International Journal of Applied Earth Observations and Geoinformation 108 (2022) 102731

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer

Yi Xiao [a], Qiangqiang Yuan [a,*], Jiang He [a], Qiang Zhang [b], Jing Sun [c], Xin Su [d], Jialian Wu [a], Liangpei Zhang [b]

[a] School of Geodesy and Geomatics, Wuhan University, Wuhan, China
[b] State Key Laboratory of Information Engineering, Survey Mapping and Remote Sensing, Wuhan University, Wuhan, China
[c] School of Resource and Environmental Sciences, Wuhan University, Wuhan, China
[d] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

ARTICLE INFO

Keywords:
Video super-resolution
Video frame interpolation
Space–time upsampling
Jilin-1 satellite video
Deep learning

ABSTRACT

Satellite video is an emerging type of earth observation tool, which has attracted increasing attention because of its application in dynamic analysis. However, most studies only focus on improving the spatial resolution of satellite video imagery. In contrast, few works are committed to enhancing the temporal resolution, and the joint spatial-temporal improvement is even less. The joint spatial-temporal enhancement can not only produce high-resolution imagery for subsequent applications, but also provide the potentials of clear motion dynamics for extreme events observation. In this paper, we propose a joint framework to enhance the spatial and temporal resolution of satellite video simultaneously. Firstly, to alleviate the problem of scale variation and scarce motion in satellite video, we design a feature interpolation module that deeply couples optical flow and multi-scale deformable convolution to predict unknown frames. Deformable convolution can adaptively learn the multi-scale motion information and profoundly complement optical flow information. Secondly, a multi-scale spatial-temporal transformer is proposed to aggregate the contextual information in long-time series video frames effectively. Since multi-scale patches are embedded in multiple heads for spatial-temporal self-attention calculation, we can comprehensively exploit multi-scale details in all frames. Extensive experiments on the Jilin-1 satellite video demonstrate that our model is superior to the existing methods. The source code is available at https://github.com/XY-boy.

1. Introduction

Nowadays, several satellites with video sensors are launched into space, which dramatically improves the temporal resolution of regional observations and provides a novel data application scenario for dynamic event monitoring. In fields like vehicle tracking (Feng et al., 2021) and flood disaster monitoring (de Alwis Pitts and So, 2017), the instantaneity and continuity of video data are incomparably superior to traditional static images. However, on the one hand, the spatial resolution of satellite video is not comparable to traditional satellite imagery. In remote imaging procedure, the spatial resolution will suffer from degradation like atmospheric scattering and extreme weather (He et al., 2021; Lanaras et al., 2018). These factors have caused a bottleneck in the improvement of spatial resolution. On the other hand, the temporal resolution of satellite video is hindered by the limited bandwidth of the sensor. Within these limitations, high quality satellite video data is urgently needed in remote sensing field (He et al., 2022; Zhang et al., 2020, 2021a). A high spatial resolution remote sensing image with rich detailed information is of great significance for subsequent application (Chen et al., 2021), such as semantic segmentation (Li et al., 2022; Ma et al., 2021; Peng et al., 2021), environmental parameters estimation (Wang et al., 2021, 2022; Zhang et al., 2021b), and land cover classification (Abid et al., 2021; Amato et al., 2021). A slow-motion video with high temporal resolution provides clear motion dynamics, which is beneficial for us to analyze the evolution of extreme and transient events (Vandal and Nemani, 2021). Therefore, it is worthwhile to improve the spatial and temporal resolution of satellite video for its broad application prospects.

Space-time video super-resolution (STVSR) aims to improve the spatial and temporal resolution of the video simultaneously (Kang et al.,

* Corresponding author.
E-mail address: yqiang86@gmail.com (Q. Yuan).

https://doi.org/10.1016/j.jag.2022.102731
Received 4 January 2022; Received in revised form 27 January 2022; Accepted 19 February 2022
Available online 25 February 2022

2020; Shi et al., 2021; Xu et al., 2021). With the development of convolutional neural networks, video space super-resolution (S-SR) and time super-resolution (T-SR) have achieved remarkable performance. So a simple idea to implement STVSR is to perform T-SR and S-SR sequentially (two-stage). However, temporal interpolation and spatial enhancement are intrinsically related and may complement each other. Separate processing makes it hard to take advantage of this property. In addition, the two-stage manner usually has a huge number of parameters and is inconvenient to deploy. One-stage means a joint improvement in spatial and temporal resolution, thus T-SR and S-SR are intrinsically facilitated. Although some one-stage STVSR methods (Dutta et al., 2021; Haris et al., 2020) have been proposed, they either involve a huge amount of parameters or fail to take into account the characteristics of satellite video imagery. Until now, the study on STVSR for satellite video is still in its infancy.

At present, related work for video super resolution can be divided into three categories:

(1) Space super-resolution: S-SR is mainly divided into single image super-resolution (SISR) and video super-resolution (VSR).

SISR recovers a high-resolution (HR) image from its corresponding low-resolution (LR) image (Shen et al., 2020). The pioneering SISR method based on deep learning is proposed by Dong et al. (2015), which uses convolutional neural network to fit the nonlinear mapping between LR space and HR space. Subsequently, Lim et al. (2017) proposed an enhanced deep super-resolution network (EDSR) to introduce both residual learning and sub-pixel convolution. Zhang et al. (2018) employed the channel attention mechanism into SISR and proposed a deep residual channel attention network (RCAN). Recently, Liang et al. (2021) came up with a model named SwinIR for SISR based on the Swin Transformer rather than convolutional neural networks.

VSR recovers one or multiple HR frames from multiple LR frames. Compared with SISR, the modeling of the spatial–temporal relationship between frames is essential to VSR. A class of typical methods (Caballero et al., 2017; Haris et al., 2019; Wang et al., 2020) use the optical flows encoded with motion information to warp the adjacent frames for explicit spatial–temporal relationship modeling. Another type of implicit alignment can be realized by deformable convolution (DConv) (Tian et al., 2018; Wang et al., 2019), which can realize adaptive alignment at the feature level. In the study of satellite video, Xiao et al. (2021) designed a multi-scale deformable convolution to achieve multi-scale motion information alignment. Currently, the research on satellite VSR is still at its primary stage.

(2) Time super-resolution: T-SR requires predicting frames that do not exist between two original frames that already exist. The mainstream methods can be broadly divided into two kinds: optical flow-based and kernel-based. The optical flow-based method (Bao et al., 2019a,b;Jiang et al., 2018; Niklaus and Liu, 2018; Xu et al., 2019) usually combines the optical flow maps linearly to estimate the latent optical flows which need to be synthesized, and finally the original frames are warped by the predicted optical flows to estimate the intermediate target frame. Recently, Sim et al. (2021) proposed a network that adopts a recursive multi-scale shared structure to learn bidirectional optical flow between two input frames and bidirectional optical flow between target and input frames. Kernel-based methods use adaptive convolution kernel to realize motion estimation and pixel synthesis. To handle complex motion, Lee et al. (2020) designed a 2D deformable spatial-adaptive scheme to break the limits of the fixed grid shape of a regular convolution kernel.

(3) Space-time super-resolution: ST-SR was first proposed by Shechtman et al. (2005). The purpose is to recover an HR and high frame rate (HFR) video from an LR and low-frame-rate (LFR) video. Different from separate T-SR and S-SR, ST-SR requires super-resolving both spatial and temporal dimensions simultaneously.

ST-SR is a highly ill-posed problem since the requirement to predict non-existent frame pixels and HR frame pixels. Previous traditional methods (Mudenagudi et al., 2010; Shahar et al., 2011; Takeda et al., 2010) use hand-crafted priors to constrain solution space, which is complicated to constrain. Benefit from deep learning, the data-driven ST-SR method has shown far superior performance to traditional methods. Haris et al. (2020) followed a three-step ST-SR strategy to reconstruct all LR frames and HR frames. Xiang et al. (2020) employed DConv to achieve the alignment between the original input frames, then aggregated two aligned features to estimate the target features, and finally captured the contextual information between the frames through the Bidirectional Deformable Long Short-Term Memory (BD-LSTM). Shi et al. (2021) proposed an unconstrained STVSR framework that realizes frame interpolation at any temporal location through the adjustable optical flow.

These one-stage methods rely on either optical flow or DConv to synthesize missing frames, which is not sophisticated enough for more challenging satellite video. Besides, these computationally complex design (LSTM or dense structure) in satellite video may introduce interference information while being inefficient (Vaswani et al., 2017). The effective spatial–temporal information fusion is critical to ST-SR (; Chen et al., 2022; Li et al., 2020, 2021), which demands us to devise valid frameworks to aggregate spatial–temporal information. Hence, we propose a lightweight and end-to-end framework based on multi-scale spatial–temporal transformer. An optical flow and multi-scale deformable convolutional deeply coupled module is designed to realize the prediction of non-existent frames. To our best knowledge, this paper the first to study the joint enhancement of the spatial and temporal resolution for satellite video.

2. Methodology

2.1. Problem formulation

Firstly, we give the formula definition of STVSR. Given a LR and LFR video frame sequence $I^{LR} = \{I_{2t-1}^{LR}\}_{t=1}^{N+1}$, the goal of STVSR is to generate their corresponding HR frames $I^{HR} = \{I_{2t-1}^{HR}\}_{t=1}^{N+1}$ as well as N missing HR frames $\{I_{2t}^{HR}\}_{t=1}^{N}$, and finally obtain a HR and HFR video sequence $I^{HR} = \{I_{t}^{HR}\}_{t=1}^{2N+1}$. As shown in Fig. 1, assuming that the network takes two existing original LR frames I_{t-1}^{LR} and I_{t+1}^{LR} as input, we need to predict the HR frame I_t^{HR} at the intermediate moment t while obtaining the HR frame I_{t-1}^{HR} and I_{t+1}^{HR} corresponding to the original LR frames.

Our framework is divided into the following parts: feature extraction, Flow-DConv deep-Coupled (FDC) feature interpolation, Multi-scale Spatial-Temporal Transformer (MSTT) and reconstruction module. First, the optical flow map $f_{1\rightarrow3}$ and $f_{3\rightarrow1}$ containing forward and backward motion information can be obtained through an optical flow estimation approach. Then, two LR features and two optical flow maps enter the FDC feature interpolation module to synthesize the missing intermediate frame features F_2^{LR}. After that, we further explores the contextual information in all frames through our MSTT module, and finally we reconstructs the super-resolved results I_1^{SR}, I_2^{SR} and I_3^{SR}. The details of each part will be explained below.

2.2. Feature extraction

For each LR frame $I_{2t-1}^{LR} \in R^{h' \times w' \times c}$, where h' and w' are the height and width of LR frames and $c = 3$ is the number of RGB channels, we utilize five residual blocks without batch normalization (BN) for feature extraction. Each residual block follows the structure of "Conv + ReLU + Conv". Finally, we get an LR features $I_{2t-1}^{LR} \in R^{h' \times w' \times 64}$. In this paper, we have $N = 3$, so after feature extraction we will obtain 4 LR features $\{F_1^{LR}, F_3^{LR}, F_5^{LR}, F_7^{LR}\}$.

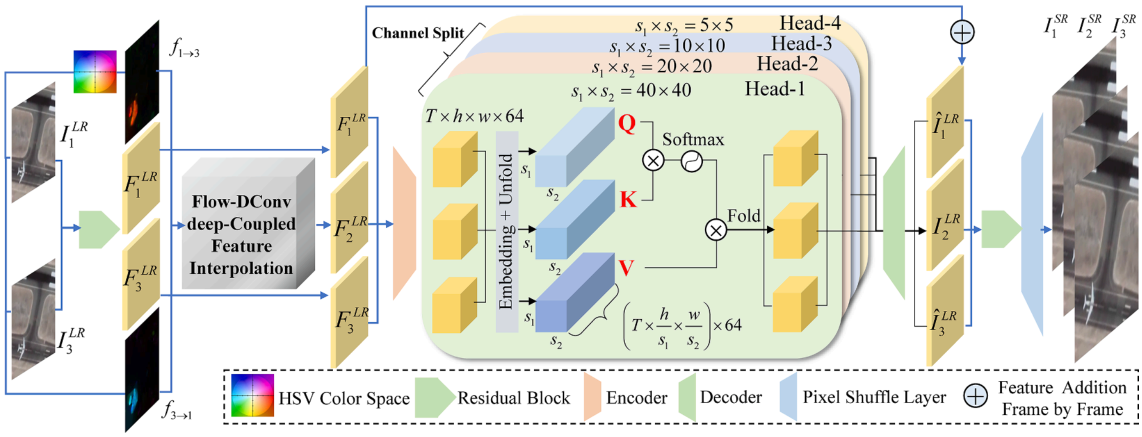


Fig. 1. Network structure diagram. Input $N+1$ LR frames, our network can get $2N+1$ HR frames in an end-to-end manner. For the convenience of explanation, we set $N=1$ here and $T=2N+1=3$ is the length of all frames.

2.3. Flow-DConv deep-Coupled feature interpolation

We believe that the motion information in satellite video learned by optical flow is insufficient since it naturally has a disadvantage in capturing the large motion (Xiang et al., 2020). Although deformable convolution has the advantage of adaptive learning, it is difficult to capture moving objects with various scales (Xiao et al., 2021). Our FDC feature interpolation (FI) module aims to make optical flow and deformable convolution mutually constrain each other by deep coupling manner, and introduce the adaptive learnability of deformable convolution to complement the information that optical flow ignores, so the latent motion information can be learned adaptively. The structure of FDC-FI is shown in Fig. 2.

(1) Optical flow estimation: Let the forward optical flow $f_{1\to3}$ warp LR feature F_3^{LR} , the feature F_1^f estimated by the forward motion information can be obtained. Similarly, we get the feature F_3^b derived from the backward optical flow $f_{3\to1}$. This process can be formulated as:

$$F_1^f = warp\left(F_3^{LR}, f_{1\to3}\right), \tag{1}$$

$$F_3^b = warp\left(F_1^{LR}, f_{3\to1}\right). \tag{2}$$

(2) Deformable convolution estimation: To accurately exploit the motion information of moving objects with variable scales in satellite video, we adopt the multi-scale residual block

(MSRB) proposed in our previous work Xiao et al. (2021) to learn sampling parameters Θ for deformable grids. This means the sampling parameters can be expressed as:

$$\Theta_1 = MSRB\left(concat\left(F_1^{LR}, F_3^{LR}\right)\right), \tag{3}$$

$$\Theta_3 = MSRB\left(concat\left(F_3^{LR}, F_1^{LR}\right)\right). \tag{4}$$

Where Θ_1 represents the sampling parameter that encodes the forward motion information, and Θ_3 is the sampling parameter that encodes the backward motion information. Now we can perform deformable convolution (DConv) (Zhu et al., 2019) under the guidance of sampling parameters to get the feature F_3^f estimated by the forward motion information and the feature F_1^b estimated by the backward motion information:

$$F_3^f = DConv\left(F_3^{LR}, \Theta_1\right), \tag{5}$$

$$F_1^b = DConv\left(F_1^{LR}, \Theta_3\right). \tag{6}$$

The deformable convolution means that the convolution sampling position has an additional offset, and the sampling grid is no longer a regular square grid. After deformable convolution, the value of position p_0 can be defined as:

$$F_i^b(p_0) = \sum_{k=1}^{K} \omega_k F_i^{LR}(p_0 + p_k + \Delta p_k). \tag{7}$$

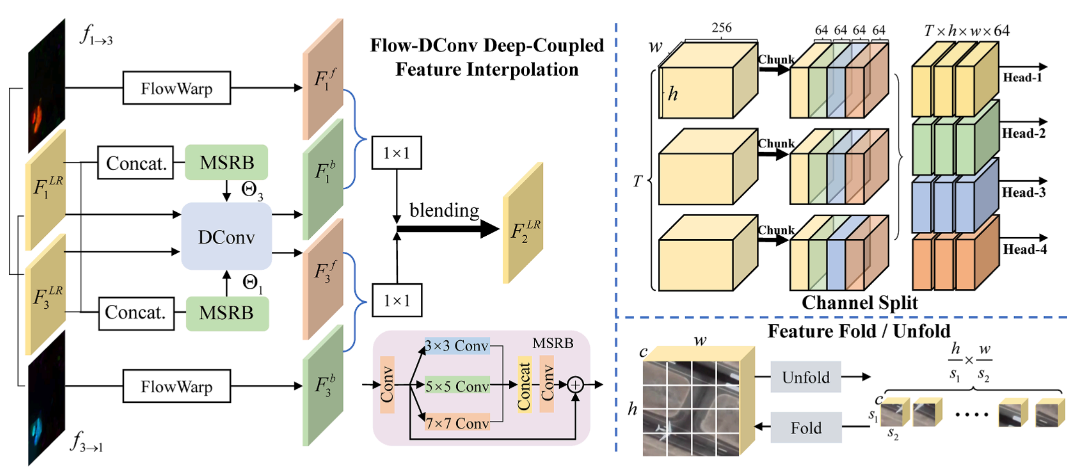


Fig. 2. The diagram of Flow-DConv Deep-Coupled Feature Interpolation module. The red features are derived from the forward motion information, and the green features are derived from the backward motion information; The channel separation operation divides the features of each frame into four heads; The lower right corner is a schematic diagram of feature unfold and fold.

Here ω_k denotes the weight of the $k-th$ sampling position in convolution grid and $\Theta_i = \{\Delta p_k\}_{k=1}^{K}$. $\Delta p_k \in R^2$ is the learned extra offset. In a 3×3 convolution grid, $K = 9$ represents the 9 sampling positions p_k centered on p_0. Similarly, in a $n \times n$ convolution grid, $K = n^2$.

(3) Deep coupled: We use 1×1 convolution to deeply couple the forward and backward estimation features. The forward and backward features are taken from the results of optical flow and DConv respectively to achieve the complementation. Finally, we blend the coupling features to the final synthesis feature F_2^{LR}. The process of deep coupling can be written as:

$$F_2^{LR} = \alpha * f_1\left(\left[F_1^f, F_1^b\right]\right) + \beta * f_2\left(\left[F_3^f, F_3^b\right]\right). \tag{8}$$

Where $*$ denotes the convolution operation, $[\cdot]$ represents the concatenation in the channel dimension, α and β are two learnable 1×1 convolutions for coupling, and f_1 and f_2 are two 1×1 convolutions used for blending.

Since the synthesized missing features will be used for subsequent SR tasks, the network will force the predicted features to be as close to the real situation as possible, so the sampling parameters of the deformable convolution will also implicitly force to approximate the real motion situation. With the constraint of optical flow, this complementarity can converge better.

Algorithm 1: Algorithm of our Flow-DConv deep-Coupled Feature Interpolation.

Input: $N+1$ LR features $I^{LR} = \{I_{2t-1}^{LR}\}_{t=1}^{N+1} = \{F_1^{LR}, F_3^{LR}, F_5^{LR}, F_7^{LR}\}$.
1 Initialization: $N = 3$, *MSRB* is Multi-Scale Residual Block, f_1, f_2, α and β are four 1×1 convolution.
2 for $i \in [1, 3, 5]$; **do**
3 *Optical flow estimation:* $F_i^f = warp(F_{i+2}^{LR}, f_{i \to i+2}), F_{i+2}^f = warp(F_i^{LR}, f_{i+2 \to i})$;
4 *Deformable Convolution estimation:* $\Theta_i = MSRB(concat(F_i^{LR}, F_{i+2}^{LR})), F_{i+2}^f = DConv(F_{i+2}^{LR}, \Theta_i); \Theta_{i+2} = MSRB(concat(F_{i+2}^{LR}, F_i^{LR})), F_i^b = DConv(F_i^{LR}, \Theta_{i+2})$;
5 *Deep coupling:* $F_{i+1}^{LR} = \alpha * f_1\left(\left[F_i^f, F_i^b\right]\right) + \beta * f_2\left(\left[F_{i+2}^f, F_{i+2}^b\right]\right)$.
6 end
Output: $2N+1$ LR features $\mathscr{F} = \{F_1^{LR}, F_2^{LR}, F_3^{LR}, F_4^{LR}, F_5^{LR}, F_6^{LR}, F_7^{LR}\}$.

2.4. Multi-scale Spatial-Temporal transformer

Transformer is initially used in the field of natural language processing (NLP), but owing to its superior temporal representation capabilities, it has now become a viable alternative to CNN in the field of computer vision (CV) (Han et al., 2020). Existing studies (Xiao et al., 2021) obtain multi-scale information through convolution kernel of multi-scale receptive fields. Different from multi-scale information learned by convolution structure, our multi-scale information is mined by multi-scale feature embedding and self-attention. Here we proposed a Multi-scale Spatial-Temporal Transformer (MSTT) to fully explore the long-term contextual information existing among frames. In vanilla transformer, the scale gap in remote sensing imagery makes it difficult to simultaneously represent multi-scale information using a single-scale design. Specifically, since moving objects only occupy very few pixels, the foreground scale is small while the background scale is large. Furthermore, the multiple types of moving objects also cause the scale variation. For this reason, we embed features into the multi-heads of the self-attention mechanism with patches of different scales to help deeply represent the multi-scale information in the satellite video. To be specific:

(1) Encoder: It is necessary to embed LR features into the embedding space for attention calculation. To save memory overhead, we first use 3×3 convolution with a stride of 2 to achieve feature down-sampling, and increase the number of feature channels to 256. Denote the feature before encoding as $\mathscr{F} = \{F_1^{LR}, F_2^{LR}, \cdots, F_T^{LR}\}$, where T is the number of frames and $F_i^{LR} \in R^{h' \times w' \times 64}, i \in [1:T]$. The encoded features $e_i = f_{encoder}(F_i^{LR}) \in R^{h \times w \times 256}$, Finally, after encoding, we have $\mathscr{F}^E = \{e_1, e_2, \cdots, e_T\} \in R^{T \times h \times w \times 256}$.

(2) Channel Split: The encoded features are matched to different heads after channel chunk operation:

$$\{e_i^1, e_i^2, e_i^3, e_i^4\} = chunk(e_i), \tag{9}$$

where $e_i^j \in R^{h \times w \times 64}$ means the part where feature e_i matches to the $j-th$ head. Finally, the features assigned to $j-th$ head are $\mathscr{F}_j^E = \left\{e_1^j, e_2^j, \cdots, e_T^j\right\} \in R^{T \times h \times w \times 64}$.

(3) Feature Embedding and Fold/Unfold: We adopt there 1×1 convolution to linearly project features in each head $\mathscr{F}_j^E$ into query (Q), key (K) and value (V) space. Note that the size of each space is still $T \times h \times w \times 64$. To facilitate the calculation of attention and accelerate the speed of inference, we need to unfold the feature into one dimension. As shown in Fig. 2, each feature can unfold $(h/s_1) \times (w/s_2)$ patches of size $s_1 \times s_2$. In the end, each space is unfolded into $N = T \times (h/s_1) \times (w/s_2)$ patches in a non-overlapping way. Similarly, we can also understand the process of feature fold.

(4) Self-attention calculation: After unfolding the features into one-dimension, now we can calculate spatial–temporal joint attention. The similarity between two patches can be transformed into matrix multiplication:

$$S_{i,j} = \frac{p_i^q \cdot \left(p_j^k\right)^T}{\sqrt{s_1 \times s_2 \times 64}}. \tag{10}$$

Where p_i^q represents the $i-th$ patch in query space, and p_j^k is the $j-th$ patch in key space.

We use softmax function along temporal axis to obtain the attention map $att_{i,j}$:

$$att_{i,j} = exp\left(S_{i,j}\right) / \sum_{n=1}^{N} exp\left(S_{i,n}\right). \tag{11}$$

So far, we have jointly considered the spatial–temporal attention weights between patch p_i^q and patch p_j^k.

By multiplying attention and value space to perform feature modulation, the corresponding position pixel in each patch can simultaneously focus on the spatial–temporal redundant information in the context. For the $i-th$ patch t_i after modulation, it can be expressed as:

$$t_i = \sum_{j=1}^{N} att_{i,j} \odot p_j^v. \tag{12}$$

Finally, the modulated value space is folded back to size $T \times h \times w \times 64$.

(5) Decoder: We use a 3×3 convolution to aggregate the results in the 4 heads, and then adopt a 3×3 deconvolution to upsample the features back to $T \times h' \times w' \times 64$. A global residual connection was designed to ensure the stability of training and force the MSTT to focus on learning the redundant information in all frame features.

2.5. Reconstruction

We use a 20-layer residual block and widely used sub-pixel convolution to achieve the final up-sampling of features. Finally, we will get a space–time super-resolved video sequence $I^{SR} = \{I_1^{SR}, I_2^{SR}, \cdots, I_T^{SR}\} \in R^{T \times rh' \times rw' \times 3}$, where r is the S-SR scale factor.

3. Experiments

3.1. Jilin-1 satellite video data setting

We use 10 Jilin-1 satellite video scenes to build our dataset: San Francisco (United States), Derna (Libya), Valencia (Spain), San Diego (United States), Tunis, Adana-01 (Turkey), Adana-02 (Turkey), Minneapolis-01 (United States), Minneapolis-02 (United States) and

Muhalag (Bahrain). Their spatial resolution is 1 m, the frame rate is 25/fps and the duration of each video is 30 s. The spatial pixel size of each video is 4096×2160 (except for 3840×2160 in San Francisco (United States)). San Diego (United States) and Minneapolis-01 (United States) scenes are selected to build our test set and the other eight scenes are used to construct the training set. Finally, we get 2646 short video clips as our training data. To construct LR and LFR input video, we eliminate even frames in each short video clip, leaving only odd frames, and use the *imresize* function in MATLAB to downsample the frame to size 160×160. Five subareas in San Diego (United States) and Minneapolis-01 (United States) are cropped respectively to make up our ten test clips. Some training and test samples are shown in Fig. 3(a) and 3(b), 3(c) shows 5 test clips extracted from San Diego (United States).

3.2. Training details

In this paper, we only focus on $4 \times$ S-SR and $2 \times$ T-SR. To ensure the fairness of comparison, we follow the setting in Haris et al. (2020) and take 4 odd-indexed frames as input, and predict 7-
frame HR and HFR sequences.

(1) Loss function: To optimize the network, we adopt a Carbonnier penalty function (Lai et al., 2017) as our loss function which can be expressed as:

$$\mathscr{L} = \sqrt{\left\| I_t^{HR} - I_t^{SR} \right\|^2 + \varepsilon^2}, \tag{13}$$

where I_t^{HR} is the ground truth frame, I_t^{SR} is the predicted SR frame and $\varepsilon = 10^{-3}$ is the empirical parameter. Under the guidance of the loss function, our network can be trained end-to-end in a supervised manner to make the predicted SR frames as close as possible to the ground truth frames.

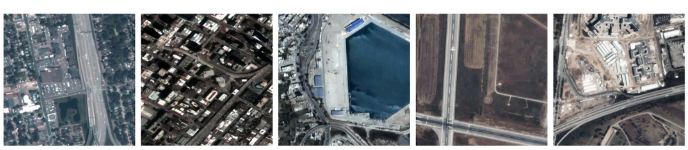
(2) Training Strategy: We use Adam as our optimizer, where $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The min-batch is set to 2, and we randomly crop the 160×160 LR frames into 80×80 patches as input. Also, we use image rotating and flipping to augment our data. The learning rate is initialized to 1×10^{-4}, and it is reduced by a factor of 10 when reaching half of the total 50 epochs. It took about 40 h on a single NVIDIA RTX 2080Ti GPU to train our model.

3.3. Comparison with state-of-the-art methods

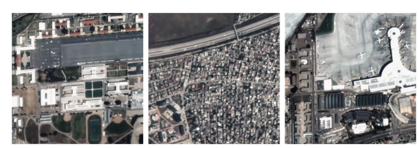
To verify the effectiveness of our method in breaking the spatial and temporal resolution barriers for satellite video, we have compared our method with several SOTA (state-of-the-art) methods in terms of quantitative and qualitative performance. First, some two-stage methods composed of T-SR and S-SR are compared. Specifically, a powerful method named XVFI (Sim et al., 2021) is chosen to achieve T-SR. Next, bicubic interpolation is selected as the baseline of S-SR. Then, two widely used SISR methods include: EDSR (Lim et al., 2017) and RCAN (Zhang et al., 2018), and two VSR approaches RBPN (Haris et al., 2019) and EDVR (Wang et al., 2019) are used to generate HR frames. Since the idea of back projection has been shown to be effective in satellite VSR (Xiao et al., 2021), we choose RBPN as a comparison method for VSR. EDVR is a SOTA VSR method, which represents the leading performance. Finally, since no one-stage ST-SR method has been proposed in the field of remote sensing, two one-stage joint ST-SR methods named STARNet (Haris et al., 2020) and ZoomingSloM (Xiang et al., 2020) are also used for comparison. ZoomingSloM is lightweight and advanced. STARNet is a more elaborate and larger model.

(1) Evaluation Metrics: Peak Signal-to-Noise Ratio (PSNR) and SSIM (Hore and Ziou, 2010) are two widely used image quality evaluation indicators in the presence of reference images. In addition, we also calculated the Root Mean Square Error (RMSE) and the Correlation Coefficient (CC). Besides, we introduced a reference-free image quality evaluation index NIQE (Mittal et al., 2012). The lower the NIQE, the more natural the image is in human vision.

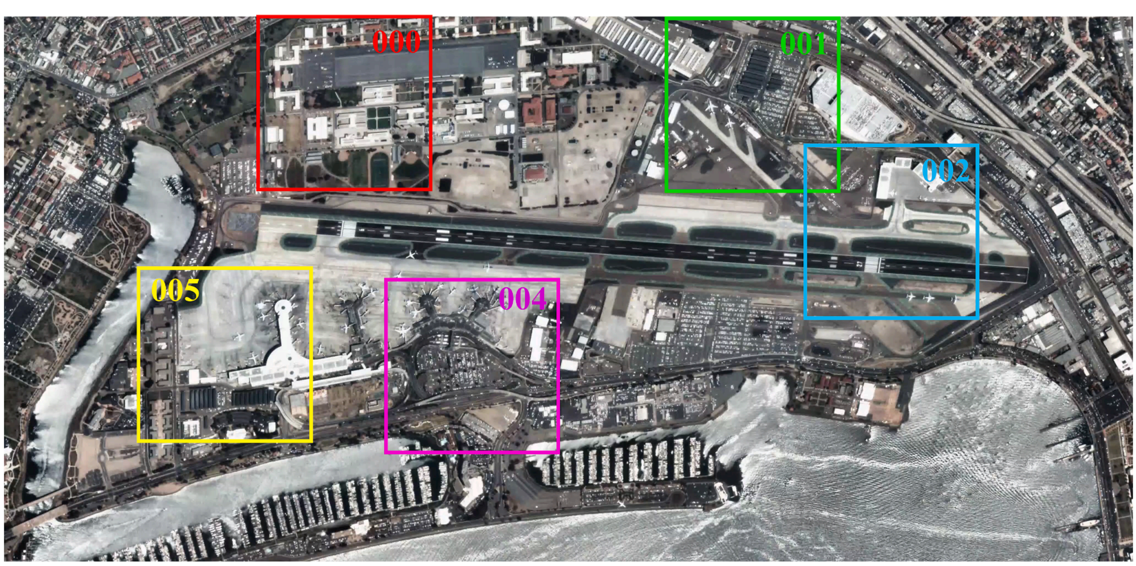
(2) Quantitative Evaluation: The average PSNR and SSIM results of each method on the 10 test videos are shown in Table. 1. Our model



(a) Training samples.



(b) Test samples



(c) San Diego (United States) and five subarea for test.

Fig. 3. Our dataset setting.

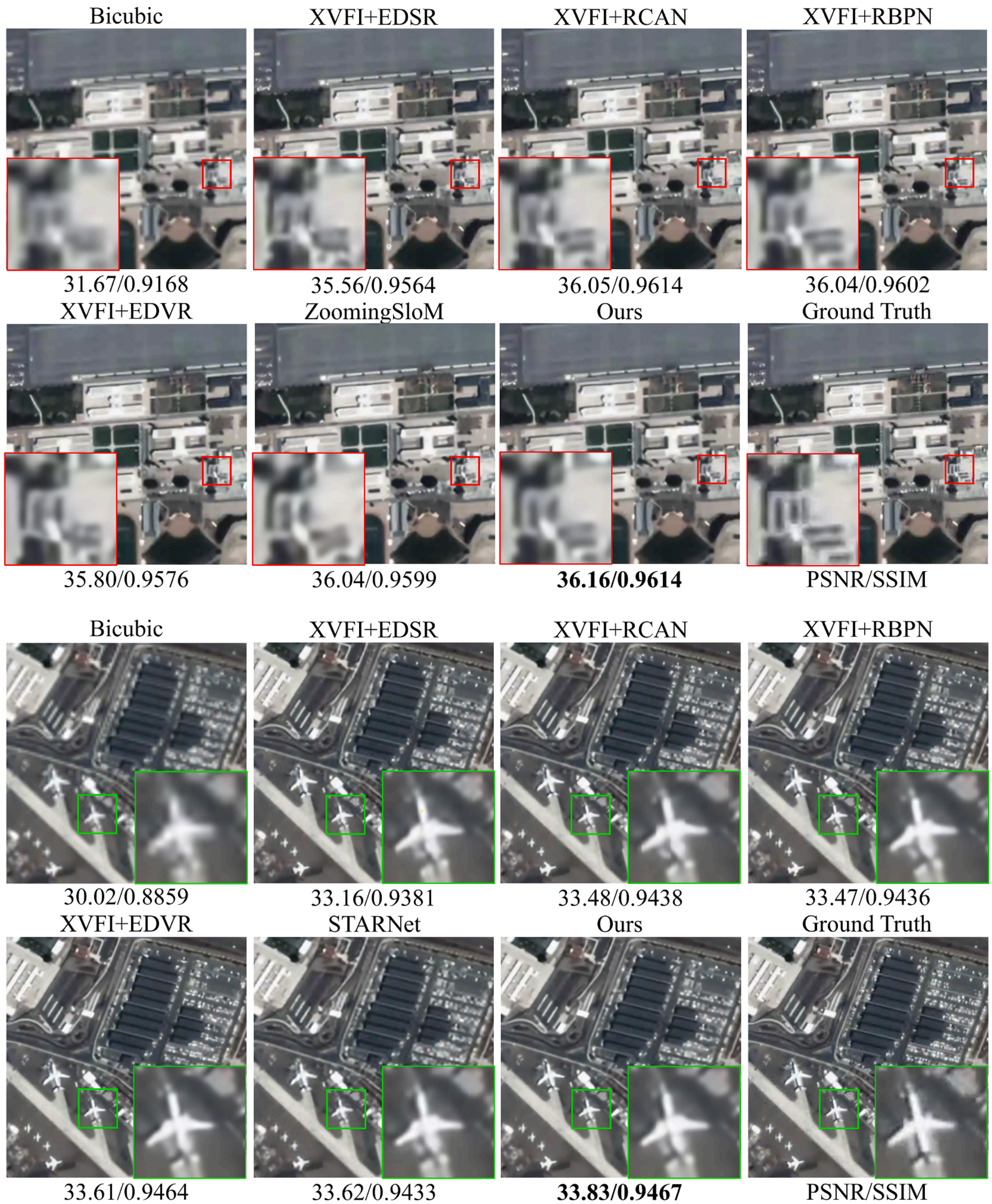


Fig. 4. × 4 SR results on test clip 000 (top) and 001 (bottom). We zoom the details for better comparison.

Table 1
The PSNR/SSIM results on 10 test video clips. **Red** and blue indicates the best and the second best performance, respectively.

Method	**Baseline**	**XVFI + SISR**		**XVFI + VSR**		**STVSR**		
	FI + Bicubic	FI + EDSR	FI + RCAN	FI + RBPN	FI + EDVR	STARNet	ZoomingSloM	Ours
Param.(M)	5.5 + 0	5.5 + 43.1	5.5 + 15.6	5.5 + 12.8	5.5 + 20.6	111.6	11.1	5.7
000	31.02/0.9089	35.09/0.9541	35.58/0.9590	35.57/0.9585	35.77/0.9598	35.55/0.9572	35.68/0.9595	**35.79/0.9603**
001	29.68/0.8794	32.83/0.9359	33.18/0.9424	33.20/0.9421	33.25/0.9433	33.25/0.9410	33.24/0.9411	**33.36**/0.9426
002	31.89/0.9230	35.56/0.9562	36.32/0.9627	36.43/0.9623	**36.80/0.9649**	36.19/0.9600	36.32/0.9616	36.49/0.9626
003	30.43/0.8961	33.45/0.9373	33.94/0.9432	33.96/0.9436	34.14/0.9448	34.05/0.9429	34.08/0.9438	**34.20/0.9449**
004	28.57/0.8628	31.57/0.9216	31.94/0.9290	32.04/0.9300	32.01/0.9299	31.93/0.9271	32.10/0.9295	**32.18/0.9308**
005	28.11/0.8448	30.71/0.9017	30.98/0.9078	31.12/0.9104	30.95/0.9073	31.19/0.9104	31.34/0.9132	**31.35/0.9133**
006	30.57/0.9003	33.93/0.9460	34.28/0.9504	34.48/0.9520	34.36/0.9511	34.41/0.9506	34.50/0.9519	**34.63/0.9530**
007	29.19/0.8714	31.59/0.9196	31.82/0.9236	31.96/0.9261	31.82/0.9237	32.16/0.9274	32.26/0.9301	**32.26/0.9294**
008	32.43/0.9243	36.20/0.9606	36.65/0.9648	36.69/0.9649	36.81/0.9654	36.56/0.9632	36.71/0.9650	**36.85/0.9654**
009	30.79/0.8984	34.01/0.9417	34.33/0.9464	34.27/0.9462	34.39/0.9469	34.40/0.9459	34.51/0.9466	**34.64/0.9490**
Avg.	30.27/0.8903	33.49/0.9375	33.90/0.9429	33.97/0.9436	34.03/0.9437	33.97/0.9426	34.07/0.9442	**34.18/0.9451**

reveals peak performance on all test sets. All of the deep-learning-based methods have significantly improved compared to the baseline. Among the two-stage methods, FI + SISR is slightly inferior to FI + VSR because using SISR in multi-temporal frames cannot utilize the redundant information between frames to enhance the spatial resolution. In the one-stage method, STARNet can achieve comparable effect with XVFI + RBPN, but it is a little bit worse than XVFI + EDVR. ZoomingSloM is slightly better than XVFI + EDVR, which proves that the joint enhancement not only has end-to-end advantage, but also reaches higher performance than existing SOTA two-stage methed.

Comprehensively evaluating model performance and efficiency, our method is ahead of the SOTA one-stage method ZoomingSloM by 0.11 dB and leads STARNet by 0.21 dB, highlighting the superiority of our framework. What's more noteworthy is that all two-stage methods have huge parameters due to the combination of two independent tasks T-SR + S-SR. Although STARNet can jointly achieve ST-SR and obtain satisfying results, the amount of parameters has reached nearly 20 times that of our method (111.6 M v.s. 5.7 M), which is evidently inefficient. The key reason is that STARNet introduces a UNet structure with a mass of parameters to enhance optical flow information for accurate missing motion information estimation. In our framework, even though the inital optical flow is rough, no additional parameters will be introduced. On the one hand, the optical flow can provide motion information estimation; on the other hand, it can be used as a constraint in the convergence of DConv to mitigate the learning pressure and make DConv focus on learning the information that the optical flow cannot pay attention to. Compared with the best two-stage method XVFI + EDVR, our model is nearly only 1/5 of its size (26.1 M v.s. 5.7 M) and has the advantage of joint ST-SR to exploit the inherent relationship between T-SR and S-SR to restrict each other and promote. The parameters of our method are even nearly equal to XVFI (5.5 M), but we can still accomplish both S-SR and T-SR simultaneously. Thanks to the multi-scale self-attention mechanism in the transformer, we can effectively mine the scale variable context information in long time series frames without introducing too many parameters.

In Table. 2, we further divided all frames into predicted frames that originally did not exist (even-indexed frames) and originally existing input frames (odd-indexed frames). In terms of even frames, the one-stage method STARNet and ZoomingSloM can achieve similar results to FI + VSR, but the performance on odd frames is lower than FI + VSR, which indicates that the one-stage method suffers from weakness of excavating enough redundant information from frame sequence. And our model not only achieved 0.16 dB ahead of ZoomingSloM when predicting unknown frames and even led the best two-stage method XVFI + EDVR by 0.17 dB on the original frame. This result not only demonstrates that our FDC feature interpolation can accurately predict missing frames, but also shows that our MSTT is able to better extract redundant information from all frames.

Table 2
We respectively calculate the average PSNR/SSIM of the original input frames (odd frame) and the non-existent frames (even frame) that needs to be predicted. **Red** and blue indicates the best and the second best PSNR/SSIM performance, respectively.

Method	Even frame		Odd frame		Overall	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM
XVFI + Bicubic	30.24	0.8902	30.30	0.8915	30.27	0.8909
XVFI + EDSR	33.49	0.9374	33.50	0.9375	33.49	0.9375
XVFI + RCAN	33.89	0.9429	33.91	0.943	33.90	0.9429
XVFI + EDVR	33.99	0.9432	34.07	0.9442	34.03	0.9437
XVFI + RBPN	33.94	0.9432	34.01	0.9441	33.98	0.9436
STARNet	33.98	0.9426	33.97	0.9425	33.97	0.9426
ZoomingSloM	34.06	0.9439	34.08	0.9445	34.07	0.9442
Ours	**34.11**	**0.9444**	**34.24**	**0.9459**	**34.18**	**0.9451**

In Table. 3, our method achieves higher fidelity, the image quality is also pleasing, which illustrates our MSTT can introduce less useless interference information and only exploit valuable context information.

(3) Qualitative Evaluation: The qualitative evaluation focuses on the texture and details of the results from a visual perspective. We have magnified the local information to better observe the details in Fig. 4. Both XVFI + EDSR and XVFI + EDVR have severe distortion, while XVFI + RCAN and XVFI + RBPN have a certain degree of blur, and the edge of buildings are not clear enough. Our method generates the most negligible blur and more texture information. The same situation can be found in test scene 001. In the XVFI + RBPN results, the wing of the aircraft was deformed. In XVFI + EDSR, XVFI + RCAN, and STARNet, the tail of the aircraft also had significant artifacts. As shown in Fig. 5, on a moving aircraft with a smaller scale in 007, only our method recovered a clear wing. Moreover, the tail of the aircraft in XVFI + EDVR is seriously distorted and mixed with the background, which was difficult to distinguish. We calculated the residual between the ground truth and the predicted frame and normalized it to [0, 1] as shown in error maps below, our results are closer to the ground truth with minimal errors, thus more detailed information is recovered. Our method can achieve good generalization in aircraft with various scales, which further proves the effectiveness of the multi-scale design in MSTT.

4. Discussions

4.1. Discussion on the coupling of optical flow and DConv

Here we need to discuss the bottleneck in using only optical flow or deformable convolution to synthesize missing frames, and these two methods can be complementary. Three experiments were designed for this purpose. Firstly, we only use the optical flow method to estimate the intermediate missing frame. Specifically, after optical flow warp operation, two warped features are blended directly. We denote this model as Flow-only. Similarly, when we only adopt DConv to synthesize intermediate frames, we blend the results of deformable convolution, and this model is denoted as D-only. To ensure a fair comparison, we use four 1×1 convolutions to achieve the blending operation, the same number of convolutions used in our FDC feature interpolation module. The third model is our FDC method, which simultaneously introduces these two methods for deep coupling complementarity. The experimental results are shown in Table. 4 and training process can be seen in Fig. 6(a). Flow-only and DConv-only can achieve similar performance. However, when they are combined, PSNR can increase by 0.1 dB and 0.8 dB, respectively, which fully demonstrates that the deep coupling idea can achieve complementary advantages for better results.

4.2. Discussion on the different strategy of coupling

Here we discuss different manners to realize coupling and prove the effectiveness of our deep coupling strategy. A direct approach is to mix the optical flow estimation result with the DConv estimation result (naive coupling). In this case, the optical flow and DConv estimation processes are essentially separate, just a shallow coupling of different results. It can be denoted as:

$$F_t^{LR} = \alpha * f_1\left(\left[F_{t-1}^f, F_{t+1}^b\right]\right) + \beta * f_2\left(\left[F_{t+1}^f, F_{t-1}^b\right]\right) \tag{14}$$

Table 3
Quantitative results of RMSE, CC and NIQE.

Method	RMSE ↓	CC ↑	NIQE ↓
XVFI + Bicubic	7.9079	0.98486	19.7265
XVFI + EDVR	5.1998	0.99299	18.4472
XVFI + RBPN	5.2150	0.99305	18.1580
STARNet	5.2099	0.99310	18.0570
ZoomingSloM	5.1987	0.99326	18.0953
Ours	**5.1054**	**0.99336**	**17.8575**

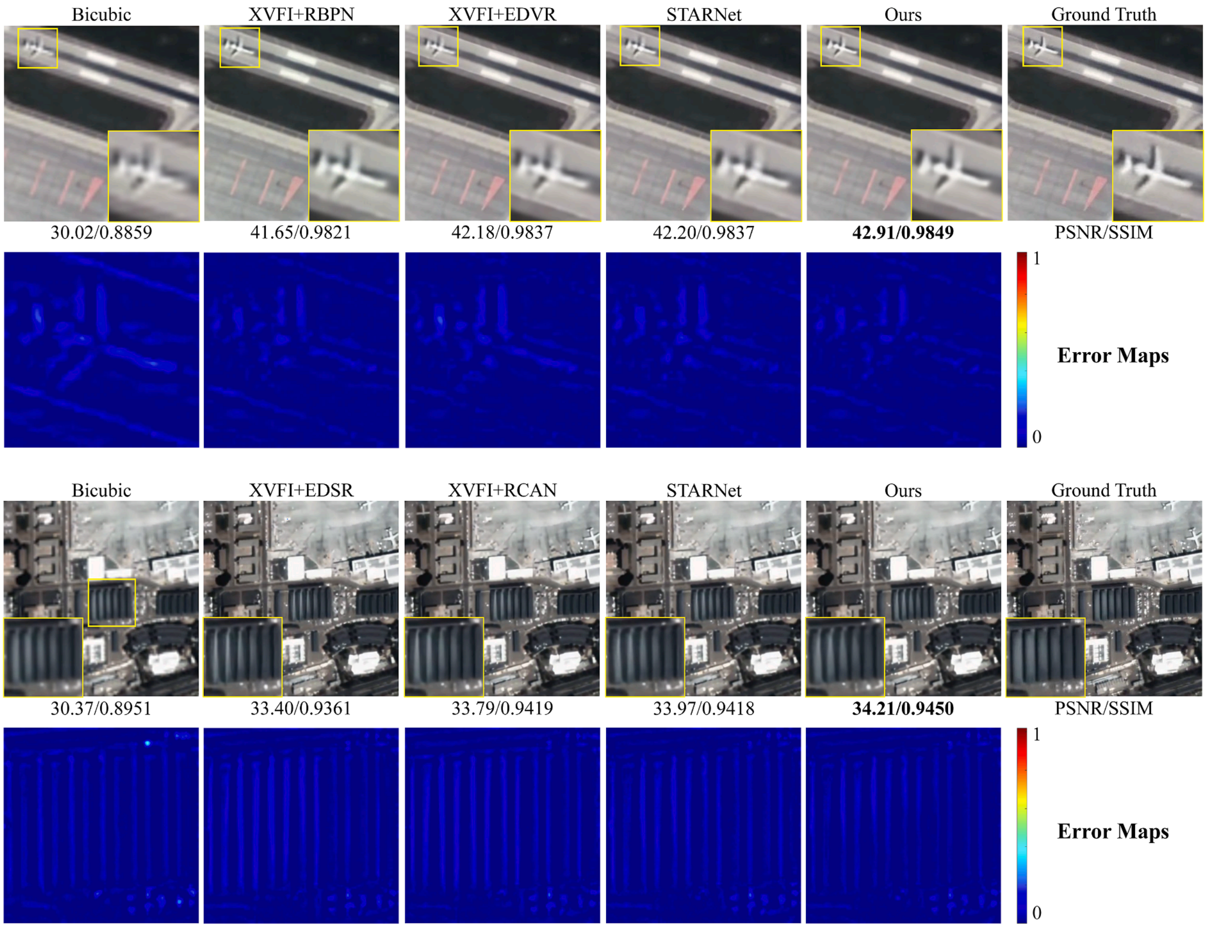


Fig. 5. × 4 SR results on test clip 007 (top) and 005 (bottom). We compute the residual maps for better comparison.

Table 4
Ablation study on the coupling of optical flow and DConv.

Model	Flow-only	DConv-only	Ours
Optical flow	√	×	√
Deformable convolution	×	√	√
PSNR (dB)	32.2759	32.2962	**32.3796**

Our method proposes the idea of deep coupling to realize the information coupling in the intermediate process of the two estimations and finally mixes the intermediate coupling results to achieve the purpose of deep coupling. In other words, this design can make the two methods complement each other earlier. The experimental results are shown in Table. 5. The parameters of the two coupling approaches remain unchanged, but the deep coupling can exceed the shallow coupling. The training process is shown in Fig. 6(b).

4.3. Discussion on multi-scale strategy

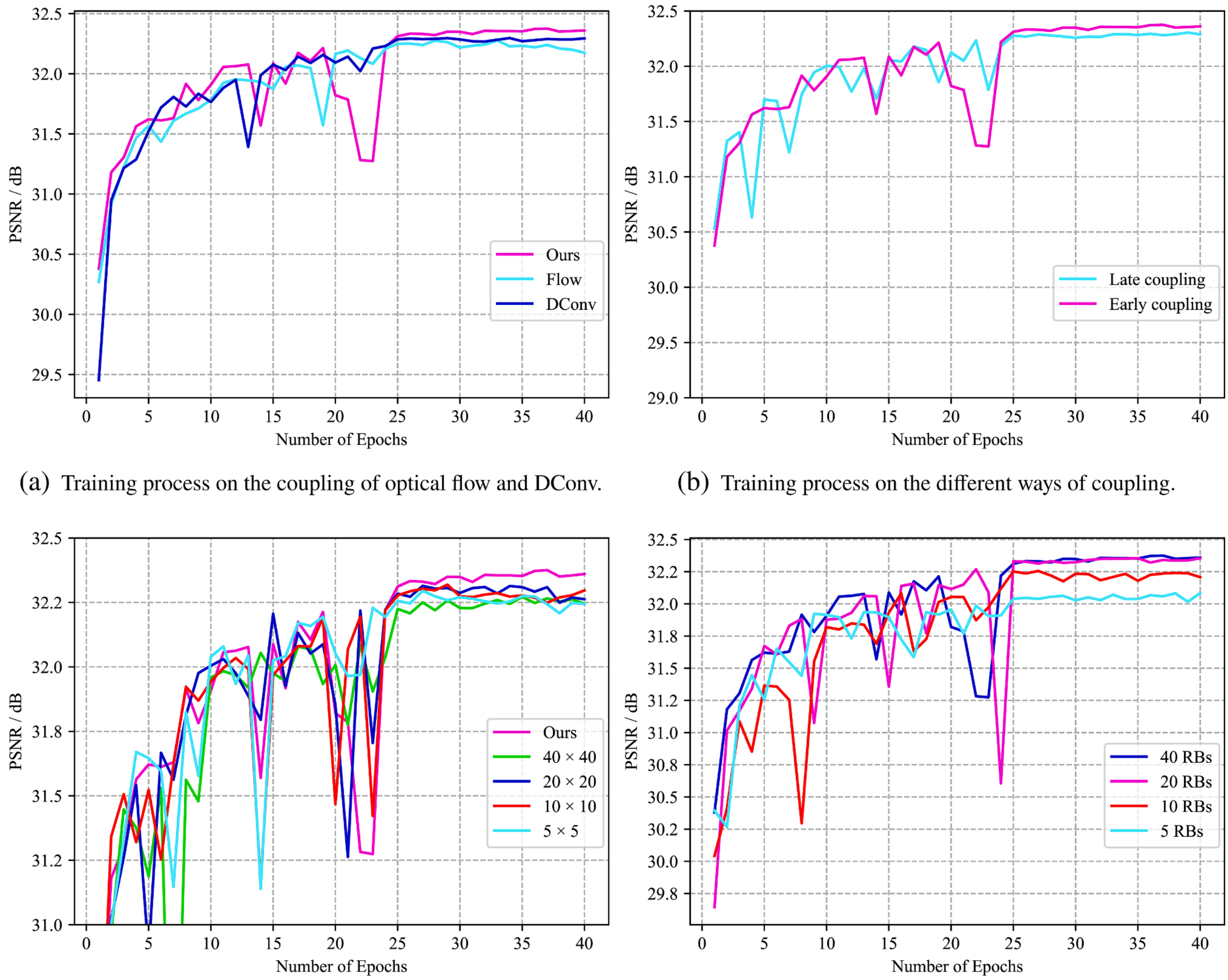
In this part, we prove that our strategy of matching multi-scale patches in multiple heads is more conducive to mining the scale variable information of satellite videos. In practice, since the transformer needs to embed features into non-overlapping patches, and the feature size in our model is encoded to 40 × 40, we set four scales of 40 × 40, 20 × 20, 10 × 10 and 5 × 5. As shown in Table. 6, if each head is assigned a fixed-scale, the network neglects multi-scale information and has limited representation ability. The result illustrates the effectiveness of our multi-scale design. The training process is drawn in Fig. 6(c).

4.4. Discussion on model efficiency

We explored the structure and efficiency of the model and proved that our method is lightweight and efficient. As shown in Fig. 6(d), we find that the performance of using 20 residual blocks is almost the same as that of 40 residual blocks. Therefore, we select 20 residual blocks for reconstruction. We also calculated FLOPs and a more intuitive indicator average processing time of each frame to better measure model complexity. The results are shown in Table. 7. Compared with the two-stage method, the complexity of our model is within an acceptable range, and the processing time is significantly faster. Compared with the one-stage method, our model has achieved performance by a great margin. Hence, we achieved the best trade-off between efficiency and performance.

4.5. Discussion the effect of ST-SR on moving object tracking

To explore the significance of SRVSR for moving object tracking, we have tracked the moving vehicle in a SkySat scene by Zhang et al. (2020). The original video size is 320 × 320 and the number of frames is 50, after STVSR, the resolution increased to 1280 × 1280 and the number of frames was 99. We calculated the average recall of all frames, and the results are shown in the Table. 8. In the original 50 frames, the recall increased by 4.7%, and for the entire 99 frames it increased by 5.3%. This illustrates that the increase in resolution and the number of frames is beneficial for moving object tracking.



(a) Training process on the coupling of optical flow and DConv. (b) Training process on the different ways of coupling.

(c) Training process on the different scales. (d) Training process on the number of residual blocks.

Fig. 6. The training process of the ablation experiment of the model.

Table 5
Ablation study by using different couple strategy.

Model	Naive Coupling	Deep Coupling (Ours)
PSNR (dB)	32.3042	**32.3796**

Table 6
Ablation study by using different patch scales.

Model	40 × 40	20 × 20	10 × 10	5 × 5	Ours
PSNR (dB)	32.2869	32.3142	32.3188	32.2951	**32.3796**

4.6. *Discussion of guidelines for satellite video ST-SR*

Although ST-SR can achieve end-to-end spatial and temporal resolution improvements, its design is still not sophisticated enough compared to the independent tasks S-SR and T-SR, which have been widely studied so far. For arbitrary factors of T-SR, some works have attempted to synthesize missing frames at any temporal location through modulation networks (Xu et al., 2021) or controlled optical flow estimation (Shi et al., 2021), which may be the guidelines of future research. For large scale of S-SR (×8, ×16), most methods have poor generalization performance when facing such a more challenging

Table 7
Model efficiency comparison. FLOPs are calculated on an HR frame size of 320 × 320. The processing time is the total time taken to complete 10 test sets divided by the total number of frames.

Method	FLOPs (G) ↓	Processing time (s) ↓	PSNR (dB) ↑
XVFI + EDSR	51.47 + 321.76	0.0549 + 0.1983	33.4944
XVFI + RCAN	51.47 + 11.75	0.0549 + 0.1755	33.9022
XVFI + EDVR	51.47 + 28.13	0.0549 + 0.1109	34.0296
XVFI + RBPN	51.47 + 276.56	0.0549 + 0.3346	33.9758
STARNet	2298.48	0.4016	33.9711
Ours	**208.01**	**0.1148**	**34.1750**

Table 8
Effects of STVSR on moving object tracking. Recall (%) is the average result of all frames.

Method	Original frame	All frames
Original video	73.80%	73.80%
After ST-SR	78.5% (↑4.7%)	79.1% (↑ 5.3%)

situation. More effort needs to be put into the large factor S-SR problem.

5. Conclusion

We present a lightweight framework that can jointly enhance the spatial and temporal resolution of satellite video in one-stage manner. A missing frame prediction module was proposed to predict the latent motion information with various scales. Under the constraints of optical flow, multi-scale deformable convolution can better converge and adaptively learn the supplementary motion information. The proposed multi-scale spatial–temporal transformer can effectively aggregate the contextual information in long-time series frames. Experiments on Jilin-1 satellite video demonstrate that our method can accurately predict the nonexistent frames and enhance the spatial resolution simultaneously.

In ongoing work, we will focus on the ST-SR under extreme events or large motion in satellite videos. In addition, owing to the insufficient data sets in real scenes, the performance of models trained on simulated degradation will severely drop in the real world, it is worth to build a real world data set that can reflect the complex degradation in satellite video.

CRediT authorship contribution statement

Yi Xiao: Methodology, Writing – original draft. **Qiangqiang Yuan:** Methodology, Supervision. **Jiang He:** Methodology. **Qiang Zhang:** Methodology. **Jing Sun:** Methodology. **Xin Su:** Methodology, Supervision. **Jialian Wu:** Methodology. **Liangpei Zhang:** Methodology, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (41922008,61971319) and the Hubei Science Foundation for Distinguished Young Scholars (2020CFA051).

References

Abid, N., Shahzad, M., Malik, M.I., Schwanecke, U., Ulges, A., Kovács, G., Shafait, F., 2021. Ucl: Unsupervised curriculum learning for water body classification from remote sensing imagery. Int. J. Appl. Earth Obs. Geoinf. 105, 102568.

Amato, G., Palombi, L., Raimondi, V., 2021. Data–driven classification of landslide types at a national scale by using artificial neural networks. Int. J. Appl. Earth Obs. Geoinf. 104, 102549.

Bao, W., Lai, W.-S., Ma, C., Zhang, X., Gao, Z., Yang, M.-H., 2019a. Depth-aware video frame interpolation. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3703–3712.

Bao, W., Lai, W.-S., Zhang, X., Gao, Z., Yang, M.-H., 2019b. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. IEEE transactions on pattern analysis and machine intelligence.

Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W., 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4778–4787.

Chen, D., Zhong, Y., Zheng, Z., Ma, A., Lu, X., 2021. Urban road mapping based on an end-to-end road vectorization mapping network framework. ISPRS J. Photogramm. Remote Sens. 178, 345–365.

Chen, Y., Shi, K., Ge, Y., Zhou, Y., 2022. Spatiotemporal remote sensing image fusion using multiscale two-stream convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 60, 1–12.

de Alwis Pitts, D.A., So, E., 2017. Enhanced change detection index for disaster response, recovery assessment and monitoring of accessibility and open spaces (camp sites). Int. J. Appl. Earth Obs. Geoinf. 57, 49–60.

Dong, C., Loy, C.C., He, K., Tang, X., 2015. Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38 (2), 295–307.

Dutta, S., Shah, N.A., Mittal, A., 2021. Efficient space-time video super resolution using low-resolution flow and mask upsampling. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 314–323.

Feng, J., Zeng, D., Jia, X., Zhang, X., Li, J., Liang, Y., Jiao, L., 2021. Cross-frame keypointbased and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos. ISPRS J. Photogramm. Remote Sens. 177, 116–130.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2020). A survey on visual transformer. arXiv preprint arXiv:2012.12556.

Haris, M., Shakhnarovich, G., Ukita, N., 2019. Recurrent back-projection network for video super-resolution. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3897–3906.

Haris, M., Shakhnarovich, G., Ukita, N., 2020. Space-time-aware multi-resolution video enhancement. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2859–2868.

He, J., Li, J., Yuan, Q., Shen, H., Zhang, L., 2021. Spectral response function-guided deep optimization-driven network for spectral super-resolution. IEEE Transactions on Neural Networks and Learning Systems.

He, J., Yuan, Q., Li, J., Zhang, L., 2022. PoNet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images. Information Fusion 80, 205–225.

Hore, A., Ziou, D., 2010. In: Image quality metrics: Psnr vs. ssim. IEEE, pp. 2366–2369.

Jiang, H., Sun, D., Jampani, V., Yang, M.-H., Learned-Miller, E., Kautz, J., 2018. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9000–9008.

Kang, J., Jo, Y., Oh, S.W., Vajda, P., Kim, S.J., 2020. In: Deep space-time video upsampling networks. Springer, pp. 701–717.

Lai, W.-S., Huang, J.-B., Ahuja, N., Yang, M.-H., 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In: In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 624–632.

Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., Schindler, K., 2018. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. ISPRS J. Photogramm. Remote Sens. 146, 305–319.

Lee, H., Kim, T., Chung, T.-Y., Pak, D., Ban, Y., Lee, S., 2020. Adacof: Adaptive collaboration of flows for video frame interpolation. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5316–5325.

Li, W., Zhang, X., Peng, Y., Dong, M., 2021. Spatiotemporal fusion of remote sensing images using a convolutional neural network with attention and multiscale mechanisms. Int. J. Remote Sens. 42 (6), 1973–1993.

Li, X., Foody, G.M., Boyd, D.S., Ge, Y., Zhang, Y., Du, Y., Ling, F., 2020. SFSDAF: An enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion. Remote Sensing of Environment 237, 111537.

Li, X., Zhang, G., Cui, H., Hou, S., Wang, S., Li, X., Chen, Y., Li, Z., Zhang, L., 2022. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. Int. J. Appl. Earth Obs. Geoinf. 106, 102638.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. Swinir: Image restoration using swin transformer. In: In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844.

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

Ma, A., Wang, J., Zhong, Y., Zheng, Z., 2021. Factseg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing.

Mittal, A., Soundararajan, R., Bovik, A.C., 2012. Making a "completely blind" image quality analyzer. IEEE Signal Process Lett. 20 (3), 209–212.

Mudenagudi, U., Banerjee, S., Kalra, P.K., 2010. Space-time super-resolution using graph-cut optimization. IEEE Trans. Pattern Anal. Mach. Intell. 33 (5), 995–1008.

Niklaus, S., Liu, F., 2018. Context-aware synthesis for video frame interpolation. In: In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1701–1710.

Peng, D., Bruzzone, L., Zhang, Y., Guan, H., He, P., 2021. Scdnet: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. Int. J. Appl. Earth Obs. Geoinf. 103, 102465.

Shahar, O., Faktor, A., Irani, M., 2011. Space-time super-resolution from a single video. In: In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3353–3360.

Shechtman, E., Caspi, Y., Irani, M., 2005. Space-time super-resolution. IEEE Trans. Pattern Anal. Mach. Intell. 27 (4), 531–545.

Shen, H., Lin, L., Li, J., Yuan, Q., Zhao, L., 2020. A residual convolutional neural network for polarimetric sar image super-resolution. ISPRS J. Photogramm. Remote Sens. 161, 90–108.

Shi, Z., Li, C., Dai, L., Liu, X., Chen, J., and Davidson, T. N. (2021). Learning for unconstrained space-time video super-resolution. arXiv preprint arXiv:2102.13011.

Sim, H., Oh, J., Kim, M., 2021. Xvfi: Extreme video frame interpolation. In: In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14489–14498.

Takeda, H., Van Beek, P., and Milanfar, P. (2010). Spatiotemporal video upscaling using motion assisted steering kernel (mask) regression. In High-Quality Visual Experience, pages 245–274. Springer.

Tian, Y., Zhang, Y., Fu, Y., and Xu, C. (2018). Tdan: Temporally deformable alignment network for video super-resolution. arXiv preprint arXiv:1812.02898.

Vandal, T.J., Nemani, R.R., 2021. Temporal interpolation of geostationary satellite imagery with optical flow. IEEE Transactions on Neural Networks and Learning Systems.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In Advances in neural information processing systems 5998–6008.

Wang, L., Guo, Y., Liu, L., Lin, Z., Deng, X., An, W., 2020. Deep video super-resolution using hr optical flow estimation. IEEE Trans. Image Process. 29, 4323–4336.

Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C., 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–10.

Wang, Y., Yuan, Q., Li, T., Zhu, L., Zhang, L., 2021. Estimating daily full-coverage near surface O3, CO, and NO2 concentrations at a high spatial resolution over China based on S5P-TROPOMI and GEOS-FP. ISPRS Journal of Photogrammetry and Remote Sensing 175, 311–325.

Wang, Y., Yuan, Q., Zhu, L., Zhang, L., 2022. Spatiotemporal estimation of hourly 2-km ground-level ozone over China based on Himawari-8 using a self-adaptive geospatially local model. Geoscience Frontiers 13 (1), 101286.

Xiang, X., Tian, Y., Zhang, Y., Fu, Y., Allebach, J.P., Xu, C., 2020. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In: In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3370–3379.

Xiao, Y., Su, X., Yuan, Q., Liu, D., Shen, H., Zhang, L., 2021. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. IEEE Trans. Geosci. Remote Sens. 1–19.

Xu, G., Xu, J., Li, Z., Wang, L., Sun, X., Cheng, M.-M., 2021. Temporal modulation network for controllable space-time video super-resolution. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6388–6397.

Xu, X., Siyao, L., Sun, W., Yin, Q., Yang, M.-H., 2019. Quadratic video interpolation. Advances in Neural Information Processing Systems 32, 1647–1656.

Zhang, Q., Yuan, Q., Li, J., Li, Z., Shen, H., Zhang, L., 2020. Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning. ISPRS Journal of Photogrammetry and Remote Sensing 162, 148–160.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y., 2018. Image super-resolution using very deep residual channel attention networks. In: In Proceedings of the European conference on computer vision (ECCV), pp. 286–301.

Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W., 2020. Fairmot: On the fairness of detection and re-identification in multiple object tracking. arXiv preprint arXiv: 2004.01888.

Zhang, Q., Yuan, Q., Li, Z., Sun, F., Zhang, L., 2021a. Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images. ISPRS Journal of Photogrammetry and Remote Sensing 177, 161–173.

Zhang, Q., Yuan, Q., Li, J., Wang, Y., Sun, F., Zhang, L., 2021b. Generating seamless global daily AMSR2 soil moisture (SGD-SM) long-term products for the years 2013–2019. Earth System Science Data 13 (3), 1385–1401.

Zhu, X., Hu, H., Lin, S., Dai, J., 2019. Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9308–9316.