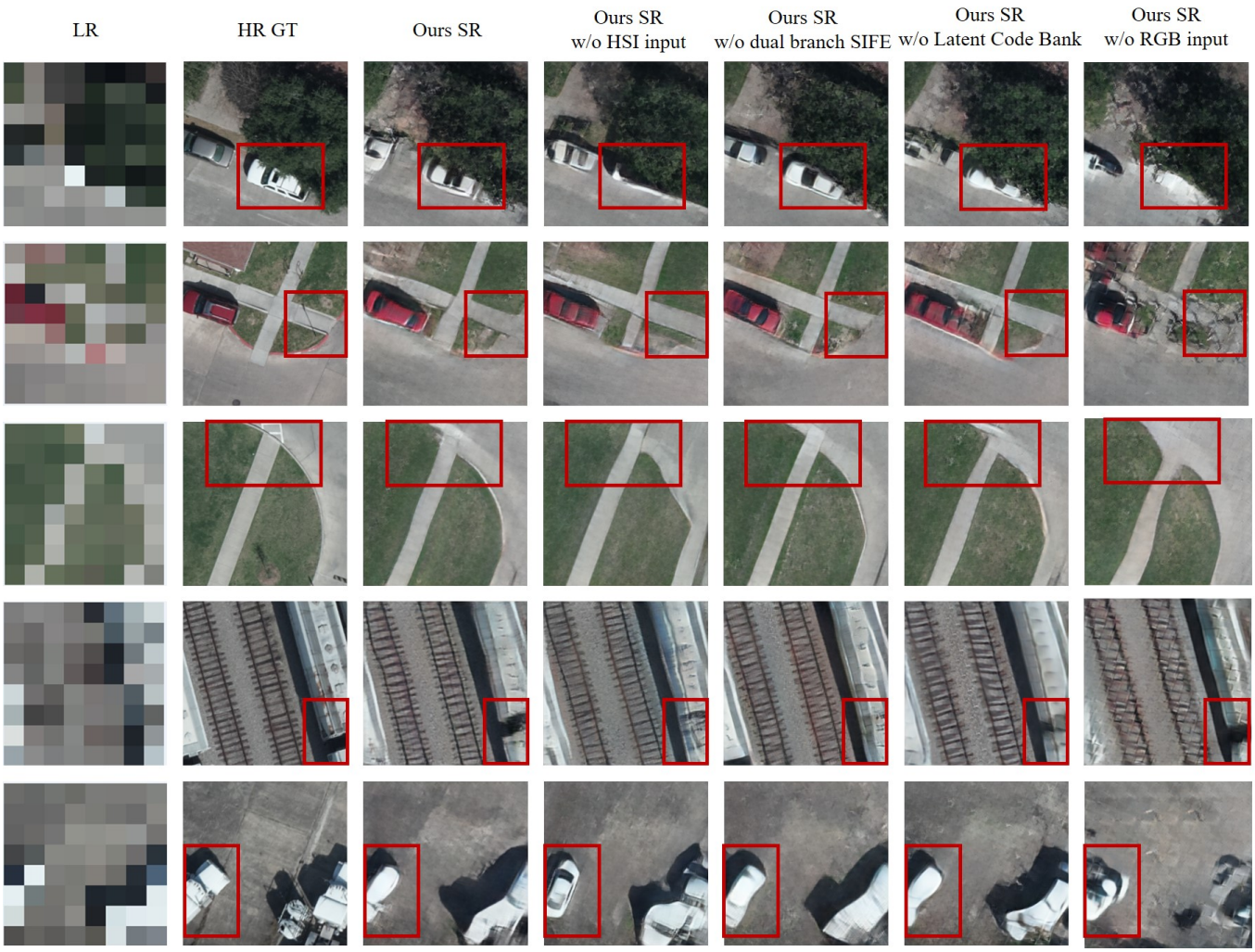

Fig. 7. Visualization results of ablation experiment. Results in the first row show the necessity of HSI in small-scale ground object disambiguation. Hyperspectral input helps the network capture car information under the shadow of trees and houses. The images in lines 2-4 show that hyperspectral input makes the boundary of ground objects more accurate. After removing HSI input, the structure of the lawn edge is inaccurate, and the segmentation of the train carriage is not constructed. The last row shows that the constraints of hyperspectral data ensure the semantic accuracy of the generated image. After removing HSI, an ordinary can is wrongly reconstructed into a car shape, causing serious semantic errors.

improve the quality and detail of the generated image and eliminate checkerboard effect. We can summarize the role of adding hyperspectral information guidance as follows:

- HSI information helps the network more effectively resolve the overlapping small-scale objects. In datasets, cars are often blocked by vegetation and house shadows. If only RGB information is used, it is difficult for even humans to distinguish cars mixed in the pixel lattice, but the addition of HSI information enables the network to capture the unique spectral information of cars and help the decoder complete the construction of small-scale ground objects.
- The addition of HSI information makes the boundary of the ground object interface more accurate and clear. Without the HSI information, the construction of the lawn shape is very inaccurate. Many curve shapes of lawn boundaries are constructed as straight lines, and some bare ground regions are also wrongly replaced by grassland.
- HSI information strengthens the semantic constraints on the generated images and avoids semantic misjudgment. When the resolution of the input image is very low, even human eyes cannot distinguish whether the white pixel grid is a white car or a white can. When we only use RGB as input, the network will produce misjudgment and construct it as a car, but after adding HSI information as a constraint, the network makes full use of the spectral specificity of different objects to avoid the seemingly reasonable but semantically wrong output.

Under spectral guidance, Ablation Experiments 2 and 3 suggest the effectiveness of the proposed network structure in information fusion and image generation. According to the comparison between the SR output of Ablation Experiment 2 and the output of our full implementation, when the dual branch structure is removed, some ground object boundaries are parsed incorrectly (such as the boundary of the train carriage), which is equivalent to the failure to make effective use of spectral information. According to the comparison

TABLE IV
IMPACT OF UNREGISTERED INPUT IMAGES.

	PSNR↑	SSIM↑	NIQE↓	LPIPS↓
The original data set	**20.80**	**0.5376**	5.56	**0.2863**
Offset 1 pixel	20.35	0.5289	5.50	0.2978
Offset 2 pixels	20.11	0.5233	5.48	0.3086
Offset 3 pixels	20.04	0.5210	5.45	0.3139
Offset 4 pixels	20.04	0.5205	**5.44**	0.3176

TABLE V
A COMPARISON ON THE MODEL PARAMETERS, FLOPS, AND INFERENCE TIME.

Method	Params	FLOPs	Inference Time
HSENet [21]	$5.9M$	$3.59G$	$0.050s$
RDN-liif [29]	$22.3M$	$92.68G$	$0.081s$
GLEAN [20]	$210.2M$	$113.51G$	$0.236s$
Ours	$234.1M$	$104.58G$	$0.224s$

between the SR effect of Ablation Experiment 3 and the original network effect, it can be found that the network generates images without the Latent Code Bank will produce a serious checkerboard effect and reduce the visual quality of the image.

In Ablation Experiment 4, by analyzing the evaluation metrics and visualization results, it can be seen that only adding hyperspectral images to try to restore high-resolution RGB images has a poor effect. We summarize the possible reasons as follows: 1) The RGB image and the hyperspectral image in the data set are slightly misaligned. When the two images are input together, our registration module can make necessary adjustments to the feature map based on attention. However, when generating RGB images only based on hyperspectral images, these alignment deviations will lead to blurred ground objects and incorrect edges. 2) The imaging mechanism of hyperspectral images and RGB images is different. The reconstruction of the RGB band based on a hyperspectral image involves the conversion of imaging style, which is not effective at present. By comparing the results of inputting only RGB images and inputting only hyperspectral images, we believe that in the proposed spectral guidance super-resolution task, RGB information mainly constrains the color of the generated image and the overall location of ground objects, while hyperspectral information mainly provides semantic auxiliary information. Both of them are necessary in our designed method.

D. Impact of unregistered input images

In this experiment, we artificially created misalignments when we preprocessed the dataset. We cause a deviation of 1 to 4 pixels between the hyperspectral image and the RGB image fed into the network, which means that the misalignment is between 12.5% and 50% of the original image. We created the deviation in both width and height dimensions. We calculate evaluation metrics PSNR, SSIM, NIQE, and LPIPS like other experiments. The evaluation value is shown in Tab. IV.

E. Computational Complexity, Parameters, and Speed

We use three different metrics to compare the computational complexity, parameters, and speed of our method with other state-of-the-art super-resolution methods. In Tab. V, we record the number of model parameters (Params), the number of floating-point operations (FLOPs), and the inference time of different models. We test on ×32 super-resolution task and execute our programs on an Nvidia GeForce GTX 1080 Ti graphics card. Compared with other methods, our method has more parameters and a longer inference time. This is because our method focuses on difficult large-factor super-resolution task, which requires more parameters and memories.

V. DISCUSSION

Although the experimental results demonstrate the effectiveness of our methods, it still has some limitations.

1) Our proposed method has a high computational cost——more model parameters, more floating-point operations, and longer inference time. So it's not worthwhile for our methods to deal with regular small-factor super-resolution tasks.
2) The ability to reconstruct sub-pixel ground objects relies on similar targets abound in the dataset, such as cars and playground track lines. However, for the targets that do not appear in the data set, the reconstruction effect of our method is poor.

VI. CONCLUSION

We propose a novel method for large-factor super-resolution of remote sensing images, which achieves up to 32x super-resolution with high visual fidelity under the premise of accurate ground object semantics. Considering the problems of sub-pixel mixing and semantic ambiguity, we propose to introduce additional hyperspectral images as input and design the spectral information fusion and extraction module (SIFE), which can effectively fuse low-resolution RGB images and hyperspectral images. We also introduce the Latent Code Bank and adversarial training to add rich details to SR results, which helps to improve image quality. Our method can produce high visual quality images with accurate semantics, clear outline, and real detailed texture. Compared with other most advanced methods, our method has achieved better results in terms of both quantitative metrics and visual quality.

REFERENCES

[1] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 2810–2825, 2021.
[2] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3232–3245, 2020.
[3] D. J. Mulla, "Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps," *Biosystems engineering*, vol. 114, no. 4, pp. 358–371, 2013.
[4] P. Shanmugapriya, S. Rathika, T. Ramesh, and P. Janaki, "Applications of remote sensing in agriculture-a review," *International Journal of Current Microbiology and Applied Sciences*, vol. 8, no. 01, pp. 2270–2283, 2019.

[5] I. Masser, "Managing our urban future: the role of remote sensing and geographic information systems," *Habitat international*, vol. 25, no. 4, pp. 503–512, 2001.

[6] A. K. Akanbi, S. Kumar, and U. Fidelis, "Application of remote sensing, gis and gps for efficient urban management plan, a case study of part of hyderabad city," *arXiv preprint arXiv:1312.4932*, 2013.

[7] J. Sanyal and X. X. Lu, "Application of remote sensing in flood management with special reference to monsoon asia: a review," *Natural Hazards*, vol. 33, no. 2, pp. 283–301, 2004.

[8] C. J. Van Westen, "Remote sensing and gis for natural hazards assessment and disaster risk management," *Treatise on geomorphology*, vol. 3, pp. 259–298, 2013.

[9] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[10] Z. Pan, J. Yu, H. Huang, S. Hu, A. Zhang, H. Ma, and W. Sun, "Super-resolution based on compressive sensing and structural self-similarity for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 9, pp. 4864–4876, 2013.

[11] B. Hou, K. Zhou, and L. Jiao, "Adaptive super-resolution for remote sensing images based on sparse representation with global joint dictionary model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2312–2327, 2017.

[12] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1920–1927.

[13] ——, "Adjusted anchored neighborhood regression for fast super-resolution," in *Asian conference on computer vision*. Springer, 2014, pp. 111–126.

[14] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3791–3799.

[15] H. Chavez-Roman and V. Ponomaryov, "Super resolution image generation using wavelet domain interpolation with edge extraction via a sparse representation," *IEEE Geoscience and remote sensing Letters*, vol. 11, no. 10, pp. 1777–1781, 2014.

[16] 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, 2015.

[17] ——, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[19] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[20] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "Glean: Generative latent bank for large-factor image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 245–14 254.

[21] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[22] M. Qin, S. Mavromatis, L. Hu, F. Zhang, R. Liu, J. Sequeira, and Z. Du, "Remote sensing single-image resolution improvement using a deep gradient-aware network with image-specific enhancement," *Remote Sensing*, vol. 12, no. 5, p. 758, 2020.

[23] B. Le Saux, N. Yokoya, R. Hänsch, and M. Brown, "Data fusion contest 2019 (dfc2019)," 2019. [Online]. Available: https://dx.doi.org/10.21227/c6tm-vw12

[24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[25] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[26] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[27] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[29] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8628–8638.

[30] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 251–260.

[31] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[32] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 606–615.

[33] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced gan for remote sensing image superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5799–5812, 2019.

[34] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 2437–2445.

[35] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[37] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[38] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[39] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, 2022, pp. 10 684–10 695.
[41] T. Shang, Q. Dai, S. Zhu, T. Yang, and Y. Guo, "Perceptual extreme super-resolution network with receptive field block," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 440–441.
[42] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.
[43] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.
[44] Y. Sugawara, S. Shiota, and H. Kiya, "Super-resolution using convolutional neural networks without any checkerboard artifacts," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 66–70.
[45] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1243–1247, 2017.
[46] J. M. Haut, M. E. Paoletti, R. Fernández-Beltran, J. Plaza, A. Plaza, and J. Li, "Remote sensing single-image superresolution based on a deep compendium model," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 9, pp. 1432–1436, 2019.
[47] Z. Pan, W. Ma, J. Guo, and B. Lei, "Super-resolution of single remote sensing image based on residual dense backprojection networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7918–7933, 2019.
[48] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5791–5800.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[50] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.
[51] T. Wang, W. Sun, H. Qi, and P. Ren, "Aerial image super resolution via wavelet multiscale convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 769–773, 2018.
[52] W. Ma, Z. Pan, J. Guo, and B. Lei, "Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3512–3527, 2019.
[53] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3633–3643, 2019.
[54] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.
[55] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 694–711.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[58] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 2018.
[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
[60] MMEditing Contributors, "MMEditing: OpenMMLab image and video editing toolbox," https://github.com/open-mmlab/mmediting, 2022.
[61] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1524–1532.
[62] 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, 2004.
[63] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

Yapeng Meng is currently working toward his bachelor's degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include deep learning, remote sensing image processing, and brain-inspired computing.

Wenyuan Li received his B.S. degree from North China Electric Power University, Beijing, China in 2017. He is currently working toward his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include self-supervised learning and remote sensing image processing.

Sen Lei received the B.S. and Ph.D. degree from the School of Astronautics, Beihang University, Beijing, China, in 2015 and 2021. He is now with AVIC Chengdu Aircraft Industrial (Group) Company Ltd.

Zhengxia Zou received his BS degree and his Ph.D. degree from Beihang University in 2013 and 2018. He is currently an Associate Professor at the School of Astronautics, Beihang University. During 2018-2021, he was a postdoc research fellow at the University of Michigan, Ann Arbor. His research interests include computer vision and related problems in remote sensing. He has published more than 20 peer-reviewed papers in top-tier journals and conferences, including TPAMI, TIP, TGRS, CVPR, ICCV, AAAI. His research was featured in more than 30 global tech media and was adopted by a number of application platforms with over 50 million users worldwide. His personal website is https://zhengxiazou.github.io/.

Zhenwei Shi (Member, IEEE) received the Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, from 2013 to 2014. He is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. He has authored or coauthored over 200 scientific papers in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters, the IEEE Conference on Computer Vision and Pattern Recognition, and the IEEE International Conference on Computer Vision. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi also serves as an Editor for the Pattern Recognition, the ISPRS Journal of Photogrammetry and Remote Sensing, the Infrared Physics and Technology, and so on. His personal website is http://levir.buaa.edu.cn/.