Super-Resolution for Remote Sensing Imagery via the Coupling of a Variational Model and Deep Learning

Jing Sun⬤, Huanfeng Shen⬤, *Senior Member, IEEE*, Qiangqiang Yuan⬤, *Member, IEEE*, and Liangpei Zhang⬤, *Fellow, IEEE*

Abstract—Image super-resolution (SR) is an effective way to enhance the spatial resolution and detail information of remote sensing images to obtain a superior visual quality. As SR is severely ill-conditioned, effective image priors are necessary to regularize the solution space and generate the corresponding high-resolution (HR) image. In this article, we propose a novel gradient-guided multiframe SR (MFSR) framework for remote sensing imagery reconstruction. The framework integrates a learned gradient prior as the regularization term into a model-based optimization method. Specifically, the local gradient regularization (LGR) prior is derived from the deep residual attention network (DRAN) through gradient profile transformation (GPT). The nonlocal total variation (NLTV) prior is characterized using the spatial structure similarity of the gradient patches with the maximum a posteriori (MAP) model. The modeled prior performs well in preserving edge smoothness and suppressing visual artifacts, while the learned prior is effective in enhancing sharp edges and recovering fine structures. By incorporating the two complementary priors into an adaptive norm-based reconstruction framework, the mixed *L1* and *L2* regularization minimization problem is optimized to achieve the required HR remote sensing image. Extensive experimental results on remote sensing data demonstrate that the proposed method can produce visually pleasant images and is superior to several of the state-of-the-art SR algorithms in terms of the quantitative evaluation.

Index Terms—Deep learning, image super-resolution (SR), local gradient regularization (LGR), nonlocal total variation (NLTV), remote sensing.

I. INTRODUCTION

WITH the development of remote sensing technology in recent years, the availability of remote sensing data

Jing Sun is with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China (e-mail: rainsunny@hotmail.com).

Huanfeng Shen is with the School of Resource and Environmental Sciences and the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: shenhf@whu.edu.cn).

Qiangqiang Yuan is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: yqiang86@gmail.com).

Liangpei Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zlp62@whu.edu.cn).

has increased, and the need for remote sensing data with high spatial and temporal resolutions has become more and more urgent in geoscientific applications [1]. However, due to the limitation of the sensors and the complexity of the imaging environment, obtaining high-resolution (HR) remote sensing images often entails high data acquisition costs [2]. Therefore, remote sensing image super-resolution (SR) has been a major research focus and is an effective way to enhance the spatial resolution of remote sensing images with an inexpensive and powerful solution. SR reconstruction technology aims to recover the HR image from one or multiple low-resolution (LR) images to meet the needs of practical applications. However, the spatial resolution and clarity of remote sensing images are degraded due to the influence of the data acquisition and transmission processes [3], [4]. There have been many proposed works on the SR of remote sensing images employing different priors or network models. According to the number of input LR remote sensing images, the conventional SR approaches can be roughly categorized into single-frame SR (SFSR) [5], [6], [7], [8] and multiframe SR (MFSR) [9], [10], [11], [12].

Compared to LR remote sensing images, HR images contain richer texture information and high-frequency details that can be lost in the process of data acquisition. The main challenge for remote sensing image SR is to recover the high-frequency details of the LR image, which are more sensitive to human perception. The early SFSR methods often employed a hand-crafted prior [13] to recover the HR remote sensing image from a single LR image. With the rapid development of deep learning, some studies [14], [15] have used remote sensing images to retrain a network designed for natural images. Due to the more complex structure of remote sensing images, compared to natural images, these methods are less effective than when applied to natural images. Subsequently, a large number of deep learning-based methods [6], [16], [17] have emerged as a preferable choice for mining a generalizable prior and intraimage information from large-scale remote sensing data and have achieved significant efficiency and generalization improvements in the SR reconstruction of remote sensing images.

Considering the unique characteristics of remote sensing images, the channel attention network was introduced by Yu et al. [18] to aggregate sufficient feature information in

each channel and achieved a superior performance. Through experiments, it has been demonstrated that this strategy can effectively preserve the spatial detail information in the recovered image. In order to boost the generalization of SR models on various remote sensing scenes, a scene-adaptive strategy [multiscale attention network (MSAN)] was employed by Zhang et al. [7] to accurately describe the structural characteristics of different scenes. Dong et al. [19] proposed an interpretable scale-propelled diffusion (ISPDiff) model for hyperspectral image (HSI) SR. Different from traditional SR methods, this model combines the underlying principles of HSI-SR with denoising diffusion probabilistic model for progressively unrolling reconstruction by making full use of spectral and spatial information at various scales from two sources in an interpretable way. In addition, according to the hierarchical distribution characteristics of remote sensing images, Xiao et al. [20] proposed a cross-scale hierarchical transformation method to effectively explore cross-scale representations in remote sensing images. More recently, a cross-sensor degradation modeling strategy was proposed for remote sensing image SR by Qiu et al. [21], which aims to bridge the gap between the images obtained by the source and target sensors. The above methods have utilized different structures for the cross-scale hierarchical characteristics and low prior information of remote sensing images, and the lack of detail in the HR space. However, the information used by the SFSR methods is limited to the spatial domain of the LR image, which does not consider the temporal information of remote sensing imagery. For remote sensing image processing, the performance of deep learning-based SFSR is limited without the use of the complementary information between temporal frames.

On the other side, the MFSR approach can super-resolve the HR image by merging the temporal subpixel information from corresponding LR images, which are usually obtained from either geostationary orbit satellite or video satellite platforms. Since sequential satellite images contain complementary spatial and temporal information, the deep learning-based MFSR methods need to simultaneously model the spatiotemporal relationship between sequential frames. To enhance the resolution of remote sensing images progressively enhanced convolutional neural network for satellite image SR (PECNN) method composed of two subnetworks was proposed by Jiang et al. [22] to finely learn the structural information and low-level features in wide scenes. For dealing with the noise in satellite images, a generative adversarial-based edge-enhancement network was proposed by Jiang et al. [23] to enhance the high-frequency edge information in satellite video. While these methods have made progress in satellite image SR, the limited spatial information restricts their ability to reconstruct more precise textures. Differing from the previous MFSR methods for remote sensing images, the multiframe video SR neural network (MVSRnet) method [10] can merge the motion information among adjacent frames and highlight the importance of extracted features with an attention mechanism. This method utilizes optical flow estimation to warp frames based on the current frame and learns the progress of the multiframe fusion from an external database. Sal-vetti et al. [24] proposed the residual attention MFSR network, which leverages feature extraction from multiple LR images of the same scene, resulting in reconstructed images with fine texture details. The experimental results demonstrated that this fusion approach can improve the quality of the reconstruction. However, the low efficiency of the spatiotemporal information fusion results in the poor generalization ability of the MFSR approach when applied to remote sensing images. More advanced methods learn the subpixel registration and achieve fusion simultaneously through a deep neural network. In order to model the spatiotemporal information collaboratively, a novel fusion strategy of temporal grouping projection and an accurate alignment module was proposed by Xiao et al. [25] for satellite video SR (VSR) to effectively alleviate the alignment difficulties. In addition, Shen et al. [1] proposed an edge-guided video SR (EGVSR) framework for satellite imagery, which can help the network focus more on the structure of ground objects and enrich the details in the SR results of remote sensing images. Most of the deep learning-based methods typically simplify the image degradation model and use bicubic interpolation downsampling to synthesize the LR images [8]. However, a model trained on simulated remote sensing images often lacks universality and generalizability when handling multiple degradations.

To handle the various unknown degradations in real-world remote sensing images, recent research has leveraged degradation estimation to reconstruct the SR image with a deep joint estimation network [26]. Nevertheless, degradation estimation methods are usually time-consuming and struggle to obtain an accurate estimation, resulting in poor SR performance due to the large estimation errors. Unlike the deep learning-based methods, which are highly dependent on training samples, the variational model-based approaches focus on designing some prior knowledge as the reconstruction constraint to regularize the super-resolved images without using any training samples [27]. These methods consider the imaging mechanism and construct the energy function according to a degradation model between the ideal image and the degraded observations, which is deterministic and theoretically reasonable [28]. Due to the rigorous theory, the variational model-based methods are often more accurate than the traditional methods and have the advantages of a strong noise reduction ability and convenient integration of spatial prior constraints.

The MFSR approach based on a variational model consists of a data fidelity term and a regularization term. Generally speaking, the data fidelity term measures the model error between the degraded observations and the ideal image, while the regularization term imposes the structural and statistical characteristics of the image itself as the model constraint to achieve a robust solution. A maximum a posteriori (MAP)-based MFSR method with $L1$ norm data fidelity and an edge-preserving Huber regularization prior was proposed by Shen et al. [29] to super-resolve multitemporal Moderate Resolution Imaging Spectroradiometer (MODIS) images. Subsequently, several variational model-based MFSR approaches [30], [31] have been proposed for remote sensing images captured with different satellite sensors and angles. Generally, the performance of the variational model-based

MFSR methods is highly dependent on the image priors, which define the different feature models of the images. Recently, the nonlocal self-similarity (NLSS) property of remote sensing images for natural scenes, where small textured regions tend to repeat multiple times at different locations inside the image, has become one of the most pervasive and powerful priors. In order to preserve fine texture details, Liu et al. [32] proposed a traditional SR framework that formulates a nonlocal regularization term with spatiotemporal-domain NLSS. However, due to the high background complexity and large target scale variation of remote sensing images, although the traditional MFSR methods can improve the image resolution to a certain extent, they still have shortcomings in high-frequency edge detail recovery and noise artifact suppression.

Motivated by the need to recover the edge sharpness, many approaches [33], [34], [35], [36] have attempted to recover the high-frequency details by modeling and estimating the image edges and gradients. With the aim of imposing appropriate edge priors, Yang et al. [34] introduced a deep edge-guided recurrent residual (DEGREE) network to reconstruct images progressively to the desired resolution. DEGREE further enhances the edge-preserving capability of the SR process, i.e., LR images and their edge plots can jointly infer the sharp edge details of the HR image during the recurrent recovery process. However, the edge priors contain only a small part of the high-frequency information, which limits the improvement in reconstruction performance. Clearly, the gradient prior is more effective in recovering sharp edges, since the lost high-frequency details, such as edges and textures, are mainly contained in the image gradient field. As the high-frequency information is mostly contained in the image gradient field, an effective SR method was proposed by Chen et al. [35] via learning-based gradient regularization and NLSS modeling. The gradient prior learned from the jointly optimized regression model was constructed as a gradient regularization term in the SR process. In addition, Song and Liu [36] trained a gradient prior learning network to regularize the image reconstruction and better recover the textural information and high-frequency components in the super-resolved image.

However, modeling the gradient field via a simple model ignores the local geometric structures of the gradients. Several SR algorithms [37], [38], [39] based on gradient profile representation have been developed to restore the edge sharpness of the reconstructed images. Since a sharp edge in an image is related to the gradient concentration perpendicular to the edge, the gradient transform is developed on the basis of statistical evaluation of the gradient profile fitting error. The LR gradient is converted to an HR gradient through statistical and parametric models, and then, the transformed gradient field is taken as the constraint for HR image estimation [38]. Nevertheless, the parametric representation models are not flexible enough, using only a few parameters to capture the diverse transformation relationships between the LR and HR gradient profiles. As a result, the obtained HR images are usually oversharpened or suffer from artifacts due to the uncertainty in the gradient field estimation. Thus, the utilization of a nonparametric transformation representation has been considered to construct a more adaptable gradient

transformation model. Li et al. [39] designed an effective image SR approach by incorporating the gradient profile prior derived from example-based gradient field estimation to enhance the edge sharpness and restore the image details in the super-resolved image. Although the learned gradient prior is more expressive than the parametric gradient profile models, the example-based architecture used in this framework has shortcomings in extracting the high-frequency features for SR reconstruction, especially in the complex detail areas.

On the one hand, these SR methods based on edge and gradient estimation are primarily tailored for natural images. Hence, in the context of remote sensing images, it is imperative to incorporate a gradient profile transformation (GPT) model in the SR process. On the other hand, it is noteworthy that, due to the diverse target objects and significant changes in the high-frequency information distribution patterns within remote sensing data, such data often exhibit highly intricate spatial and hierarchical distributions in both the local and global features. Consequently, when designing the regularization term for remote sensing image SR, consideration should be given to incorporate both the local texture details and global structure information. According to the spatial distribution of the image features, image priors can be divided into two main categories: nonlocal priors and local priors. A nonlocal prior reveals the interscale and intrascale redundant recurrence of small image patches, based on the self-similarity of the image patches. The nonlocal total variation (NLTV) prior is a promising and significant prior characterized by the internal spatial structure similarity of the image patches. A local prior can be formulated as various smoothness models based on the assumption of local smoothness. The smoothness models perform well in recovering smooth regions, but high-frequency details may be smoothed out. Due to the different emphases on the different characteristics of images, the nonlocal and local priors have their own merits and drawbacks in remote sensing image SR. By assembling multiple complementary priors, it is possible to make a tradeoff between artifact reduction and detail preservation, suppressing the displeasing artifacts well and preserving the fine details.

Inspired by these works, we propose a gradient-guided MFSR method that employs a variational model with embedded learning based on local gradient regularization (LGR) and an NLTV prior. By exploiting both the learned LGR prior and the modeled NLTV prior, the proposed method benefits from the complementary properties of the learning-based and model-based SR approaches. More specifically, a deep residual attention network (DRAN) is trained to learn the horizontal and vertical gradients. The HR gradient profile is estimated from the learned gradient field via GPT, which is used as the LGR constraint during the SR process to recover fine image details and enhance edge sharpness. Benefiting from the development of nonlocal means and cross-scale self-similarity methods, the internal spatial similarity of the gradient blocks is used to characterize the NLTV prior to suppress image noise and artifacts. In addition, to ensure the noise robustness and edge smoothness of the reconstructed image, a reliable and flexible adaptive fidelity norm model is established to suppress the artifacts that may be produced by the local

gradient prior. By incorporating the two complementary priors into an adaptive norm-based reconstruction framework, the HR estimate can be obtained via the alternating direction method of multipliers (ADMM) algorithm. The proposed method has the following benefits.

1) We propose to utilize the DRAN to learn the latent gradient field of the desired HR remote sensing image. The learned gradient serves as an LGR prior in the SR process via GPT.

2) The internal spatial similarity of gradient blocks is used to characterize the NLTV prior. This modeled prior is constructed as a regularization term to constrain the nonlocal features of the remote sensing imagery.

3) We propose an effective gradient-guided MFSR framework for remote sensing imagery by incorporating the learned LGR prior and the modeled NLTV prior. This framework combines the model-based and learning-based methods in a simple manner. The ADMM algorithm is exploited to effectively optimize the proposed minimization problem.

4) The proposed MFSR method takes advantage of the complementary modeled and learned priors, the artifacts in the super-resolved remote sensing image can be removed, and the fine structures and sharp edges can be well recovered. The extensive experiments validate the superior performance of the developed MFSR scheme.

The remainder of this article is organized as follows. Section II introduces the proposed MFSR framework based on modeled and learned priors. In Section III, the experimental results are provided, including extensive comparisons and analyses. Finally, our conclusions are drawn in Section IV.

II. METHODOLOGY

For the model-based MFSR methods, SR is usually considered as an ill-posed inverse problem defined by the reconstruction constraint and image prior. Clearly, the more accurate the model is, the better the SR performance will be. The image prior defines the different characteristic models of the images and plays a pivotal role in the quality of a reconstructed HR image. In this section, we propose a gradient-guided MFSR method combining the model-based NLTV prior with the learning-based LGR prior to significantly improve the performance of remote sensing image SR. On the one hand, the initial HR gradient is learned from the DRAN. The final HR gradient profile with higher sharpness is then estimated from the learned initial gradient field via GPT and serves as the LGR constraint during the SR process. The novel LGR prior is good at fine detail recovery and edge sharpness enhancement. On the other hand, the NLTV prior is essentially a nonlocal and internal image-based prior, which performs well in preserving edge smoothness and suppressing reconstruction artifacts. Benefiting from the complementary properties of the modeled and learned priors, HR remote sensing images are reconstructed with a better objective and subjective quality.

The model-based MAP method incorporates the two complementary prior constraints to estimate the desired HR image by minimizing an objective function of the posterior prob-

ability. In addition, to deal with the mixed noise and/or complicated model error in remote sensing image SR, a regularized framework with an adaptive fidelity norm is employed. Both the adaptive norm fidelity term and the joint regularization term guarantee the robustness of the proposed MFSR framework. The basic formula of the proposed MFSR method for remote sensing images is as follows:

$$\hat{z} = \arg\min_{z} \left\{ \sum_{k}^{K} \underbrace{\|y_k - \mathrm{DBM}_k z\|_p^p}_{\text{fidelity term}} + \alpha \underbrace{P_{\mathrm{ext}}(\nabla z)}_{\text{learned prior}} + \beta \underbrace{P_{\mathrm{int}}(\nabla z)}_{\text{modeled prior}} \right\}$$

(1)

where z is the ideal HR remote sensing image required to be reconstructed, and p is the fidelity norm value in the interval $[1, 2]$. y_k is the kth observed LR image, where $k = 1, 2, \ldots, K$, with K being the number of LR images. B represents the blur matrix, including the sensor blur, optical blur, and atmospheric turbulence. M_k is the motion matrix, and the relative movement parameters can be estimated using the subpixel shifts of the multiple LR images. D denotes the downsampling matrix. α and β represent the regularization parameters, which provide a tradeoff between the data fidelity term and the regularization term. The fidelity term measures the reconstruction error to ensure that pixels in the reconstructed HR image are close to the real values. The model-based and learning-based priors are combined with model features (such as nonlocal similarity) and learned features (such as the learned gradient features) to regularize the solution space of the HR image.

The flowchart of the proposed gradient-guided MFSR method is demonstrated in Fig. 1. This framework contains three main parts: 1) learning-based gradient estimation and LGR term construction through GPT; 2) nonuniform interpolation (NUI) initialization and NLTV regularization term construction; and 3) energy function optimization and HR image updating. The details of these parts are introduced in Sections II-A–II-C.

A. Gradient-Learning Network

The main parts of remote sensing images belong to the low-frequency components, which are a comprehensive measure of the image intensity. Meanwhile, the high-frequency components correspond to the parts of the image that change dramatically, i.e., the edges or noise and details of the image. Aimed at better reconstructing the texture information and high-frequency components of remote sensing images, the effective integration of subpixel complementary information between image sequences and the reliable prediction of the high-frequency components should be considered simultaneously. To achieve an HR remote sensing image with more fine details and sharp edges, we introduce an external gradient prior as a regularization constraint to recover the high-frequency details. An effective gradient regularization prior needs to be established on accurate gradient information in the SR reconstruction process, but the gradient of the LR image will not have sharp details, so it is necessary to use the existing gradient information to reconstruct the missing gradient information.
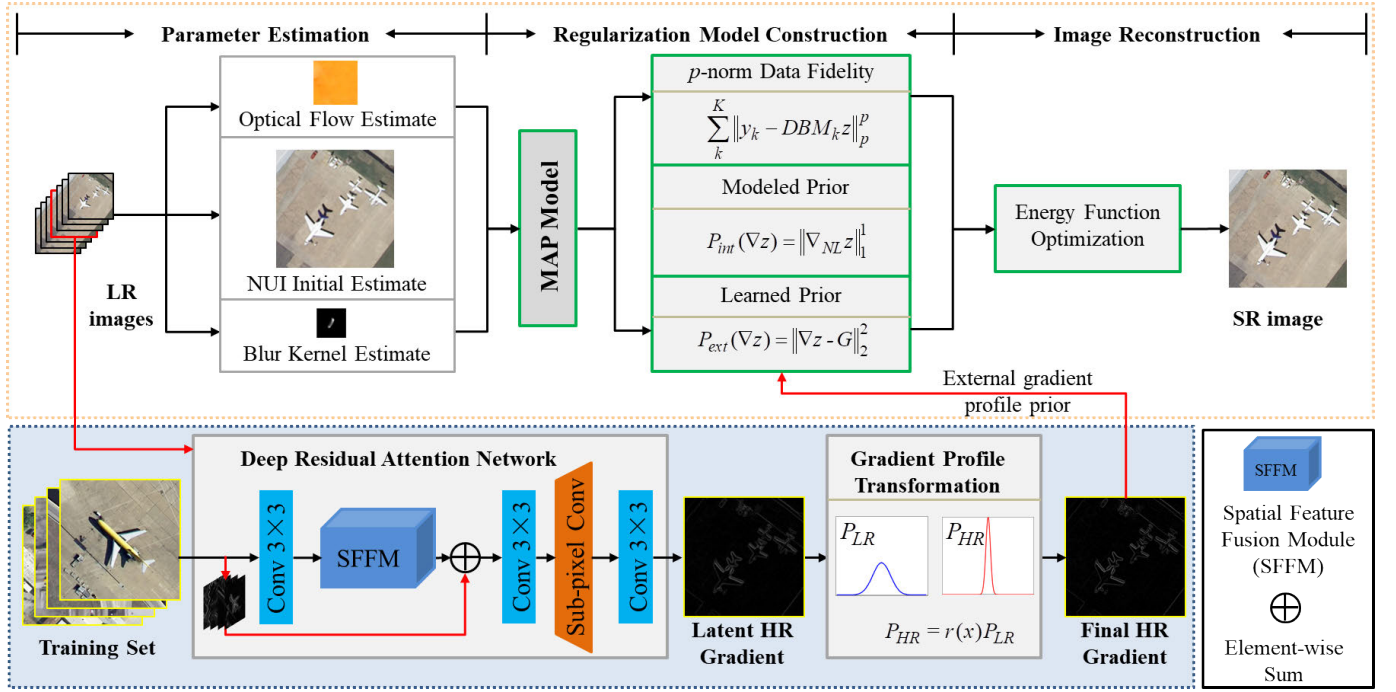
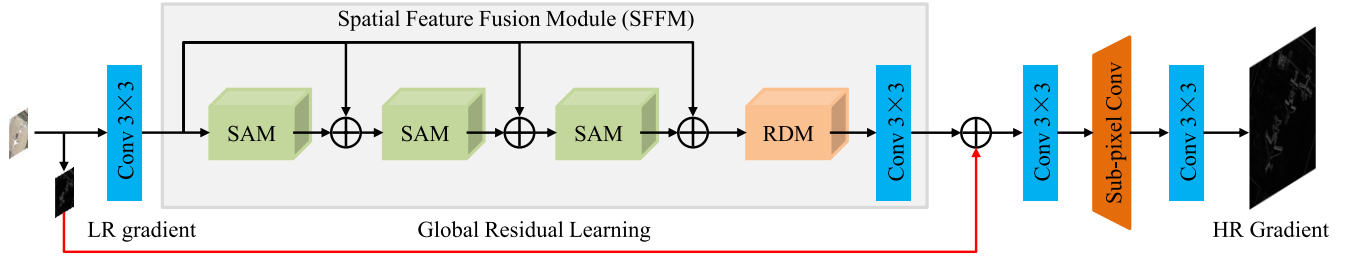Fig. 1. Flowchart of the gradient-guided MFSR algorithm.

Fig. 2. Architecture of the gradient-mapping DRAN.

Differing from the variational model-based approaches, the deep learning-based methods introduce external training datasets to capture image features, without the need for explicit expression of the regularization term. In addition, the deep learning-based approaches are more capable of information extraction and feature fusion with an end-to-end learning framework. Therefore, considering the excellent learning capacity of convolutional neural networks, the DRAN is constructed and trained for latent gradient prior learning. As shown in Fig. 2, the gradient-mapping network contains a spatial feature fusion module (SFFM), one upsampling layer, and several convolutional layers. The attention mechanism is applied in the SFFM, which consists of three spatial attention modules (SAMs), one residual dense module (RDM), and a convolutional layer.

In the proposed framework, the DRAN is used to predict a latent HR gradient field that has accurate and clear contours. Due to the high-frequency characteristics of image gradients, the mapping from LR gradient to HR gradient is actually high-frequency to high-frequency mapping. During image degradation, the high-frequency components are destroyed and are more unstable than the low-frequency components.

Therefore, learning the HR gradient prior directly from the LR gradient is difficult for the network as only the high-frequency information is given. Conversely, it is easier to learn from the LR image itself, which can also provide reliable low-frequency information. In the proposed approach, the proposed DRAN takes the LR image and the corresponding horizontal and vertical gradients as an input.

The discrete gradient operators $[-1/2, 0, 1/2]$ and $[-1/2, 0, 1/2]^{\mathrm{T}}$ are separately employed to extract the horizontal gradient g_y^h and vertical g_y^v gradient of the LR image y, respectively. Note that the binary strategy is eliminated to avoid the appearance of false edges and missing image features. The LR image y and the gradient g_y^h and g_y^v are then taken as the input of the network. The shallow feature F_0 is first extracted from the input LR image y and can be formulated as $F_0 = f_{\mathrm{Conv3}}(y)$, where f_{Conv3} denotes a convolution operation with the kernel size of 3×3.

1) Spatial Attention Module: The SAM is able to keep the good properties of the original input features, not only emphasizing the more important features but also suppressing the less useful features. Since the high-frequency features are more important for the HR gradient reconstruction, the
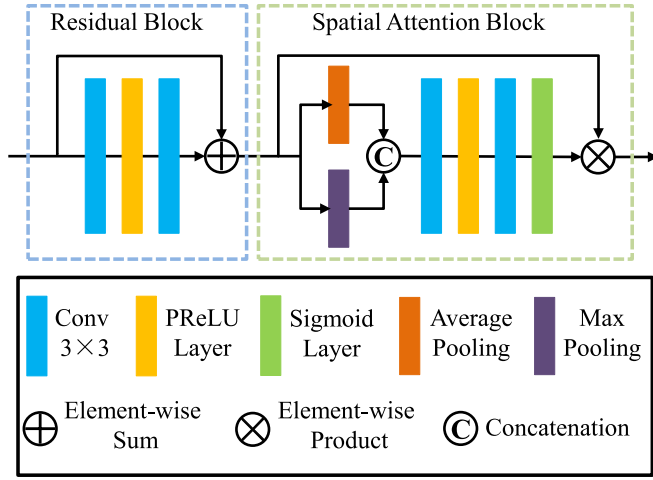
Fig. 3. Architecture of the SAM.

Fig. 4. Architecture of the RDM.

spatial attention mechanism is employed to further enhance the gradient information. The structure of the SAM is displayed in Fig. 3, which consists of one residual block (RB) and a spatial attention block (SAB).

The initial feature F_0 is fed into a stack of SAMs to make full use of the shallow layer information and capture the long-range dependencies. The deep feature map of the nth SAM can be expressed as follows:

$$F_{S_n} = f_{S_n}(F_{S_{n-1}}) + W_S F_0 \qquad (2)$$

where f_{S_n} denotes the function of the nth SAM. $F_{S_{n-1}}$ and F_{S_n} represent the input and output feature maps of the nth SAM, respectively. W_S is a learnable parameter, and "+" is the elementwise addition operation.

In the SAM, the initial feature F_0 is first processed by an RB that consists of two convolutional layers with the kernel size of 3×3 and a parametric rectified linear unit (PReLU) activation function [40]. The output feature of the RB can be written as follows:

$$F_{R_n} = f_{\text{Conv3}}(\rho(f_{\text{Conv3}}(F_{S_{n-1}}))) \qquad (3)$$

where $\rho(\cdot)$ denotes the PReLU activation function, and F_{R_n} represents the intermediate feature of the nth SAM. The obtained intermediate feature F_{R_n} is then fed into an SAB, which uses the spatial attention mechanism to make the model focus on the high-frequency region of the image.

In the SAB, both maximum and average pooling operations are employed simultaneously to squeeze the spatial information of the input feature map. The weights of the spatial attention mechanism can be expressed as follows:

$$W_{\text{SA}} = \sigma(f_{\text{Conv3}}(\rho(f_{\text{Conv3}}([P_{\text{avg}}(F_{R_n}); P_{\text{max}}(F_{R_n})])))) \qquad (4)$$

where $[\cdot; \cdot]$ represents the concatenation operation; P_{avg} and P_{max} denote the average pooling operation and the maximum pooling operation, respectively; and $\sigma(\cdot)$ represents the sigmoid activation function. The attention map is then used to modulate the intermediate features. The modulated features F_{S_n} can be obtained by multiplying the intermediate features
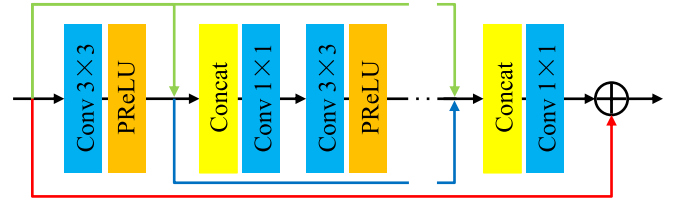
F_{R_n} with the spatial weight map W_{SA}, which can be written as follows:

$$F_{S_n} = F_{R_n} \otimes W_{\text{SA}} \qquad (5)$$

where $\otimes$ denotes the elementwise multiplication.

Finally, the modulated features F_{S_N} of the last SAB and the initial feature F_0 are connected by a skip connection structure to generate the final output feature F_S of the last SAM.

2) Residual Dense Module: To fully utilize the hierarchical features, the final enhanced feature F_S from the SAM is then passed into an RDM. The RDM combines the residual structure and dense connections to produce improved features and encourage feature reuse, as shown in Fig. 4.

In the RDM, each layer aggregates the feature maps from all the preceding layers. More specifically, to generate the feature maps effectively, the input for each convolutional layer is the concatenation of the outputs from all the previous convolutional layers. A 1×1 convolutional layer $f_{\text{Conv1}}(\cdot)$ is used for feature pooling and dimension reduction after the concatenated feature map. Thus, the concatenated feature map is merged along the channel dimension to produce the merged feature map F_{C_m}. The concatenate module can be written as follows:

$$F_{C_m} = f_{\text{Conv1}}(f_{\text{Concat}}(F_{D_1}, F_{D_2}, \dots, F_{D_m})) \qquad (6)$$

$$F_{D_{m+1}} = \rho(f_{\text{Conv3}}(F_{C_m})) \qquad (7)$$

where F_{C_m} denotes the output feature maps of the mth concatenation operation f_{Concat}. F_{D_m} and $F_{D_{m+1}}$ represent the input and output feature maps of the $(m + 1)$th convolutional layer, respectively. Finally, the merged features F_{D_m} containing more representational information can be obtained through the RDM.

3) Residual Reconstruction Module: The residual reconstruction module mainly consists of one upsampling layer and several convolutional layers. The output feature of the last RDM is passed through a convolutional layer, followed by global residual learning, which is aimed at facilitating the feature representation and preserving the base performance of the network. As stated before, the extracted gradient of the LR image y is denoted as $g = [g_y^h, g_y^v]$, which is connected with the final merged feature F_D by the global residual learning. The output of the global residual learning is denoted as F_G, which is formulated as follows:

$$F_G = f_{\text{Conv3}}(F_D) + g. \qquad (8)$$

In the proposed approach, the common subpixel convolutional layer is employed as the postupsampling layer. At the

end of the network, the final latent HR gradient $\hat{g} = [\hat{g}_y^h, \hat{g}_y^v]$ can be obtained through the following formula:

$$\hat{g} = f_{\text{Conv3}}\big(f_{\text{Up}}(f_{\text{Conv3}}(F_G))\big) \tag{9}$$

where $f_{\text{Up}}(\cdot)$ represents the subpixel convolutional layer.

Finally, the end-to-end gradient-mapping network is optimized with a certain loss function. An $L1$ regularization term is constructed to constrain the training of the gradient-learning network (LG), so that it can generate a gradient field with more accurate details and provide a gradient prior for the SR process. The gradient-learning loss function is defined as follows:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\big\| f_{\text{DRAN}}\big(y_i, g_{y_i}; \theta\big) - g_{z_i}\big\|_1^1 \tag{10}$$

where $f_{\text{DRAN}}(\cdot)$ denotes the whole end-to-end mapping function and θ represents the parameter of the network.

The proposed network has 29 convolutional layers. Since the DRAN mainly transforms the spatial distribution of edges and textures in the LR and HR gradient maps, the designed lightweight LG can capture the structural dependency and generate accurate gradient maps.

B. Gradient Profile Transformation

Since gradients reveal the local variations in image intensity, we chose to learn the gradient-mapping function from LR image to HR image via the above DRAN. However, modeling the gradient field via the lightweight LG only considers the marginal distributions and ignores the local geometric structures of the gradients. In this section, we describe how the deep learning-based gradient-mapping network is combined with GPT to estimate the HR gradient profiles, which further make up an HR gradient image to provide the gradient field prior for the proposed MFSR method.

The gradient spatial information determines the texture details and edge information of an image. The GPT approach is introduced to enhance the useful gradient spatial information and reconstruct the HR gradient image with more accurate and sharp details. The gradient profile is a feature describing the spatial layouts of the edge gradients and is defined as a 1-D profile of the gradient magnitudes along the gradient direction of an edge pixel [38]. In brief, the gradient profile prior is a parametric distribution that describes the shape and sharpness of the image gradients.

The generalized Gaussian distribution (GGD) model is applied to model the normalized gradient profile and is defined as follows:

$$h\big(x; \sigma, \lambda\big) = \frac{\lambda\alpha(\lambda)}{2\sigma\Gamma\big(\frac{1}{\lambda}\big)}\exp\left\{-\Big[\alpha(\lambda)\Big|\frac{x}{\sigma}\Big|\Big]^\lambda\right\} \tag{11}$$

where $\Gamma(\cdot)$ is the gamma function and σ represents the sharpness of the gradient profile. $\alpha(\lambda) = (\Gamma(3/\lambda)/\Gamma(1/\lambda))^{1/2}$ is the scaling factor, which makes the second moment of GGD equal to σ^2. λ is the shape parameter, which controls the overall shape of the distribution. As indicated in [38], the value of $\lambda = 1.6$ represents a suitable generic model for the gradient profiles sharpness and is also independent of the

image resolution. Based on this property, it can be deduced that the profile sharpness is the only parameter that necessitates further investigation.

The HR gradient field is estimated by transforming the LR gradient field with the gradient profile model. Intuitively, the HR gradient profile p_{HR} can be derived through multiplying the LR gradient profile p_{LR} by the transform ratio $r(d)$. The ratio between the two gradient profiles is computed as follows:

$$\begin{aligned} r(d) &= \frac{h\big(d; \sigma_{\text{HR}}, \lambda\big)}{h\big(d; \sigma_{\text{LR}}, \lambda\big)} \\ &= \frac{\sigma_{\text{LR}}}{\sigma_{\text{HR}}}\cdot\exp\left\{-\left(\frac{\alpha(\lambda)\cdot d}{\sigma_{\text{HR}}}\right)^\lambda + \left(\frac{\alpha(\lambda)\cdot d}{\sigma_{\text{LR}}}\right)^\lambda\right\} \end{aligned} \tag{12}$$

where d is the curve distance to the edge pixel along the gradient profile. Sun et al. [37] suggested a function for sharpness enhancement: $\sigma_{\text{HR}} = \sigma_{\text{LR}}(1 - e^{-\mu\sigma_{\text{LR}}})$.

Subsequently, GPT is performed on the learned gradient field, based on the gradient profile sharpness mapping between the LR image and the HR image, in order to obtain the final HR gradients. Consequently, the final HR gradient field $G(x)$ is directly estimated from the learned gradient field $\hat{g}(x)$ by the following formula:

$$G(x) = r(x)\cdot\hat{g}(x). \tag{13}$$

The gradient field transformation is performed for each pixel rather than each gradient profile. The reason for this is that the new gradient is required at each pixel grid, while a pixel grid may not fit the gradient profile exactly. If each gradient profile is transformed, many subpixel gradient values will be generated, and it will be difficult to determine the gradient magnitude from the subpixel gradient. In contrast, the gradient field transformation for each pixel is more straightforward and easy to implement. Finally, the estimated HR gradient is taken as a gradient constraint in the subsequent image reconstruction to produce a better HR image.

C. Adaptive Norm Reconstruction Model

In the proposed MFSR approach, the fidelity term guarantees that the solution accords with the degradation process of remote sensing images, while the regularization term enforces the desired property of the output HR image. In this section, we introduce the proposed MFSR algorithm in detail. In the traditional MFSR methods, the initial HR estimation is obtained by performing bicubic or bilinear interpolation on the reference LR image. However, if the quality of the reference LR image is poor, the quality of the initial HR estimation will be unsatisfactory. Based on MAP estimation, we exploit NUI initialization to obtain the initial HR estimation for the image SR.

1) Adaptive Norm Data Fidelity: The fidelity term is mainly responsible for dealing with motion outliers, image noise, and model errors. For the complicated types of noise and model error in remote sensing image SR, both the $L1$ norm and $L2$ norm have their advantages and drawbacks. To deal with different types of noise and blur, we employ adaptive p-norm data fidelity [41] to select the error norm adaptively, instead

of using fixed $L1$ and $L2$ norms. The adaptive p-norm data fidelity is formulated as follows:

$$D(y, z) = \sum_{k}^{K} \|y_k - \text{DBM}_k z\|_p^p. \qquad (14)$$

The norm value of p reflects the proportional relationship of the different distributions in the mixed error model. According to the distribution models of different types of noise, the optimal p-norm is determined adaptively, based on a generalized likelihood ratio test (GLRT), which is formulated as follows:

$$p = \begin{cases} 1, & 0 < \gamma \leq 0.112 \\ a\tan(b\gamma + c) + d, & 0.112 < \gamma \leq 0.798 \\ 2, & 0.798 < \gamma < 1 \end{cases} \qquad (15)$$

where $\gamma = \sigma_L/\sigma_G$ is the ratio between the Gaussian and Laplacian distributions. σ_G and σ_L are the variance of the Gaussian noise and impulse noise distribution, respectively. As described in [41], the model parameters a–d are obtained from a large number of experimental statistics.

For each remote sensing image, the noise vector is computed according to the image observation model. The two variances σ_G and σ_L of the noise can then be computed to obtain the corresponding γ value. According to (15), we can adaptively determine the optimal p-norm value for each test image. For the p-norm model, the Lagrange approximation method can be flexibly applied to solve it.

2) Reconstruction Model With Combined Prior: Given that SR is regarded as an ill-posed minimization problem, it is essential to implement a specific regularization to constrain the solution space. In the proposed approach, we construct the reconstruction framework by coupling the NLTV prior and the LGR prior, which, respectively, consist of the reconstruction constraint in the image domain and the gradient domain.

Benefiting from the development of NLSS methods, we integrate the nonlocal property into the algorithm framework by considering the spatial similarity of the gradient patches. The nonlocal similarity information can be obtained by clustering the similar patterns throughout a remote sensing image, from which a nonlocal regularization term is formulated to enhance the texture details of the reconstructed image. Mathematically, we let z_i denote the image patch at location i, and the nonlocal gradient operator at the pixel location is defined as follows:

$$\nabla_{\text{NL}} z = \sum_{i \in \Omega(z)} \sum_{j \in \Omega(i)} \left((z_j - z_i)\sqrt{w_{i,j}}\right) \qquad (16)$$

where $\Omega(z)$ is the index set for all the pixels of z, and $\Omega(i)$ denotes the index set for similar patches of z_i. $w_{i,j}$ represents the similarity weight between patches z_i and z_j, which is defined by

$$w_{i,j} = \exp\left(-\frac{G(z_j - z_i)}{2\eta^2}\right) \qquad (17)$$

where η is a parameter set to the standard deviation of the image noise. $G(\cdot)$ is the Gaussian filter function to reduce the influence of noise on the weight. To effectively exploit the nonlocal redundancy within the remote sensing imagery, the NLTV prior is formulated as follows:

$$P_{\text{int}}(\nabla z) = \|\nabla_{\text{NL}} z\|_1^1. \qquad (18)$$

In the variational embedded learning MFSR method, the NLTV regularization term is exploited to enhance the edge information and suppress artifacts in the smooth regions of the super-resolved image. The other prior is the LGR prior, derived from the GPT, which considers the distribution of the image gradients along local image structures. The gradient constraint requires that the gradient field of the recovered HR image should be consistent with the transformed HR gradient field. Thus, the LGR prior can be formulated by

$$P_{\text{ext}}(\nabla z) = \|\nabla z - G\|_2^2. \qquad (19)$$

Using this LGR constraint, the gradient profile of the reconstructed HR image is encouraged to have the desired statistics learned from the training images. By coupling the nonlocal reconstruction constraint and local gradient domain constraint in the MAP framework, the performance of the gradient-guided MFSR method is further improved. Inserting (18) and (19) into (1), the objective function of the proposed gradient-guided MFSR method can be rewritten as

$$\hat{z} = \arg\min_z \left\{ \sum_k^K \|y_k - \text{DBM}_k z\|_p^p + \alpha\|\nabla z - G\|_2^2 + \beta\|\nabla_{\text{NL}} z\|_1^1 \right\}. \qquad (20)$$

In the MAP-based MFSR algorithm, the adaptive norm fidelity term is exploited to ensure robustness in dealing with different noise types. The combined prior is used to enhance the edge information of the image and suppress the artifacts in the smooth regions. Typically, the geometric registration and the blur can be estimated from the input data and used with the SR model to reconstruct the super-resolved image. In this work, the warping matrix M and blur matrix B are computed with the optical flow approach [42] and the blur kernel estimation method [43], respectively. Finally, the combined SR problem is solved via the ADMM algorithm.

3) Energy Function Optimization: Since p is an arbitrary value in an interval, the conventional linear optimization methods cannot be employed directly. In the proposed approach, the iteratively reweighted norm (IRN) method [44] is employed for the linearization of the fidelity term. The main idea of the IRN algorithm is to transform the indefinite norm problem into an equivalent $L2$ norm problem by selecting a suitable weight coefficient matrix, so that the objective function can be optimized by a conventional solution method. In order to simplify the problem, DBM_k can be regarded as a system matrix A_k. Introducing the idea into the energy function in (20), at iteration $n + 1$, the solution $\hat{z}^{n+1}$ can be obtained by

$$\hat{z}^{n+1} = \arg\min_z \left\{ \sum_k^K \left\|W_{z^n}^{1/2}(y_k - A_k z)\right\|_2^2 + \alpha\|\nabla z - G\|_2^2 + \beta\|\nabla_{\text{NL}} z\|_1^1 \right\} \qquad (21)$$

where $W_{z^n}^{1/2}$ is the weight coefficient matrix, which is defined as follows:

$$W_{z^n}^{1/2} = \text{diag}\left(\varphi_\varepsilon \left(y_k - A_k z^n\right)\right) \qquad (22)$$

with

$$\varphi_\varepsilon(x) = \begin{cases} |x|^{p-2}, & \text{if } |x| > \varepsilon \\ \varepsilon^{p-2}, & \text{else} \end{cases} \qquad (23)$$

where ε is a small positive number to guarantee global convergence, which is fixed as 10^{-5} in the proposed approach.

Based on the variable splitting minimization approach, the ADMM algorithm [45] is adopted to solve the hybrid $L1$- and $L2$-regularized minimization problem in (21). The auxiliary variable b representing z is introduced to decouple the $L1$ and $L2$ parts. The objective function can be rewritten as

$$\hat{z}^{n+1} = \arg\min_z \left\{ \begin{array}{c} \sum_k^K \left\| W_{z^n}^{1/2}(y_k - A_k z) \right\|_2^2 \\ + \alpha\|\nabla z - G\|_2^2 + \beta\|\nabla_{\text{NL}} b\|_1^1 \end{array} \right\}$$
$$\text{s.t. } b = z. \qquad (24)$$

By transforming (24) to generate an unconstrained problem with the augmented Lagrangian algorithm, it can be rewritten as follows:

$$\hat{z}^{n+1} = \arg\min_z \left\{ \begin{array}{c} \sum_k^K \left\| W_{z^n}^{1/2}(y_k - A_k z) \right\|_2^2 + \alpha\|\nabla z - G\|_2^2 \\ + \tau u^T(z-b) + \dfrac{\tau}{2}\|z - b\|_2^2 + \beta\|\nabla_{\text{NL}} b\|_1^1 \end{array} \right\}$$
$$(25)$$

where τ is a penalty parameter and u is a Lagrangian multiplier. To tackle the optimal minimization problem in (25), an iterative algorithm is designed to alternately minimize z, b, and u independently, by fixing the other variables. The subproblem optimization can be rewritten as follows:

$$z^{n+1} = \arg\min_z \left\{ \begin{array}{c} \sum_k^K \left\| W_{z^n}^{1/2}(y_k - A_k z) \right\|_2^2 + \alpha\|\nabla z - G\|_2^2 \\ + \tau (u^n)^T(z - b^n) + \dfrac{\tau}{2}\|z - b^n\|_2^2 \end{array} \right\}$$
$$(26)$$

$$b^{n+1} = \arg\min_z \left\{ \beta\|\nabla_{\text{NL}} b\|_1^1 + \tau u^T(z-b) + \frac{\tau}{2}\|z - b\|_2^2 \right\}$$
$$(27)$$

$$u^{n+1} = u^n + z^{n+1} - b^{n+1}. \qquad (28)$$

For the z subproblem in (26), a preconditioned conjugate gradient method is used to quickly find the optimal value. The solution of b in (27) is obtained using a simple shrinkage operation based on [46]. According to (28), u can then be determined directly from the given variables z and b. By iteratively optimizing z, b, and u, the optimal solution of (21) can be obtained in a fast and stable manner.

III. EXPERIMENTAL RESULTS

The performance of the proposed gradient-guided MFSR method was evaluated in several experiments. Initially, the proposed method was applied on the test datasets and compared with some of the state-of-the-art SR approaches. At the same time, simulation experiments on noisy images were further conducted to verify the robustness of the proposed method to noise. In addition, the effectiveness of the two incorporated regularization terms was validated using different extended versions of the proposed method. The detailed steps are presented in the following sections.

A. Experimental Setup

1) Degradation Models: In the experiments, synthetic data with the ground truth were used to quantitatively analyze the proposed method, as well as make a fair comparison with the other methods at a scale factor of $4\times$. For each HR image from the test sets, we generated a set of $K = 16$ images with different subpixel shifts applied before further degradation. The 16 HR images were then blurred by a 3×3 isotropic Gaussian kernel with standard deviation 1. Finally, the row and column of the blurred image were downsampled by a factor of 4. Additionally, in the robustness analysis described in Section III-C, the LR images were further contaminated by additive white Gaussian noise with a variance of 0.005.

In the MFSR method, the central frame of the LR sequence was chosen as the reference frame. The regularization parameters α and β were determined empirically based on numerous experiments to produce the best performance. Since minimizing the objective function by the preconditioned conjugate gradient method usually converges within 30 iterations, the maximum iteration number was set to 30 in the proposed method.

2) Dataset and Training Settings: We chose the University of California (UC) Merced dataset [47], which is a collection of remote sensing images with a relatively high spatial resolution (0.3 m/pixel), to evaluate the proposed method. As shown in Fig. 5, the UC Merced dataset contains 21 different scene categories, with each class comprising 100 images of 256×256 in the red, green and blue (RGB) space. The UC Merced dataset was used as the experimental data and divided into two disjoint parts. The first part contained images in the ten categories of airplane, baseball diamond, buildings, dense residential, freeway, harbor, intersection, mobile home park, storage tanks, and tennis court. We first selected 900 images from the ten categories, made up of 90 images for each class, to build the training dataset. The remaining 100 images formed test set I. The second part included images in the remaining 11 classes of agricultural, beach, chaparral, forest, golf course, medium residential, overpass, parking lot, river, runway, and sparse residential, which made up test set II. The reason for this division was to measure the generalization ability of the proposed method. The images in test set I were similar to those in the training set, while the images in test set II were quite different from those in the training set. Therefore, the performance under these test sets could better validate the generalization ability of the proposed method.

In the DRAN training phase, we exploited the above degradation models without subpixel shifts to simulate the LR input images. In total, 10% of the training samples were randomly selected as the validation set for the model selection, and the
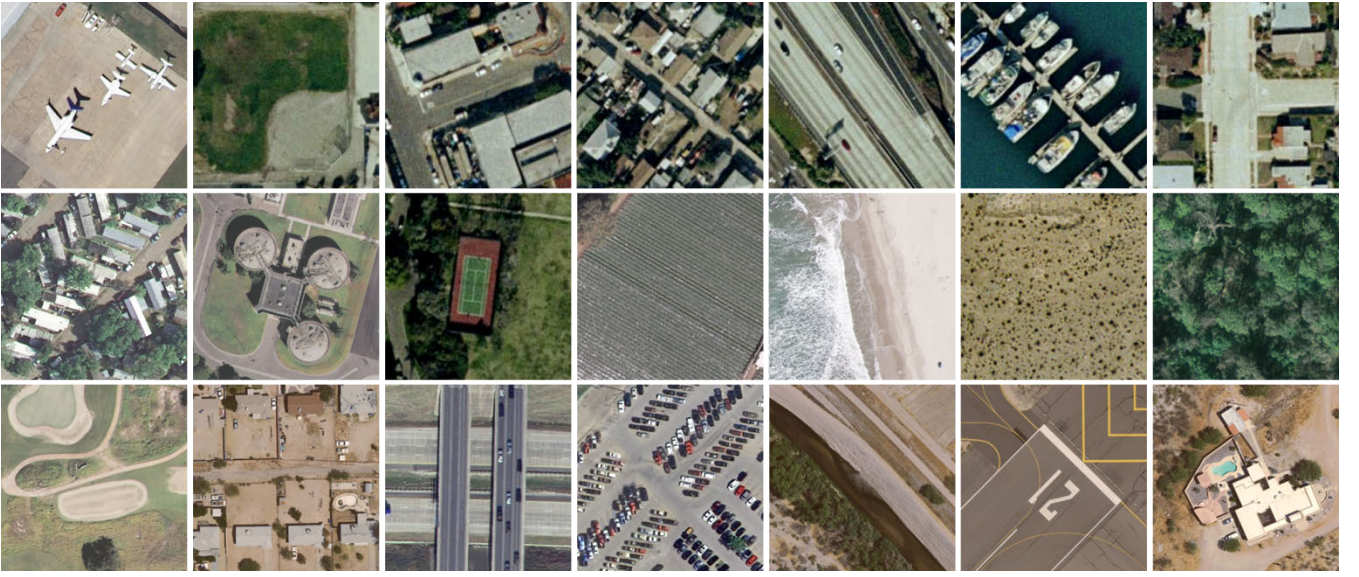
Fig. 5.   Examples of the 21 different scene categories in the UC Merced dataset.

other 90% were used for the training. The factor of the residual scaling was set to 0.2. The weights of the network were initialized based on the method in [40]. In each mini-batch, 32 degraded LR images with a patch size of 48 × 48 were provided as inputs for the model, and the corresponding HR image served as the ground truth for calculating the loss. The models were optimized using the ADAM optimizer [48] with momentum parameters $\beta1 = 0.9$, $\beta2 = 0.999$, and $\varepsilon = 10^{-8}$. The initial learning rate was set to 10^{-4} and then decreased by half every ten epochs. A total of 100 epochs were used for training the models since more epochs did not bring further improvements. All the experiments were implemented using the Caffe framework and MATLAB R2022a on a Nvidia RTX GPU.

3) Comparison Baselines: To comprehensively compare the proposed method with the state-of-the-art SR methods and classical SR methods, the comparison baseline methods were made up of bicubic interpolation, two deep learning-based VSR methods (the recurrent back-projection network (RBPN) [49] and BasicVSR [50]), a spatially weighted bilateral total variation method (SWBTV) [51], and a joint prior-based method (L0RIG) [52]. For the two deep learning-based approaches, we used the training data to properly retrain these models with the specific training settings in their corresponding articles. Since the human visual system is more sensitive to the luminance component than the chrominance components, we converted the color RGB frames to the YCbCr color space and reconstructed only the luminance component with the proposed algorithm. Bicubic interpolation was used for the other components.

4) Evaluation Metrics: Image enhancement or visual quality improvement can be subjective because the perception of better image quality can vary from person to person. To assess the image quality of the SR reconstructed results, two classical evaluation criteria—the peak signal-to-noise ratio (PSNR/dB) and the structural similarity index measure

(SSIM)—were chosen to measure the performance of the different SR methods [53]. The higher the quantitative measure, the better the quality of the reconstructed image. Since no ground-truth HR image was available for the experiments on the real sequences, we introduced no-reference image evaluation metrics—the natural image quality evaluator (NIQE) [54] and the perception-based image quality evaluator (PIQE) [55]—to further evaluate the quality of the real image SR results. Smaller values of NIQE and PIQE indicate better SR results. In addition, to give a fair comparison, we cropped the image boundary pixels before evaluation. The best results in Table I are highlighted in red/bold, while the second-best results are in green/underlined.

B. Synthetic Data Experiments

In this section, both the quantitative and qualitative simulation experiment results are given to confirm the performance of the proposed method. Table I provides a quantitative performance comparison of the various methods in 4× enlargement for test sets I and II. The visual comparisons of some of the SR results are provided in Figs. 6–8 to perceptually compare the performance of the proposed method, and the ground-truth HR images are given for reference.

A comparison of the average PSNR and SSIM results for each scene class is provided in Table I. It can be seen that the proposed method achieves the highest PSNR and SSIM values. As is well known, with the significant progress of deep learning, the deep learning-based VSR approaches have greatly improved the performance on synthetic LR frames. It can be seen from the result that except for the agricultural category on ×4 enlargement, the proposed method outperforms the compared state-of-the-art BasicVSR methods. Compared to the state-of-the-art VSR method (BasicVSR), the proposed method achieves a better performance and is slightly superior on average for both test sets I and II, which shows the generalization ability of the model. Based on the overall results, these two methods are then followed

TABLE I
AVERAGE PSNR (dB) AND SSIM RESULTS OF EACH CLASS IN 4× SR

Data	Bicubic	Learning-based		Model-based		
		RBPN [49]	BasicVSR [50]	SWBTV [51]	L0RIG [52]	Proposed
Airplane	25.815/0.751	28.439/0.832	28.803/0.857	28.014/0.822	28.737/0.840	**28.981/0.872**
Baseball diamond	30.358/0.778	32.397/0.829	32.775/0.844	32.028/0.821	32.605/0.836	**32.958/0.861**
Buildings	21.461/0.661	24.115/0.795	24.484/0.814	3.703/0.782	24.337/0.807	**24.632/0.827**
Dense residential	3.355/0.672	26.002/0.795	26.308/0.818	25.575/0.784	26.179/0.802	**26.475/0.834**
Freeway	26.207/0.675	28.355/0.777	28.732/0.798	28.069/0.773	28.593/0.789	**28.965/0.812**
Harbor	17.379/0.691	19.927/0.854	20.108/0.870	19.835/0.848	20.097/0.867	**20.179/0.883**
Intersection	24.343/0.692	26.178/0.790	26.601/0.801	25.966/0.781	26.466/0.795	**26.737/0.816**
Mobile home park	21.841/0.663	24.869/0.794	25.135/0.805	24.407/0.783	24.978/0.799	**25.292/0.822**
Storage tanks	23.549/0.698	25.685/0.795	26.043/0.815	25.388/0.790	25.955/0.804	**26.265/0.825**
Tennis court	28.073/0.780	30.420/0.857	30.801/0.881	30.114/0.852	30.588/0.865	**30.998/0.894**
Average test I	24.238/0.706	26.639/0.812	26.979/0.830	26.310/0.804	26.854/0.820	**27.148/0.845**
Agricultural	25.169/0.433	25.993/0.493	**26.210/0.510**	25.806/0.485	26.055/0.495	26.196/0.509
Beach	33.145/0.839	34.436/0.876	34.797/0.887	34.205/0.870	34.668/0.879	**34.964/0.896**
Chaparral	24.082/0.646	25.632/0.742	26.053/**0.768**	25.292/0.735	25.901/0.751	26.102/0.763
Forest	26.267/0.603	27.177/0.679	27.534/0.694	26.946/0.672	27.421/0.687	**27.703/0.705**
Golf course	30.718/0.769	32.380/0.815	32.745/0.827	32.012/0.807	32.556/0.823	**32.875/0.835**
Medium residential	23.462/0.653	25.909/0.765	26.204/0.781	25.500/0.755	26.081/0.771	**26.397/0.793**
Overpass	23.124/0.642	25.716/0.759	26.260/0.780	25.462/0.746	26.009/0.765	**26.346/0.791**
Parking lot	19.168/0.607	20.782/0.737	21.149/0.779	20.403/0.727	20.962/0.762	**21.320/0.793**
River	26.473/0.658	27.499/0.726	27.841/0.741	27.287/0.714	27.745/0.731	**27.968/0.755**
Runway	26.556/0.719	29.896/0.806	30.515/0.835	29.735/0.798	30.380/0.819	**30.733/0.846**
Sparse residential	26.158/0.671	28.018/0.749	28.389/0.769	27.766/0.742	28.285/0.755	**28.552/0.782**
Average test II	25.847/0.658	27.585/0.741	27.972/0.761	27.310/0.732	27.824/0.749	**28.114/0.771**

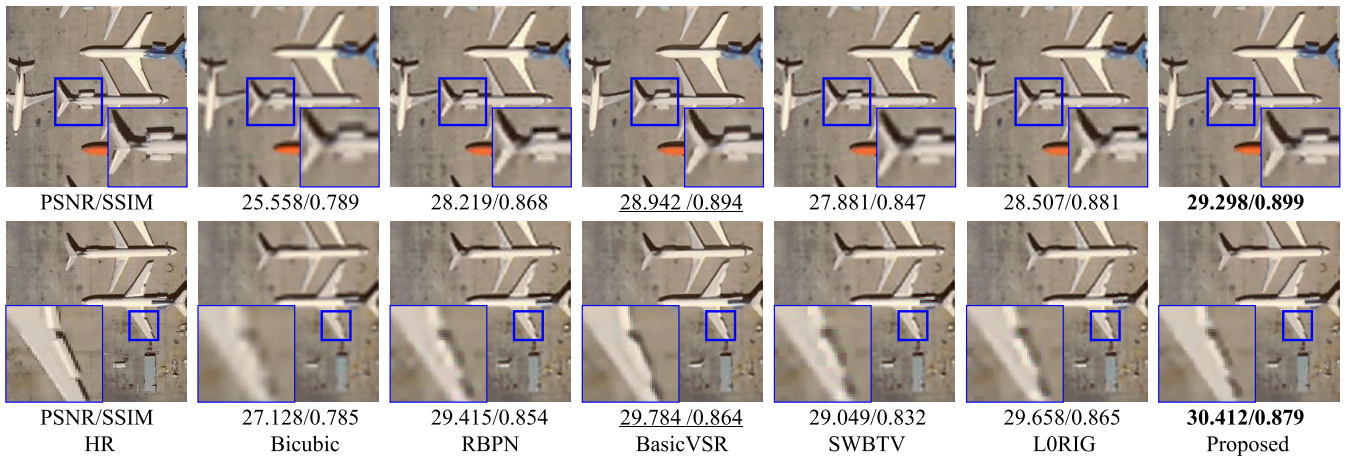| PSNR/SSIM	25.558/0.789	28.219/0.868	28.942 /0.894	27.881/0.847	28.507/0.881	**29.298/0.899**
PSNR/SSIM	27.128/0.785	29.415/0.854	29.784 /0.864	29.049/0.832	29.658/0.865	**30.412/0.879**
HR	Bicubic	RBPN	BasicVSR	SWBTV	L0RIG	Proposed

Fig. 6.   Visual results for "airplane90" and "airplane91" in test set I using the various methods in 4× SR.

by L0RIG, RBPN, and SWBTV. The model-based methods (L0RIG and SWBTV) achieve superior results in the remote sensing image SR, due to the effective priors as well as the global reconstruction constraint. It should be noted that some classes, such as baseball diamond, and golf course, present a relatively high PSNR due to the fact that the images associated with these classes are characterized by a greater degree of smoothness than those of the other classes. Among these categories, images in the harbor category have the lowest PSNR of 20.179 dB with the proposed method, which is still better than the other methods. Nevertheless, in test set II, the evaluation metrics of agricultural and chaparral categories by the BasicVSR method are better than that of the proposed method. Since the overall texture of these two classes presents
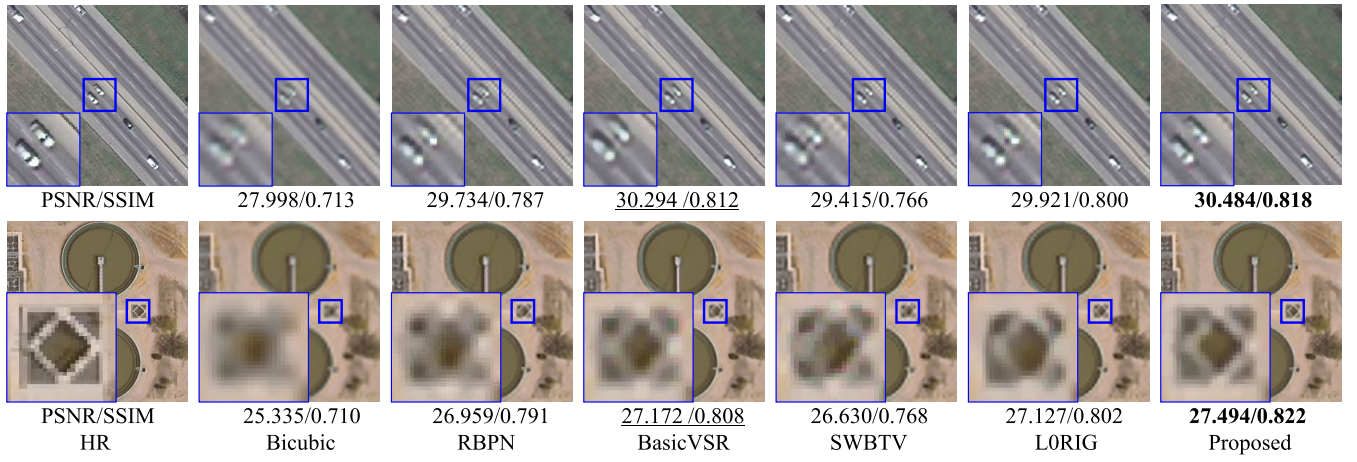
PSNR/SSIM	27.998/0.713	29.734/0.787	<u>30.294 /0.812</u>	29.415/0.766	29.921/0.800	**30.484/0.818**
PSNR/SSIM	25.335/0.710	26.959/0.791	<u>27.172 /0.808</u>	26.630/0.768	27.127/0.802	**27.494/0.822**
HR	Bicubic	RBPN	BasicVSR	SWBTV	L0RIG	Proposed

Fig. 7. Visual results for "freeway23" and "storagetanks98" in test set I using the various methods in 4× SR.

PSNR/SSIM	21.703/0.766	25.240/0.860	<u>26.225 /0.886</u>	24.871/0.841	25.591/0.874	**26.603/0.891**
PSNR/SSIM	23.982/0.650	26.122/0.753	<u>26.538 /0.764</u>	25.770/0.728	26.345/0.864	**26.888/0.792**
HR	Bicubic	RBPN	BasicVSR	SWBTV	L0RIG	Proposed
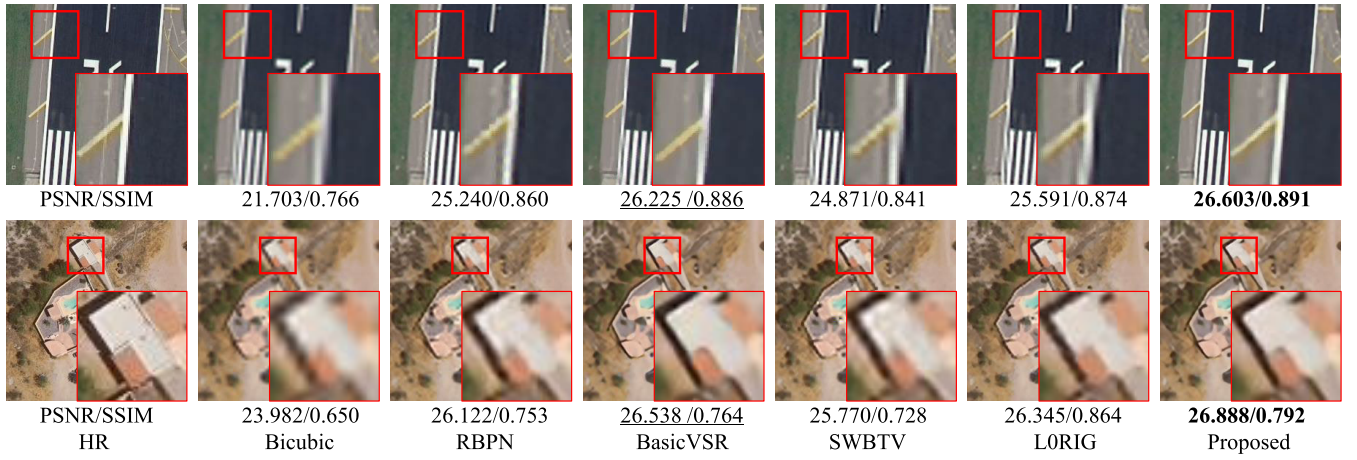
Fig. 8. Visual results for "runway50" and "sparseresidential92" in test set II using the various methods in 4× SR.

a fine and irregular distribution, it is not conducive to the operation of gradient profile optimization and similar patches searching in the regularizing term of the proposed model. The experimental results indicate that in most cases, since texture details and sharp edges are essential in all ground features, the proposed method with the combined constraint achieves notable SR performance improvements for the remote sensing images in each class.

To compare the visual qualities, some of the SR results for the test set with a scale factor of 4× are presented in Figs. 6–8. For a better comparison from a subjective perspective, the rectangles show the zoomed regions of the corresponding images to compare the qualitative performance of the different methods. The visual effect basically agrees with the objective results. The experimental results show that the HR remote sensing images recovered by the proposed method exhibit sharper edges and clearer contours, compared with the state-of-the-art SR methods. This is consistent with the proposed local and nonlocal regularization terms, which are capable of providing the HR gradient information and nonlocal similar textures for the SR reconstruction model.

Specifically, Fig. 6 gives the reconstructed HR images for the "airplane90" and "airplane91" images generated by the

various methods. The reconstructed images of the proposed method have a superior visual quality, as the intricate details in the original images are well preserved. In addition, the close-up images illustrate that the proposed method effectively reconstructs the clear wing edge, whereas the other algorithms suffer from varying degrees of blurring and displeasing artifacts. From the visual comparison in Fig. 7, the proposed algorithm reconstructs more accurate results and preserves finer texture details in the super-resolved image, such as the windscreen of the car. In comparison to the other algorithms, the proposed method demonstrates a superior visual performance, effectively alleviating the blurring artifacts and recovering more details in a manner that is more faithful to the ground truth. Given that the images in test set II are quite distinct from those in the training set, this dataset can be used to verify the generalization ability of the proposed method.

Fig. 8 presents the recovered images in 4× enlargement for the "runway50" and "sparseresidential92" images in test set II obtained by the various methods. It can be observed that the proposed method is superior in its ability to generate sharper edges, as evidenced by the traffic index line on the right side of the road. In contrast, the reconstruction artifacts and unnatural structures are still noticeable in the outcomes

TABLE II
AVERAGE PSNR (dB) AND SSIM RESULTS FOR A NOISY IMAGE IN 4× SR

Metric	Bicubic	RBPN [49]	BasicVSR [50]	SWBTV [51]	L0RIG [52]	Proposed
PSNR	24.028	24.893	25.257	25.029	25.426	**25.934**
SSIM	0.686	0.749	0.776	0.763	0.792	**0.807**

| HR PSNR/SSIM | Bicubic 27.022/0.770 | RBPN 29.647/0.863 | BasicVSR 30.081/0.882 | SWBTV 29.911/0.875 | L0RIG 30.226 /0.894 | Proposed **30.407/0.901** |
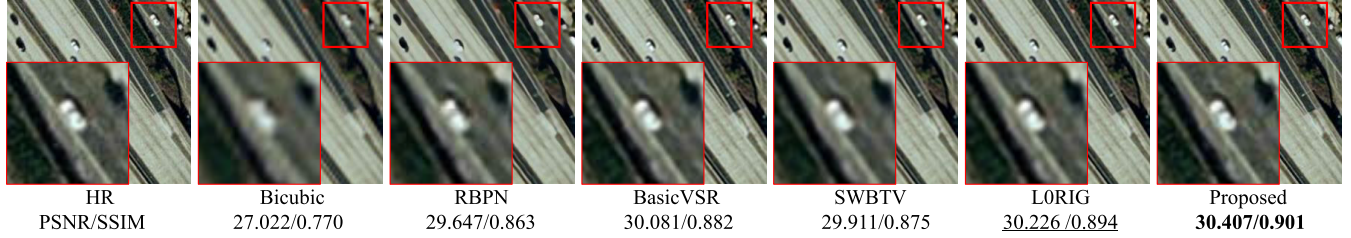
Fig. 9.   Visual results for "freeway00" in test set I under Gaussian noise in 4× SR.

of the aforementioned comparison methods, particularly in the edge regions. In fact, the performance of the proposed method is in consonance with our initial hypothesis: the NLTV regularization term is apt in preserving edge smoothness and suppressing visual artifacts, while the LGR constraint is beneficial in recovering sharp edges and fine structures.

Overall, the proposed method achieves the best performance among all the methods in terms of PSNR/SSIM values and visual quality in remote sensing image SR. The average PSNR/SSIM gains on the whole test set over the RBPN, BasicVSR, SWBTV, and L0RIG methods are 0.519 dB/0.032, 0.155 dB/0.012, 0.821 dB/0.040, and 0.292 dB/0.023, respectively. The experimental results demonstrate that the proposed method, which combines the constraints of LGR and NLTV, allows for the reliable recovery of high-frequency details while preserving strong edges and contours with minimal artifacts.

C. Robustness Analysis

Since image noise makes the SR problem more challenging, we analyzed and justified the robustness of the proposed method with regard to noise. Extra additive white Gaussian noise of variance 0.005 was added to corrupt the LR images in test set I. The other parameters were the same as in the previous experiments. Due to the limited space, we only report the average PSNR/SSIM scores of each method in Table II. For the visual quality comparison, the SR results for the noisy "freeway00" image are presented in Fig. 9. Details of the output reconstructed images are given for better illustration.

In terms of the objective performance, the proposed method outperforms the others in terms of PSNR and SSIM results in the quantitative evaluation. Generally speaking, learning-based approaches are susceptible to noise, resulting in a significant drop in their performance on noisy images. Specifically, in the experiment with a noise variance of 0.005, the average PSNR result of the proposed method is 25.934 dB, which is better than that of RBPN, BasicVSR, SWBTV, and L0RIG, by 1.041, 0.677, 0.905, and 0.508 dB, respectively. Fig. 9 illustrates the comparative performances of the various methods in an enlarged area within the red boxes. It can be observed that the comparison learning-based methods fail to effectively

suppress the noise artifacts. While RBPN also reproduces fine details well, it introduces unpleasant noise artifacts. The SWBTV method based on bilateral total variation partially suppresses image noise and preserves edges but exhibits visual artifacts around the edge regions. As the partial enlargement shows, L0RIG demonstrates an optimal balance between noise removal and edge preservation but fails to recover the lost fine details. The BasicVSR algorithm produces undesirable edge artifacts, including the appearance of artificial edges on flat surfaces and an inadequate suppression of image noise in the textured areas. In contrast, the proposed method demonstrates a superior performance, exhibiting clear high-frequency details and fewer ringing artifacts, as shown in Fig. 9. It is noteworthy that the distorted content, such as the windscreen of the car, can be finely restored with the proposed approach. Compared to the other methods evaluated on noisy images, the proposed method performs favorably due to its learned gradient constraint, which enhances edge sharpness while suppressing jaggy artifacts along high-frequency structures.

In conclusion, based on both the qualitative and quantitative analyses, the results consistently demonstrate that the proposed method exhibits a superior SR performance in accurately recovering high-frequency information that closely resembles the ground-truth image. By leveraging a combined prior, the proposed approach effectively enhances edge sharpness and preserves the intricate texture details of the recovered image. Consequently, it significantly outperforms the comparison methods by a substantial margin in terms of PSNR, thereby further validating its superiority in reconstructing textures and enhancing edges. Moreover, the comparative evaluation against other state-of-the-art SR approaches confirms the robustness and effectiveness of the proposed method in terms of noise corruption.

D. Real-Data Experiments

In this section, we describe how we employed real sequences of remote sensing images to further evaluate the robustness of the proposed method. The three sequences of real LR remote sensing images were extracted from the Jilin No. 1 satellite video dataset [1], each consisting of 16 consecutive
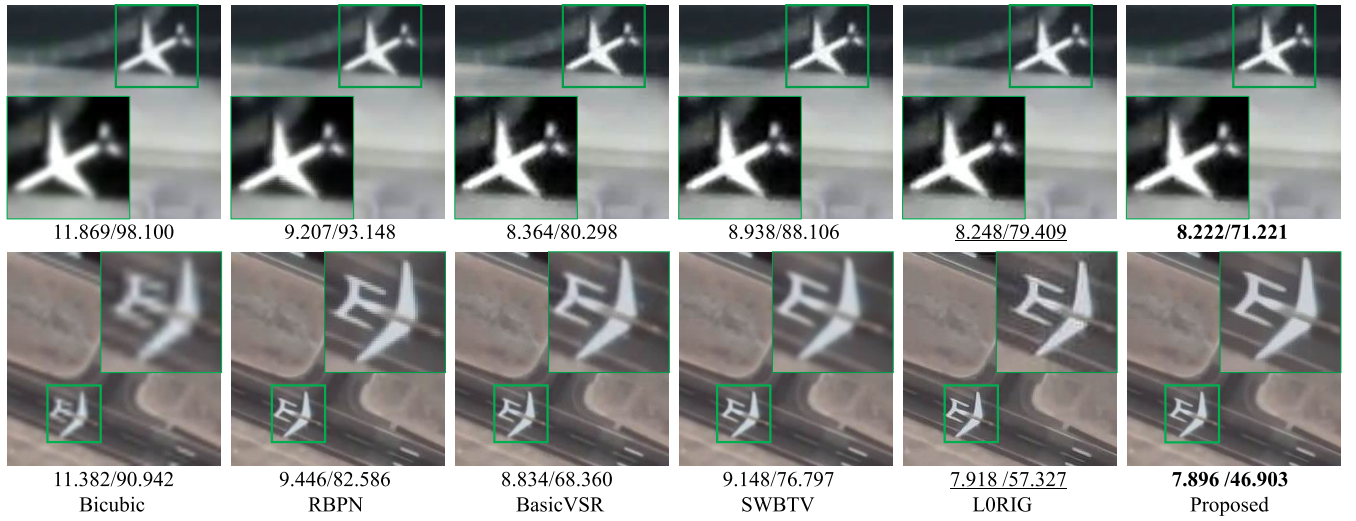
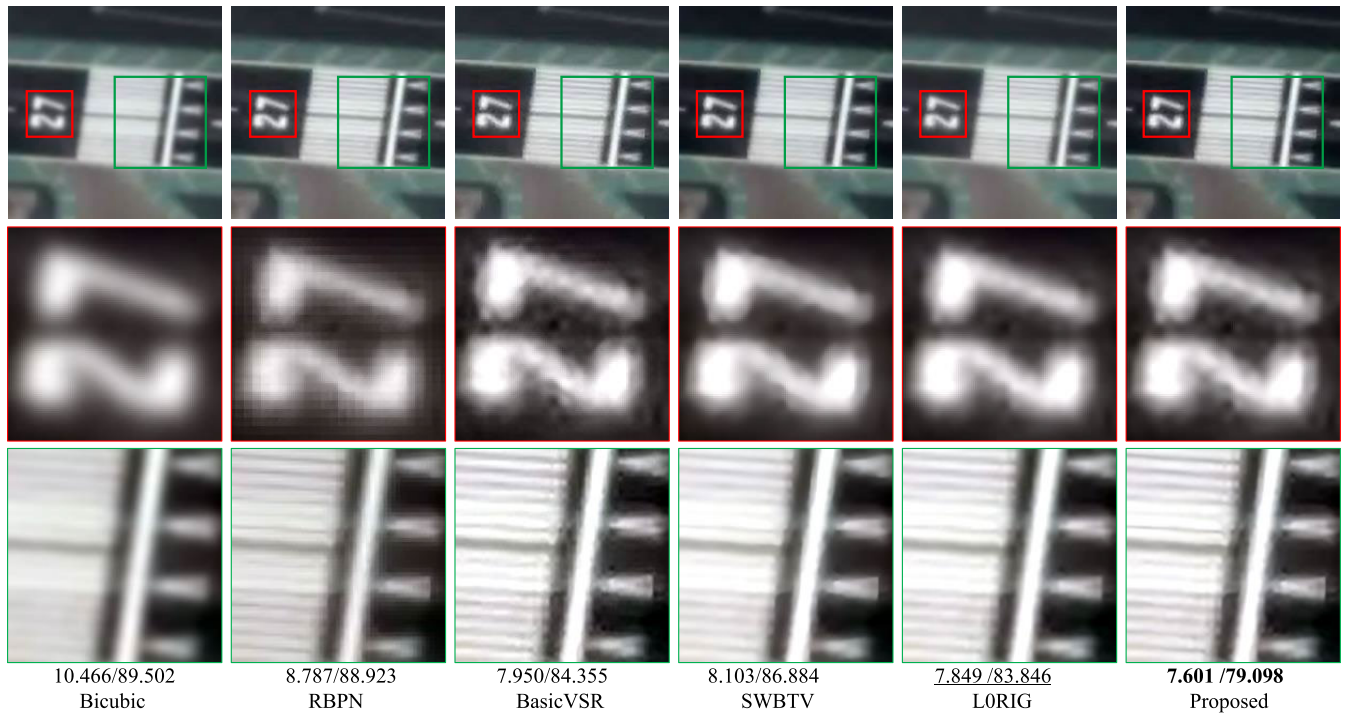| 11.869/98.100	9.207/93.148	8.364/80.298	8.938/88.106	8.248/79.409	**8.222/71.221**
11.382/90.942	9.446/82.586	8.834/68.360	9.148/76.797	7.918 /57.327	**7.896 /46.903**
Bicubic	RBPN	BasicVSR	SWBTV	L0RIG	Proposed

Fig. 10. Visual results for the "airplane" scene in the Jilin-1 satellite imagery obtained by the different methods in 4× SR, where the numbers below refer to the NIQE and PIQE values of the corresponding results.

10.466/89.502	8.787/88.923	7.950/84.355	8.103/86.884	7.849 /83.846	**7.601 /79.098**
Bicubic	RBPN	BasicVSR	SWBTV	L0RIG	Proposed

Fig. 11. Visual results for the "runway" scene in the Jilin-1 satellite imagery obtained by the different methods in 4× SR, where the numbers below refer to the NIQE and PIQE values of the corresponding results.

frames with a fixed size of 256 × 256. The resolution of the Jilin-1 imagery (1.12 m/pixel) is within the resolution range of our test set, which is 1.2 m/pixel (0.3 × 4) for a scale factor of 4. Since no ground-truth HR image was available for the experiments on the real sequences, the no-reference image evaluation metrics of NIQE [54] and PIQE [55] were employed to further evaluate the quantitative quality of the real image sequence SR results. Visual comparisons of the results for the Jilin-1 satellite imagery obtained by the different methods are shown in Figs. 10 and 11. In each image, we have marked a local region of interest and displayed its zoomed version separately alongside the image.

The experimental results on the real image sequences show that the proposed method yields a superior performance in both the objective metrics and visual quality. For a real-world image, the downsampling kernel is typically unknown and complicated, which significantly affects the performance of the SR methods. Nevertheless, the proposed method can produce visually pleasant images and effectively suppress the errors caused by noise, registration, and poor estimation of unknown point spread function (PSF) kernels.

In the case of the "airplane" scene depicted in Fig. 10, the proposed method generates superior results with a diminished prevalence of jagged and ringing artifacts, particularly
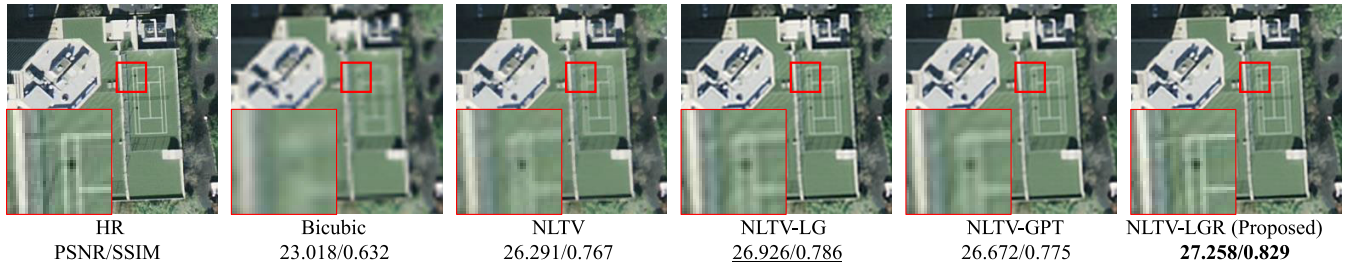
HR	Bicubic	NLTV	NLTV-LG	NLTV-GPT	NLTV-LGR (Proposed)
PSNR/SSIM	23.018/0.632	26.291/0.767	<u>26.926/0.786</u>	26.672/0.775	**27.258/0.829**

Fig. 12.   Visual results for "tenniscort28" with different extended versions of the proposed method in 4× SR.

in instances where motion estimation errors are minimal. The recurrent neural network and explicit motion estimation enable the RBPN method to fuse the multiframe information, which shows its great ability for satellite VSR, as well as the BasicVSR method. The learning-based method shows enhanced edge sharpness for manufactured objects but also generates false textures and undesired visual artifacts in the SR results. The model-based methods (L0RIG and SWBTV) achieve superior results in the real satellite VSR, due to the effective priors as well as the global reconstruction constraint. While SWBTV is effective at noise suppression, it is obviously inferior to the proposed method in terms of detail preservation and sharp edge recovery. For the "runway" scene in Fig. 11, due to the serious loss of object information and the lack of any reasonable reference objects, all the comparison SR methods show limitations for the small static objects. From the visual results, it can be seen that better edge regions and texture regions are reconstructed by the proposed method, compared with the state-of-the-art SR methods. In contrast, BasicVSR tends to generate more artifacts at the edges, and the result of RBPN suffers from visible ghosting artifacts and is seriously affected by the blur effects.

In summary, the proposed method can improve the reconstruction results for Jilin-1 satellite imagery greatly, in both the visual effects and quantitative results. The experimental results demonstrate that the proposed method is superior to many of the learning-based and model-based MFSR methods, especially in the edge and texture regions. These results also indicate that the proposed model is more robust than the other methods.

E. Effectiveness of the Regularization Terms in the Proposed Method

The proposed method incorporates two regularization terms: the LGR and NLTV priors. Moreover, the LGR prior uses the learned gradient and HR profile transformation to, respectively, address the issues of edge ambiguity and noise sensitivity. To validate the effectiveness of both the local and nonlocal regularization terms in the proposed framework, ablation experiments were performed on the images in test set I, with three extended versions of the proposed method. The three extended versions were as follows: NLTV (only the NLTV prior was used), NLTV-LG (the gradient regularization term was constructed with the DRAN), and NLTV-GPT (the gradient regularization term was constructed with the GPT). The proposed framework—NLTV-LGR—incorporates

TABLE III
AVERAGE PSNR (dB) AND SSIM RESULTS OF THE DIFFERENT EXTENDED VERSIONS IN 4× SR

Method	Bicubic	NLTV	NLTV-LG	NLTV-GPT	NLTV-LGR
NLTV	×	√	√	√	√
LG	×	×	√	×	√
GPT	×	×	×	√	√
PSNR	24.238	26.706	<u>26.954</u>	26.881	**27.148**
SSIM	0.706	0.821	<u>0.831</u>	0.825	**0.845**

both the NLTV and LGR priors, where the LGR prior is constructed with both an LG and GPT. The bicubic interpolation method was taken as a benchmark for comparison. The average PSNR and SSIM scores for the different extended versions are listed in Table III. It can be observed that the complete version (NLTV-LGR) achieves the best performance, indicating that the combination of NLTV and LGR priors is significantly beneficial in enhancing the quality of super-resolved images. Moreover, for the perceptual quality comparison, the super-resolved results of the different extended versions for the "tenniscort28" test image are shown in Fig. 12.

1) Effectiveness of the LG and GPT Schemes: In the proposed method, the estimated HR gradient field is taken as the prior term to constrain the SR reconstruction process. The gradient regularization term can be constructed by either an LG or GPT. A comparison was made between NLTV-LG and NLTV-GPT with the original NLTV prior to verify the effectiveness of the LG and GPT schemes. According to the results presented in Table III and Fig. 12, NLTV-LG and NLTV-GPT can produce sharper edges than NLTV. The experimental results indicate that both the developed LG and the GPT have their own respective advantages. However, as a whole, the LG term is more effective than the GPT.

2) Effectiveness of the LGR Prior: Compared with the other two gradient regularization terms (LG and GPT), the recovered high-frequency details and edges of the proposed NLTV-LGR are more accurate and much clearer. For instance, in the lower left zoomed regions in Fig. 12, the edges in the results of NLTV-LG and NLTV-GPT are severely blurred or distorted, whereas the edges of the proposed method are much clearer and more similar to the original HR edges. This comparison also verifies the complementary properties of the LG and GPT. While their respective contributions differ, these

TABLE IV
AVERAGE RUNNING TIME (s) ON TEST SET I IN 4× SR

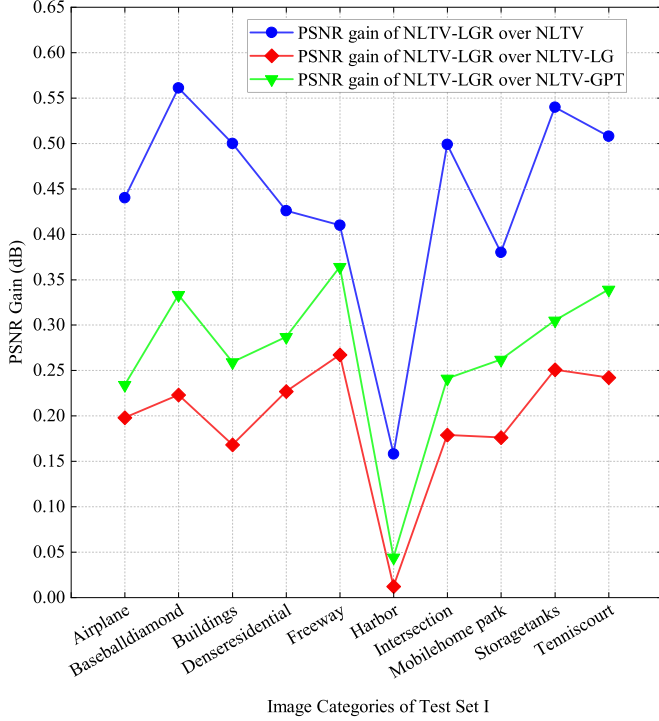| Method	RBPN [49]	BasicVSR [50]	SWBTV [51]	L0RIG [52]	NLTV	NLTV-LG	NLTV-GPT	Proposed
Time (s)	18.953	5.799	46.583	58.832	77.782	84.065	129.922	147.917

Fig. 13. Comparison of the SR performance between different extended versions for each image category in test set I.

Fig. 14. Average PSNR scores versus running times on test set I.

components are mutually reinforcing. The quantitative and qualitative results demonstrate that the estimated LGR prior can assist in the generation of visually pleasing SR results with clearer textures for remote sensing imagery.

3) Effectiveness of Combining NLTV With LGR: Fig. 13 illustrates the average PSNR gains of the proposed NLTV-LGR method over the three other extended versions for each image category in test set I. By coupling the NLTV and LGR priors in the proposed framework, the proposed method achieves a 0.442 dB/0.024 improvement over the use of a separate NLTV prior. With regard to the visual quality, the blur effects remain evident in the result obtained with only the NLTV prior, especially along the edges. In contrast, the LGR prior has advantages in reconstructing accurate edge and texture details. Consequently, the combined prior can produce sharp edges and fine structures, while simultaneously suppressing the reconstruction artifacts well. This demonstrates that the proposed method can fully exploit the complementary advantages of the two priors and can achieve more reliable super-resolved images with fewer artifacts and sharper edges.

F. Empirical Study on Regularization Parameters

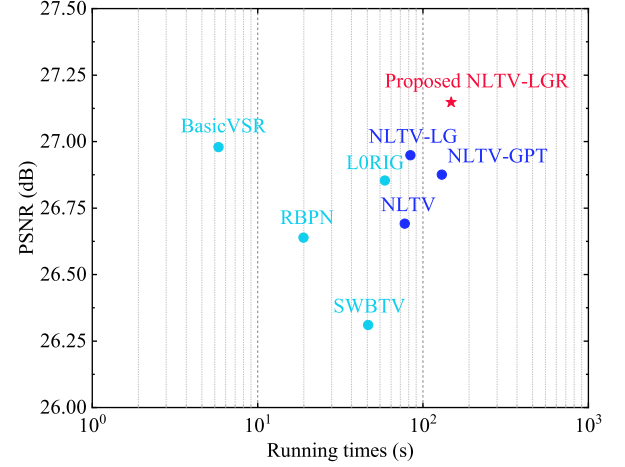There are two main regularization parameters α and β that are crucially important to the SR performance of the proposed method. The selection of regularization parameter has always been a headache for image SR. To make a fair comparison between the different methods, in the case of a small volume of experimental data, the regularization parameters can be determined manually by trying a series of values and selecting the ones with the highest PSNR or the best visual effect. In the experiments conducted in this study, if the regularization parameters had been selected manually, it would have been very time-consuming and tedious since there were more than 1000 images in the test set. In order to set the gradient regularization parameter α and the NLTV regularization parameter β reasonably, we adopted the adaptive parameter selection method [56] to obtain the approximate optimal regularization parameters in most cases of the simulation experiments.

G. Discussion on Computational Cost

By roughly analyzing the proposed method, the following four parts are the main computational cost: 1) estimation of the guidance gradient field with GPT; 2) parameter estimation of the motion and blur kernel; 3) similar patch search for the NLTV prior; and 4) inner iterative optimization with ADMM.

To comprehensively evaluate the computational cost of the proposed method, the average PSNR values versus running times of the proposed method and the other comparison methods on the images of test set I are shown in Fig. 14. All the experiments were conducted with MATLAB R2022b on an Intel Core i7 3.5-GHz CPU. Owing to their inner iterative optimization process, the model-based methods are more time-consuming than the learning-based methods, as illustrated in Table IV. In contrast, the deep learning-based methods are extremely fast in the SR process due to their natural ability for massively parallel computing. However, in this

comparison, the complex and time-consuming training process of the learning-based methods has not been taken into account. Specifically, the average running time of the proposed method is approximately 147.917 s, which is considerably slower than that of the comparison methods. Therefore, the computational cost is indeed a drawback of the proposed approach, although it significantly enhances the SR performance for remote sensing images. It is an effective way to further speed up the proposed method by implementing it in C++ code.

IV. Conclusion

In this article, we have introduced a framework that combines the advantages of both deep learning-based and variational model-based approaches. The gradient-guided MFSR method based on NLTV and LGR priors simultaneously exploits the interframe aliasing information and external feature learning. Specifically, the DRAN is employed to learn the latent HR gradient, which is then constructed as a local reconstruction constraint through the GPT for recovering fine image details. The well-designed network is capable of mining high-frequency spatial information that can be used to assist in restoring sharp edges in the HR image. Meanwhile, the NLTV prior is incorporated into the proposed MFSR framework as a global regularization term, which can effectively utilize the nonlocal spatial similarity within remote sensing imagery, thereby enhancing the texture details and suppressing image artifacts. Bycombining the two complementary priors within the adaptive norm-based reconstruction framework, the mixed $L1$ and $L2$ regularization minimization problem is solved via an iterative optimization algorithm.

Both the simulation and real-data experiments on remote sensing images illustrated that the proposed framework works more effectively than the other comparison SR methods, especially in the restoration of fine details. The proposed method can achieve superior evaluation metrics and visual results with higher fidelity and richer high-frequency information. In addition, we have demonstrated that the combination of the local and nonlocal regularization terms can further enhance the SR performance by imposing a local texture detail and global structure information constraint on remote sensing SR. In our future work, we will further examine the coupling of the variational model-based and the deep learning-based MFSR methods in order to bring out their respective advantages.

References

[1] H. Shen, Z. Qiu, L. Yue, and L. Zhang, "Deep-learning-based super-resolution of video satellite imagery by the coupling of multiframe and single-frame models," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612114, doi: 10.1109/TGRS.2021.3121303.

[2] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Sci. Rev.*, vol. 232, Sep. 2022, Art. no. 104110.

[3] Y. Xiao, Q. Yuan, Q. Zhang, and L. Zhang, "Deep blind super-resolution for satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5516316, doi: 10.1109/TGRS.2023.3291822.

[4] D. Valsesia and E. Magli, "A novel rate control algorithm for onboard predictive coding of multispectral and hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6341–6355, Oct. 2014, doi: 10.1109/TGRS.2013.2296329.

[5] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 305–319, Dec. 2018, doi: 10.1016/j.isprsjprs.2018.09.018.

[6] Z. Pan, W. Ma, J. Guo, and B. Lei, "Super-resolution of single remote sensing image based on residual dense backprojection networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7918–7933, Oct. 2019, doi: 10.1109/TGRS.2019.2917427.

[7] S. Zhang, Q. Yuan, J. Li, J. Sun, and X. Zhang, "Scene-adaptive remote sensing image super-resolution using a multiscale attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4764–4779, Jul. 2020, doi: 10.1109/TGRS.2020.2966805.

[8] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023, doi: 10.1016/j.inffus.2023.03.021.

[9] Y. Xiao et al., "Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial–temporal transformer," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, Apr. 2022, Art. no. 102731, doi: 10.1016/j.jag.2022.102731.

[10] Z. He, J. Li, L. Liu, D. He, and M. Xiao, "Multiframe video satellite image super-resolution via attention-based residual learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605015, doi: 10.1109/TGRS.2021.3072381.

[11] Z. He and D. He, "A unified network for arbitrary scale super-resolution of video satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8812–8825, Oct. 2021, doi: 10.1109/TGRS.2020.3038653.

[12] H. Bai and J. Pan, "Self-supervised deep blind video super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 4641–4653, Jul. 2024, doi: 10.1109/TPAMI.2024.3361168.

[13] Z. Pan et al., "Super-resolution based on compressive sensing and structural self-similarity for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4864–4876, Sep. 2013, doi: 10.1109/TGRS.2012.2230270.

[14] Y. Luo, L. Zhou, S. Wang, and Z. Wang, "Video satellite imagery super resolution via convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2398–2402, Dec. 2017, doi: 10.1109/LGRS.2017.2766204.

[15] A. Xiao, Z. Wang, L. Wang, and Y. Ren, "Super-resolution for 'Jilin-1' satellite video imagery via a convolutional network," *Sensors*, vol. 18, no. 4, p. 1194, 2018, doi: 10.3390/s18041194.

[16] X. Dong, X. Sun, X. Jia, Z. Xi, L. Gao, and B. Zhang, "Remote sensing image super-resolution using novel dense-sampling networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1618–1633, Feb. 2021, doi: 10.1109/TGRS.2020.2994253.

[17] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615611, doi: 10.1109/TGRS.2021.3136190.

[18] Y. Yu, X. Li, and F. Liu, "E-DBPN: Enhanced deep back-projection networks for remote sensing scene image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5503–5515, Aug. 2020, doi: 10.1109/TGRS.2020.2966669.

[19] W. Dong, S. Liu, S. Xiao, J. Qu, and Y. Li, "ISPDiff: Interpretable scale-propelled diffusion model for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5519614, doi: 10.1109/TGRS.2024.3407967.

[20] Y. Xiao, Q. Yuan, J. He, and L. Zhang, "Remote sensing image super-resolution via cross-scale hierarchical transformer," *Geo-Spatial Inf. Sci.*, vol. 27, no. 6, pp. 1914–1930, Dec. 2023, doi: 10.1080/10095020.2023.2288179.

[21] Z. Qiu, H. Shen, L. Yue, and G. Zheng, "Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network," *ISPRS J. Photogramm. Remote Sens.*, vol. 199, pp. 226–241, May 2023, doi: 10.1016/j.isprsjprs.2023.04.016.

[22] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "A progressively enhanced network for video satellite imagery superresolution," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1630–1634, Nov. 2018, doi: 10.1109/LSP.2018.2870536.

[23] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019, doi: 10.1109/TGRS.2019.2902431.

[24] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sens.*, vol. 12, no. 14, p. 2207, Jul. 2020, doi: 10.3390/RS12142207.

[25] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610819, doi: 10.1109/TGRS.2021.3107352.

[26] H. Liu and Y. Gu, "Deep joint estimation network for satellite video super-resolution with multiple degradations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621015, doi: 10.1109/TGRS.2022.3163790.

[27] C. Ren, X. He, Y. Pu, and T. Q. Nguyen, "Enhanced non-local total variation model and multi-directional feature prediction prior for single image super resolution," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3778–3793, Aug. 2019, doi: 10.1109/TIP.2019.2902794.

[28] H. Shen, M. Jiang, J. Li, C. Zhou, Q. Yuan, and L. Zhang, "Coupling model- and data-driven methods for remote sensing image restoration and fusion: Improving physical interpretability," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 231–249, Jun. 2022, doi: 10.1109/MGRS.2021.3135954.

[29] H. Shen, M. K. Ng, P. Li, and L. Zhang, "Super-resolution reconstruction algorithm to MODIS remote sensing images," *Comput. J.*, vol. 52, no. 1, pp. 90–100, Jan. 2009, doi: 10.1093/comjnl/bxm028.

[30] F. Li, X. Jia, D. Fraser, and A. Lambert, "Super resolution for remote sensing images based on a universal hidden Markov tree model," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1270–1278, Mar. 2010, doi: 10.1109/TGRS.2009.2031636.

[31] J. Ma, J. Cheung-Wai Chan, and F. Canters, "An operational superresolution approach for multi-temporal and multi-angle remotely sensed imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 110–124, Feb. 2012, doi: 10.1109/JSTARS.2011.2182505.

[32] H. Liu, Y. Gu, T. Wang, and S. Li, "Satellite video super-resolution based on adaptively spatiotemporal neighbors and nonlocal similarity regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8372–8383, Dec. 2020, doi: 10.1109/TGRS.2020.2987400.

[33] J. Zhao, Z. Chen, Q. M. J. Wu, X. Li, L. Cai, and K. Zhu, "Improved edge-guided network for single image super-resolution," *Multimedia Tools Appl.*, vol. 81, no. 1, pp. 343–365, Jan. 2022, doi: 10.1007/S11042-021-11429-3.

[34] W. Yang et al., "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5895–5907, Dec. 2017, doi: 10.1109/TIP.2017.2750403.

[35] H. Chen, X. He, L. Qing, and Q. Teng, "Single image super-resolution via adaptive transform-based nonlocal self-similarity modeling and learning-based gradient regularization," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1702–1717, Aug. 2017, doi: 10.1109/TMM.2017.2688920.

[36] Q. Song and H. Liu, "Deep gradient prior regularized robust video super-resolution," *Electronics*, vol. 10, no. 14, p. 1641, Jul. 2021, doi: 10.3390/electronics10141641.

[37] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, "Gradient profile prior and its applications in image super-resolution and enhancement," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1529–1542, Jun. 2011, doi: 10.1109/TIP.2010.2095871.

[38] Q. Yan, Y. Xu, X. Yang, and T. Q. Nguyen, "Single image superresolution based on gradient profile sharpness," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3187–3202, Oct. 2015, doi: 10.1109/TIP.2015.2414877.

[39] T. Li, X. Dong, and H. Chen, "Single image super-resolution incorporating example-based gradient profile estimation and weighted adaptive p-norm," *Neurocomputing*, vol. 355, pp. 105–120, Aug. 2019, doi: 10.1016/j.neucom.2019.04.051.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[41] H. Shen, L. Peng, L. Yue, Q. Yuan, and L. Zhang, "Adaptive norm selection for regularized image restoration and super-resolution," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1388–1399, Jun. 2016, doi: 10.1109/TCYB.2015.2446755.

[42] J. Chen, Z. Cai, J. Lai, and X. Xie, "Fast optical flow estimation based on the split Bregman method," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 664–678, Mar. 2018, doi: 10.1109/TCSVT.2016.2615324.

[43] L. Chen, F. Fang, T. Wang, and G. Zhang, "Blind image deblurring with local maximum gradient prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1742–1750, doi: 10.1109/CVPR.2019.00184.

[44] P. Rodriguez and B. Wohlberg, "Efficient minimization method for a generalized total variation functional," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 322–332, Feb. 2009, doi: 10.1109/TIP.2008.2008420.

[45] S. Boyd et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011, doi: 10.1561/2200000016.

[46] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via L_0 gradient minimization," in *Proc. SIGGRAPH Asia Conf.*, 2011, pp. 1–12.

[47] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 270–279, doi: 10.1145/1869790.1869829.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[49] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3892–3901, doi: 10.1109/CVPR.2019.00402.

[50] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4945–4954.

[51] X. Liu, L. Chen, W. Wang, and J. Zhao, "Robust multi-frame super-resolution based on spatially weighted half-quadratic estimation and adaptive BTV regularization," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4971–4986, Oct. 2018, doi: 10.1109/TIP.2018.2848113.

[52] J. Sun, Q. Yuan, H. Shen, J. Li, and L. Zhang, "A single-frame and multi-frame cascaded image super-resolution method," *Sensors*, vol. 24, no. 17, p. 5566, Aug. 2024, doi: 10.3390/s24175566.

[53] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[54] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013, doi: 10.1109/LSP.2012.2227726.

[55] N. Venkatanath, D. Venkatanath, M. Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. 21st Nat. Conf. Commun. (NCC)*, Feb. 2015, pp. 1–6, doi: 10.1109/NCC.2015.7084843.

[56] Q. Yuan, L. Zhang, H. Shen, and P. Li, "Adaptive multiple-frame image super-resolution based on U-curve," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3157–3170, Dec. 2010, doi: 10.1109/TIP.2010.2055571.

Jing Sun received the B.S. degree from the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, in 2015, where he is currently pursuing the Ph.D. degree.

His research interests focus on image super-resolution reconstruction and remote sensing image quality improvement and deep learning.
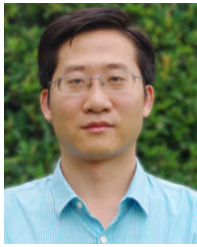
Huanfeng Shen (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

He is currently a Distinguished Professor with Wuhan University, where he is also the Dean of the School of Resource and Environmental Sciences. He was/is the Principal Investigator (PI) of two projects supported by the National Key Research and Development Program of China and six projects supported by the National Natural Science Foundation of China. He has authored or co-authored more than 150 peer-reviewed international journal articles, where over 60 appeared in IEEE JOURNALS and published four books as a chief editor. His research interests include remote sensing image processing, multisource data fusion, and intelligent environmental sensing.

Dr. Shen is a fellow of the Institution of Engineering and Technology (IET), an Education Committee Member of the Chinese Society for Geodesy Photogrammetry and Cartography, and a Theory Committee Member of the Chinese Society for Geospatial Information Society. He was a recipient of the First Prize in Natural Science Award of Hubei Province in 2011, the First Prize in Nature Scientific Award of China's Ministry of Education in 2015, and the First Prize in Scientific and Technological Progress Award of Chinese Society for Geodesy Photogrammetry and Cartography in 2017. He is also a Senior Regional Editor of the *Journal of Applied Remote Sensing* and an Associate Editor of *Geography*, *Geo-Information Science*, and *Journal of Remote Sensing*.

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is a Professor. He has published more than 90 research articles, including more than 70 peer-reviewed articles in international journals, such as *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTION ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was a recipient of the Youth Talent Support Program of China in 2019, the Top-Ten Academic Star of Wuhan University in 2011, and the Recognition of Best Reviewers of the IEEE GRSL in 2019. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He is an associate editor of five international journals and has frequently served as a referee for more than 40 international journals for remote sensing and image processing.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently a "Chang-Jiang Scholar" Chair Professor appointed by the Ministry of Education of China at the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He was a Principal Scientist for the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He has published more than 700 research articles and five books. He is the Institute for Scientific Information (ISI) highly cited author. He is the holder of 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes of the IEEE GRSS 2014 Data Fusion Contest. His students have been selected as the winners or finalists of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He is the Founding Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter. He also serves as an associate editor or an editor for more than ten international journals. He is currently serving as an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.