# Deep-Learning-Based Super-Resolution of Video Satellite Imagery by the Coupling of Multiframe and Single-Frame Models

Huanfeng Shen, *Senior Member, IEEE*, Zhonghang Qiu, *Student Member, IEEE*, Linwei Yue, and Liangpei Zhang, *Fellow, IEEE*

***Abstract*— Image super-resolution (SR) is an effective solution to the limitation of the spatial resolution of video satellite images, which is caused by the degradation and compression in the imaging phase. For the processing of satellite videos, the commonly employed deep-learning-based single-frame SR (SFSR) framework has limited performance without using complementary information between the video frames. On the other side, the multiframe SR (MFSR) can utilize temporal subpixel information to super-resolve the high-resolution (HR) imagery. However, although deeper and wider deep learning network provides powerful feature representations for SR methods, it has always been a challenge to accurately reconstruct the boundaries of ground objects in video satellite images. In this article, to address these issues, we propose an edge-guided video SR (EGVSR) framework for video satellite image SR, which couples the MFSR model and the edge-SFSR (E-SFSR) model in a unified network. The EGVSR framework is composed of an MFSR branch and an edge branch. The MFSR branch is used to extract the complementary features from the consecutive video frames. Concurrently, the edge branch acts as an SFSR model to translate the edge maps from the low-resolution modality to the HR one. At the final SR stage, the DBFM is built to focus on the promising inner representations of the features of the two branches and fuse them. Extensive experiments on video satellite imagery show that the proposed EGVSR method can achieve superior performance compared to the representative deep-learning-based SR methods.**

***Index Terms*— Deep learning, edge prior, super-resolution (SR), video satellite.**

Manuscript received July 22, 2021; revised August 18, 2021; accepted October 15, 2021. Date of publication October 18, 2021; date of current version February 8, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB2102900 and in part by the National Natural Science Foundation of China under Grant 41801263. *(Corresponding author: Linwei Yue.)*

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

Zhonghang Qiu is with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China (e-mail: qiu_zh@whu.edu.cn).

Linwei Yue is with the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China (e-mail: yuelw@cug.edu.cn).

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).

Digital Object Identifier 10.1109/TGRS.2021.3121303

## I. Introduction

With the rapid development of satellite imaging technology, the need for remote sensing data with high spatial and temporal resolutions has become more and more urgent in geoscientific applications. Compared with traditional satellite imaging techniques, video satellites (e.g., the SkySat series [1], Jilin-1 series, and OVS-1 series) are new kinds of earth observation remote sensing satellites that can perform dynamic earth observation. Video satellites have a distinct advantage in temporal resolution and, thus, play an important role in monitoring dynamic objects, such as vehicles, aircraft, and ships [2]. However, the spatial resolution and clarity of video satellite imagery are degraded due to the influence of the data acquisition and transmission process [3], [4]. Therefore, it is necessary to apply super-resolution (SR) reconstruction technology, which can maintain the high temporal resolution of video satellite images and achieve a high spatial resolution.

The image SR refers to the technique of reconstructing a high-resolution (HR) image by processing one or multiple low-resolution (LR) images [5]. Reconstruction-based SR methods that are developed from the frequency domain [6] to the spatial domain [7]–[9] have advantages in recovering clear details but encounter complicated calculations and performance degradation when the scale factor is large. Learning-based SR methods, such as neighborhood embedding [10]–[12], random forest [13], and sparse coding [14]–[16] SR methods, introduce external training datasets to capture image features and combine the acquired prior knowledge to learn the mapping relationship between the LR and HR images. Since Dong *et al.* [17] proposed an SR reconstruction algorithm (denoted as SRCNN) based on a simple three-layer convolutional neural network (CNN), which was the pioneering work introducing a CNN into SR, many deep-learning-based SR methods [18]–[24] have been proposed and achieved advanced performances on various public SR benchmark datasets compared with the traditional SR methods. In addition to focusing on the network architecture and loss functions, a number of methods [25]–[29] have embedded prior knowledge of the images into the CNN network to effectively assist with image SR reconstruction.

It should be noted that the first work in SR reconstruction originated from the processing of multitemporal Landsat remote sensing images [30]. However, in the subsequent

development of SR technology, most of the aforementioned SR methods have been designed for use with natural images, and only a small number of methods have been designed for remote sensing applications. In 2007, Shen *et al.* [31] proposed a maximum *a posteriori* (MAP)-based SR method with an edge-preserving Huber prior and L1 norm data fidelity to super-resolve multitemporal MODIS images. Subsequently, several multiframe SR (MFSR) methods [32], [33] have been proposed for remote sensing images captured in different angles and dates. More recently, studies have begun to address MFSR problems with deep learning methods in the context of remote sensing. For example, Kawulok *et al.* [34] adopted several CNNs to perform SR on the multiple input satellite images followed by a postfusion process. Recent promising methods (e.g., DeepSUM [35] and HighRes-net [36]) have also achieved excellent performances in super-resolving multitemporal PROBA-V images.

Compared to the MFSR case, single-frame SR (SFSR) algorithms benefit from requiring only a single image as the input and have attracted increased attention in remote sensing applications [37]–[39]. Among the existing studies, the frameworks based on variational regularization [40], geostatistical downscaling [41]–[44], and deep learning [45], [46] are the most common solutions. As for the processing of video satellite images with high resolution, most studies have focused on edge enhancement and employed a deep-learning-based SFSR framework. For instance, Jiang *et al.* [47] proposed an edge-enhancement GAN that introduces an adversarial strategy for dealing with the noise in satellite images. Moreover, several SFSR methods [48]–[50] that consider the difference of the scale and content between satellite images and natural images have also been proposed to improve the resolution of video satellite images. Although these methods have made progress in the performance of video satellite image SR, the limited spatial information in a single input image restricts their ability to reconstruct more accurate textures. The MFSR approaches proposed in [51]–[53] utilize temporal sequences of satellite video frames with complementary information to super-resolve the HR imagery and achieve superior performance, but few studies focus on preserving and enhancing the boundaries of small objects. Satellite videos provide sequential frames that contain complementary spatial information and temporal information. Furthermore, the reconstruction of the contours and edges of ground features is of great concern in super-resolving video satellite data as this is widely used in fine-scale earth monitoring applications. To better reconstruct the natural textural information and high-frequency components, the efficient fusion of the subpixel supplementary information among the video sequences and paying special attention to the image edges should be considered simultaneously.

In this article, to address the above issues, we introduce an edge prior to guiding the MFSR reconstruction phase, and we refer to the proposed method as the edge-guided video SR (EGVSR) framework. The EGVSR framework consists of two branches. On the one hand, a primary MFSR network is constructed to capture the spatiotemporal features of the input satellite video sequences. With a subpixel convolutional layer, the intermediate features are upsampled to the HR space. On the other hand, in the edge branch, we first utilize an edge operator to extract the LR edge map from the central LR video frame. Next, an edge-SFSR (E-SFSR) model, namely, the edge-enhancement network, is constructed to reconstruct the HR edge map from its LR version, and the HR edge features from the edge branch are finally used as an edge prior to guiding the SR process. Note that we integrate the intermediate-level features of the MFSR branch, which are pivotal to the recovery of edge maps, into the edge branch by the developed dual-branch fusion module (DBFM). The DBFM is again used at the end of the EGVSR framework to better fuse the output features of the two branches. Finally, visually pleasing and high-quality super-resolved results can be obtained through the overall EGVSR framework.

In summary, the contributions of our work are threefold.

1) We propose a unified framework, namely, EGVSR for video satellite imagery SR reconstruction by coupling the MFSR and E-SFSR models. The MFSR model in the MFSR branch mines the spatiotemporal information, and the E-SFSR model in the edge branch enhances the resolution of the coarse edge map. As a result, the proposed model can achieve a better performance than the representative deep-learning-based SFSR and MFSR methods
2) We propose an edge branch to introduce edge priors to guide the reconstruction of the main branch during the training phase, which can help the network focus more on the structure of ground objects and enrich the details in the SR results.
3) We propose a DBFM that is used in the edge branch and at the end of the EGVSR network to fuse the features with different representations of the two branches. A joint loss function that is made up of SR loss and edge-preserving loss is also utilized to impose restriction on the SR results.

The rest of this article is organized as follows. In Section II, we describe the proposed EGVSR framework in detail. The experimental results are given in Section III. Discussion and an analysis of the proposed method are given in Section IV. In Section V, we conclude this article.

## II. Methodology

The overall framework of EGVSR is shown in Fig. 1. Given $2N + 1$ consecutive LR satellite video frames $(I_{t-N}^{\mathrm{LR}}, \ldots, I_t^{\mathrm{LR}}, \ldots, I_{t+N}^{\mathrm{LR}})$ as the input of EGVSR, our target is to reconstruct the HR image of the central frame. First, video SR is performed for the multiple video frames in the MFSR branch. Adjacent LR frames are first fed into a feature extraction module consisting of several residual blocks (RBs). The Pyramid, Cascading, and Deformable (PCD) alignment module is then applied to implicitly align the extracted features. Next, spatial and temporal attention maps are calculated for these features in the temporal and spatial attention (TSA) fusion module, so as to better fuse the features. The fused features are then passed through a stack of residual dense blocks (RDBs) and subpixel convolutional layers. Meanwhile, the central frame is fed into the edge branch to first obtain

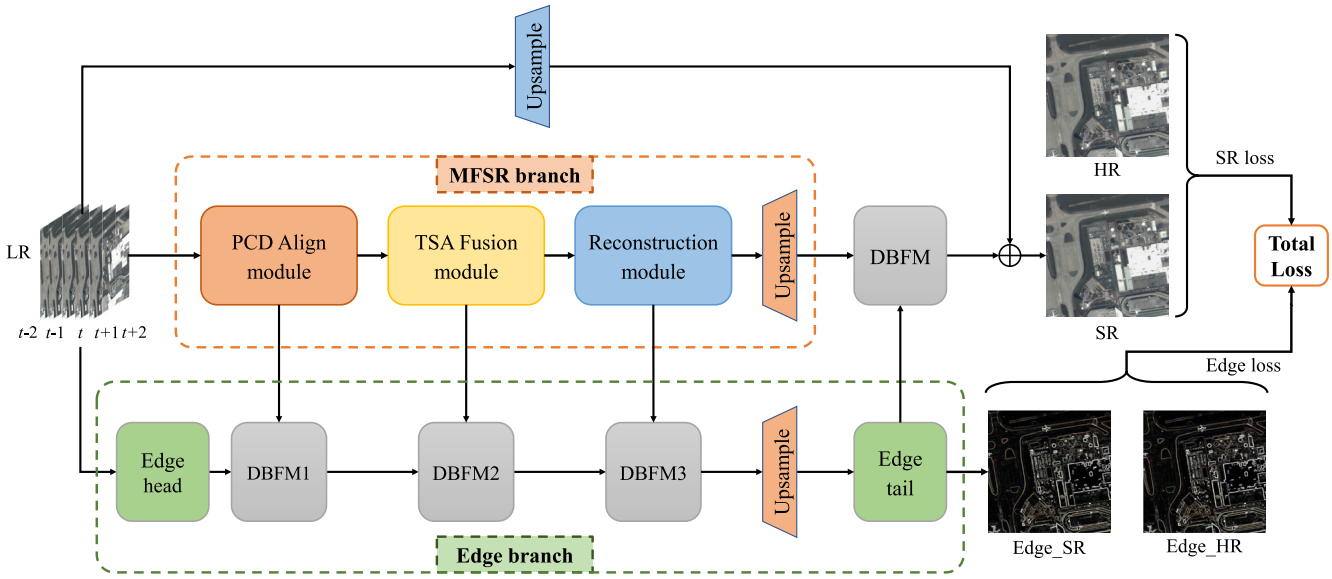



Fig. 1. The architecture of the proposed EGVSR framework. We use three input frames as an illustrative example. The MFSR branch contains three major functional modules: the PCD align model, the TSA fusion module, and the deep reconstruction module. The edge branch contains an edge extraction operation and an edge-enhanced network to generate the HR edge map. The DBFM in EGVSR is used to fuse the features of the two branches.

a rough edge map. The LR edge map is then passed into the edge-enhanced network to generate a sharp and clear edge map. In particular, during the edge reconstruction phase, we use the DBFM to incorporate the intermediate-level features of the MFSR branch that contains spatial information and temporal information into the edge branch to further promote the performance of the edge branch. The DBFM is also used at the end of the network to fuse the features output by the two branches. The SR features are then added to the upsampled LR image to obtain the final SR reconstruction result. In the rest of this section, the architecture of the EGVSR framework, and the MFSR and SFSR coupled strategy are described in detail.

### *A. Multiframe Super-Resolution Branch*

The MFSR branch is the primary network for super-resolving multiple frames in the EGVSR architecture, which can be roughly partitioned into three substructures, namely, the PCD alignment module, the TSA fusion module, and the deep reconstruction module, in which the PCD alignment module and TSA fusion module are based on the excellent work by Wang *et al.* [54] on a video restoration framework with enhanced deformable networks (EDVRs). In the following, the three major modules are further explained.

*1) PCD Alignment Module:* In video satellite imagery, the subpixel displacement between different frames is not very large. The main motion is in fact caused by moving objects, such as vehicles and aircraft. Thus, it would be suboptimal to use optical flow estimation for the alignment of the satellite video frames. We, therefore, use the very effective PCD alignment module from EDVR to align the satellite video frames without explicit motion estimation. We define the number of sampling locations in a convolutional kernel as $k$, and the weight and offset for the $k$th location are denoted as $\omega_k$ and $p_k$. Therefore, in a common $3 \times 3$ convolutional kernel, $k$ is 9, and $p_k \in \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$. Offsets are added to the conventional regular original sampling grid in the deformable convolution [55], [56], which allows the convolutional kernels to implement irregular sampling. The operation of the deformable convolution is given as follows:

$$F_{t+i}^{\text{align}}(p_0) = \sum_{k=1}^{K} \omega_k F_{t+i}(p_0 + p_k + \Delta p_k) \tag{1}$$

where $F_{t+i}$ and $F_{t+i}^{\text{align}}$ denote the input features and aligned features, respectively, $i \in [-N, N]$. $\Delta p_k$ is the learnable offset for the $k$th location, which can be calculated from the features of the reference feature and the neighboring frame

$$\Delta P_i = f_{\text{op}}\left(\left[F_{t+i}, F_t\right]\right) \tag{2}$$

where $\Delta P_i$ is the set of $\Delta p_k$, [.,.] represents the concatenate operation, and $f_{\text{op}}(\cdot)$ represents the function consisting of a convolutional layer to predict the offsets. In the PCD module, all the operations are performed on features that are extracted from a group of RBs at the beginning of the EGVSR framework. Pyramidal processing and cascading refinement are adopted in the PCD module. As shown in Fig. 2, there are three-level pyramids of feature representation in the PCD module, i.e., $L = 3$. The feature $F_{t+i}^{l}$ at the $l$th level is downsampled by the strided convolution filters with a factor of 2 to obtain the feature $F_{t+i}^{l+1}$ at the $(l+1)$th level. The alignment features and offsets at the $(l+1)$th level are, in turn, upsampled for the prediction of the alignment features and offsets at the $l$th level

$$\Delta P_{t+i}^{l} = f_{of}\left(\left[f_{\text{op}}\left(\left[F_{t+i}^{l}, F_t^{l}\right]\right), \left(\Delta P_{t+i}^{l+1}\right)^{\uparrow 2}\right]\right) \tag{3}$$

$$\left(F_{t+i}^{\text{align}}\right)^{l} = f_g\left(\left[f_{\text{DConv}}\left(F_{t+i}^{l}, \Delta P_{t+i}^{l}\right), \left(\left(F_{t+i}^{\text{align}}\right)^{l+1}\right)^{\uparrow 2}\right]\right) \tag{4}$$

where $f_{\text{DConv}}$ is the deformable convolution described in (1), [.,.] represents the concatenation operation, $(\cdot)^{\uparrow 2}$ refers to the

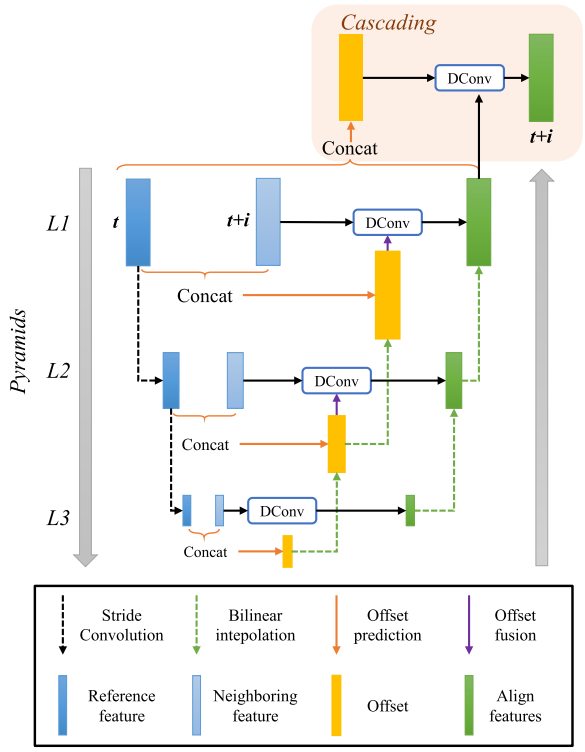



Fig. 2. PCD alignment module with PCD convolution.

upsampling by a scale factor of 2, $f_{of}(\cdot)$ represents the function with a convolution layer to fuse the offsets at the $l$th level with the upsampled offsets at the $(l+1)$th level, respectively, and $f_g(\cdot)$ represents the function with a convolutional layer to fuse the aligned features. Finally, the PCD module can achieve image alignment with higher accuracy by this coarse-to-fine approach.

*2) TSA Fusion Module:* Different neighboring frames have different importances in restoring the reference frame. Therefore, after aligning the features of each neighboring frame, we adopt an attention mechanism to supply different weights for the aligned features in the spatial and temporal dimensions. The structure of the TSA fusion module is shown in Fig. 3. For each aligned frame $F_{t+i}^{\text{align}}$, $i \in [-N:N]$, a temporal attention heat map is calculated to represent its similar distance $d$ with the reference frame feature

$$d\left(F_{t+i}^{\text{align}}, F_t^{\text{align}}\right) = \sigma\left(f_{c1}\left(F_{t+i}^{\text{align}}\right)^{\mathrm{T}} f_{c2}\left(F_t^{\text{align}}\right)\right) \tag{5}$$

where $f_{c1}(\cdot)$ and $f_{c2}(\cdot)$ represent the operation of simple convolutional filters, and $\sigma(\cdot)$ denotes the sigmoid activation function. We perform elementwise multiplication on the temporal attention maps and the original aligned features $F_{t+i}^{\text{align}}$. Next, we use an extra convolutional layer to fuse these attention-modulated features $\tilde{F}_{t+i}^{\text{align}}$

$$\tilde{F}_{t+i}^{\text{align}} = F_{t+i}^{\text{align}} \otimes d\left(F_{t+i}^{\text{align}}, F_t^{\text{align}}\right) \tag{6}$$

$$F_{\text{fusion}} = \text{Conv}\left(\left[\tilde{F}_{t-N}^{\text{align}}, \ldots, \tilde{F}_t^{\text{align}}, \ldots, \tilde{F}_{t+N}^{\text{align}}\right]\right) \tag{7}$$

where $\otimes$ represents the elementwise multiplication and $\text{Conv}(\cdot)$ denotes a convolutional layer.

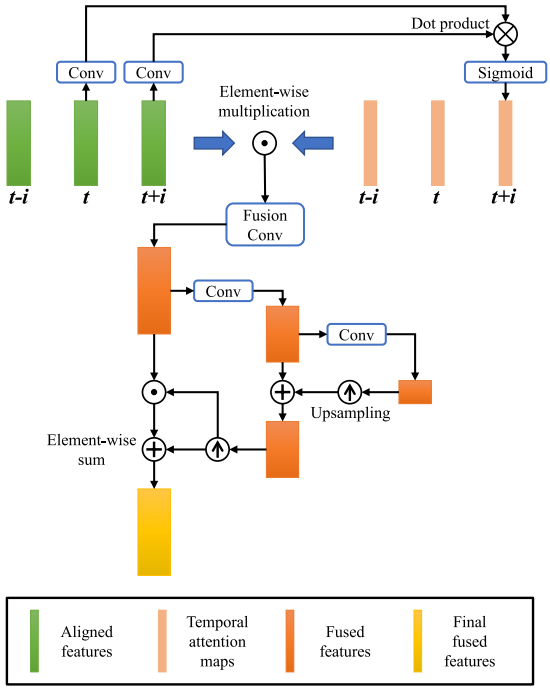



Fig. 3. TSA fusion module with TSA.

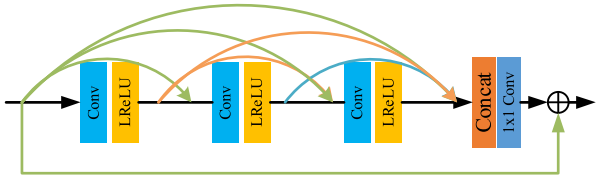



Fig. 4. Architecture of the RDB.

We then calculate the spatial attention mask from the fusion features and use a pyramid design to increase the range of the receptive field of the attention map. Finally, we perform elementwise multiplication and addition on the features and the spatial attention mask to obtain the fused features.

*3) Deep Reconstruction Module:* The final fused feature $F_{\text{fusion}}$ output from the TSA module is then fed into the following deep SR reconstruction module. The deep SR reconstruction module mainly consists of stacked RDBs. It can be seen in Fig. 4 that the RDBs combine residual networks and dense connections [57]. Benefiting from this, the RDBs can encourage feature reuse and strengthen feature propagation to obtain a better restoration quality. Finally, the intermediate features containing more representational information can be obtained through the deep reconstruction module.

### *B. Edge Branch*

The low-frequency component of the image describes the main part of the image and is a comprehensive measure of the intensity of the entire image. The high-frequency components correspond to the sharply changing parts of the image, i.e., the edges or noise and details of the image. In image restoration tasks, the texture details corresponding to the high-frequency components are often more difficult to recover. As one of the most informative natural image priors, we introduce an edge prior to regularizing the restoration process. Therefore,

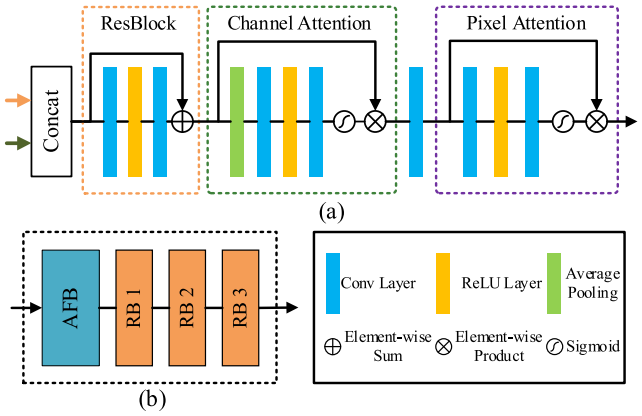



Fig. 5. Architecture of the DBFM: (a) AFB and (b) structure of the DBFM, consisting of the AFB and three RBs.

we construct an SFSR subnetwork as a branch in EGVSR to guide the modeling of the edge map of the image.

*1) LR Edge Information Extraction:* The Sobel operator, which has low computational complexity and easy implementation, is utilized to extract the corresponding edge map from image $I$. The calculation can be represented by the following formula:

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & 2 & 1 \end{bmatrix} \tag{8}$$

$$I_{\text{edge}} = M(I) = \sqrt{(S_x * I)^2 + (S_y * I)^2} \tag{9}$$

where $S_x$ and $S_y$ stand for the Sobel template in the horizontal and vertical directions, respectively, and $M(\cdot)$ represents the operation to obtain the edge map, which can retain accurate edge information. Note that the binarization strategy is eliminated to avoid the appearance of false edges and the loss of image features [33]. By setting the template of the Sobel operator as the fixed kernel of the convolutional layer, the structure information extraction operation becomes a built-in component of the network.

At the head of the edge branch, we use $M(\cdot)$ in (9) to obtain the LR edge map from the middle input frame $I_t^{\text{LR}}$.

*2) Dual-Branch Fusion Module:* In the proposed EGVSR framework, the edge branch encodes rich structural information, while the MFSR branch contains spatial information and temporal information. The output features of the two branches contain different feature representations, so simply concatenating them and then fusing them with a convolutional layer would be suboptimal. Therefore, we designed a DBFM consisting of an attention fusion block (AFB) and three RBs, so as to pay more attention to the representative information in the concatenated features of the MFSR branch and edge branch.

As shown in Fig. 5(a), we first utilize the basic AFB containing the channel attention (CA) and pixel attention (PA) mechanisms to deal with the different types of features. The features output by the MFSR branch and edge branch are first concatenated in the channel direction and then passed through an RB (denoted as $F_S$). We then introduce the CA mechanism to assign different weights to the channels in $F_S$. The weights

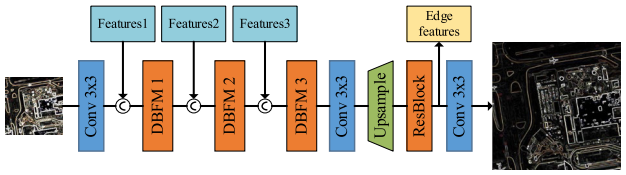



Fig. 6. Architecture of the edge-enhanced network.

of the CA mechanism can be expressed as follows:

$$W_{\text{CA}} = \sigma(\text{Conv}(\rho(\text{Conv}(P(F_S))))) \tag{10}$$

where $P(\cdot)$ represents the global average pooling operation and $\rho$ is the rectified linear unit (ReLU). We perform elementwise multiplication on $F_S$ and $W_{\text{CA}}$

$$F_{\text{rw1}} = F_S \otimes W_{\text{CA}}. \tag{11}$$

Considering that the edges and textures in the images are uneven, the PA mechanism is used to make the model focus on the high-frequency image region. We utilize a $1 \times 1$ convolutional layer to change the shape of $F_{\text{rw1}}$ (the output of the CA) from $2N \times H \times W$ to $N \times H \times W$, denote it as $F_{\text{rw1}}^*$, and then feed it into the PA module. The PA module can be expressed as follows:

$$W_{\text{PA}} = \sigma\left(\text{Conv}\left(\rho\left(\text{Conv}\left(F_{\text{rw1}}^*\right)\right)\right)\right). \tag{12}$$

We then elementwise multiply $F_{\text{rw1}}^*$ and $W_{\text{PA}}$ to obtain the reweighted features

$$F_{\text{rw2}} = F_{\text{rw1}}^* \otimes W_{\text{PA}}. \tag{13}$$

Finally, the reweighted features pass through a cascade of three RBs to obtain the final fused features. The overall structure of the DBFM is shown in Fig. 5(b).

*3) Edge-Enhanced Network:* The edge map extracted from the LR image does not have sharp details and may not serve as an effective prior to guiding the SR process. Therefore, we construct an edge-enhanced network to predict an edge map that has accurate and clear outlines from the extracted LR one. As shown in Fig. 6, the edge-enhanced network contains three DBFMs, one RB, and several convolutional layers. The well-designed MFSR branch is able to mine the spatial information and temporal information, which can be used to assist with the recovery of edges in the HR image. In the MFSR branch, the features of multiple frames are transferred from the shallow layers to the deep layers through the PCD, TSA, and reconstruction modules. Therefore, the features of different levels output from the three modules of the MFSR branch are incorporated into the edge branch to further improve the performance of the edge branch. The DBFM is then adopted to fuse the features of the two branches. Finally, the edge features generated by the next-to-last convolutional layer serve as the edge prior and are integrated into the MFSR branch. In addition, the SR edge map is output at the end of the edge branch to calculate the loss. Since the edge branch mainly performs the spatial distribution transformation of the edges and textures in the LR and HR edge maps, the designed lightweight edge-enhancement network can capture the structural dependency and generate accurate edge maps.

### C. MFSR and E-SFSR Coupled Mechanism

*1) Integration of the MFSR and Edge Branches:* In the proposed EGVSR framework, we first utilize the MFSR branch and edge branch to extract features from the input satellite video frames, respectively. The feature extraction of the two branches in the first stage can be expressed as

$$f_{\mathrm{MFSR}} = F_{\mathrm{MFSR}}\left(I_{t-N}^{\mathrm{LR}}, \ldots, I_t^{\mathrm{LR}}, \ldots, I_{t+N}^{\mathrm{LR}}\right) \tag{14}$$

$$f_{\mathrm{edge}} = F_{\mathrm{Edge}}\left(I_t^{\mathrm{LR}}\right) \tag{15}$$

where $F_{\mathrm{MFSR}}(\cdot)$ and $F_{\mathrm{Edge}}(\cdot)$ denote the MFSR and edge enhancement process, respectively. $f_{\mathrm{MFSR}}$ represents the deep features output by the MFSR branch. $f_{\mathrm{edge}}$ represents the edge features output by the next-to-last layer of the edge branch. At the end of the network, the two branches converge. We utilize the DBFM with an attention mechanism to fuse $f_{\mathrm{MFSR}}$ and $f_{\mathrm{edge}}$ for the final reconstruction

$$f_{\mathrm{SR}} = F_{\mathrm{DBFM}}\left(\left[f_{\mathrm{MFSR}}, f_{\mathrm{edge}}\right]\right) \tag{16}$$

where $F_{\mathrm{DBFM}}$ denotes the DBFM and $f_{\mathrm{SR}}$ represents the reconstructed features. We adopt global residual learning in EGVSR to reduce the burden of network training. Thus, $f_{\mathrm{SR}}$ is added to the upsampled LR image to obtain the final SR result

$$I_t^{\mathrm{SR}} = \left(I_t^{\mathrm{LR}}\right)^{\uparrow s} \oplus f_{\mathrm{SR}} \tag{17}$$

where $(\cdot)^{\uparrow s}$ refers to the bicubic upsampling operation with a scale factor of $s$, $\oplus$ denotes the elementwise sum operation, and $I_t^{\mathrm{SR}}$ is the final SR reconstruction result for the input central LR image $I_t^{\mathrm{LR}}$ at time $t$.

*2) Objective Function:* Most of the previous methods learn the nonlinear mapping between the LR image and the corresponding HR image and use a common loss function (e.g., pixelwise L1 loss and L2 loss) to guide the model optimization. As shown in Fig. 1, the EGVSR outputs an SR result and an edge map. Therefore, we use two loss terms $L_{\mathrm{SR}}$ and $L_{\mathrm{edge}}$ for the SR reconstruction result and edge map, respectively. First, to minimize the difference between the SR reconstruction result and the corresponding HR image, we use the robust Charbonnier loss function [58]

$$L_{\mathrm{SR}} = \sqrt{\left\|I_t^{\mathrm{HR}} - I_t^{\mathrm{SR}}\right\|^2 + \gamma^2} \tag{18}$$

where $\gamma$ is set to $1 \times 10^{-3}$, and $I_t^{\mathrm{HR}}$ and $I_t^{\mathrm{SR}}$ are the SR result and HR ground truth, respectively.

Second, at the end of the edge branch, an HR edge map that has the same dimension as the HR image is generated. An $L1$ regularization term is constructed to constrain the training of the edge branch so that it can generate an edge map with more accurate details and provide an edge prior for the super-resolving process. The edge-preserving loss $L_{\mathrm{edge}}$ is defined as

$$L_{\mathrm{edge}} = \left\|M\left(I_t^{\mathrm{HR}}\right) - E\left(I_t^{\mathrm{LR}}\right)\right\|_1 \tag{19}$$

where $M(\cdot)$ represents the operation of extracting edge map, $M(I_t^{\mathrm{HR}})$ represents the ground-truth HR edge map extracted from the HR image, $E(\cdot)$ denotes the edge branch, and $E(I_t^{\mathrm{LR}})$ is the SR edge map reconstructed from $I_t^{\mathrm{LR}}$.

Finally, the SR loss $L_{\mathrm{SR}}$ and edge-aware loss $L_{\mathrm{edge}}$ form the complete loss $L_{\mathrm{total}}$ and the coupled network composed of the two branches can achieve end-to-end training. $L_{\mathrm{total}}$ is defined as follows:

$$L_{\mathrm{total}} = L_{\mathrm{SR}} + \lambda * L_{\mathrm{edge}} \tag{20}$$

where $\lambda$ is a tradeoff parameter to balance the two loss terms.

## III. Experiments

### A. Data Preparation

In this study, we conducted experiments on two available video image datasets, namely, Jilin-1 video satellite[1] imagery and OVS-1 video satellite[2] imagery. For the Jilin-1 video satellite datasets with a resolution of 0.92 m, we extracted 9000 video clips, each consisting of seven consecutive frames with a fixed resolution of $160 \times 160$. These video clips cover a variety of urban and natural scenarios with diverse objects, 90% of which were used for the training, and the remaining video clips were regarded as the validation data. As shown in Fig. 7(a), several scenarios in the Jilin-1 datasets with representative surface coverage types, such as airports, industrial, and intersections, were used to build the test set. The video clip of each scene in the test set was made up of 30 consecutive frames with a size of $400 \times 400 \times 3$.

In order to verify the performance of the proposed method in processing data with different spatial resolutions and levels of degradation, OVS-1 video satellite images, for which the frame rate was 20 frames/s and the spatial resolution was 1.98 m, were used for the real-data experiments. As shown in Fig. 7(b), video clips of four scenes in the dataset were selected as the real-data test set. The size of each video frame in the four video clips was $120 \times 120 \times 3$, and no downsampling operation was performed.

### B. Implementation Details

We performed horizontal and vertical flipping and 90° rotation to achieve data augmentation and then used bicubic interpolation to downsample the frames in each video clip by a factor of 4. During the training phase, five consecutive downsampled frames with a patch size of $40 \times 40$ in the LR video clip were extracted as input for the model, and the corresponding frames in the HR video clip were selected as the ground truth. The batch size was set to 16. Note that the input and output of the network were both three-channel RGB images.

All the experiments were conducted using PyTorch 1.1 and Python 3.6.2 on an NVIDIA RTX 2080 GPU. We initialized the weights of the convolutional layers using the method proposed in [59]. The Adam optimizer [60] with momentum parameters $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\varepsilon = 10^{-8}$ was employed for the optimization. Meanwhile, we set the tradeoff parameter $\lambda$ to 0.1. The initial learning rate was set to $1 \times 10^{-4}$ and was then halved after every 100 epochs. Typically, 150 epochs were sufficient for training the models since more epochs do not always mean further improvement.

[1]Jilin-1 video satellite imagery is available at http://charmingglobe.com/

[2]OVS-1 video satellite imagery is available at https://www.myorbita.net/

(a) (b)

Fig. 7. Sample frames of the two satellite video test datasets. (a) Sample frames of representative scenes in the Jilin-1 test dataset, i.e., airport, overpass, parking lot, industrial, downtown, intersection, residential, building, freeway, storage tank, runway, and school. (b) Sample frames of representative scenes in the OVS-1 test dataset, i.e., Real_scene1, Real_scene2, Real_scene3, and Real_scene4.

TABLE I

QUANTITATIVE COMPARISON ON THE JILIN-1 VIDEO SATELLITE TEST SETS FOR 4× SR, WHERE THE BOLD FONT INDICATES THE BEST PERFORMANCE

| Input | | Single Frame | | | | Multiple Frames | | |
|---|---|---|---|---|---|---|---|---|
| Test scene | Method<br>Scale | Bicubic<br>PSNR/SSIM | VDSR<br>PSNR/SSIM | SRGAN<br>PSNR/SSIM | RCAN<br>PSNR/SSIM | SOFVSR<br>PSNR/SSIM | EDVR<br>PSNR/SSIM | Ours<br>PSNR/SSIM |
| Airport | ×4 | 32.309/0.904 | 36.070/0.947 | 36.391/0.945 | 36.849/0.952 | 37.040/0.953 | 37.391/0.956 | **37.625/0.958** |
| Overpass | ×4 | 36.578/0.942 | 39.740/0.966 | 39.397/0.963 | 40.633/0.970 | 40.704/0.970 | 40.684/**0.972** | **40.800/0.972** |
| Parking lot | ×4 | 33.203/0.906 | 35.657/0.941 | 35.942/0.944 | 36.802/0.953 | 36.769/0.952 | 36.923/0.955 | **37.627/0.962** |
| Industrial | ×4 | 34.310/0.935 | 37.960/0.963 | 38.261/0.961 | 39.655/0.969 | 39.766/0.969 | 39.735/0.969 | **39.968/0.970** |
| Downtown | ×4 | 31.772/0.888 | 35.167/0.935 | 35.296/0.936 | 36.045/0.944 | 36.139/0.945 | 36.593/0.950 | **36.657/0.951** |
| Intersection | ×4 | 36.282/0.937 | 39.319/0.963 | 39.029/0.961 | 40.243/0.967 | 40.237/0.967 | 40.361/**0.968** | **40.487/0.968** |
| Residential | ×4 | 33.558/0.889 | 35.780/0.926 | 35.69/0.928 | 36.585/0.938 | 36.483/0.936 | 36.808/0.941 | **36.833/0.942** |
| Buildings | ×4 | 32.830/0.907 | 37.159/0.952 | 37.0/0.951 | 38.317/0.961 | 38.414/0.961 | 38.445/0.962 | **38.740/0.964** |
| Freeway | ×4 | 34.913/0.947 | 39.182/0.971 | 39.194/0.969 | 40.230/0.974 | 40.375/0.975 | 40.578/0.975 | **40.955/0.977** |
| Storage tank | ×4 | 36.096/0.960 | 41.354/0.980 | 40.568/0.977 | 42.072/0.981 | 42.452/0.982 | 42.244/**0.982** | **42.598/0.982** |
| Runway | ×4 | 39.928/0.971 | 43.257/0.983 | 42.526/0.98 | 44.412/0.985 | 44.474/0.985 | 44.270/0.984 | **44.754/0.985** |
| School | ×4 | 33.008/0.910 | 36.774/0.949 | 37.081/0.951 | 38.030/0.958 | 37.945/0.958 | 38.252/0.960 | **38.600/0.962** |
| Average | ×4 | 34.567/0.925 | 38.118/0.956 | 38.031/0.956 | 39.156/0.963 | 39.233/0.963 | 39.357/0.964 | **39.637/0.966** |

### *C. Simulated Experiments*

In the simulated experiments, the test LR video frames were synthesized by applying the bicubic interpolation with a scaling factor of 4 to the test video clips collected from the Jilin-1 video satellite data. To validate the overall performance of the proposed method, we compared the EGVSR method with several representative SR methods. These methods included three SFSR methods, i.e., VDSR [61], RCAN [62], and SRGAN [63], and two MFSR methods, i.e., SOFVSR [64] and EDVR [54]. For a fair comparison, all the models were retrained using the Jilin-1 video satellite datasets. For the evaluation, we adopted the peak signal-to-noise ratio (PSNR) and the structural similarity measure (SSIM) [65]. The total PSNR and SSIM values of a video clip were calculated by averaging the PSNRs/SSIMs of all the frames.

The quantitative results in terms of PSNR and SSIM are listed in Table I, where it can be seen that the three video SR methods, including EGVSR, perform better than the other methods in terms of PSNR and SSIM. The proposed EGVSR method overperforms all the other methods in terms of PSNR and SSIM. In certain scenes, such as parking lots and freeways that contain rich image content, EGVSR surpasses all the other methods by a large margin in terms of PSNR. The results for all the test sets show that the proposed EGVSR method obtains the best performance, which proves that the method has a strong reconstruction capability for video satellite imagery.

We also present some visual results in Figs. 8–10 to investigate how the methods perform in terms of visual quality. For a better comparison of the visual results, we provide zoomed-in views within the yellow box region and their corresponding reconstruction mean error (ME) maps, which displays obvious distinctions between the different methods. It can be seen that EGVSR can recover clear and sharp edges and is faithful to the ground truth. For example, in the overpass video clip in Fig. 8, the other methods blur the two white vehicles driving side by side, and only EGVSR can distinguish these two objects. As shown in Fig. 9, most of the methods produce blurring artifacts. Although SOF-VSR and EDVR can reproduce the small white cars beside the trains, they still oversmooth the trains, and it is clear that EGVSR can recover the frames with more detailed patterns. Because many vehicles exist in the parking lot scene, the frames contain abundant high-frequency information. The proposed EGVSR method can reconstruct a result that is closest to the ground-truth frame and, therefore, greatly outperforms the other methods by a large margin in terms of PSNR, which further demonstrates the superiority of the proposed method in reconstructing textures and details of adjacent ground objects. Moreover, for the texture of the four neighboring trucks in Fig. 10, only EGVSR can separate the four trucks and recover clear textures, while the other methods suffer from varying degrees of ambiguity. Meanwhile, in the ME maps in Figs. 8–10, the whiter the pixel, the smaller the error, and the better the SR result. It can be observed that the proposed EGVSR can effectively preserve the textures. In general, the other methods show limitations in recovering consecutive objects and small moving objects, whereas the

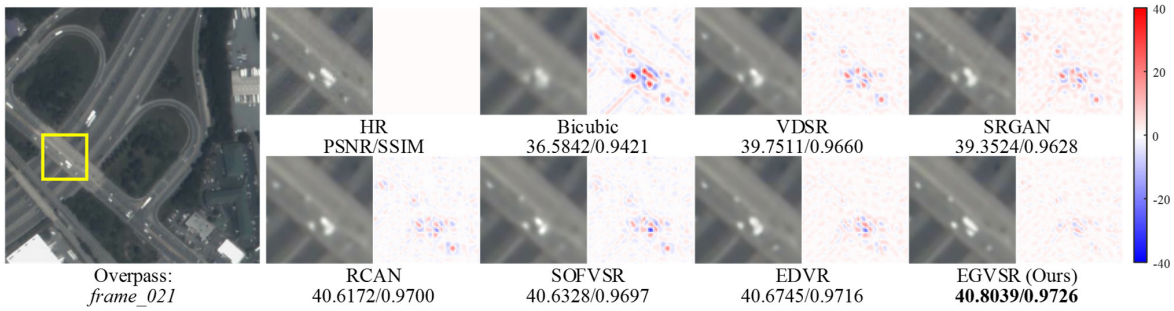



Fig. 8. Visual results obtained using the different SR methods on the 21st frame of the overpass scene with a scale factor of 4. The yellow box area is zoomed in for better visualization, and the corresponding reconstruction error image is shown on the right-hand side. The legend color bar is shown on the far right.

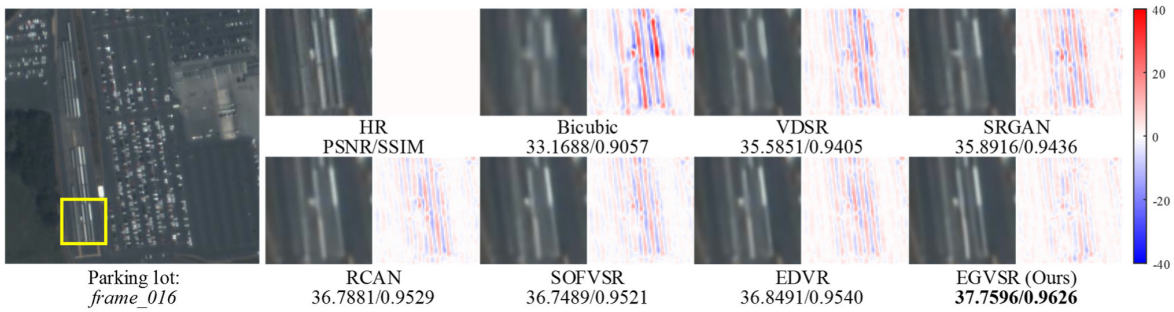



Fig. 9. Visual results obtained using the different SR methods on the 16th frame of the parking lot scene with a scale factor of 4. The yellow box area is zoomed in for better visualization, and the corresponding reconstruction error image is shown on the right-hand side. The legend color bar is shown on the far right.

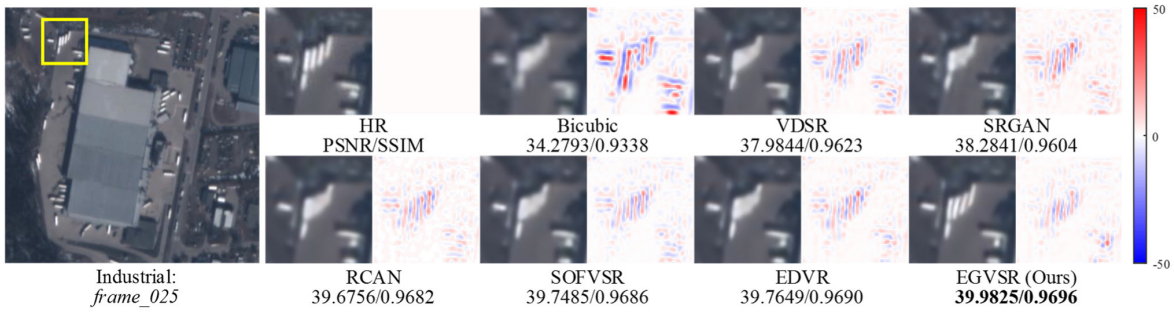



Fig. 10. Visual results obtained using the different SR methods on the 25th frame of the industrial scene with a scale factor of 4. The yellow box area is zoomed in for better visualization, and the corresponding reconstruction error image is shown on the right-hand side. The legend color bar is shown on the far right.

reconstructed results of EGVSR can not only show sharp image contours but also contain finer details.

### D. Real-Data Experiments

To further evaluate the robustness of the proposed EG-VSR method on real scenes, we conducted another group of experiments on the test video clips collected from the OVS-1 video satellite data. In real-world SR experiments, the degradation factors in the test video clips are unknown, so we directly feed observed video frames instead of downsampled LR frames as the input. In addition, we introduce the average gradient (AG) [66] to further evaluate the quality of the reconstruction results since there are no corresponding HR images for reference. The calculation of AG is given as follows:

$$\mathrm{AG} = \frac{1}{(H-1)(W-1)} \sum_{x} \sum_{y} \frac{|G(x,y)|}{\sqrt{2}} \tag{21}$$

where $H$ and $W$ are the height and the width of the image, respectively, and $G(\cdot)$ is the gradient vector of the image. The AG is often used to assess image clarity because it reflects the small detail contrast and texture variation characteristics in the imagery. Generally speaking, the larger the AG, the clearer the reconstruction images.

As shown in Table II, EGVSR obtains better quantitative results than the other compared methods in all the test scenes. The SRGAN and RCAN methods are single-image SR methods, which obtains poorer results than the other video SR methods. We take RCAN, SOFVSR, and EDVR into account for the qualitative comparison. The visual results of the different methods are displayed in Fig. 11. In the Real_scene1 sample image, the proposed method reconstructs two small white objects, while the other methods blur one of the objects. For the white roof in the Real_scene3 sample image, most of the methods produce blurred and distorted

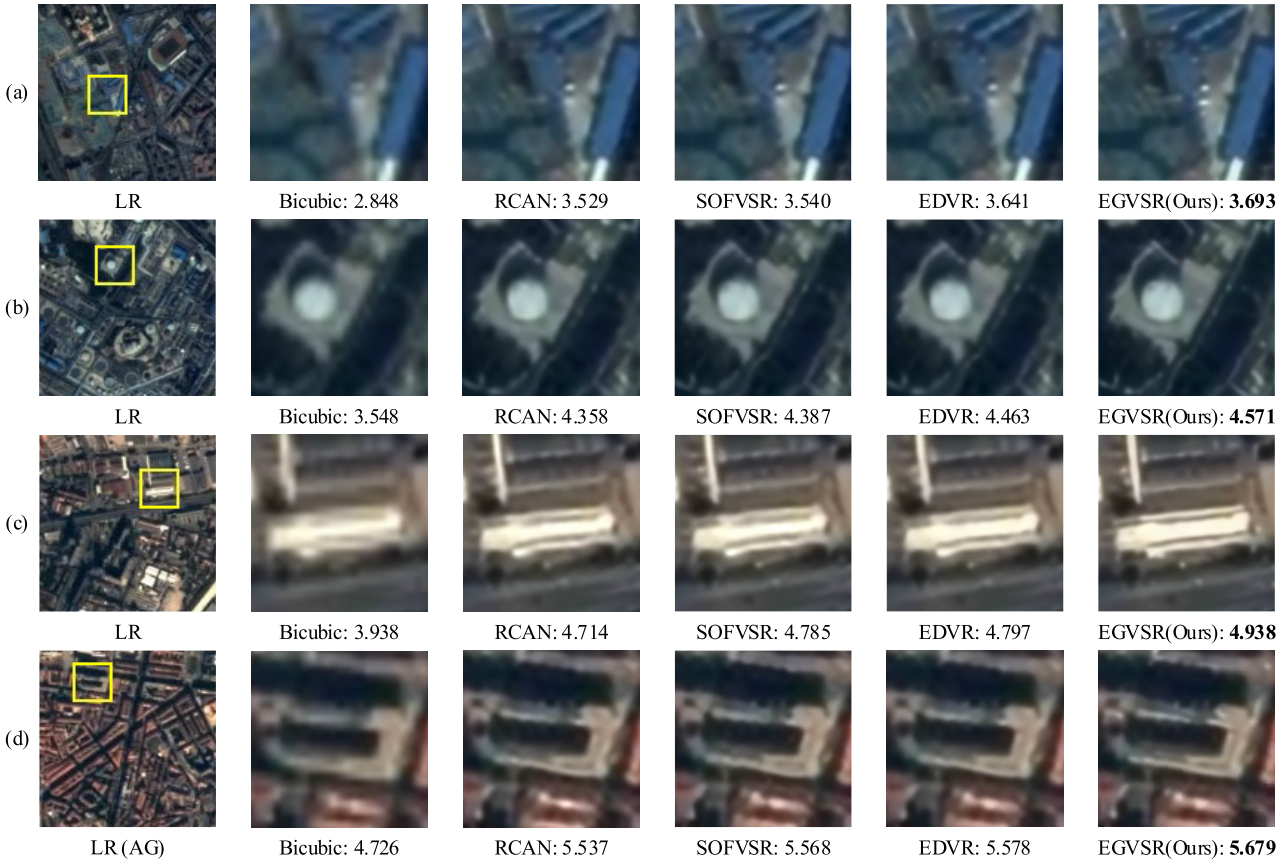



Fig. 11. Results obtained on the OVS-1 video satellite imagery with a scale factor of 4. (a) Super-resolved results for Real_scene1. (b) Super-resolved results for Real_scene2. (c) Super-resolved results for Real_scene3. (d) Super-resolved results for Real_scene4.

TABLE II

RESULTS OF THE DIFFERENT METHODS ON THE OVS-1 TEST SET WITH THE SCALE FACTOR OF 4

| Method<br>Metric | Bicubic<br>(AG) | VDSR<br>(AG) | SRGAN<br>(AG) | RCAN<br>(AG) | SOFVSR<br>(AG) | EDVR<br>(AG) | EGVSR<br>(AG) |
|---|---|---|---|---|---|---|---|
| Real_scene1 | 2.846 | 3.251 | 3.441 | 3.526 | 3.538 | 3.636 | **3.683** |
| Real_scene1 | 3.544 | 4.134 | 4.231 | 4.345 | 4.384 | 4.46 | **4.572** |
| Real_scene3 | 3.924 | 4.477 | 4.519 | 4.694 | 4.712 | 4.732 | **4.902** |
| Real_scene4 | 4.706 | 5.262 | 5.163 | 5.472 | 5.517 | 5.539 | **5.643** |
| Average | 3.755 | 4.306 | 4.339 | 4.509 | 4.537 | 4.591 | **4.700** |

textures, but the proposed method shows a better performance in recovering clearer textures on the roof. Overall, the proposed EGVSR method is capable of reconstructing images with sharper edges and more details than the other methods.

## IV. DISCUSSIONS AND ANALYSIS

### *A. Exploring the Effectiveness of the Two Branches*

The effectiveness of the proposed EGVSR was proven by the experiments described in Sections III-C and III-D. To give a more transparent explanation of the effectiveness of the proposed model, we conducted further additional investigations. First, we conducted an ablation study on the edge branch to explore its importance and performance. In the proposed method, the edge branch serves as a part of the EGVSR framework to provide an image edge prior for reconstructing high-quality images. We, therefore, removed it from the EGVSR framework and only used the MFSR branch for inference, which is denoted as EGVSR_NEDGE.

TABLE III

ABLATION STUDY ON THE EDGE BRANCH IN THE EGVSR FRAMEWORK

| Model | EGVSR_NEDGE | EGVSR |
|---|---|---|
| Edge prior | × | ✓ |
| mPSNR | 39.396 | 39.637 |
| mSSIM | 0.965 | 0.966 |

We conducted experiments on EGVSR and EGVSR_N-EDGE for a quantitative and qualitative comparison. As shown in Table III, the quantitative evaluation results of the full model after removing the edge branch drop by 0.24 dB. As shown in Fig. 12, the green roof super-resolved by EGVSR_NEDGE is blurry, while the images recovered by EGVSR have more detailed textures. The quantitative and qualitative results demonstrate that the edge prior provided by the edge branch can help the model generate visually pleasing SR results with clearer textures.

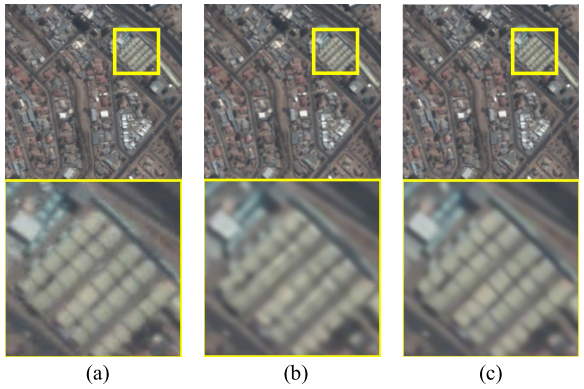



Fig. 12. Visual comparison of the SR reconstruction results obtained on the residential scene from the test set: (a) HR image; (b) SR reconstruction result obtained by EGVSR_NEDGE, which only has the MFSR branch; and (c) SR reconstruction result obtained by EGVSR.

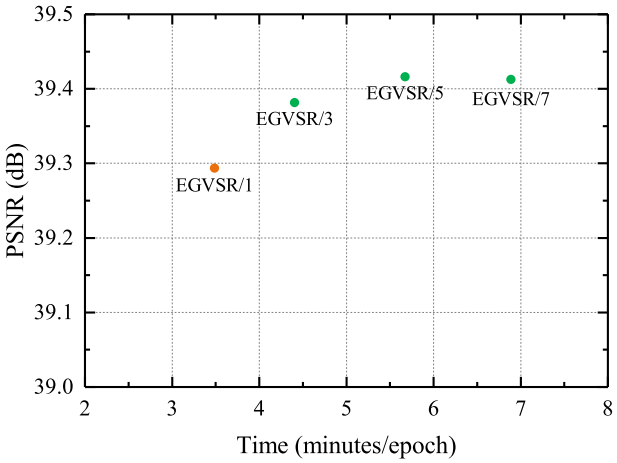



Fig. 13. Effect of the length of the input frames for 4× SR with EGVSR. EGVSR/($n$): EGVSR takes $n$ frames as input.

In EGVSR, the MFSR branch is used to deal with satellite video sequences. In order to explore the influence of complementary information between the different video frames on the reconstruction, we also evaluated EGVSR with different lengths of temporal sequences, i.e., different numbers of input frames $n \in \{1, 3, 5, \text{and } 7\}$. The corresponding trained models are denoted as EGVSR/1, EGVSR/3, EGVSR/5, and EGVSR/7, respectively. The performance and training time per epoch of each model, as measured on the Jilin-1 video satellite imagery, is shown in Fig. 13. Compared with EGVSR/1, the different input frames in EGVSR/3 possess additional complementary information, thereby improving the SR performance. By adding more frames, the performance of EGVSR/5 increases by roughly 0.05 dB over EGVSR/3. However, the performance improvement is marginal when the length of the temporal sequences is longer than 5, and the performance of EGVSR/5 is even better than that of EGVSR/7, which uses seven neighboring frames. The deployment speed decreases with a longer temporal sequence length. Therefore, we set the number of input frames as five in the proposed model by making a tradeoff between the performance and speed.

TABLE IV
ABLATION STUDY ON DBFM IN THE EGVSR FRAMEWORK

| Model | EGVSR_NDBFM | EGVSR |
|---|---|---|
| DBFM | × | ✓ |
| mPSNR | 39.566 | 39.637 |
| mSSIM | 0.966 | 0.966 |

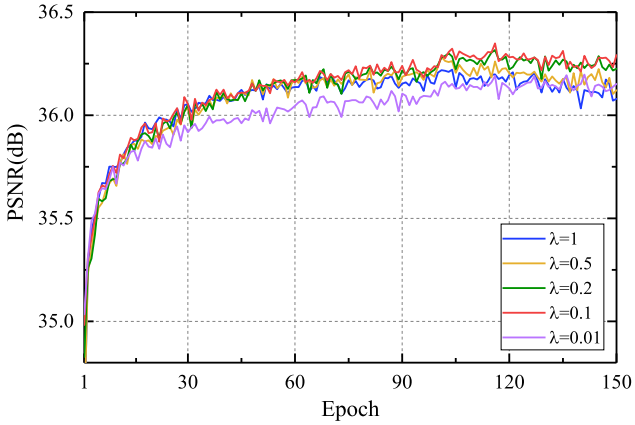



Fig. 14. Selection of a different $\lambda$. When $\lambda = 0.1$, the model achieves the best results.

### B. Effectiveness of the Dual-Branch Fusion Module

The features in the MFSR branch and the edge features in the edge branch contain different feature representations. In EGVSR, we use the DBFM to merge the features from the two branches. In this section, we describe the ablation experiments conducted to analyze the contributions of the DBFM. We replaced the DBFM with common RBs. The trained model is referred to as EGVSR_NDBFM. As shown in Table IV, the performance of EGVSR_NDBFM drops by 0.07 dB. It is often difficult to obtain further improvements in a very deep network, but we can still obtain an improvement by introducing the attention mechanism into the module, which demonstrates the effectiveness of the DBFM in fusing the features from different branches.

### C. Investigation on the Hyperparameter $\lambda$

The total loss of EGVSR consists of a SR loss and an edge loss, among which we use the hyperparameter $\lambda$ to balance the proportion of the two parts. If $\lambda$ is too big, the edge branch will dominate and affect the reconstruction result, which leads to degraded performance. On the contrary, if $\lambda$ is too small, edge guidance does not work.

In this section, we conduct multiple sets of ablation experiments to find the suitable $\lambda$. The SR result fidelity term deserves a relatively larger weight than the edge prior modeling term for obtaining better reconstruction results; thus, setting $\lambda$ in the range of 0.1–1 gives slightly better performance. The impact of different values of $\lambda$ on the model performance is displayed in Fig. 14. According to the experimental results, our model is robust to different values of $\lambda$ in a wide range. It can be seen that our model achieves the best performance when $\lambda = 0.1$. Therefore, we finally set $\lambda = 0.1$ in our model. Although we have found a suitable $\lambda$ through multiple experiments, we hope to find the optimal $\lambda$

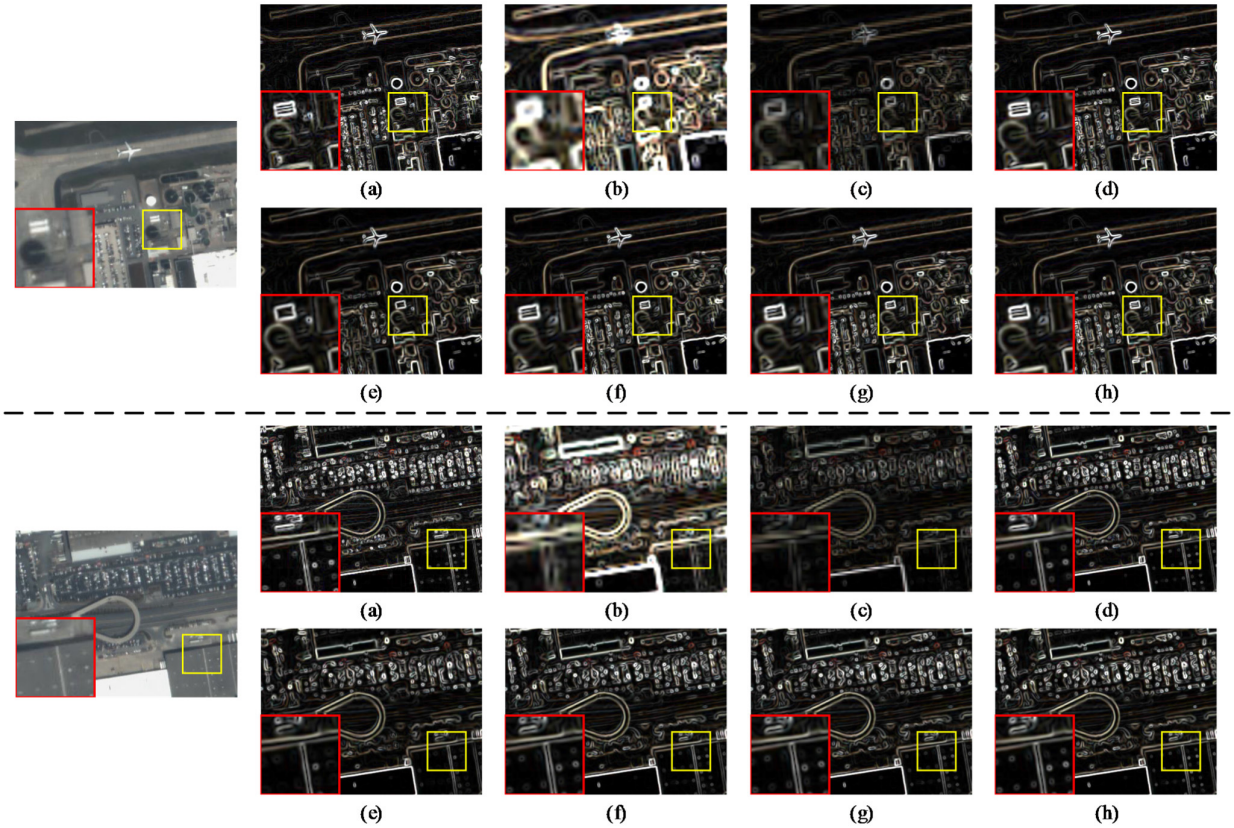



Fig. 15. Visual comparison of the generated edge maps: (a) ground truth obtained using the Sobel operator to directly extract the edge map from the HR image; (b)–(d) edge maps generated by Sobel + bicubic, bicubic + Sobel, and the edge-enhanced network, respectively; and (e)–(h) edge maps output by EGVSR_woFeatures, EGVSR_wFeature1, EGVSR_wFeature2, and EGVSR_wFeature3, respectively.

by introducing dynamic learning strategies and achieve better results in future work.

### *D. Effectiveness of the Edge-Generation Capability*

In the edge branch, we utilize the Sobel operator to extract coarse edge maps from the LR images and utilize the edge-enhanced network to predict the HR edge map so as to provide an edge prior for the SR reconstruction. In order to validate the capability of generating an accurate edge map through the edge-enhanced network, we compared the proposed edge-enhanced network with two other generation strategies, i.e., Sobel + bicubic and bicubic + Sobel. It can be seen in Fig. 15(b) that the edge map extracted from the LR counterparts contains thick lines after the bicubic interpolation. The edge map extracted from the upsampled image has blurred edges [see Fig. 15(c)]. In particular, the edge maps generated by Sobel + bicubic and bicubic + Sobel both lose texture details. The CNN-based edge-enhanced network can model the spatial translation between the LR and HR edge maps. As shown in Fig. 15(d), the edge-enhanced network successfully recovers an edge map with sharp details, which is very similar to the ground truth.

Furthermore, as shown in Fig. 6, several intermediate features from the MFSR branch are incorporated into the edge branch. To investigate the influence of these features, we separately incorporated them into the edge branch and train the network under the same conditions to obtain corresponding models, which are denoted as EGVSR_wFeature1, EGVSR_wFeature2, and EGVSR_wFeature3, respectively. We also trained a model that removed all the features and denoted them as EGVSR_woFeatures. The HR edge maps reconstructed by these models are displayed in Fig. 15(e)–(h), and the PSNR curves of different models during the training phase are plotted in Fig. 16. It is clear that the edge map [see Fig. 15(e)] output by the EGVSR_woFeatures shows fewer textures than those of the other three models.

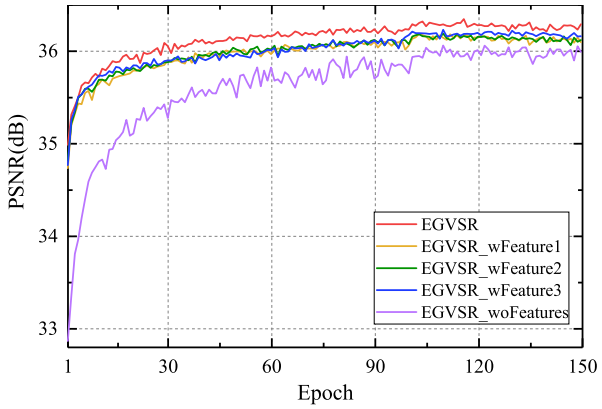



Fig. 16. PSNR curves of different models using features from the MFSR branch.

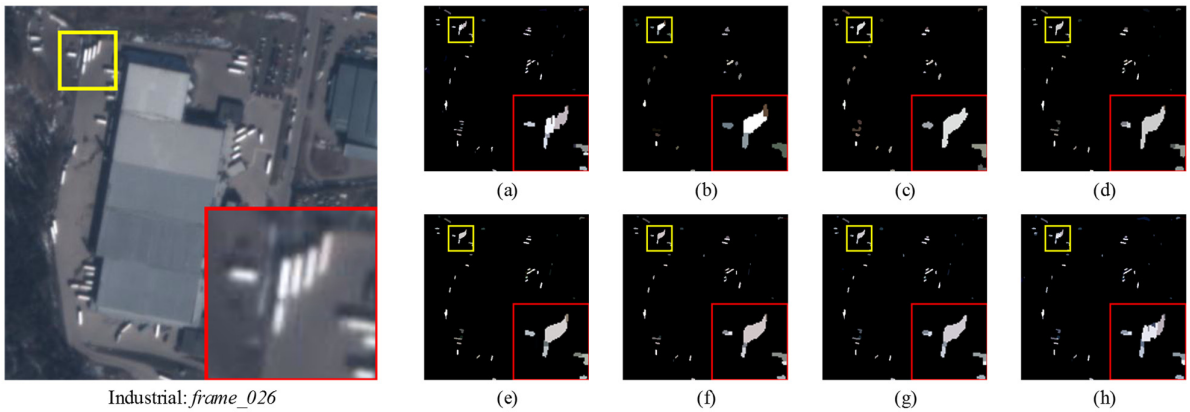



Fig. 17. Visual comparison of the ground features extracted from the SR results of the different methods: (a) HR image, (b) Bicubic, (c) VDSR, (d) SRGAN, (e) RCAN, (f) SOFVSR, (g) EDVR, and (h) proposed method.

Similarly, it can be seen from Fig. 16 that the PSNR curves of EGVSR_wo-Features start with the lowest PSNR value and oscillate more obviously than other models. For the texture of the cylinders on the roof in the second scene in Fig. 15, EG-VSR_wFeature3 can preserve relatively more textures than EGVSR_wFeature1 and EGVSR_wFeature2. As shown in Fig. 16, the PSNR curve of EGVSR_wFeature3 is also relatively higher than those of EGVSR_wFeature1 and EGV-SR_wFeature2, which indicates that the high-level features output by the deep reconstruction module are relatively more effective than the features of the other modules in improving the performance of the edge branch. Specifically, we incorporate the features output by the three modules in the MFSR branch into the edge branch in the proposed EGVSR, which achieves the clearest edge map [see Fig. 15(d)] and the highest PSNR curve (see Fig. 16).

### *E. Experiments in Ground Feature Extraction*

In order to further explore the effect of the proposed EGVSR on object extraction performance, we conducted a group of feature extraction experiments on video satellite imagery. First, the image segmentation algorithm based on edge detection was used to segment the image at multiple scales, and then, the full lambda-schedule algorithm [67] was used to fuse adjacent small patches with spectral and spatial feature information. The complete extraction process was implemented using the Segment Only Feature Extraction Workflow in ENVI5.3, and the parameters were set to the same values as those used in the processing of the different SR reconstruction results.

We performed target extraction for the vehicles in the image and display the results in Fig. 17. The extraction results can be identified by the color and shape of the patch. We display zoomed-in results in the red boxes for better comparison and visualization. Most of the methods extract the two adjacent vehicles in the lower right corner of the red box into a whole patch and only in the results of the two video SR methods, and the proposed method can the two vehicles be distinguished. It can be seen from Fig. 17(h) that the patch of the four consecutive cars in the middle of the red boxes has two characteristic colors of gray and white, which is similar to the result in Fig. 17(a), while the other methods treat the four cars as a whole. Overall, the small target extraction results obtained from the SR reconstruction results of the proposed EGVSR method are closest to the ground truth, which demonstrates that the proposed method has advantages in reconstructing accurate edges and textures.

## V. Conclusion

In this article, we have proposed an MFSR and E-SFSR coupled network for video satellite image SR. The proposed EGVSR framework is made up of an MFSR branch and an edge branch. In the MFSR branch, the shallow features of the multiple input satellite video frames are first aligned implicitly and then fused with spatial and temporal attention mechanisms. We then utilize a reconstruction module to further super-resolve the features to obtain an intermediate result. Meanwhile, we apply the edge branch to predict an HR edge map from the central frame. The features from the two branches are input into the DBFM, which can select and focus on the important parts of the features for the final SR reconstruction. The extensive experimental results showed that the proposed EGVSR can recover accurate and clear details, resulting in an improvement of both the SR accuracy and visual effect. The visual comparisons with video satellite images demonstrated the effectiveness of the proposed edge-guided strategies. The comparison with several representative SFSR and MFSR methods further demonstrated the outstanding performance of the proposed EGVSR. Although the current model and strategy have achieved good results, further improvements, such as embedding a more advanced edge extraction module in our proposed framework, could be made to handle the images with unknown degradation.

## References

[1] K. Murthy, M. Shearn, B. D. Smiley, A. H. Chau, J. Levine, and M. D. Robinson, "SkySat-1: Very high-resolution imagery from a small satellite," *Proc. SPIE*, vol. 9241, Oct. 2014, Art. no. 92411E.

[2] 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.

[3] D. Valsesia and P. T. Boufounos, "Universal encoding of multispectral images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4453–4457.

[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.

[5] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Process.*, vol. 128, pp. 389–408, Nov. 2016.

[6] S. Rhee and M. G. Kang, "Discrete cosine transform based regularized high-resolution image reconstruction algorithm," *Opt. Eng.*, vol. 38, no. 8, pp. 1348–1356, 1999.

[7] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 6, no. 11, pp. 1715–1726, 1989.

[8] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 996–1011, Jun. 1996.

[9] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1646–1658, Dec. 1997.

[10] X. Gao, K. Zhang, D. Tao, and X. Li, "Image super-resolution with sparse neighbor embedding," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3194–3205, Jul. 2012.

[11] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2012, pp. 1–10.

[12] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun./Jul. 2004, pp. 1–8.

[13] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3791–3799.

[14] X. Lu, H. Yuan, P. Yan, Y. Yuan, and X. Li, "Geometry constrained sparse coding for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1648–1655.

[15] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.

[16] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[18] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 391–407.

[19] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1664–1673.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[24] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2014, pp. 2672–2680.

[25] 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.

[26] H. Liu, Z. Fu, J. Han, L. Shao, S. Hou, and Y. Chu, "Single image super-resolution using multi-scale deep encoder–decoder with phase congruency edge map guidance," *Inf. Sci.*, vol. 473, pp. 44–58, Jan. 2019.

[27] F. Fang, J. Li, and T. Zeng, "Soft-edge assisted network for single image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 4656–4668, 2020.

[28] F. Fang, J. Li, Y. Yuan, T. Zeng, and G. Zhang, "Multilevel edge features guided network for image denoising," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3956–3970, Sep. 2021.

[29] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7769–7778.

[30] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," *Adv. Comput. Vis. Image Process.*, vol. 1, no. 2, pp. 317–339, 1984.

[31] 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, 2009.

[32] 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.

[33] J. Ma, J. C.-W. 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.

[34] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, "Deep learning for multiple-image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1062–1066, Jun. 2020.

[35] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3644–3656, May 2020.

[36] M. Deudon *et al.*, "HighRes-net: Recursive fusion for multi-frame super-resolution of satellite imagery," 2020, *arXiv:2002.06460*. [Online]. Available: http://arxiv.org/abs/2002.06460

[37] S. Kanakaraj, M. S. Nair, and S. Kalady, "SAR image super resolution using importance sampling unscented Kalman filter," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 2, pp. 562–571, Feb. 2018.

[38] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe, "Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training," *Remote Sens.*, vol. 10, no. 3, p. 394, Mar. 2018.

[39] K. Zheng *et al.*, "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2487–2502, Mar. 2021.

[40] J. Hu, Y. Ge, Y. Chen, and D. Li, "Super-resolution land cover mapping based on multiscale spatial regularization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2031–2039, May 2015.

[41] P. C. Kyriakidis, "A geostatistical framework for area-to-point spatial interpolation," *Geograph. Anal.*, vol. 36, no. 3, pp. 259–289, 2004.

[42] A. Boucher, P. C. Kyriakidis, and C. Cronkite-Ratcliff, "Geostatistical solutions for super-resolution land cover mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 272–283, Jan. 2008.

[43] Q. Wang, W. Shi, P. M. Atkinson, and Y. Zhao, "Downscaling MODIS images with area-to-point regression Kriging," *Remote Sens. Environ.*, vol. 166, pp. 191–204, Sep. 2015.

[44] Q. Wang, W. Shi, P. M. Atkinson, and E. Pardo-Iguzquiza, "A new geostatistical solution to remote sensing image downscaling," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 386–396, Jan. 2016.

[45] 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.

[46] T. Lu, J. Wang, Y. Zhang, Z. Wang, and J. Jiang, "Satellite image super-resolution via multi-scale residual deep neural network," *Remote Sens.*, vol. 11, no. 13, p. 1588, Jul. 2019.

[47] 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.

[48] 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.

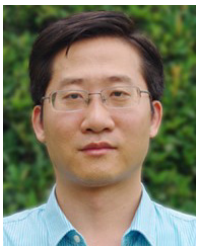
[49] Z. Wang, K. Jiang, P. Yi, Z. Han, and Z. He, "Ultra-dense GAN for satellite imagery super-resolution," *Neurocomputing*, vol. 398, pp. 328–337, Jul. 2020.

[50] J. Wu, Z. He, and L. Zhuo, "Video satellite imagery super-resolution via a deep residual network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 2762–2765.

[51] 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.

[52] 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.

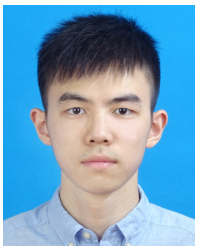
[53] Z. He, D. He, X. Li, and J. Xu, "Unsupervised video satellite super-resolution by using only a single video," *IEEE Geosci. Remote Sens. Lett.*, early access, Dec. 11, 2020, doi: 10.1109/LGRS.2020.3040972.
[54] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1954–1963.
[55] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[56] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.
[57] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2472–2481.
[58] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 624–632.
[59] 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.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[61] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.
[62] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 286–301.
[63] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[64] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using HR optical flow estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 4323–4336, 2020.
[65] Z. Wang, 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, no. 4, pp. 600–612, Apr. 2004.
[66] A. A. Chen, X. Chai, B. Chen, R. Bian, and Q. Chen, "A novel stochastic stratified average gradient method: Convergence rate and its complexity," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.
[67] D. J. Robinson, N. J. Redding, and D. J. Crisp, "Implementation of a fast algorithm for segmenting SAR imagery," Electron. Res. Lab., Salisbury, SA, Australia, Tech. Rep. DSTO-TR-1242, Jan. 2002.



**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.

In 2007, he joined the School of Resource and Environmental Sciences (SRES), Wuhan University, where he is currently a Luojia Distinguished Professor and an Associate Dean. He was or is a 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 over 100 research papers in peer-reviewed international journals. His research interests include remote sensing image processing, multisource data fusion, and intelligent environmental sensing.

Dr. Shen is also a Council Member of the China Association of Remote Sensing Application, 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 is also a member of the Editorial Board of the *Journal of Applied Remote Sensing, Geography, and Geo-Information Science*.



**Zhonghang Qiu** (Student Member, IEEE) received the B.S. degree from the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Resource and Environmental Sciences, Wuhan University, Wuhan.

His research interests include remote sensing image processing, image super-resolution reconstruction, and deep learning.

**Linwei Yue** received the B.S. degree in geographic information systems and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2012 and 2017, respectively.

She is currently an Associate Professor with the School of Geography and Information Engineering, China University of Geosciences, Wuhan. Her research interests include multisource remote sensing data fusion and hydrological applications.

**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 the 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 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 is currently a Chair Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and remote sensing (LIESMARS), Wuhan University. He has authored or coauthored than 700 research papers and five books. He is the Institute for Scientific Information (ISI) Highly Cited Author. He holds 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is also 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 also the Founding Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter. He is also an associate editor or an editor for more than ten international journals. He is also an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.