A Lightweight Recurrent Aggregation Network for Satellite Video Super-Resolution

Han Wang, Shengyang Li, and Manqi Zhao

***Abstract*—Intelligent processing and analysis of satellite video has become one of the research hotspots in the representation of remote sensing, and satellite video super-resolution (SVSR) is an important research direction, which can improve the image quality of satellite video. However, existing approaches for SVSR often underutilize a notable advantage inherent to satellite video, the presence of extensive sequential imagery capturing a consistent scene. Presently, the majority of SVSR methods merely harness a limited number of adjacent frames for enhancing the resolution of individual frames, thus resulting in suboptimal information utilization. In response, we introduce the recurrent aggregation network for satellite video superresolution (RASVSR). This innovative framework leverages a bidirectional recurrent neural network to propagate extracted features from each frame across the entire video sequence. It relies on an alignment method based on optical flow and deformable convolution (DCN) to realize the alignment of the features, and a temporal feature fusion module to realize effective feature fusion over time. Notably, our research underscores the positive influence of employing lengthier image sequences in SVSR. In the context of RASVSR, with better alignment and fusion, we make the perceptual field of each frame spanning 100 frames of the video, thus, acquiring richer information, and information between different images can be complementary. This strategic approach culminates in superior performance compared with alternative methods, as evidenced by a noteworthy 1.15 dB improvement in PSNR, with very few parameters.**

***Index Terms*—Attention mechanism, recurrent neural network (RNN), satellite video, video super-resolution (VSR).**

I. Introduction

SATELLITE video is a new type of ground observation method emerging in recent years, which has gained wide attention and application nowadays [1]. A large number of video satellites are already in orbit, such as Jilin-1 and SkySat, which can realize gaze observation of the ground. Compared with the traditional single-frame remote sensing, satellite video contains time domain information, has much larger data volume, and has important application value in the fields of transportation, resources, security, environment, etc. [2], but due to hardware limitations, such as sensors, its spatial resolution and image quality are lower compared with remote sensing images. Satellite video superresolution (SVSR) can improve the image quality of satellite video to achieve better results on high-level tasks, such as satellite video object detection [3], object tracking [4], [5], and object segmentation. SVSR can also be used for data compression and restoration to save transmission bandwidth.

Superresolution (SR) is a very classical low-level task in computer vision, which utilizes low-resolution images to reconstruct high-resolution images for the purpose of improving image quality. Its being a pathological problem, one input will correspond to multiple reasonable outputs, so SR can be formulated as a distribution estimation problem conditional on the input image [6]. Based on the number of images utilized, SR can be classified into single image superresolution (SISR) and multiple image superresolution (MISR). Among them, SISR has the longest history and the methods can be categorized into prediction models, edge-based methods, image statistical mehtods, patch-based methods, and deep learning methods [7]. Most of the latest algorithms rely on data-driven deep learning models to reconstruct the details required for SR, which automatically learn the relationship between inputs and outputs directly from the data, and have outperformed other traditional methods [8]. Since the information lost by image degradation cannot be retrieved in a single frame, MISR was created, which utilizes the information from multiple frames to reconstruct a single image, and benefits to information fusion, higher reconstruction accuracy can usually be achieved [9]. Image SR has been widely used in remote sensing, such as fusing hyperspectral images to obtain high-resolution images, and deep learning-based methods are proven to be effective when used on remote sensing data [10], [11].

Video super-resolution (VSR) requires SR of each frame in a video and imposes requirements on the coherence of the image. Broadly speaking, VSR can be regarded as an extension of image SR and can be processed frame by frame using the SISR algorithm. However, in practice, the results of using image SR algorithms to process video are hardly satisfactory because it may bring artifacts and lagging, which leads to unwanted temporal incoherence within frames [12], so most VSR methods are designed based on MISR. Video has richer information, and by utilizing this redundant information, VSR has a higher upper bound than simple image SR, but video has an additional temporal dimension compared with images, so designing SR

Manuscript received 25 August 2023; revised 2 October 2023; accepted 6 November 2023. Date of publication 13 November 2023; date of current version 29 November 2023. This work was supported by Key Deployment Program of the Chinese Academy of Sciences under Grant KGFZD-145-23-18. *(Corresponding author: Shengyang Li.)*

The authors are with the Technology and Engineering Center for Space Utilization, Key Laboratory of Space Utilization, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Aeronautics and Astronautics, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wanghan221@mails.ucas.ac.cn; shyli@csu.ac.cn; zhaomanqi19@csu.ac.cn).

Our source code and dataset SAT-MTB-VSR are available at https://github.com/Alioth2000/RASVSR

Digital Object Identifier 10.1109/JSTARS.2023.3332449

algorithms for it is more challenging. In order to better utilize this temporal information, researchers often introduce frame alignment to remove the effects of object or background movement in video, and the commonly used alignment methods are optical flow-based motion estimation, deformable convolution, etc. There are also many methods that do not perform alignment, and unlike methods with alignment, methods without alignment tend to use 3-D convolution or nonlocal networks to extract and fuse features directly.

Compared with normal videos, first, satellite videos have lower spatial resolution and more lack of texture information, second, the view range of satellite videos is much larger than that of normal videos, and the information density is greater, third, the scales and velocities of the moving objects in satellite videos vary greatly, which all bring more challenges to SR. Therefore, although VSR-based on deep learning has made great development in recent years, these generic methods are not suitable to be directly applied to satellite videos, so there are many recent researches specifically focusing on SVSR. Satellite video also has some advantages that can be utilized in SR, as the scene is more fixed, there is no significant difference between each frame, so more valid information can be utilized for each scene. Making full use of the multiframe information of the video is the key to SVSR, most of the previous SVSR methods utilize the target frame and adjacent 2–6 frames to fuse and reconstruct a high-resolution image, without utilizing the information of more frames. Simply increasing the number of frames will significantly increase the computational complexity, making it difficult for these methods to utilize more information from the video.

Image or feature alignment of neighboring frames is a key and difficult problem in SVSR. Due to the movement of ground objects, such as vehicles, airplanes, ships, and the change of view angle brought by the movement of the satellite itself, the direct fusion will introduce errors, which will result in the degradation of the effect. Therefore, most current methods introduce alignment operations, which help to accurately find missing information in neighboring frames. There are two types of alignment operations: image alignment and feature alignment. Image alignment is usually performed using optical flow methods, which have better results for motions of small scale, such as the movement of ground objects, but are not as effective for large scale changes, such as background movement. Feature alignment now usually uses deformable convolution network (DCN) [13] and has achieved very good performance [14], but the training of DCN is unstable [15]. Neither optical flow nor deformable convolution is directly suitable for satellite videos due to the fact that satellite videos have both moving ground objects and moving background, as well as their motion features are even less distinct.

Although SVSR introduces an alignment operation, the error of alignment cannot be eliminated, and at the same time, due to the object motion in the timing of the different regions of the occlusion, it will cause the lack of information in the relevant region, the direct fusion of multiframe information is prone to poor results, and may even be fused into the wrong information resulting in the degradation of the effect [16]. In addition, many of the current SR methods use a large number of parametric neural networks, allowing the neural network to memorize the mapping from low to high resolution, which makes the computational complexity high, and does not make full use of the advantages of the video to enhance the use of the available information.

To solve the above problems, we propose recurrent aggregation network for satellite video superresolution (RASVSR), a concise and lightweight recurrent neural network that focuses on sufficiently aggregating the information in the video to achieve good SR results with very few parameters. We use a bidirectional recurrent network to propagate the feature extracted by neutral network of the entire video sequence, so that the reconstruction of each frame can utilize the information of all frames. We found in our study that the length of the sequence utilized in SVSR has a significant effect on the results, and we believe that with good alignment and fusion, the use of longer sequences means that more information can be utilized. In order to fuse feature from different frames in the propagation, we propose temporal feature fusion module (TFF), an approach based on an attentional mechanism that enhances the utilization of critical information and reduces the impact caused by erroneous information. At the same time, the attention mechanism captures the key information in the image and can generate images that are more compatible with human visual perception [17]. We also use a DCN-based alignment method, but incorporate an optical flow method to assist DCN in order to address the difficulty of DCN training.

The main contributions of this article are as follows.

1) A lightweight bidirectional recurrent neural network, RASVSR, is proposed for SVSR, which achieves good performance with very few parameters by propagating and aggregating information across the entire video sequence, and we demonstrate experimentally that longer sequences can achieve better SR results.
2) We propose TFF module for extracting valid information and eliminating erroneous information as features propagate through the sequence. And we introduce a feature alignment method that combines the optical flow method and DCN to achieve better alignment and thus be able to utilize longer sequence information.
3) We create a dataset for SVSR, SAT-MTB-VSR, which is a subset of SAT-MTB [2], produced from the raw video of Jilin-1, covering a wide range of scenes, and is currently the largest publicly available dataset in the field.

The rest of this article is organized as follows. In Section II, we introduce the existing works related to VSR and SVSR. The details of our proposed method will be presented in Section III. In Section IV, we present our dataset, implementation details and experimental results, and analyze the results. Finally, Section V concludes this article.

II. Related Work

A. Video Super Resolution

Many methods for VSR have been proposed so far, including traditional methods and deep learning-based methods. Schultz and Stevenson [18] proposed a novel observation model based

on motion compensated subsampling, which estimates motion by affine modeling. Liu and Sun [19] proposed a Bayesian approach to achieve an adaptive VSR by simultaneously estimating the motion, blur kernel, and noise level. Ma et al. [20] proposed an EM framework to guide residual blur estimation and high-resolution image reconstruction. However, the results of these traditional methods are still hardly satisfactory and have been largely replaced by deep learning methods. Since superresolution convolutional neural networks (CNN) [21] first used deep learning in the field of SR, a large number of deep learning-based image and VSR methods have emerged, and CNN, adversarial generative networks (GAN), recurrent neural networks (RNN), etc., have been widely used for SR. These methods can be broadly categorized into two groups: methods without alignment and methods with alignment. Representative methods without alignment include dynamic upsampling filters (DUF) [22], which is able to generate DUF and a residual image to avoid explicit motion compensation, ensuring the temporal consistency of the reconstructed image. In addition, RBPN [23] integrated the spatial and temporal contexts of consecutive video frames using a recurrent encoder–decoder module to fuse multiframe information into conventional SISR. Representative methods with alignment include TDAN [14], which first introduced DCN into VSR to calculate offsets between target and neighboring frames, and warp the neighboring frames according to the offsets to align them with the target frames. EDVR [24] goes a step further by invoking DCN in a multiscale way to achieve a more accurate alignment. Basic VSR and IconVSR proposed by Chan et al. [25] advanced in both speed and reconstruction quality with a more concise network and proposed to divide the VSR into four steps: propagation, alignment, aggregation, and upsampling. The subsequent BasicVSR++ [26] achieves another improvement in performance by enhancing the propagation and alignment operations.

B. Satellite Video Super Resolution

SVSR techniques appeared relatively late and are still in their infancy, but researchers have proposed many deep learning-based methods. Alignment using DCN is a very common operation in SVSR, Zhang et al. [27] is one of the earliest to utilize multiframe images of satellite video for SR. They employs a combination of a single-frame and multiframe network, where the multiframe network comes from the classical generic VSR network EDVR, which utilizes a DCN for feature alignment. The method proposed by Ni et al. [28] also uses DCN for alignment and proposes a scale-adaptive feature extraction module as well as an upsampling module that enables arbitrary magnification. Xiao et al. [29] proposed a novel fusion strategy for temporal grouped projections as well as a DCN-based multiscale residual alignment module, and Xiao et al. [30] achieved simultaneous SR for both time and space in a single network that predicts unknown frames by coupling optical flow and multiscale deformable convolution. The method of He et al. [31], on the other hand, uses the optical flow method for alignment, where the image is first upsampled before going through an attention-based residual network to obtain the final high-resolution image. There are also methods that use dual-stream networks, Liu and Gu's method [32] has two subnetworks, one branch predicts the high-resolution image and the other predicts the blur kernel and is coupled by a cross-task feature fusion module. Its alignment is based on patch matching in the feature space, which is more stable than using the optical flow. The method proposed by Shen et al. [33] adds an edge branch to EDVR that will simultaneously predict high-resolution edge maps and the features of both branches are fused at the end of the network.

The authors in [34] and [35] also propose SVSR methods without alignment which directly use 3-D convolution for feature extraction and fusion. The authors proposed a network with arbitrary SR scale in [34], which achieves image upsampling by subpixel convolution and Bicubic. He et al. [35] split the objective function of the degenerate model into two suboptimization problems and propose to fuse deep learning and model-based approaches for SVSR for the first time. In addition to this, there are unsupervised learning approaches in the field of SVSR [36], which consists of a down-sampling network and an up-sampling network and does not require LR-HR training pairs. Wang and Sertel [37], on the other hand, apply GAN to SVSR and introduced an attention module to improve the generative capability.

III. Proposed Methods

A. Overall Architecture of RASVSR

Fig. 1 shows the overall architecture of our network, which is a bidirectional recurrent neural network that consists of three main parts, a feature extraction part consisting of a residual network [38], an information propagation and aggregation part consisting of an alignment module, a residual network, and a TFF module, and finally a reconstruction part for outputting high-resolution images.

Feature extraction: The feature extraction part consists of five residual blocks of the same structure, each containing two 3×3 convolutional layers and a ReLU activation layer, which does not contain a BN layer. In this part, we only extract the shallow features and do not downsample. The feature map is noted as f_t^1

$$f_t^1 = \mathrm{RB}_5\left(x_t\right) \tag{1}$$

where t represents the tth frame in the video sequence, RB_5 represents five residual blocks, and the extracted feature map is $f_t^1 \in \mathbb{R}^{C\times H\times W}$, $C = 64$. One of the reasons why our method is lightweight enough is that the feature extraction part does not employ a deeper network, and experimentally we found that extracting deeper features is not necessary in SVSR. Also, the channel of the feature map is 64, which significantly reduces the number of parameters. *Information propagation and aggregation:* There are two stages in the information propagation and aggregation part, one stage passes the information forward along time and the other stage passes the information backward, which allows the perceptual field of each frame to be expanded to the whole sequence. In this stage, there is also no downsampling, and the feature maps obtained in the forward and backward directions are f_t^2 and $f_t^3 \in \mathbb{R}^{C\times H\times W}$, respectively. Also, we use second-order propagation, which means that f_t^2 and f_t^3 are

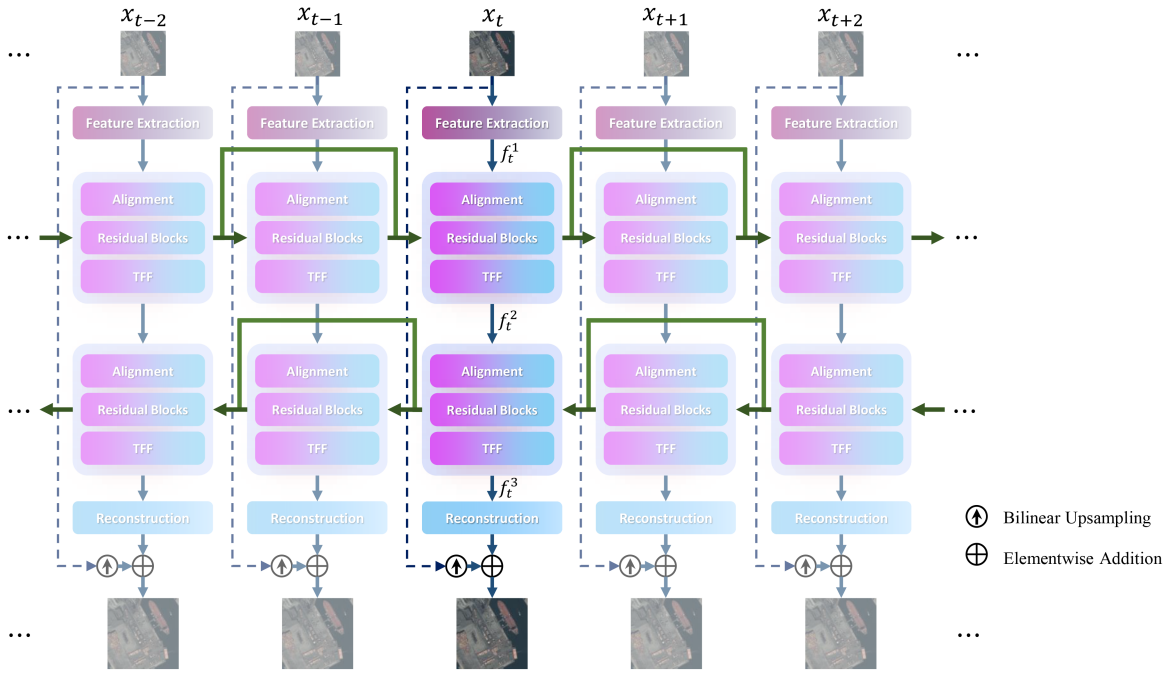


Fig. 1. Overall network structure of RASVSR.

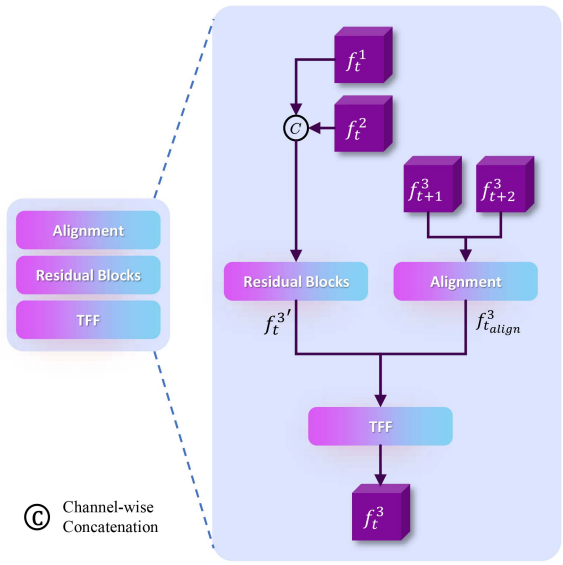


Fig. 2. Data stream of propagation and aggregation part during backward feature propagation. This section consists of three modules, which are residual blocks, alignment module, and TFF module.

passed forward and backward twice. Fig. 2 shows the specific structure and information flow of the backward propagation stage of this part, and the forward propagation stage is similar. First, the residual blocks are used to further extract the features, the backward phase is

$$f_t^{3\prime} = \mathrm{RB}_7\left(C\left(f_t^1, f_t^2\right)\right) \tag{2}$$

where C represents the concatenation operation in the channel dimension, and after concatenation the number of its input channel is 128. The forward propagation phase has no concatenation, and the output feature maps are all of 64 channels, for the forward propagation phase

$$f_t^{2\prime} = \mathrm{RB}_7\left(f_t^1\right). \tag{3}$$

In addition to this, the network structure is the same for both forward and backward propagation, and the following are all examples of backward propagation. The features of $t+1$ and $t+2$ frames are aligned to obtain $f_{t_{\mathrm{align}}}^3 \in \mathbb{R}^{C\times H\times W}$ in backward propagation

$$f_{t_{\mathrm{align}}}^3 = \mathrm{Align}\left(f_{t+1}^3, f_{t+2}^3\right) \tag{4}$$

where Align represents the alignment operation. Finally, the final feature map of the stage are obtained after fusing the features by the TFF module

$$f_t^3 = \mathrm{TFF}\left(f_t^{3\prime}, f_{t_{\mathrm{align}}}^3\right). \tag{5}$$

Reconstruction: In the reconstruction part, the feature maps f_t^1, f_t^2, and f_t^3 are all used to reconstruct the high-resolution image by convolution and pixel shuffle, and we use $2\times$ of pixel shuffle to achieve a $4\times$ SR.

B. TFF Module

The performance enhancement of RASVSR relies heavily on improvements in feature fusion. Unlike other previous VSR methods that also employ RNN [16], [25], [26], [39], we propose a temporal and spatial attention-based feature fusion mechanism to take advantage of RNN's strength in feature propagation. Due to the effects caused by occlusion, motion blur, change in view angle, and change in illumination in satellite video, direct feature fusion after alignment is also ineffective. The weights redistributed by the attention mechanism can help the network to pick up the focused information and cut down the erroneous

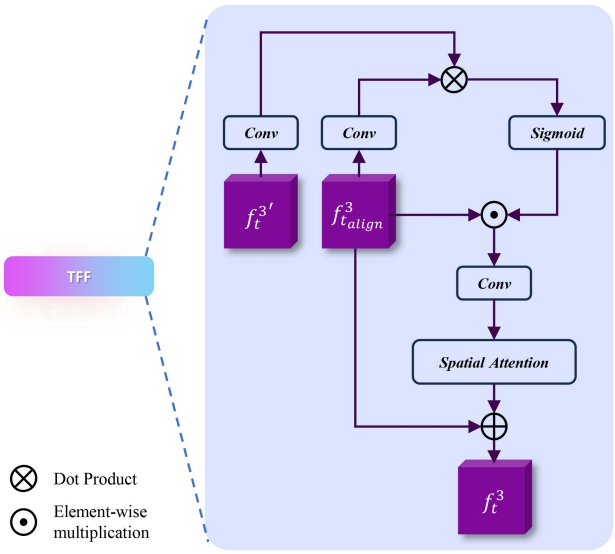


Fig. 3. Structure of TFF module.

information, which can lead to the effective utilization of longer sequences. Due to the effects caused by occlusion, motion blur, change in view angle, and change in illumination in satellite video, direct feature fusion after alignment is also ineffective. The weights redistributed by the attention mechanism can help the network to pick up the focused information and cut down the erroneous information, which can lead to the effective utilization of longer sequences.

The specific structure of the TFF module is shown in Fig. 3, first, we introduce the attention mechanism in the time dimension. For the feature map $f_t^{3'}$ extracted from the current frame and the feature map $f_{t_{\text{align}}}^3$ obtained from the alignment of the subsequent frames, their similarity distances after embedding by convolution are first calculated

$$d\left(f_t^{3'}, f_{t_{\text{align}}}^3\right) = \text{Sigmoid}\left(\text{Conv}_1\left(f_t^{3'}\right) \otimes \text{Conv}_2\left(f_{t_{\text{align}}}^3\right)\right) \tag{6}$$

where Conv stands for convolution operation and $\otimes$ for dot product. After that the aligned feature map is multiplied pixel by pixel with the similarity distance and passed through a convolutional layer to get the final temporal attention processed feature map

$$f_{t_{a1}}^3 = \text{Conv}_3\left(f_{t_{\text{align}}}^3 \odot d\left(f_t^{3'}, f_{t_{\text{align}}}^3\right)\right) \tag{7}$$

where $\odot$ stands for pixel-by-pixel multiplication.

The fused feature $f_{t_{a1}}^3$ is also processed using spatial attention, which is constructed in the same way as in [40] to enhance information, such as texture. The final obtained feature map is

$$f_t^3 = \text{SA}\left(f_{t_{a1}}^3\right) + f_{t_{\text{align}}}^3 \tag{8}$$

where SA stands for spatial attention operations.

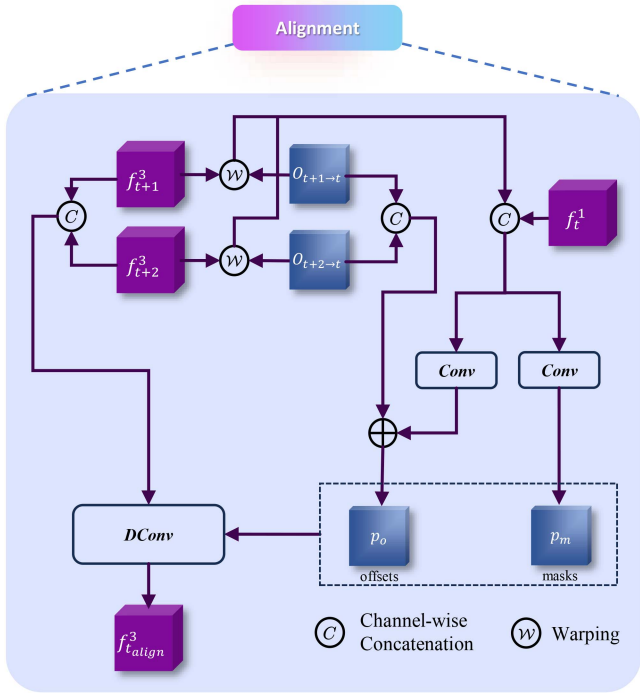


Fig. 4. Structure of alignment module.

C. Alignment Module

Alignment is very important for VSR and RASVSR performs alignment operations at the feature level and uses both optical flow and DCN methods. A very lightweight deep learning method SpyNet [41] is used in our method to compute optical flow maps, the optical flow network is used to compute offsets on the RGB image, we use $O_{t+1\to t}$ to represent the optical flow map from frame t+1 to frame t. In order to reduce the computation amount, our method only computes the optical flow maps of the two adjacent frames, the second order optical flow maps, such as $O_{t+2\to t}$, is then obtained by warping the first order optical flow maps.

However, the optical flow maps are not directly used for warping images or features, but assist the deformable convolution for feature alignment, and the deformable convolution we use is DCNv2 [42]. As shown in Fig. 4, its offsets are obtained by adding the offsets provided by the optical flow and the computed residuals of the feature map after warping. The optical flow is first used to prealign the first-order and second-order propagation of the feature maps

$$\widehat{f_{t+1}^3} = \mathcal{W}\left(f_{t+1}^3, O_{t+1\to t}\right) \tag{9}$$

$$\widehat{f_{t+2}^3} = \mathcal{W}\left(f_{t+2}^3, O_{t+2\to t}\right) \tag{10}$$

where $\mathcal{W}$ represents the warping operation. In order to reduce the computational effort, the alignment operations of the first-order and second-order features are performed together at the same time, which is realized by concatenating them in the channel dimension. The prealigned feature maps are convolved to compute the residuals of the optical flow and the masks of the DCN,

so that the offsets and masks of the DCN are, respectively

$$p_o = C\left(O_{t+1\to t}, O_{t+2\to t}\right) + \text{Conv}_o\left(C\left(f_t^1, \widehat{f_{t+1}^3}, \widehat{f_{t+2}^3}\right)\right) \tag{11}$$

$$p_m = \text{Conv}_m\left(C\left(f_t^1, \widehat{f_{t+1}^3}, \widehat{f_{t+2}^3}\right)\right). \tag{12}$$

Finally, the aligned features are obtained by DCN

$$f_{t_{\text{align}}}^3 = \text{DConv}\left(C\left(f_{t+1}^3, f_{t+2}^3\right) \mid p_o, p_m\right) \tag{13}$$

where DConv is deformable convolution with 128 input channels and 64 output channels.

D. Reconstruction Module

The reconstruction part utilizes feature maps f_t^1, f_t^2, and f_t^3. First, three feature maps are concatenated and then further fused and extracted through the residual blocks to get the final features

$$f_t = \text{RB}_5\left(C\left(f_t^1, f_t^2, f_t^3\right)\right) \tag{14}$$

where $f_t \in \mathbb{R}^{C\times H\times W}$ remains at 64 channels. Then, the number of channel is increased and the size is raised by two convolutions and two pixel shuffles

$$f_{t(\text{HR})} = \text{PS}\left(\text{Conv}_{R2}\left(\text{PS}\left(\text{Conv}_{R1}\left(f_t\right)\right)\right)\right) \tag{15}$$

where PS stands for pixel shuffle operation. Here, convolution doubles the number of channels and pixel shuffle halves the number of channels and doubles the feature map size. Finally, two more convolutions are performed to obtain the residual $\bar{y}_t$ of the high-resolution image with channel number of 3. Except for the last convolution, all the convolutions here use the Leaky ReLU activation function. Then the final predicted high resolution image $\widehat{y}_t$ for frame t is

$$\widehat{y}_t = \bar{y}_t + \text{Bicubic}\left(x_t\right) \tag{16}$$

where Bicubic stands for bicubic interpolation. The original image is interpolated to improve the resolution by four times and then added with $\bar{y}_t$ to get the final predicted image.

E. Loss Function

We chose Charbonnier Loss [43] as our loss function, which can better handle outliers and performs better than L2 loss on SR tasks [44]. Its computational formula is

$$\mathcal{L}\left(y_t, \widehat{y}_t\right) = \sqrt{\left\|y_t - \widehat{y}_t\right\|^2 + \epsilon^2} \tag{17}$$

where y_t is the ground truth of the high resolution image and ϵ is a constant which we set to 10^{-6}.

IV. Experiments

A. Dataset

Due to the lack of a publicly available large-scale dataset in the field of SVSR, we use the original videos of Jilin-1 to produce a satellite video dataset SAT-MTB-VSR and make it publicly available, which is a subset of the satellite video multitasking dataset SAT-MTB. The dataset is cropped from 18 videos captured by the Jilin-1 video satellite, covering a wide range of terrains, such as cities, docks, airports, suburbs, forests, and deserts, with a resolution of about 1 m. And the videos contain dynamic scenes, such as moving cars, airplanes, trains, and ships, which test the ability of the VSR method to deal with moving targets of different sizes and speeds. At the same time, due to the motion of the satellite, the video contains changes in viewing angle and lighting.

We crop out 431 videos, each of which is 100 consecutive frames, of which 413 are used as the training set and 18 as the validation set, and all the 18 validation sets are from different original videos, while the size of the images is 640×640. These images are downsampled $4\times$ by bicubic interpolation to get 160×160 low-resolution images, thus obtaining the LR-HR training pairs. Some examples of the dataset are shown in Fig. 5

B. Implementation Details

Our approach is built using Pytorch and borrows code from the open-source toolkit BasicSR [45]. We use 2 Nvidia Titan RTX GPUs to train our network, utilizing a total of 48 GB GPU memory. For training, we set the batch size to 4, the length of the video sequences to 60, and all images are randomly cropped to 256×256 for training, as well as using image flipping and rotating for data enhancement. Adam [46] is used as the optimizer for training, and the learning rate is set to 10^{-4}, while we also use the Cosine Annealing Warm Restart [47] strategy to adjust the learning rate. The training is set to take a total of 10^5 iterations, which is about ten epochs, and completing a full training takes about three days on our device. The number of epochs for training is less because the video data itself is more redundant. For the optical flow network SpyNet, we use pretrained weights whose parameters are fixed for the first 5×10^3 iterations of training, after which they are optimized with a 0.25x learning rate.

When evaluating the quality of SR images, we mainly use PSNR and SSIM as metrics. PSNR is calculated based on the mean square error, which measures the ratio of the maximum possible power of the signal to the power of the noise that affects the quality of the signal, the higher the value the better. SSIM measures the degree of similarity of images, which is based on a perceptual model, with higher values being better. In addition, we use the no-reference image metric NIQE [48], which constructs a set of features for measuring image quality and calculates the difference in the distribution of images over them, with lower values being better for image quality.

C. Ablation Studies

Ablation studies on the SAT-MTB-VSR dataset have been performed to demonstrate the correctness and effectiveness of our method and concept. First, we analyze the effectiveness of the alignment module and the TFF module, which are the core of our method to be able to exploit long sequences. Then, we validate the advantages of utilizing long sequences in SVSR, our method benefits from being able to utilize the whole sequence information for SR of each frame, so that only few parameters and a small amount of computation are required.

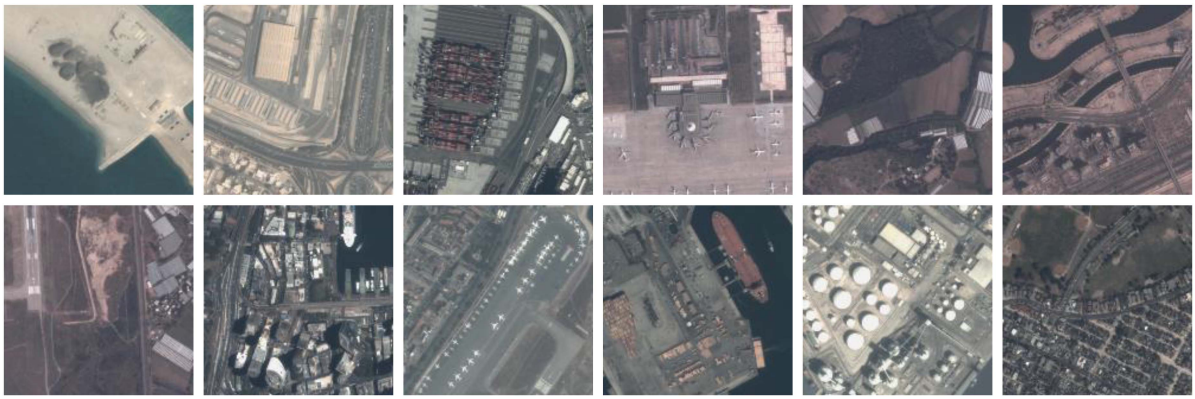


Fig. 5. Examples of SAT-MTB-VSR dataset.

TABLE I
ABLATION STUDY OF ALIGNMENT MODULE AND TFF MODULE

Alignment	TFF	PSNR	SSIM
✗	✓	23.345	0.6189
✓	✗	38.162	0.9511
✓	✓	**39.930**	**0.9670**

The bold entities indicate the best result.

TABLE II
ABLATION STUDY ON USING DIFFERENT SEQUENCE LENGTH AND BATCH SIZE FOR TRAINING

Sequence length	batch size	PSNR	SSIM
15	16	37.217	0.9459
30	8	39.291	0.9635
60	4	**39.930**	**0.9670**

The bold entities indicate the best result.

Effectiveness of alignment module and TFF module: The results of the ablation study are shown in Table I. Alignment operations are necessary and important in VSR, and for RNN methods that utilize the entire video, not performing alignment will accumulate too much error information, leading to complete unavailability. We attempted to remove the alignment module completely, and the training became extremely unstable, with the PSNR reaching a maximum of only 23.345 dB, which is even far worse than the SISR method. This is because RASVSR is very sensitive to the alignment operation, as longer input sequences imply greater image variations. Benefiting from the good alignment of the optical flow plus DCN, the utilization of longer sequences is achieved, which would otherwise lead to the continuous accumulation of error information, resulting in a serious degradation of the results. The enhancement of feature aggregation by the TFF module brings further performance improvement, we replace the TFF module with three residual blocks for the aggregation of timing information, and the PSNR and SSIM drop to 38.162 and 0.9511 dB, respectively, which means that the TFF module delivers a PSNR improvement of 1.768 dB and an SSIM improvement of 0.0159 dB.

Effectiveness of long sequence usage for training: The biggest difference between RASVSR and other current SVSR methods is its ability to utilize information from longer sequences. Most of the previous SVSR methods only utilize 3–7 frames of images to SR one image, whereas our method can achieve utilization of all 100 frames of our video. Experiments have demonstrated that increasing the length of the training and testing sequences has a positive effect on the results of SVSR. The first is the sequence length used for training. There is a limit to the number of images in a batch due to the memory constraints of the GPU, so there is a tradeoff between batch size and sequence length. We experimented with different batch sizes and sequence lengths,

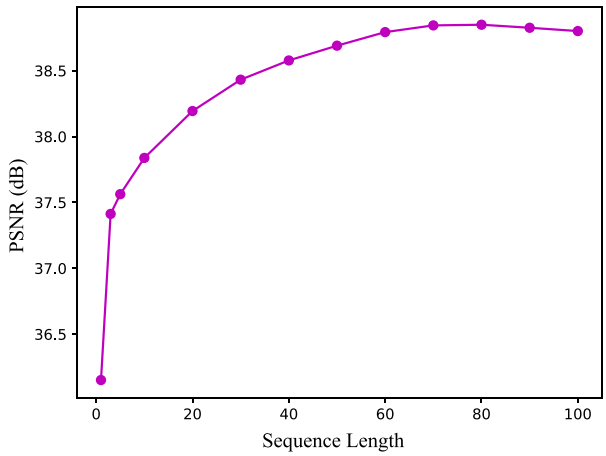


Fig. 6. PSNR of the first frame for inference using sequences of different lengths.

and all 100 frames were input for testing. The results are given in Table II, where we find that increasing the sequence length is more effective than increasing the batch size. The models trained using 15 and 30 frame sequences do not perform as well as 60 frames. We believe that because there is not enough sequence length, the model is unable to learn how to aggregate information from long sequences, which results in the model not aligning and fusing features well enough when using 100 frames of sequences for inference in the test.

Effectiveness of long sequence usage for inference: In order to prove that the model achieves information aggregation and utilization of long sequences, we use different sequence lengths for inference in our tests, and then calculate the PSNR and SSIM on the first frame of the video. As given in Table III and Fig. 6, there is a significant improvement in both PSNR and SSIM as

TABLE III
ABLATION STUDY ON USING DIFFERENT SEQUENCE LENGTH FOR INFERENCE

Sequence length	1	3	5	10	20	30	40	50	60	70	80	90	100
PSNR	36.152	37.413	37.563	37.838	38.194	38.432	38.578	38.690	38.793	38.844	**38.849**	38.826	38.801
SSIM	0.9360	0.9474	0.9488	0.9511	0.9543	0.9563	0.9576	0.9586	0.9596	0.9601	**0.9602**	0.9601	0.9599

The bold entities indicate the best result.

TABLE IV
COMPARISON WITH OTHER METHODS

Methods	Bicubic	EDVR [24]	MSDTGP [29]	BasicVSR [25]	IconVSR [25]	BasicVSR++ [26]	ours
PSNR (dB)	35.583	37.824	37.827	38.027	38.351	38.780	**39.930**
SSIM	0.9280	0.9533	0.9534	0.9535	0.9566	0.9607	**0.9670**
NIQE	7.7373	7.2666	7.1752	7.1126	7.0827	6.9408	**6.8273**
Frames	1	5	5	100	100	100	100
Params (M)	/	20.6	14.2	6.3	8.7	7.3	**5.4**
Runtime (ms)	/	233	332	**119**	127	136	126
Year	/	2019	2022	2021	2021	2022	2023

The bold entities indicate the best result.

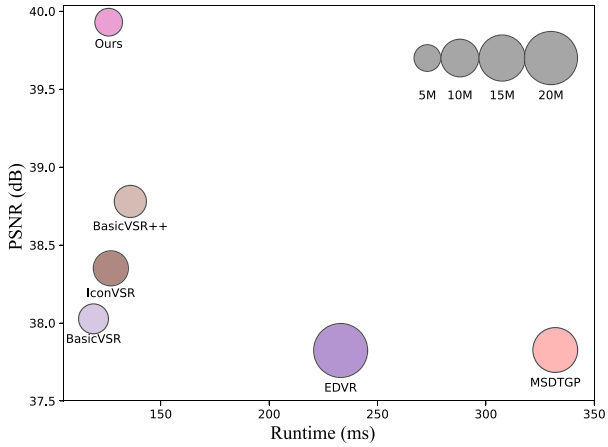


Fig. 7. Speed and performance comparison. The size of the circles indicates the amount of parameters of these models.

the length of the used sequences increases, but there is a drop when the sequence length is over 80, which is due to the fact that even with the TFF and alignment module, the error cannot be completely eliminated, and too long sequences lead to the accumulation of error information, which results in a decrease in effectiveness.

D. Comparison With State-of-the-Art Methods

We compare our method with several representative VSR methods, including EDVR, MSDTGP, BasicVSR, IconVSR, and BasicVSR++, as given in Table IV and Fig. 7, where the runtime is the time to process 1 frame on a single Nvidia Titan RTX. All methods are retrained on the SAT-MTB-VSR dataset and tested on the validation set with 4× SR. All of these methods are designed for video, but the number of frames utilized for SR an image is different, the specific data can be given in Table IV, and since BasicVSR, IconVSR, and BasicVSR++ are also RNN-based methods, they are also able to utilize information from the entire video sequence. EDVR and MSDTGP, on the other hand, only utilize five frames, and at each time each image is processed, they need to reextract the features of the adjacent frames, and such repeated extraction of features greatly increases the computational effort. The RNN-based method extracts features only once for each image, which has a great advantage in terms of speed.

Our method achieves the best in all three metrics, PSNR, SSIM, and NIQE, and it still has the smallest number of parameters, which is only 5.4 M. As a comparison, the EDVR has a parameter of 20.6 M. In terms of PSNR, our method outperforms the second-place BasicVSR++ by 1.15 dB, which is a great advantage. For SSIM, our method is still the highest at 0.9670, which is 0.0063 higher than the second place. On the NIQE metric, which is more in line with human intuition, our method still outperforms the other methods at 6.8273. In terms of speed, our method takes only 126 ms to process 1 frame, which is also advantageous and only slightly slower than BasicVSR.

We also provide a comparison of the images after SR, as shown in Fig. 8. For the airport runway in (a), our method restores the most clear and sharp figure. In (b), our method most clearly and accurately restores the area of white lines. For the white vehicles in the lower right and upper left corners in (c), our method reconstructs an image that can identify the boundaries of the different vehicles instead of blurring them into a single piece, and in some methods, there is an error in the direction of the stripes.

E. Real-World Experiment

Our method demonstrates excellent performance on the SR of satellite videos, but most of the current research and our study are based on manually downsampling images to get simulated low-resolution images to build LR-HR training pairs, not real-world SR. Real-world SR, also called blind SR, i.e., SR of images for which the degradation mode is not known, is a difficult problem in the field of SR. We also tested real-world SR on RASVSR, using it directly on raw satellite video that has not been downsampled. In order for the network to learn high-resolution remote sensing information, we pretrained RASVSR using the AID dataset [49], which is a higher resolution remote sensing image dataset. We found that pretraining using higher resolution remote

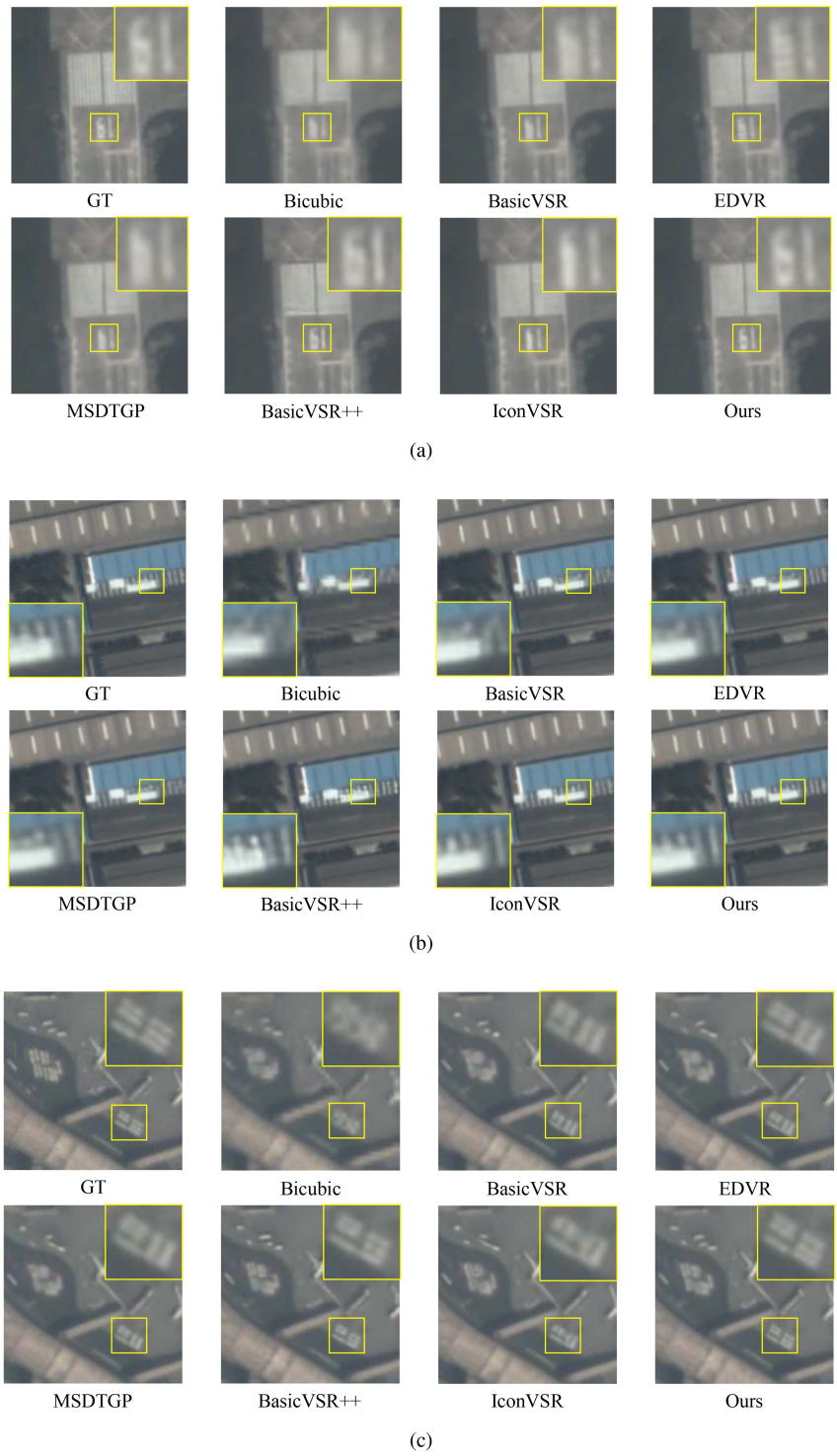


Fig. 8. Visualization of test results on SAT-MTB-VSR dataset.

sensing images can marginally improve performance at real-world SR. Benefiting from the fuller use of video information, our method also has better performance at real-world SR, as shown in Fig. 9, where our method yields the smoothest image with the cleanest colors and clearer vehicle boundaries in this parking lot scene. However, there is still a long way to go for real-world SVSR, and we believe that the most important thing is to innovate in the training data and training method of the network, although our method can obtain more information on the video, but it is hard to learn the way of image degradation in the real world, and there is still a lot of room for improvement.

F. Discussion

The continuous imaging of fixed scenes by satellite video does facilitate SR, and with good alignment and fusion

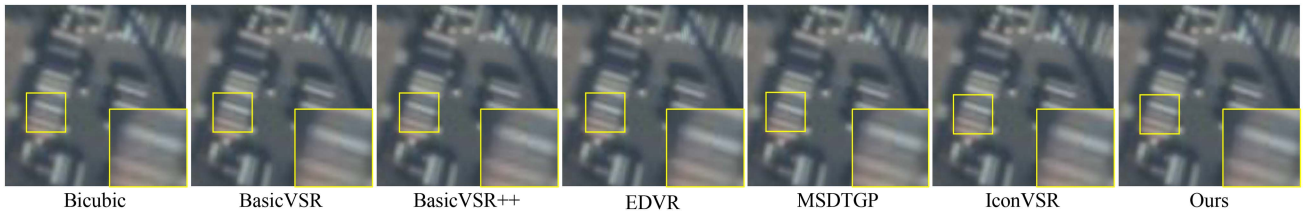


Fig. 9. Visualization of real-world experienment.

operations there is a richer set of information to utilize. Using longer sequences for SR as much as possible can bring more improvements at less cost, which has often been ignored in the past. However, there are some drawbacks to using RNN, such as the fact that the training and inference of the network takes up a lot of video memory, even if we have a small number of network parameters, which leads to higher device requirements to run the network. In addition, we believe that the SR effect in a real environment is still the first issue that needs to be considered for SVSR afterward. Our method can be improved in terms of training approach, training data and network structure to accommodate real-world SR.

V. Conclusion

This article proposes a deep learning-based SR method for satellite video, RASVSR, which is designed for the characteristics of satellite video scenes that are fixed and possess temporal information, and focuses on achieving the aggregation and utilization of long sequence information. First, we propose a bidirectional RNN-based information propagation structure so that the features of each image frame can be propagated to the whole sequence after feature extraction. We then propose an optical flow and DCN-based alignment method and a TFF module to enable features to be aligned and properly utilized in propagation, reducing the accumulation of misinformation. We also release a SVSR dataset for subsequent studies. Sufficient experiments have proved that utilizing the information on more frames has positive significance for SVSR, and thanks to the RNN-based network structure and specially designed information aggregation method, RASVSR achieves the best SR effect with very few parameters and computations. In the future, improving the SR effect of deep learning methods for real-world satellite video is a problem worth exploring.

References

[1] S. Li et al., "Recent advances in intelligent processing of satellite video: Challenges, methods, and applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6776–6798, Jul. 2023.

[2] S. Li et al., "A multi-task benchmark dataset for satellite video: Object detection, tracking, and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–21, May 2023.

[3] C. Xiao et al., "DSFnet: Dynamic and static fusion network for moving object detection in satellite videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Oct. 2021.

[4] M. Zhao, S. Li, S. Xuan, L. Kou, S. Gong, and Z. Zhou, "SatSOT: A benchmark dataset for satellite video single object tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, Jan. 2022.

[5] S. Liu, P. Chen, and M. Woźniak, "Image enhancement-based detection with small infrared targets," *Remote Sens.*, vol. 14, no. 13, pp. 1–19, 2022.

[6] J. Yoo, S.-H. Lee, and N. Kwak, "Image restoration by estimating frequency distribution of local patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6684–6692.

[7] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 372–386.

[8] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2020.

[9] 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.

[10] C. Deng, X. Luo, and W. Wang, "Multiple frame splicing and degradation learning for hyperspectral imagery super-resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8389–8401, Sep. 2022.

[11] J. Hu, Y. Tang, and S. Fan, "Hyperspectral image super resolution based on multiscale feature fusion and aggregation network with 3-D convolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5180–5193, Sep. 2020.

[12] H. Liu et al., "Video super-resolution based on deep learning: A comprehensive survey," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 5981–6035, 2022.

[13] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[14] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3360–3369.

[15] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Understanding deformable alignment in video super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 973–981.

[16] B. N. Chiche, A. Woiselle, J. Frontera-Pons, and J.-L. Starck, "Stable long-term recurrent video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 837–846.

[17] X. Liu, S. Chen, L. Song, M. Woźniak, and S. Liu, "Self-attention negative feedback network for real-time image super-resolution," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, pp. 6179–6186, 2022.

[18] 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.

[19] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[20] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu, "Handling motion blur in multi-frame super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5224–5232.

[21] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[22] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3224–3232.

[23] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3897–3906.

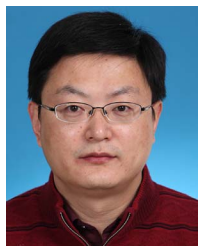
[24] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1954–1963.

[25] K. C. 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.*, 2021, pp. 4947–4956.

[26] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR: Improving video super-resolution with enhanced propagation and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5972–5981.
[27] S. Zhang, Q. Yuan, and J. Li, "Video satellite imagery super resolution for 'JILIN-1'via a single-and-multi frame ensembled framework," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2731–2734.
[28] N. Ni, H. Wu, and L. Zhang, "Deformable alignment and scale-adaptive feature extraction network for continuous-scale satellite video super-resolution," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 2746–2750.
[29] 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, pp. 1–19, Sep. 2021.
[30] 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, 2022, Art. no. 102731.
[31] 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, pp. 1–15, May 2021.
[32] H. Liu and Y. Gu, "Deep joint estimation network for satellite video super-resolution with multiple degradations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, Mar. 2022.
[33] 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, pp. 1–14, Oct. 2021.
[34] 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.
[35] Z. He, X. Li, and R. Qu, "Video satellite imagery super-resolution via model-based deep neural networks," *Remote Sens.*, vol. 14, no. 3, p. 749, 2022.
[36] 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.*, vol. 19, pp. 1–5, Dec. 2020.
[37] P. Wang and E. Sertel, "Multi-frame super-resolution of remote sensing images using attention-based GAN models," *Knowl.-Based Syst.*, vol. 266, 2023, Art. no. 110387.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[39] D. Fuoli, S. Gu, and R. Timofte, "Efficient video super-resolution through recurrent latent space propagation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 3476–3485.
[40] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 606–615.
[41] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4161–4170.
[42] 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.*, 2019, pp. 9308–9316.
[43] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. 1st Int. Conf. Image Process.*, 1994, pp. 168–172.
[44] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.
[45] X. Wang, L. Xie, K. Yu, K. C. Chan, C. C. Loy, and C. Dong, "BasicSR: Open source image and video restoration toolbox," 2022. Accessed: Jan. 5, 2023. [Online]. Available: https://github.com/XPixelGroup/BasicSR
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv:1412.6980.
[47] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–16.
[48] 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.
[49] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

Han Wang received the B.E. degree in electrical engineering from Chongqing University, Chongqing, China, in 2022. He is currently working toward the Ph.D. degree in computer applied technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision, with a focus on applications of super-resolution and object tracking.



Shengyang Li received the Ph.D. degree in computer application technology from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently a Professor with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. His research interests include machine learning in remote sensing image interpretation, deep learning in satellite videos processing and analysis, intelligent image processing, analysis and understanding for space utilization, and space scientific Big Data modeling and analysis.

Manqi Zhao received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2019. He is currently working toward the Ph.D. degree in computer applied technology with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing.

His research interests include satellite video, unmanned aerial vehicle video, and conventional video analysis, with focus on object tracking.