ISPRS Journal of Photogrammetry and Remote Sensing 214 (2024) 179–192

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

RFE-VCR: Reference-enhanced transformer for remote sensing video cloud removal

Xianyu Jin, Jiang He, Yi Xiao, Ziyang Lihe, Xusi Liao, Jie Li, Qiangqiang Yuan *

School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei, China

ARTICLE INFO

Keywords:
Deep learning
Satellite videos
Video inpainting
Cloud removal

ABSTRACT

As a novel data source for earth observation, satellite video can provide large-scale temporal information for dynamic monitoring. However, the cloud occlusion prevents satellite video from continuous and seamless observation of the earth's surface. We propose the first satellite video cloud removal model RFE-VCR to approach this problem. In RFE-VCR, an efficient strategy of taking distant frames into training period is applied. A reference enhance block based on gated aggregation layers is proposed to explore the complementary information hidden in distant frames. A bidirectional local enhance block using deformable convolution is improved for feature refinement. Moreover, a decoupled temporal-spatial transformer is utilized for long-distance dependence modeling. Simulative and real experiments on Jilin-1 satellite videos demonstrate that our proposed network can achieve remarkable performance in video cloud removal task, as well as sensitive object hiding and high-reflection removal. More dynamic results of our experiments can be found at https://xyjin99.github.io/RFE-VCR/.

1. Introduction

Satellite video is a novel data source for earth observation, which has drawn more and more attention among researchers in recent years. With its high temporal resolution and large scale of land coverage, satellite video is beneficial to multiple long-term analysis tasks like dynamic monitoring (Xuan et al., 2019; Pan et al., 2022), object detection (Zhang et al., 2021), and even military usage. However, the quality of satellite video could be damaged easily by cloud occlusion, which is a common atmospheric phenomenon in earth observation. As shown in Fig. 1, while the 1500 × 1500 frame covers approximately 2.25 km^2 on the ground, about a quarter frame is covered by thick clouds or influenced by thin clouds, which inevitably limits the capability of video satellites to seamlessly record high-quality data of the earth surface. Dropping those areas affected by clouds directly might be an option, yet will result in a large amount of data lost and discarded. Hence, a cloud removal method for satellite video is urgently needed.

A potential way to conduct video cloud removal (VCR) is to extend existing cloud removal methods to satellite videos, which can take advantage of the property in remote sensing (RS) fields. However, the gaps among algorithms and data sources make it inappropriate for VCR. In RS, most methods either apply an inpainting-like method to remove clouds within one single satellite image (Enomoto et al., 2017; Li et al., 2019, 2020a; Xu et al., 2022; He et al., 2023b; Guo et al., 2023), or utilize multiple images to guide the restoration of one certain image (Sarukkai et al., 2020; Oehmcke et al., 2020; Zhang et al., 2020; Wang et al., 2023; Zheng et al., 2023). The former, denoted as RS image generation methods, can only be applied to satellite videos frame by frame, which lack temporal information utilization. The latter, denoted as sequence-to-image RS methods, which are also known as multi-temporal cloud removal methods, can aggregate information from several different satellite images by fully exploring the highly correlated spatial–spectral–temporal information (Zheng et al., 2023). When extending those methods to satellite videos, we need to choose other frames and regard them as extra temporal imagery for each frame in videos, which is time-consuming. A few other methods consider cloud removal as a sequence-to-sequence task (Zhang et al., 2021; Ebel et al., 2022; Peng et al., 2022; Zhao et al., 2023; Stucker et al., 2023). However, these methods are mostly based on multi-temporal datasets, which are multi-spectral and multi-source. Besides, the temporal resolution of those datasets is relatively low (5 days) so the change of cloud occlusion can be distinguished easily between adjacent temporal images. Nevertheless, there are only three basic channels RGB in satellite videos, thus the auxiliary information provided by other spectra or other data sources like synthetic aperture radar (SAR) images is not available. In addition, the temporal resolution of satellite videos (0.04 s) is much higher than multi-temporal data, which indicates that variation between adjacent frames could be tiny and hard to be utilized.

* Corresponding author.
E-mail addresses: jin_xy@whu.edu.cn (X. Jin), yqiang86@gmail.com (Q. Yuan).

https://doi.org/10.1016/j.isprsjprs.2024.06.010
Received 14 March 2024; Received in revised form 1 June 2024; Accepted 13 June 2024
Available online 24 June 2024

Fig. 1. Example scenario of cloud occlusion in satellite videos.

Fig. 2. The difference of distant frames between natural videos and satellite videos.

In summary, it is hard for existing RS methods to be transferred to VCR due to the inherent data gap as well as the lack of architecture exploration.

Another solution is resorting to computer vision (CV) inpainting methods for cloud removal, which can be roughly classified as image inpainting and video inpainting. Image inpainting methods (Guo et al., 2021; Zheng et al., 2022; Jain et al., 2023) try to fill the corrupted area by utilizing latent spatial correspondence from images, yet without the exploration of temporal dependence. Simply applying those methods in satellite video frames will result in spatial–temporal inconsistency. As for video inpainting (Zou et al., 2021; Zeng et al., 2020; Liu et al., 2021b; Li et al., 2022), although those methods are often conducted on RGB and high-temporal-resolution videos, the data gap between satellite and natural videos could decrease the performance of models. As we can see in Fig. 2, the natural and satellite videos have disparate purposes. Natural videos are object-oriented, which have a tracking manner of observation. The interested object is usually centered in the frame while the surrounding pixels and background always change rapidly. In that case, motions in local frames are abundant while the information from distant frames is hard to be utilized due to the large difference between frames. Thus, some algorithms would restrict information exchange within local adjacent frames to model interframe motion effectively and avoid interference from distant frames. However, there is a different situation in satellite videos. As it overlooks from the sky, satellite video observes in a sit-and-stare manner, making its background features hold still during the whole video. This different characteristic will result in three challenges in the algorithm. Firstly, the optical flow modeling which is often used in video inpainting has no effect. As shown in Fig. 3, the flow-warp operation will generate an occlusion at the same location in different frames and the interframe complementation will be lost. Secondly, the utilization of distant video frames becomes vital. Taking the complementary information of distant frames into account correctly can benefit the restoration of occluded background. Thirdly, the static background of satellite videos makes temporal modeling more important than the spatial. As available correspondence may exist in distant frames, the related features could be explored along temporal dimension for corrupted areas in advance, then spatial correspondence can be captured for feature hallucination and enhancement.

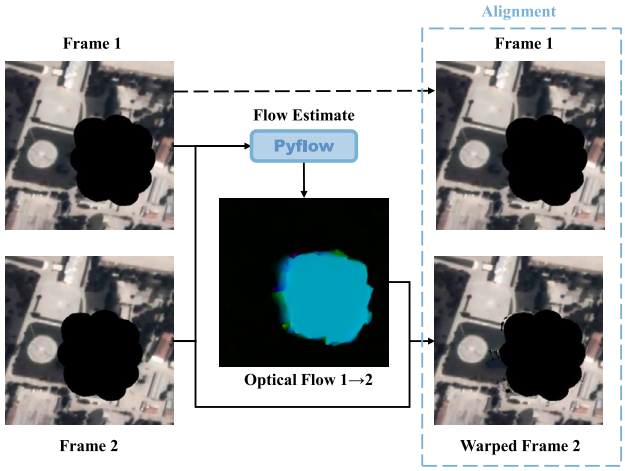


Fig. 3. The flow-warp operation generates masks at the same location in different frames. The optical flows are acquired using Pyflow in RBPN (Haris et al., 2019).

Based on the detailed analysis above, we proposed the first satellite video cloud removal architecture named Reference Frame Enhance VCR (RFE-VCR). To mask full use of abundant correlative information in distant frames, RFE-VCR takes them as reference frames, which dramatically improves the model performance. We denote the simple and effective strategy as distant frame training strategy (DFTS). Besides, a reference enhance block (REB) consisting of gated aggregation layer (GAL) is proposed to perform directional feature compensation from reference features to local features. Convolution layers are utilized in GAL to get the learnable mask for both features, which can efficiently exploit complementary information between frames. A modified bidirectional local enhance block (BLEB) is followed to perform feature propagation and aggregation within local frames. No optical-flow operations are used for either supervision or feature alignment in this method. Only deformable convolution is applied to fuse feature implicitly. Further, a decoupled temporal-spatial transformer (DTST) is used to model long-distance correspondence in a temporal-to-spatial order, paying more attention to temporal information meanwhile reducing the huge computation of spatial–temporal modeling in the vanilla transformer.

The contributions of this work are summarized as follows:

(1) Introducing distant frames as extra information, the proposed RFE-VCR utilizes complementary features among reference frames. With the simple but effective strategy, RFE-VCR gains

huge performance improvement in cloud removal task. Such a strategy has robustness for not only RFE-VCR but also other video inpainting methods in VCR task.

(2) A reference enhance block REB is proposed by utilizing GAL to mask different temporal features, conducting efficient feature compensation. Besides, a bidirectional local enhance block BLEB is modified according to the sit-and-stare observation of satellite videos. Furthermore, a decoupled temporal-spatial transformer is applied for modeling long-distance correspondence at a lower computation cost and laying more emphasis on temporal dependence capture.

(3) As the first satellite video cloud removal model, the proposed RFE-VCR achieves remarkable performance in both simulative experiments and real experiments on Jilin-1 satellite videos.

(4) The proposed model can also solve the high reflection removal and tiny object shelter tasks without additional fine-tuning.

The rest of this paper is organized as follows. Section 2 introduces the related work to VCR tasks. The proposed methodology is described in Section 3 and the experimental results on Jilin-1 satellite video dataset are provided in Section 4. Section 5 discusses the potential extensibility and limitations of the proposed model. Finally, the conclusions and prospects of our work are summarized in Section 6.

2. Related work

VCR aims at recovering cloud-free areas from corresponding cloudy videos, which is a typical ill-posed problem. As deep learning-based methods have made great progress in solving ill-posed problems in the fields of CV (Liu et al., 2021c; Ren et al., 2022; Wu et al., 2022) and RS (He et al., 2022, 2023a; Xiao et al., 2024b) in recent years, we mainly focus on these methods in this paper. As far as we know, there are no existing methods for satellite video cloud removal. The most related work can be classified into three main categories: RS multi-temporal cloud removal methods, CV image inpainting methods, and CV video inpainting methods. We reviewed these methods as below.

2.1. Multi-temporal cloud removal

Multi-temporal cloud removal methods search to utilize a sequence of satellite images for cloud removal. Some methods take a sequence of images as input and recover one image, denoted as sequence-to-image methods. Among them, a few approaches add SAR images as auxiliary information, making full use of the ability of SAR imaging to penetrate clouds. Sebastianelli et al. (2022) translated a SAR image into an optical image, providing a reference for cloudy optical inputs at another time. Ebel et al. (2022) proposed a model using a SAR sequence aligned to the optical cloudy image sequence to predict a single target image. Others conduct cloud removal on a specific image using multi-temporal optical images. Those methods either build a Unet-based architecture (Sarukkai et al., 2020) as well as a TimeGate strategy (Oehmcke et al., 2020) or use ResNet (Sarukkai et al., 2020; Zhang et al., 2020) and attention mechanism (Zhang et al., 2018; Yang et al., 2022) to incorporate the spatiotemporal information from the temporal-stacked image feature. Although these methods gained impressive performance on multi-temporal data, however, when applied to satellite videos, those sequence-to-image methods would have to be executed multiple times to recover the whole video sequence. Moreover, how to select other frames as multi-temporal information needs to be considered carefully and systematically, which is out of our research.

Another series of multi-temporal cloud removal methods is sequence-to-sequence models, which have been rarely studied in recent years. Ebel et al. (2022) utilized a batch of multi-temporal SAR images, performed translation from cloudless SAR to optical images, and reconstructed the corresponding optical sequences. Peng et al. (2022) proposed a recurrent neural network equipped with a two-layer gated recurrent unit (GRU) to learn the mapping from a SAR sequence to an image sequence, mainly focusing on rice field pixels. Zhao et al. (2023) designed a multi-modal attention mechanism to fuse SAR and optical image sequences with an adversarial training strategy to accomplish multi-temporal cloud removal. All these above-mentioned methods require SAR images, as radio frequency signals in the ultrahigh- or superhigh-frequency band can penetrate clouds and capture information which is unavailable in the visible spectrum. In other words, those methods introduce cloud-free images from another data source for help. There is only one method that does not require SAR images (Stucker et al., 2023). The proposed model U-TILISE consists of a spatial encoder and decoder for feature encoding and decoding of each temporal image. It also utilizes an attention-based temporal encoder to model temporal dependence and conduct information interchange to implicitly capture spatial–temporal patterns. However, when sequence-to-sequence methods are applied to RGB satellite videos, there is no available source of corresponding SAR images so those SAR-required methods might lose effectiveness. Moreover, the high temporal resolution of satellite videos and tiny motion between frames make it hard to model spatial–temporal discrepancy.

2.2. Image inpainting

Image inpainting aims at recovering the damaged areas inside the image meanwhile maintaining its overall consistency. Impressive progress has been made in image inpainting thanks to the rapid development of deep-learning. Pathak et al. (2016) first introduced a GAN-based adversarial loss to make the inpainted results more realistic. Then (Yu et al., 2018) proposed a two-stage framework with a novel contextual attention layer to explicitly attend on related feature patches at distant spatial locations. EdgeConnnet (Nazeri et al., 2019) utilized image edges to guide reasonable structure generation. Partial convolution (Liu et al., 2018) and gated convolution (Yu et al., 2019) are employed to make the vanilla convolution more effective in free-form inpainting tasks. Then (Zeng et al., 2021) further improved (Yu et al., 2019) by applying iterative filling and confidence estimation to refine the textures. Recently, Guo et al. (2021) proposed a dual generation framework for both textual and structure reconstruction. Zheng et al. (2022) employed a transformer to directly capture long-distance dependence with an attention-aware layer. Inspired by the high-frequency fast fourier convolution, Jain et al. (2023) achieved a remarkable visual quality in both structure generation and repeating texture synthesis. Although great progress has been made in image inpainting, these algorithms can only take latent spatial features within an image into account but ignore the interframe information exploration when directly applied to videos. The lack of temporal feature aggregation brings temporal inconsistency and causes severe flickering artifacts.

2.3. Video inpainting

The goal of video inpainting is to fill up the corrupted holes and missing areas with continuous and reliable context in a video. Different from image inpainting methods, video inpainting methods can utilize redundant interframe information to generate temporally consistent results. In recent years, great progress has been made in deep learning-based video inpainting methods, which can be roughly divided into three classes: 3D convolution-based, attention-based, and flow-based methods. 3D convolution-based methods like (Wang et al., 2019) extended 2D convolution from image inpainting to three-dimension data, applying 3D convolution to solve the spatial–temporal modeling in videos. Variants like 3D gated convolution (Chang et al., 2019a) and temporal shift module (Chang et al., 2019b; Zou et al., 2021) are also introduced to improve the performance or reduce the complexity. Attention-based methods utilize attention for context matching (Lee et al., 2019) and aggregation (Hu et al., 2020), or make good use

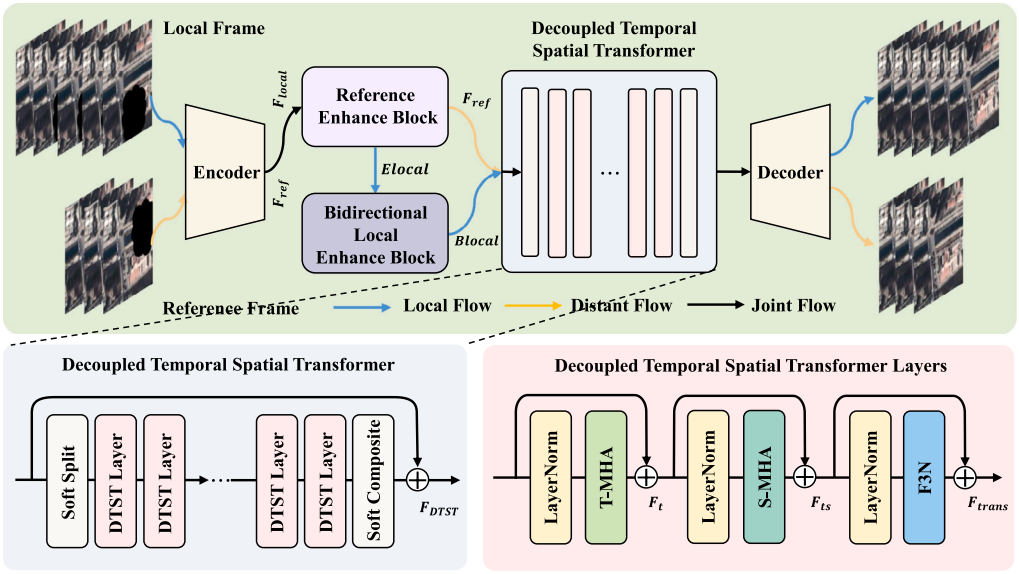


Fig. 4. Overall structure of our proposed network. Blue and yellow arrow denotes local and distant feature flow respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of transformer variants for long-distant correspondence modeling (Oh et al., 2019; Liu et al., 2021b; Zeng et al., 2020; Liu et al., 2021a). As for flow-based methods, which introduce optical flows between frames for interframe motion modeling, they often use optical flows to align frames (Kim et al., 2019; Li et al., 2020b, 2022), guide pixel propagation (Xu et al., 2019; Gao et al., 2020) or perform as the guidance of its transformer-based backbone (Zhang et al., 2022). Focusing on training strategy, some algorithms introduced internal learning into video inpainting (Zhang et al., 2019; Ouyang et al., 2021). They do not need extra datasets for training, but learn internal correspondence inside the specific video.

Although video inpainting is the most related work to our satellite video cloud removal task, most of them do not pay attention to distant frames. Some of the above methods (Liu et al., 2021c; Wu et al., 2021, 2022; Ren et al., 2022) train their models on the consecutive frames extracted from video sequences, restricting the information in local clips of videos. Other methods either introduce randomly sampled frames into the training stage (Lee et al., 2019; Chang et al., 2019a; Zou et al., 2021; Zhang et al., 2022; Kang et al., 2022), or sample frames randomly or consecutively at a 50% chance (Liu et al., 2021a; Zeng et al., 2020; Liu et al., 2021b), still lacking the explicit utilization of distant information. E2FGVI (Li et al., 2022) explicitly employs distant frames during training. However, it treats distant frames equally with local frames, only inserting them into the transformer step as an assistance of local frames. Besides, the widely used flow-based operations might be invalid in satellite videos. In this work, we proposed a distant frame training strategy and reference enhancement block to enforce the model to concentrate more on the complementary information hidden in those distant frames.

3. Methodology

3.1. Overview

Fig. 4 shows the overall structure of our proposed RFE-VCR. Given a cloud-free video sequence $\{X_t \in R^{3\times H\times W} \mid t = 1 \dots l + r\}$ consists of $l+r$ frames, we denoted the former l frames $\{X_1, \dots, X_l\}$ as local frames and the latter r frames $\{X_{l+1}, \dots, X_{l+r}\}$ as reference frames. We utilize the binary masks $\{M_t \in R^{1\times H\times W} \mid t = 1 \dots l + r\}$ to mask cloud areas, producing a sequence of corrupted frames. Our goal is to recover the masked areas in the whole sequence using both local and reference frames.

Taking l continuous local frames and r separate reference frames as input, RFE-VCR first uses a context encoder to encode all video frames to a low-resolution latent feature space for efficient computation. Secondly, reference features are delivered into the reference enhance block together with local features, where directional compensation is made from each reference feature to local features for long-distant feature complementation. After enhanced by reference features, local features are further performed feature propagation and aggregation in bidirectional local enhance block. Then the refined local features and reference features are concatenated and fed into the decoupled temporal-spatial transformer, which models the global dependence. Finally, RFE-VCR utilizes a decoder to decode the global-enhanced feature from the transformer for deep feature reconstruction and produce the final cloud-free sequence $\{\hat{X}_t \in R^{3\times H\times W}\}$. The combination of X_t and $\hat{X}_t$ using M_t is fed to a discriminator network for visually pleasant results.

3.2. Distant frame training strategy

As shown in Fig. 2, the distant frames far from local frames provide extra complementary information, which is helpful for recovering occluded areas under clouds. Therefore we apply this simple and effective strategy in both training and testing period. In training stage, we randomly select continuous local frames and separate reference frames. We concatenate these frames along the time dimension and feed them into the network. In testing stage, we evenly extract several frames from the video sequence according to the number of video frames as distant frames. We regard them as reference for the whole video and concatenate them with a batch of local frames in a sliding window to conduct cloud removal. During testing period, reference frames are discarded and only local reconstruction results are retained, as any reference frame could be one of those local frames in a particular sliding window of the whole sequence.

There are some similar strategies used in video inpainting (Liu et al., 2021a; Zeng et al., 2020; Liu et al., 2021b; Li et al., 2022) and video segmentation (Lao et al., 2023). However, most of them only apply the strategy in testing period but rarely in training stage. For instance, STTN (Zeng et al., 2020), FFM (Liu et al., 2021b), and DSTT (Liu et al., 2021a) utilize 50% continuous local frames and 50% randomly

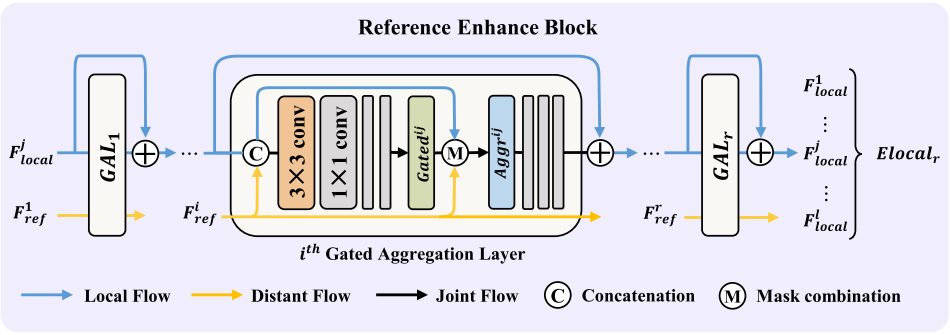


Fig. 5. Schematic of REB module with its component GAL. We take jth local frame for example. Same operation is conducted on other local frames in REB module and form $Elocal$ eventually.

selected frames during training. Although their strategy implicitly involves information from distant frames, the utilization of reference features is inadequate. Another algorithm (Li et al., 2022) utilizes distant frames explicitly in the transformer blocks during training, yet treats them equally with local frames. Nevertheless, the importance of distant frames should be emphasized in cloud removal tasks. As satellite video observes in a sit-and-stare manner and holds the background features still, we argue that reference frames are non-negligible in VCR and prove our speculation by experiments.

3.3. Reference enhance block

In satellite videos, the extra information hidden in reference frames should be explored and utilized sufficiently. Thus we propose a novel module named REB to conduct feature enhancement between local and reference features.

After embedded into the same feature space by the encoder, both local features F_{local} and reference features F_{ref} are fed into REB for directional feature compensation. As shown in Fig. 5, the REB consists of r gated aggregation layers for each reference feature $\left\{ F^i_{ref} \mid i \in 1 \dots r \right\}$ at different frame i. The ith GAL concatenates F^i_{ref} with local feature $\left\{ F^j_{local} \mid j \in 1 \dots l \right\}$ of each frame along the channel dimension, then utilizes four convolutions and a $tanh$ function to produce feature $Gated^{ij}$, which performs as a flexible mask to gate the information between local and reference features. The aggregated feature $Aggr^{ij}$ is fed into three convolution layers for feature refinement. In this way, after conducting feature fusion for each local frame, GAL outputs the reference-enhanced local features and adds them back to the input local features in the form of residuals. By delivering separate reference features into the GAL, the module digs out abundant information from distant frames. The REB step can be formulated as follows:

$$\begin{aligned} Elocal_i &= Elocal_{i-1} + GAL_i(F^i_{ref}, F_{local}), i \in 1 \dots r \\ Elocal_0 &= F_{local} \end{aligned} \tag{1}$$

where GAL_i denotes the ith gated aggregation layer for the ith reference frame and $Elocal_i$ denotes the enhanced feature output from GAL_i. In each GAL, the input reference feature F^i_{ref} interacts with each local feature F^j_{local} and achieves feature fusion by convolution and element-wise sum:

$$Gated^{ij} = f_{conv*4}([F^i_{ref}, F^j_{local}]) \tag{2}$$

$$Aggr^{ij} = (1 - Gated^{ij}) * F^i_{ref} + Gated^{ij} * F^j_{local} \tag{3}$$

$$F^j_{local} = f_{conv*3}(Aggr^{ij}) \tag{4}$$

where f_{conv*n} denotes n convolution layers, $j \in 1 \dots l$ denotes the frame index of local features. $Gated^{ij}$ and $Aggr^{ij}$ are the feature used for mask and gated features respectively.

3.4. Bidirectional local enhance block

Apart from exploring abundant information within distant frames, we also pay close attention to the information interaction among local features. Inspired by Chan et al. (2021, 2022), we propose a more effective module named BLEB for feature propagation and aggregation in VCR tasks, as shown in Fig. 6. Estimating optical flows could be pivotal in video inpainting tasks, where the motion information in the background helps track the shifting pixels and guiding the frame alignment. However, as for VCR task, in which background pixels mostly stay static during frames, the usage of optical flows becomes invalid and even harmful to the interframe information exaction. Therefore, we discard flow-based operation and utilize deformable convolution (DCN) operation in our BLEB module.

Obtaining the enhanced local feature $Elocal$ from REB, we generate an initial feature F_{prop} for both forward and backward propagation respectively. Take the backward propagation for example, we first concatenate F^{back}_{prop} with the current feature $Elocal^l$ along channel dimension, then stack several convolution layers to learn the offsets Δp and mask Δm for deformable convolution on F^{back}_{prop} as follows:

$$\Delta p, \Delta m = f_{conv}([F^{back}_{prop}, Elocal^l]) \tag{5}$$

$$F^{back}_{prop} = \sum_{k=1}^{n^2} w(p_k) \cdot F^{back}_{prop}(p + p_k + \Delta p_k) \cdot \Delta m_k \tag{6}$$

where p_k denotes the kth sampling offset in a standard convolution with kernel size $n \times n$, Δp_k is the additional learned offset and Δm_k is the modulation coefficient.

Subsequently, the spatially enhanced feature is concatenated with the current feature and fed into two convolution layers for interframe feature aggregation, which is further added back to F^{back}_{prop} in a residual form. The obtained feature is denoted as current refined feature $Back^l$. As it contains the information from both current feature and propagation feature, $Back^l$ can be regarded as a new propagation feature in the next time step:

$$Back^l = F^{back}_{prop} + f_{conv*2}([F^{back}_{prop}, Elocal^l]) \tag{7}$$

$$F^{back}_{prop} = Back^l \tag{8}$$

At the next time step, the new feature is further concatenated with the next current feature $Elocal^{l-1}$, where the same feature fusion above is performed to generate $Back^{l-1}$. The step-by-step process continues until the propagation reaches the first local frame and produces the refined feature $Back^1$:

$$F^{back}_{prop}, Back^i = backprop_i(F^{back}_{prop}, Elocal^i) \tag{9}$$

where $i = l \dots 1$ denotes the reversed index of local frames and $backprop_i$ denotes the ith step of backward propagation.

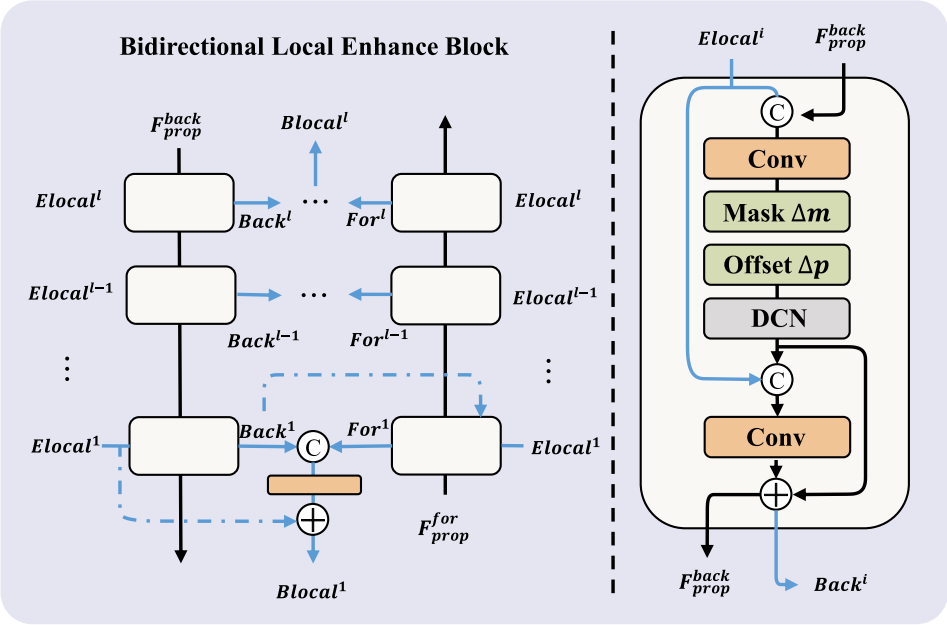


Fig. 6. Schematic of BLEB module with its backward propagation period.

The forward propagation step shares the symmetrical flow with backward propagation, which differs in the opposite direction from $Elocal^1$ to $Elocal^l$. Besides, the refined features in backward propagation are also utilized as extra information during the forward step, which can be formulated as follows:

$$For^j = F^{for}_{prop} + f_{conv*2}([F^{for}_{prop}, Elocal^j, Back^j]) \tag{10}$$

$$F^{for}_{prop} = For^j \tag{11}$$

where $j = 1 \dots l$ denotes the index of local frames. F^{for}_{prop} and For^j represent the propagation feature and refined jth feature in forward propagation respectively.

After getting the propagation features from two directions, BLEB finally employs a convolution as feature fusion. A global skip connection is also adopted for residual learning:

$$Blocal^h = Elocal^h + f_{conv}([For^h, Back^h]) \tag{12}$$

where $h = 1 \dots l$ denotes the index of local frames. By accomplishing the bidirectional propagation and aggregation, features within local frames are further enhanced.

3.5. Decoupled temporal-spatial transformer

To better explore the long-distance dependence in images, vision transformer (Dosovitskiy et al., 2020; Liang et al., 2021; Xiao et al., 2024a) is widely used in multiple fields. However, the huge computation cost makes transformer inefficient for high-resolution images, and even more so for videos. Thus, we employed a variant of vanilla transformer named decoupled temporal-spatial transformer (DTST) for VCR. The similar structure has been utilized in high-level tasks like video understanding (Bertasius et al., 2021) and low-level tasks like video inpainting (Liu et al., 2021a), in which the space–time factorization has been proven effective for reducing computation cost and improving accuracy. We argue that the decoupled structure for separate temporal and spatial modeling is also valid in VCR tasks. As there is little motion in satellite video background, the temporal correspondence can be competent when the clouds move fast between frames. When the masked area is large or the cloud motion is slow between frames, in which case some areas could be corrupted all over time, the proposed module resorts to spatial reference to generate continuous context. Such a structure alleviates the computation burden of transformer by preventing irrelevant spatial locations from long-distant correspondence calculation and makes the module concentrate more on the related areas across time domain. The decoupled temporal-spatial transformer is shown in Fig. 4.

After getting the reference-enhanced and local-enhanced features $Blocal$, we concatenate them with the original reference features F_{ref} along the time dimension, then utilize Soft Split in Liu et al. (2021b) to perform overlapped patch embedding and generate embedded feature $F_{embed} \in R^{B\times T\times h\times w\times m}$, where h and w denote the number of embedded patches in height and width dimension and m denote the feature dimension.

$$F_{embed} = SoftSplit([Blocal, F_{ref}]) \tag{13}$$

The embedded feature is further delivered into several DTST Layers (DTSTL), where two groups of self-attention are conducted for separate temporal-spatial modeling. In each DTSTL, F_{embed} is firstly reshaped into $R^{(B\times h\times w)\times T\times m}$ and passed through LayerNorm (LN) and an attention layer $atten_t$ for temporal modeling. Then the aggregated feature is reshaped again into $R^{(B\times T)\times(h\times w)\times m}$ and fed to the next group of LN and attention layer $atten_s$ for spatial modeling. A feed-forward network F3N in Liu et al. (2021b) is also employed for feature fusion. The process can be expressed as:

$$F^i_{trans} = DTSTL_i(F^{i-1}_{trans}) \tag{14}$$

where i denoted the ith DTSTL and F^0_{trans} equals to F_{embed}. In each DTSTL, two modules for temporal and spatial attention are conducted:

$$F_t = F^{i-1}_{trans} + reshape(atten_t(LN(reshape(F^{i-1}_{trans})))) \tag{15}$$

$$F_{ts} = F_t + reshape(atten_s(LN(reshape(F_t)))) \tag{16}$$

$$F^i_{trans} = F_{ts} + F3N(LN(F_{ts})) \tag{17}$$

where $atten_t$ and $atten_s$ denote temporal and spatial attention layers. In each attention layer, linear layers are used to calculate query/key/value vectors q, k, v and multi-head self-attention is performed:

$$q, k, v = Linear(F_{in}) \tag{18}$$

$$Attention(q,k,v) = Softmax(\frac{q(k)^T}{\sqrt{m}})v \tag{19}$$

$$F_{atten} = Linear(reshape(Attention(q,k,v))) \tag{20}$$

At last, the Soft Composite operation (Liu et al., 2021b) takes the output of the last transformer layer and regroups the aggregated feature into the same size as the original feature $F_{DTST} \in R^{B\times T\times C\times H\times W}$, where H, W, C are the height, width and channel of the input feature of transformer respectively.

$$F_{DTST} = SoftComposite(F_{trans}) \tag{21}$$

3.6. Loss function

We employ two widely used loss functions to train our model. One is the reconstruction loss measuring the L_1 distance between the restored video $\hat{X}$ and the ground truth video X using the mask M to distinguish the hole and valid area:

$$L_{hole} = \frac{\left\| M \odot (\hat{X} - X) \right\|_1}{\|M\|_1} \tag{22}$$

$$L_{valid} = \frac{\left\| (1-M) \odot (\hat{X} - X) \right\|_1}{\|1-M\|_1} \tag{23}$$

We also apply a GAN loss for realistic content generation in synthetic videos, which has been proven effective in video inpainting tasks (Li et al., 2022). The T-PatchGAN based discriminator is utilized to focus on the local and global features of all temporal neighbors, where the adversarial loss of the video cloud removal generator is formulated as:

$$L_{adv} = -E_{z\sim P_{\hat{X}}(z)}[D(z)] \tag{24}$$

D denotes the discriminator network, where the training object is:

$$L_D = E_{x\sim P_X(x)}[ReLU(1-D(x))] + E_{z\sim P_{\hat{X}}(z)}[ReLU(1+D(z))] \tag{25}$$

where $\hat{X}$ and X denote the restored video and the ground truth video respectively, z and x are frames belonging to data distribution of $\hat{X}$ and X. ReLU denotes the activation function Rectified Linear Unit.

4. Experiment

4.1. Dataset setting

In this section, we introduce the used dataset and experimental details first. Then we compared our proposed RFE-VCR with several state-of-the-art (SOTA) methods from the field of both CV and RS. Further, a series of comprehensive ablation studies are conducted to verify the effectiveness of our proposed strategy and modules.

4.2. Settings

4.2.1. Dataset

We use the ground truth data of the Jilin-1 video super-resolution dataset in Xiao et al. (2021) as our training dataset, which includes 189 satellite video clips cropped from eight videos. Each video clip has 100 frames with 640*640 spatial resolution. Following Li et al. (2022), we generate moving irregular masks over video clips to simulate the motion of clouds during training. As for evaluation, we cropped 6 scenes of 320*320 spatial resolution from the original testset of Jilin-1 dataset, which includes static objects like buildings and roads, as well as tiny moving objects like cars. The cloud-shaped mask in Zhang et al. (2020) is used to randomly generate evaluation mask sequences for each clip, which contains different sizes, moving directions and speed.

Table 1
Quantitative comparisons with SOTA models on our satellite videos dataset. ↑ indicates higher is better and ↓ indicates low is better.

Models	Accuracy				Efficiency	
	PSNR ↑	SSIM ↑	LPIPS ↓	VFID ↓	Param (M)	GFLOPs
Ctsdg	24.71	0.8985	0.0611	0.8861	52.15	459.95
TFill	23.34	0.8836	0.0750	0.9039	64.18	113.31
FcF	26.39	0.9295	0.0439	0.6903	70.34	201.30
DSTT	32.85	0.9788	0.0464	0.3937	34.53	257.04
STTN	35.00	0.9920	0.0160	0.2412	16.56	632.64
FFM	36.99	0.9928	0.0101	0.1431	36.59	408.01
E2FGVI	38.13	0.9939	0.0097	0.1224	41.12	397.21
U-TILISE	19.71	0.8427	0.1108	0.7211	**0.88**	**80.66**
RFE-VCR	**40.67**	**0.9955**	**0.0076**	**0.0968**	20.43	243.46

4.2.2. Implement details

Our model takes five consecutive local frames and three extra reference frames as input, which are randomly sampled from the original clips and rescaled to 320*320 spatial resolution using bicubic interpolation. We also apply randomly flipping and rotation for data augmentation in training period. The batch size is set to be 2 and the total training iterations is 150 000. We train our model using the Adam optimizer (Kingma and Ba, 2014) with the momentum $\beta_1 = 0$ and $\beta_2 = 0.99$. The initial learning rate is set to 0.0001 and decays to 1/10 every 50 000 iterations. All experiments are conducted on the Pytorch framework using one NVIDIA RTX 3090 GPU.

4.3. Comparison

We compare our proposed model with eight state-of-the-art methods, including three image inpainting methods, four video inpainting methods along with the only one optical sequence-to-sequence cloud removal method U-TILISE (Stucker et al., 2023). Ctsdg (Guo et al., 2021), TFill (Zheng et al., 2022), and FcF (Jain et al., 2023) are selected as image inpainting methods. STTN (Zeng et al., 2020), FFM (Liu et al., 2021b), DSTT (Liu et al., 2021a), and E2FGVI (Li et al., 2022) are selected as video inpainting methods. For fair comparison, we retrained those methods carefully on our satellite video dataset and tested them under the same condition with our model. We choose PSNR, SSIM, LPIPS and VFID as our evaluation metrics. Specifically, PSNR and SSIM are widely used in low-level image restoration tasks. LPIPS measures the perception distance between two frames. VFID is used to evaluate the perceptual similarity of two input videos as in Li et al. (2022).

4.3.1. Quantitative results

Quantitative comparison on our satellite video testset is reported in Table 1. The image inpainting methods conduct cloud removal on each video frame independently, which ignore the temporal information and produce poor results. Video inpainting methods like DSTT (Liu et al., 2021a), STTN (Zeng et al., 2020) and FFM (Liu et al., 2021b) do not employ the distant frames during training explicitly, which results in inferior performance, even though their testing phase includes extra frames. The SOTA method E2FGVI (Li et al., 2022), which is the champion of NTIRE 2022 video inpainting challenge, can achieve comparable results with ours. However, the data characteristic gap between video inpainting and cloud removal tasks makes E2FGVI lack the full exploration for distant frames, which hinders the potential of their model. As for the remote sensing method, U-TILISE (Stucker et al., 2023) shows poor capability of removing clouds in satellite videos. We speculate the tiny model parameters limit its performance. Besides, they are conducted on multi-temporal cloud removal tasks originally, in which the temporal resolution is about 5 days. The high temporal resolution in satellite videos may be tough for the model as the complementary information in consecutive local frames could be compressed. On the contrary, our proposed RFE-VCR surpasses all other methods on the four metrics with a relatively light-weight parameter. With the elaborate design to fit the satellite video characteristics, RFE-VCR shows superior performance in video cloud removal task.

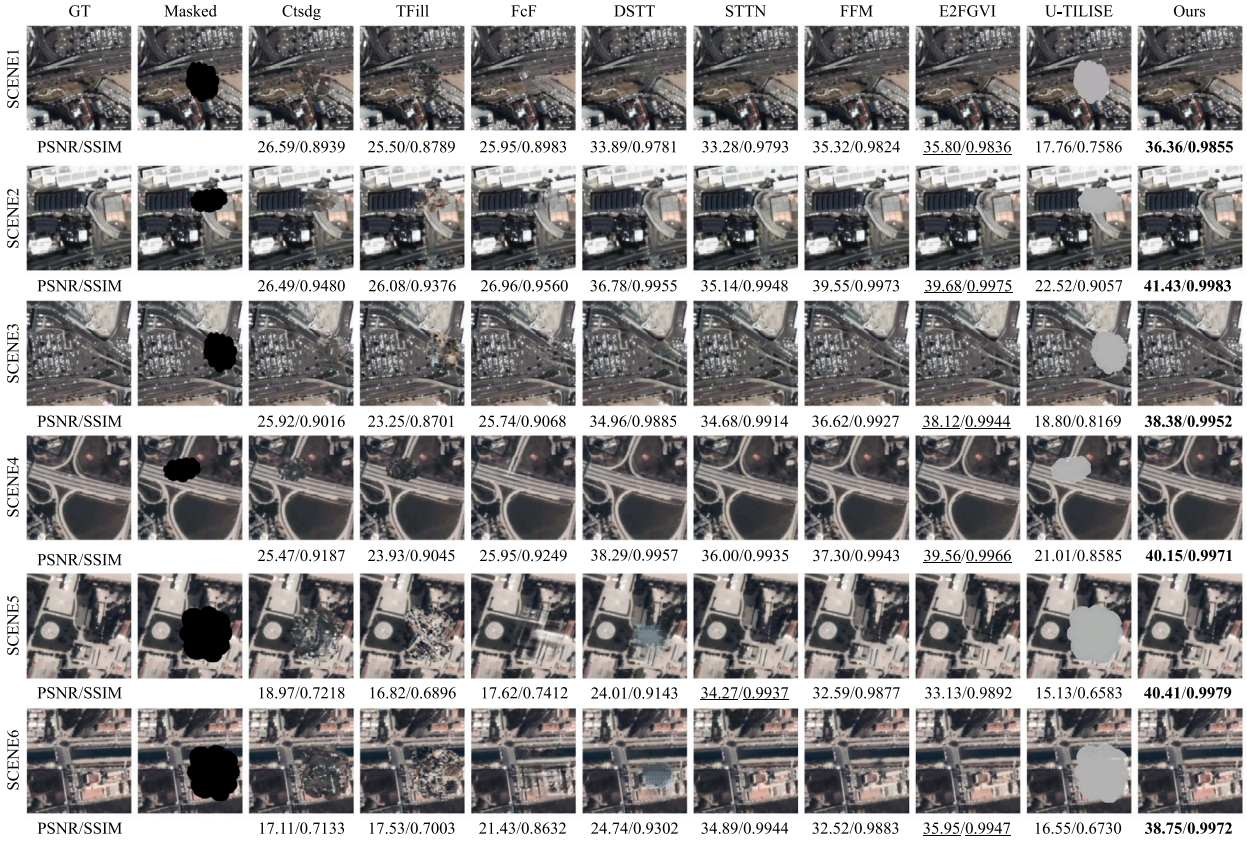


Fig. 7. Visual comparisons with SOTA models on our satellite videos dataset.

4.3.2. Qualitative results

Visual results are selected randomly from each test scene for qualitative comparison. As we can see in Fig. 7, image inpainting methods Ctsdg (Guo et al., 2021) and TFill (Zheng et al., 2022) fail to recover the occluded areas under clouds due to the lack of temporal information utilization. Although FcF (Jain et al., 2023) can generate relatively better results, the inherent disadvantage of image-based methods prevents it from revealing the correct structure in video frames, such as static roads under clouds in SCENE4 and buildings in SCENE5. As for video inpainting methods, the reconstruction results of DSTT (Liu et al., 2021a) contain blur and artifact, especially in the widely occluded areas of SCENE5 and SCENE6. STTN (Zeng et al., 2020) and E2FGVI (Li et al., 2022) can recover visually pleasant results relatively while FFM (Liu et al., 2021b) generates distorted texture around the cloudy area, representing as the bending roads in SCENE4 and the corrupted riverbank in SCENE5. Besides, U-TILISE (Stucker et al., 2023) can only recover a small part of the background at the edges of the cloud. Although it has gained remarkable performance in multi-temporal cloud removal task, the data gap along with the tiny capacity of its network make U-TILISE hard to be extended to VCR. Our method, on the contrary, can generate coherent context and structure which benefits from the efficient DFTS, REB, and BLEB modules. Further visual results of the temporal profile are shown in Fig. 8. The smoother the profile is, the more consecutive video the model generates. Overall, our proposed model can recover temporal consistent results compared to other methods, which demonstrates the effectiveness of our model.

4.3.3. Real-world results

We conduct real-world experiments on the Jilin-1 satellite videos and show the generalization ability of our model in Fig. 9. We roughly depict the cloud mask manually in the first frame and propagate it over the whole video clip by calculating its moving speed, as the accurate real-world cloud mask is not available and hard to be acquired. Then we utilize our proposed method to remove clouds. Although the mask is not accurate enough, our model can see through clouds and generate visually consecutive results. As most algorithms cannot achieve satisfactory performance, we only display the results of the second and third best algorithms FFM and E2FGVI in Table 1 for comparison. Some parts of the thick clouds seem not to be removed adequately in FFM, which results in blurry reconstructed scenes. We speculate this phenomenon might be caused by the implicit utilization of distant frames. As E2FGVI employs the distant frames explicitly, the clouds are removed more precisely and the results are sharper and clearer. Moreover, our proposed model not only takes distant frames into consideration explicitly using DFTS, but also utilizes a reference enhance block to fully explore the hidden information between frames. We further exhibit the temporal profiles of the reconstruction results in Fig. 10, where the cloud removal results of RFE-VCR are smoother and more consecutive than other video inpainting methods. Overall, our RFE-VCR is efficient, relatively light-weight, and well-generalized compared to other methods.

4.4. Ablation study

We conduct our ablation study starting from a basic backbone, which contains an encoder, a deep feature extraction module, and a decoder. The encoder and the decoder are stacked convolution layers, along with the dense skip-connection in the former. We conduct a comprehensive ablation study to verify the effectiveness of our proposed strategy and modules.

4.4.1. Distant frame training strategy

We choose two widely used backbones, 3D convolution (denoted as 3D Conv) and transformer as our deep feature extraction module.

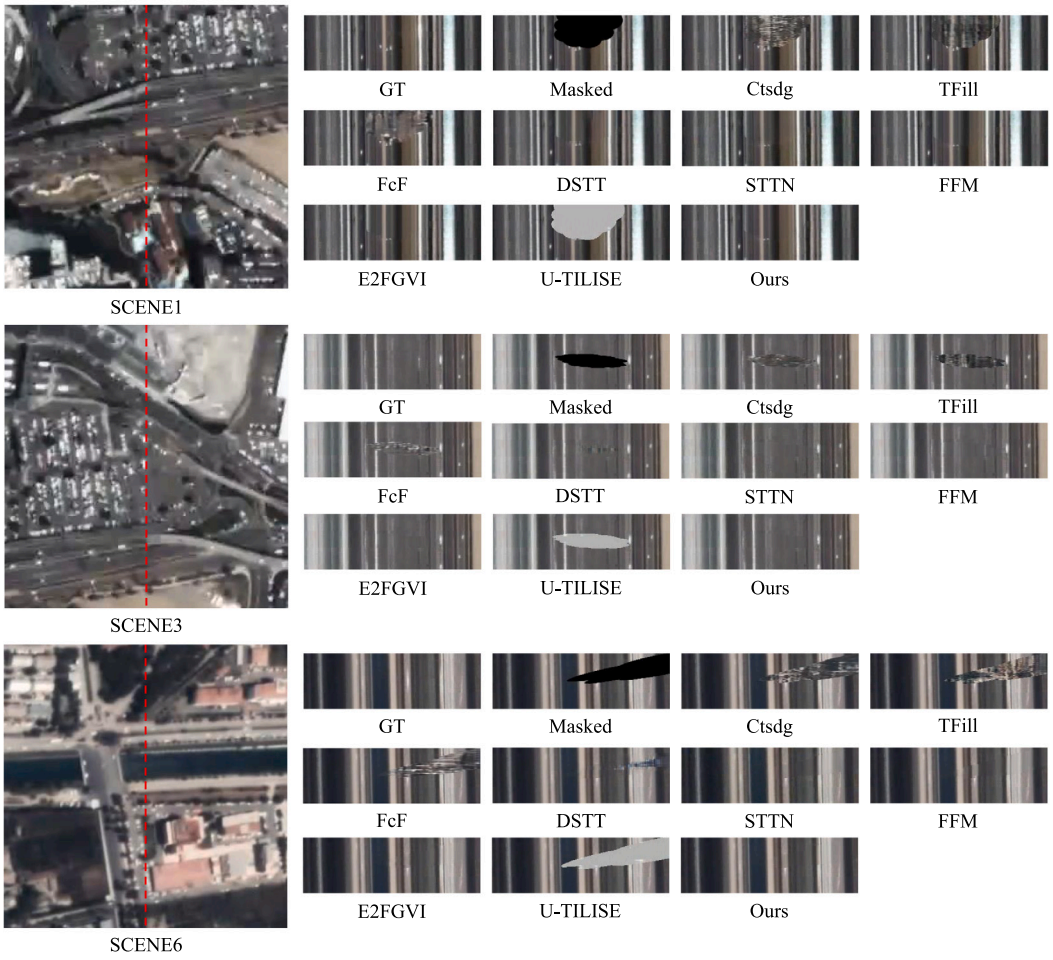


Fig. 8. Qualitative results of the temporal profiles. Red dashed line denotes the profile position. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

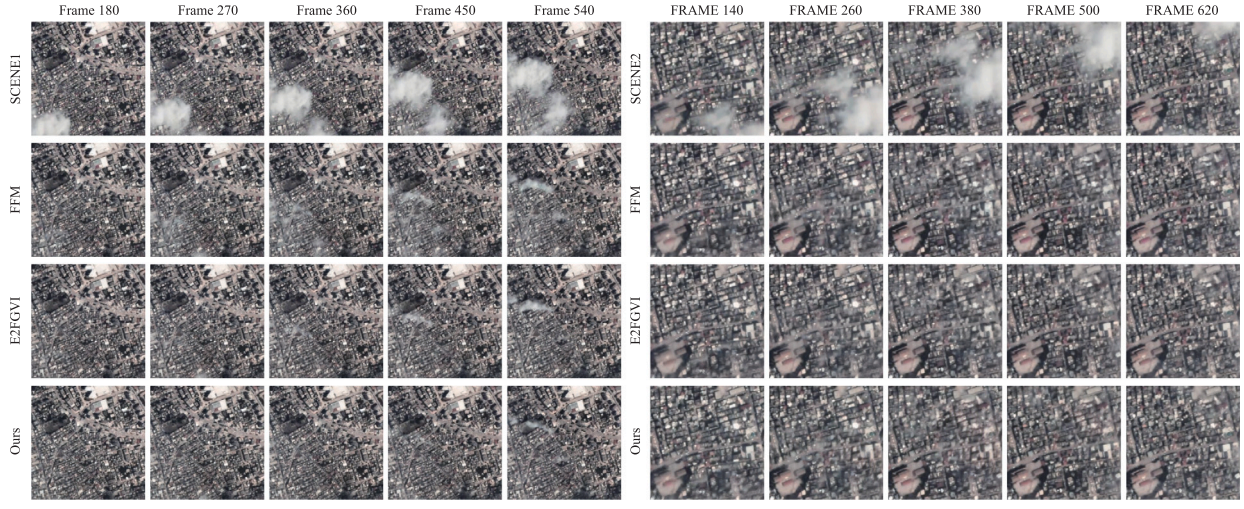


Fig. 9. Visual comparisons with SOTA models on our satellite videos dataset.

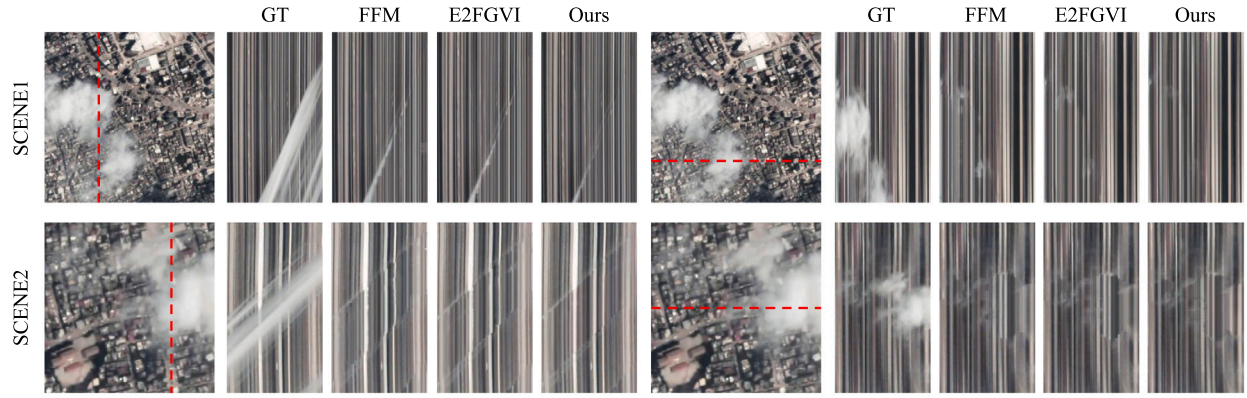


Fig. 10. Qualitative results of the temporal profiles in real-world experiments. Red dashed line denotes the profile position. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Ablation study of the proposed DFTS and DTST. Three distant frames are used in DFTS. Focal Transformer indicates temporal focal transformer used in Li et al. (2022). We choose the Decoupled TS Transformer as our DTST in the following experiments.

Type	Backbone	PSNR (dB)	
		w/o DFTS	w/ DFTS
CNN	3D Conv	26.17	31.14
Transformer	Spatial	25.86	25.86
	Joint ST	26.21	34.32
	Decoupled TS	26.64	**36.29**
	Decoupled ST	–	35.28
	Temporal	–	35.62
	Focal	–	34.71

Table 3
Ablation study of DFTS for other video inpainting methods STTN (Zeng et al., 2020) and DSTT (Liu et al., 2021a) in VCR task.

Models	Strategy	Metrics			
	DFTS	PSNR ↑	SSIM ↑	LPIPS ↓	VFID ↓
STTN	–	35.00	0.9920	0.0160	0.2412
	✓	36.72	0.9933	0.0190	0.2034
DSTT	–	32.85	0.9788	0.0464	0.3937
	✓	34.38	0.9863	0.0381	0.2886

We design three variants of the vanilla transformer, the one conducts self-attention on the joint spatial–temporal dimension (denoted as Joint ST), the one only conducts self-attention over the spatial dimension (denoted as $Spatial$), and our implemented Decoupled Temporal-Spatial transformer (denoted as Decoupled TS). Four models are trained using five consecutive frames, with or without DFTS using three extra distant frames.

As shown in Table 2, when the DFTS is not applied, none of those variants can successfully recover the corrupted areas, the training is failed and resulting in poor performance. However, with the introduction of distant frames, we observe that those who utilize temporal information improve their performance dramatically. We suspect that as the background of satellite videos often holds still over time, the motion within consecutive frames is too tiny to be explored so that interframe dependence cannot be used effectively. DFTS utilizes distant frames during training, which actually introduces extra motion information. By deploying DFTS, the interframe corresponding modeling is eased and models can make use of the temporal information better. Besides, we also verify the effectiveness of DFTS for other video inpainting methods in VCR task. Two models STTN (Zeng et al., 2020) and DSTT (Liu et al., 2021a) are trained with their original training strategy and our DFTS. The quantitative comparisons are shown in Table 3. It can be seen that with the explicit utilization of distant frames, other video inpainting methods can also gain performance improvement, indicating the robustness of DFTS for not only RFE-VCR but also other video inpainting methods in VCR, thanks to the additional compensatory information such a strategy introduces.

4.4.2. Decoupled temporal-spatial transformer

We further verify the effectiveness of the implemented DTST in Table 2. 3D convolution along with several variants of the transformer are conducted for comparison, which learn self-attention correspondence on: joint spatial–temporal (Joint ST), only spatial dimension ($Spatial$), only temporal dimension ($Temporal$) and spatial–temporal focal transformer ($Focal$) used in E2FGVI (Li et al., 2022). As for the decoupled temporal-spatial transformer, we designed two variants for opposite orders of decoupled modeling, where Decoupled TS denotes conducting self-attention first temporally then spatially and Decoupled ST denotes the reversed version. Experimental results show that Decoupled TS has the best modeling capability for spatial–temporal information, which is in line with our expectations. The Joint ST models correspondence within all spatial–temporal patches, which will inefficiently introduce unconcerned patches into calculation and increase the computational burden. $Spatial$ generates unsatisfying results due to the lack of modeling temporal information. $Focal$ is inferior to DST because its spatial–temporal window partition strategy loses sight of global correspondence. Besides, $Temporal$ is superior to $Spatial$ and Decoupled TS is a little better than Decoupled ST, indicating that the temporal dependence is more important than spatial correspondence in VCR.

4.4.3. Reference enhance block

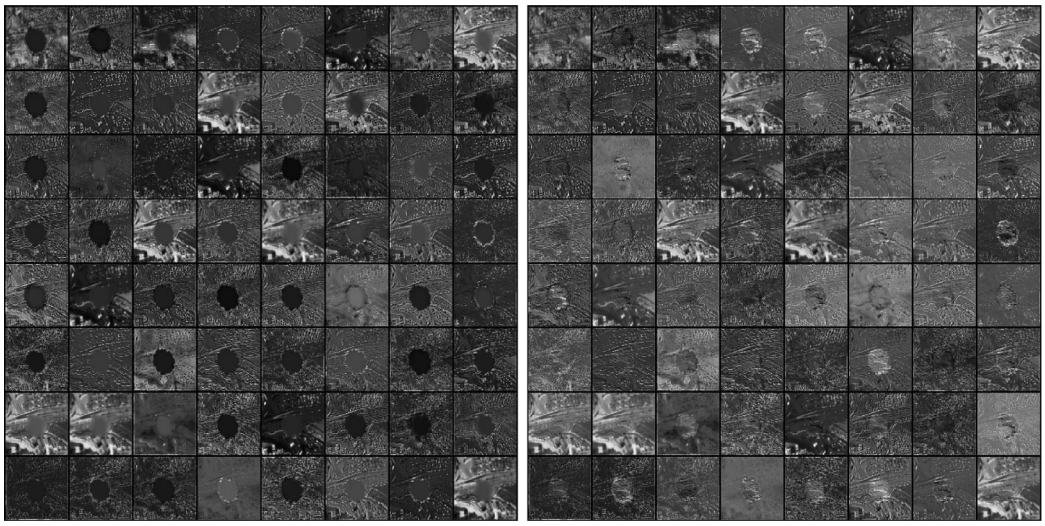
The effectiveness of our REB is verified in Table 4. We design two variants for comparison, one directly uses convolution layers to fuse local and reference features, denoted as REB_{conv} and another directly applies the ground truth mask to balance the fusion between features, denoted REB_{mask}. Table 4 shows that by learning the flexible gating feature for fusion, our proposed REB can conduct better feature exploration and fusion between frames. We also visualize the input and output features of REB. As shown in Fig. 11, information from distant frames is utilized in REB to accomplish feature complementation and enhancement.

4.4.4. Bidirectional local frame enhanced block

Two variants using flow-warp operation ($BLEB_{flow}$) and vanilla convolution ($BLEB_{conv}$) to replace the deformable convolution after concatenating the propagation feature and the current feature are conducted. As shown in Table 4, $BLEB_{flow}$ gained the lowest performance, indicating the efficacy loss in flow warping operations.

Table 4
Ablation study of our REB and BLEB modules. We choose DTST along with DFTS of three distant frames as the baseline.

Reference frame enhance block			Bidirectional local frame enhance block			PSNR
REB_{conv}	REB_{mask}	REB	$BLEB_{conv}$	$BLEB_{flow}$	$BLEB$	
–	–	–	–	–	–	36.29
✓	–	–	–	–	–	33.07
–	✓	–	–	–	–	33.60
–	–	✓	–	–	–	36.48
–	–	✓	✓	–	–	36.72
–	–	✓	–	✓	–	36.56
–	–	✓	–	–	✓	**36.80**



(a) 2nd input feature of REB (b) 2nd output feature of REB

Fig. 11. Visualization of the input and output features of REB.

The convolution-based $BLEB_{conv}$ shows lower performance than deformable convolution layers. By implementing our simplified but efficient BLEB, features within local frames are further explored and enhanced.

4.4.5. Parameter sensitivity analysis

We also conduct parameter sensitivity analysis for several hyper parameters of our proposed model, including the number of the transformer layers along with the number of local and distant frames used during training. All experiments prove the effectiveness and robustness of our proposed model.

Number of transformer layers. We build our RFE-VCR using different numbers of transformer layers, as shown in Fig. 12(a). The performance of our model increases continuously as the transformer block gets deeper.

Number of local frames. We train our model using different numbers of local frames, as shown in Fig. 12(b). The number of local frames used in training does not have a great influence to the model performance, as our BLEB module can handle arbitrary number of local frames and aggregate them efficiently. Broadly speaking, five consecutive local frames used in training result in the best model performance.

Number of distant frames. The number of distant frames determines the number of GAL layers in REB, which influences the model parameters. Thus we train variants of RFE-VCR using different numbers of distant frames. As shown in Fig. 12(c), when the number of distant frames increases from 1 to 5, the model performance keeps rising till saturation. However, when we use 7 frames as extra frames in DFTS, the performance slightly decreases. We speculate that the dense sample of distant frames in video sequence makes it more difficult for models to distinguish the useful information from several reference features.

Table 5
Mask combination strategy. $MaskComb$ denotes the mask combination operation used in Li et al. (2022).

$MaskComb$	PSNR (dB)	SSIM	VFID
–	36.80	0.9930	0.2126
✓	40.67	0.9955	0.0967

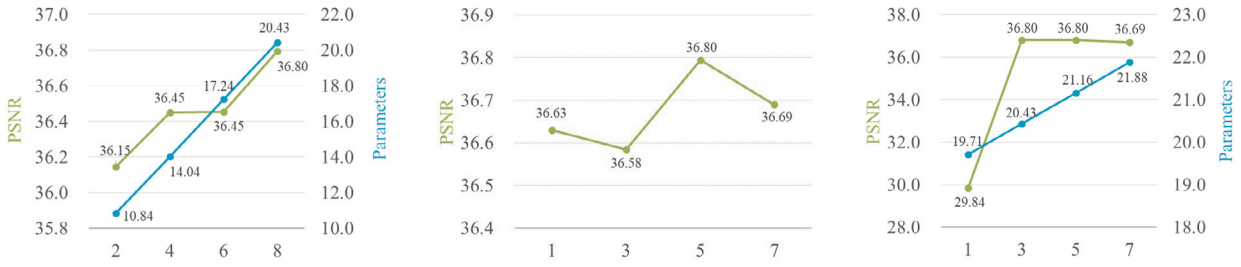
We ultimately choose three distant frames in DFTS as it can already achieve satisfactory performance.

4.4.6. Testing strategies

We observe that several testing strategies are applied in both image and video inpainting tasks Ctsdg (Guo et al., 2021), TFill (Zheng et al., 2022), FcF (Jain et al., 2023), STTN (Zeng et al., 2020), FFM (Liu et al., 2021b), E2FGVI (Li et al., 2022). For fair comparison, we adopt the most widely used testing strategy on our model, which utilizes the mask sequence M_t to combine the ground truth X_t and the synthetic results $\hat{X}_t$. As shown in Table 5, the model performance further gains.

5. Discussion and limitations

We investigate our model for other potential usages. For example, the tasks of hiding sensitive military objects like planes. We attempt to extend the object removing usage of video inpainting methods to satellite videos. Hence, we utilize SAM in Yu et al. (2023) to get the mask of the plane in the video clips and use our RFE-VCR to remove it, as shown in Fig. 13. Besides, for the high-reflection situations which often occur in satellite videos, we manually draw the mask of high-reflection areas of a 150-frame video clip, and utilize RFE-VCR to remove those undesired areas. Fig. 13 shows that our model also has



(a) Number of transformer layers (b) Number of Local frames (c) Number of distant frames

Fig. 12. Parameter sensitivity analysis of (a) Number of transformer, (b) Number of local frames, (c) Number of distant frames.

Fig. 13. Potential capability of our proposed model to hide objects and remove high-reflection areas.

the high-reflection removal capability dispensed with designing and training a brand-new model.

However, as our model is the first attempt for satellite video cloud removal task, still there are some drawbacks and limitations which can be improved in future work. For instance, the necessity of inputting masks makes the model inconvenient when dealing with real-world tasks like cloud or reflection removal. The clouds covering the earth contain thin and thick components, which are actually difficult to be distinguished from the background, especially the thin clouds. When the mask is not accurate, although cloud removal still can be conducted and generate overall visually pleasant results like in Fig. 13, it is easy to produce chromatic aberration alongside the mask edges. A similar situation exists in the real-world high-reflection removal task, and could be tougher as the high-reflection transformation during frames is more irregular than clouds. However, to our acknowledge, there are no existing video inpainting methods or VCR methods that could remove clouds without masks, due to the inherent characteristic of the specific task. As shown in Fig. 14, methods fail to remove clouds without given masks. So in this work, our compromised solution is manually drawing rough or accurate masks for clouds and high-reflections, which is very time-consuming. We also tried SAM for segmenting them but resulted in unsatisfactory masks. Maybe there would be a mask-free model or a segmentation method of high accuracy in the future to solve the aforementioned drawbacks. A potential solution is firstly conducting video cloud detection in satellite videos to obtain the accurate cloud masks of each frame, then conducting VCR in a successive pipeline. The video cloud detection method might be an extension of RS image cloud detection methods at the temporal dimension, taking deep consideration of the characteristics of satellite videos and we are working on it too.

6. Conclusion

As a common atmospheric phenomenon in earth observation, cloud occlusion often occurs in satellite videos. To solve the problem, we have proposed REF-VCR, the first deep learning-based model to deal

Fig. 14. Reconstruction results with and without given masks.

with the satellite video cloud removal task. An efficient training strategy DFTS along with a reference enhancement block are proposed to make full use of complementary information among distant frames. A bidirectional local enhance block is modified according to the sit-and-stare observation of satellite videos, in which we abandon flow-relevant operations and focus on feature fusion and enhancement by deformable convolution. Moreover, a decoupled temporal-spatial transformer is employed for long-distance correspondence modeling. Experiments have shown the effectiveness, robustness, extensibility and generalization ability of our proposed model. In addition, the proposed RFE-VCR can be further extended to the high-reflection removal and tiny object shelter tasks without fine-tuning. However, as the first attempt for satellite video cloud removal task, our work still has some limitations. For instance, masks are needed in the real-world tasks, which can be difficult and inaccurate. In future work, we will continue exploring more effective solutions to improve the quality of satellite videos.

CRediT authorship contribution statement

Xianyu Jin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jiang He:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yi Xiao:** Writing – review & editing, Resources, Funding acquisition. **Ziyang Lihe:** Writing – review & editing, Investigation. **Xusi Liao:** Investigation, Formal analysis. **Jie Li:** Supervision. **Qiangqiang Yuan:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

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.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China 42230108 and 423B2104, and the Fundamental Research Funds for the Central Universities under Grant 2042024kf0020 and 2042023kfyq04.

References

Bertasius, G., Wang, H., Torresani, L., 2021. Is space-time attention all you need for video understanding? In: ICML. volume 2, p. 4.

Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C., 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4947–4956.

Chan, K.C., Zhou, S., Xu, X., Loy, C.C., 2022. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5972–5981.

Chang, Y.-L., Liu, Z.Y., Lee, K.-Y., Hsu, W., 2019a. Free-form video inpainting with 3d gated convolution and temporal patchgan. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9066–9075.

Chang, Y.-L., Liu, Z.Y., Lee, K.-Y., Hsu, W., 2019b. Learnable gated temporal shift module for deep video inpainting. arXiv preprint arXiv:1907.01131.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Ebel, P., Xu, Y., Schmitt, M., Zhu, X.X., 2022. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., Kawaguchi, N., 2017. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 48–56.

Gao, C., Saraf, A., Huang, J.-B., Kopf, J., 2020. Flow-edge guided video completion. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer, pp. 713–729.

Guo, Y., He, W., Xia, Y., Zhang, H., 2023. Blind single-image-based thin cloud removal using a cloud perception integrated fast Fourier convolutional network. ISPRS J. Photogramm. Remote Sens. 206, 63–86.

Guo, X., Yang, H., Huang, D., 2021. Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14134–14143.

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

He, J., Li, J., Yuan, Q., Shen, H., Zhang, L., 2022. Spectral response function-guided deep optimization-driven network for spectral super-resolution. IEEE Trans. Neural Netw. Learn. Syst. 33 (9), 4213–4227.

He, J., Yuan, Q., Li, J., Xiao, Y., Liu, D., Shen, H., Zhang, L., 2023a. Spectral super-resolution meets deep learning: achievements and challenges. Inf. Fusion 97, 101812.

He, J., Yuan, Q., Li, J., Xiao, Y., Zhang, L., 2023b. A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection. ISPRS J. Photogramm. Remote Sens. 204, 131–144.

Hu, Y.-T., Wang, H., Ballas, N., Grauman, K., Schwing, A.G., 2020. Proposal-based video completion. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. Springer, pp. 38–54.

Jain, J., Zhou, Y., Yu, N., Shi, H., 2023. Keys to better image inpainting: Structure and texture go hand in hand. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 208–217.

Kang, J., Oh, S.W., Kim, S.J., 2022. Error compensation framework for flow-guided video inpainting. In: European Conference on Computer Vision. Springer, pp. 375–390.

Kim, D., Woo, S., Lee, J.-Y., Kweon, I.S., 2019. Deep video inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5792–5801.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lao, J., Hong, W., Guo, X., Zhang, Y., Wang, J., Chen, J., Chu, W., 2023. Simultaneously short-and long-term temporal modeling for semi-supervised video semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14763–14772.

Lee, S., Oh, S.W., Won, D., Kim, S.J., 2019. Copy-and-paste networks for deep video inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4413–4421.

Li, W., Li, Y., Chen, D., Chan, J.C.-W., 2019. Thin cloud removal with residual symmetrical concatenation network. ISPRS J. Photogramm. Remote Sens. 153, 137–150.

Li, Z., Lu, C.-Z., Qin, J., Guo, C.-L., Cheng, M.-M., 2022. Towards an end-to-end framework for flow-guided video inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17562–17571.

Li, J., Wu, Z., Hu, Z., Zhang, J., Li, M., Mo, L., Molinier, M., 2020a. Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion. ISPRS J. Photogramm. Remote Sens. 166, 373–389.

Li, A., Zhao, S., Ma, X., Gong, M., Qi, J., Zhang, R., Tao, D., Kotagiri, R., 2020b. Short-term and long-term context aggregation network for video inpainting. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer, pp. 728–743.

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

Liu, R., Deng, H., Huang, Y., Shi, X., Lu, L., Sun, W., Wang, X., Dai, J., Li, H., 2021a. Decoupled spatial-temporal transformer for video inpainting. arXiv preprint arXiv:2104.06637.

Liu, R., Deng, H., Huang, Y., Shi, X., Lu, L., Sun, W., Wang, X., Dai, J., Li, H., 2021b. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14040–14049.

Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B., 2018. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 85–100.

Liu, R., Weng, Z., Zhu, Y., Li, B., 2021c. Temporal adaptive alignment network for deep video inpainting. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 927–933.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M., 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212.

Oehmcke, S., Chen, T.-H.K., Prishchepov, A.V., Gieseke, F., 2020. Creating cloud-free satellite imagery from image time series with deep learning. In: Proceedings of the 9th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data. pp. 1–10.

Oh, S.W., Lee, S., Lee, J.-Y., Kim, S.J., 2019. Onion-peel networks for deep video completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4403–4412.

Ouyang, H., Wang, T., Chen, Q., 2021. Internal video inpainting by implicit long-range propagation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14579–14588.

Pan, J., Gu, Y., Li, S., Gao, G., Wu, S., 2022. Intrinsic satellite video decomposition with motion target energy constraint. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544.

Peng, T., Liu, M., Liu, X., Zhang, Q., Wu, L., Zou, X., 2022. Reconstruction of optical image time series with unequal lengths SAR based on improved sequence–sequence model. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Ren, J., Zheng, Q., Zhao, Y., Xu, X., Li, C., 2022. Dlformer: Discrete latent transformer for video inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3511–3520.

Sarukkai, V., Jain, A., Uzkent, B., Ermon, S., 2020. Cloud removal from satellite images using spatiotemporal generator networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1796–1805.

Sebastianelli, A., Puglisi, E., Del Rosso, M.P., Mifdal, J., Nowakowski, A., Mathieu, P.P., Pirri, F., Ullo, S.L., 2022. PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Stucker, C., Garnot, V.S.F., Schindler, K., 2023. U-TILISE: A sequence-to-sequence model for cloud removal in optical satellite time series. arXiv preprint arXiv:2305.13277.

Wang, C., Huang, H., Han, X., Wang, J., 2019. Video inpainting by jointly learning temporal structure and spatial details. In: Proceedings of the AAAI Conference on Artificial Intelligence. volume 33, pp. 5232–5239.

Wang, J.-L., Zhao, X.-L., Li, H.-C., Cao, K.-X., Miao, J., Huang, T.-Z., 2023. Unsupervised domain factorization network for thick cloud removal of multi-temporal remotely sensed images. IEEE Trans. Geosci. Remote Sens..

Wu, Z., Sun, C., Xuan, H., Zhang, K., Yan, Y., 2022. Divide-and-conquer completion network for video inpainting. IEEE Trans. Circuits Syst. Video Technol..

Wu, Z., Zhang, K., Xuan, H., Yang, J., Yan, Y., 2021. Dapc-net: Deformable alignment and pyramid context completion networks for video inpainting. IEEE Signal Process. Lett. 28, 1145–1149.

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. 60, 1–19.

Xiao, Y., Yuan, Q., Jiang, K., He, J., Lin, C.-W., Zhang, L., 2024a. TTST: A top-k token selective transformer for remote sensing image super-resolution. IEEE Trans. Image Process. 33, 738–752. http://dx.doi.org/10.1109/TIP.2023.3349004.

Xiao, Y., Yuan, Q., Jiang, K., Jin, X., He, J., Zhang, L., Lin, C.-W., 2024b. Local-global temporal difference learning for satellite video super-resolution. IEEE Trans. Circuits Syst. Video Technol. 34 (4), 2789–2802. http://dx.doi.org/10.1109/TCSVT.2023.3312321.

Xu, R., Li, X., Zhou, B., Loy, C.C., 2019. Deep flow-guided video inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3723–3732.

Xu, Z., Wu, K., Wang, W., Lyu, X., Ren, P., 2022. Semi-supervised thin cloud removal with mutually beneficial guides. ISPRS J. Photogramm. Remote Sens. 192, 327–343.

Xuan, S., Li, S., Han, M., Wan, X., Xia, G.-S., 2019. Object tracking in satellite videos by improved correlation filters with motion estimations. IEEE Trans. Geosci. Remote Sens. 58 (2), 1074–1086.

Yang, X., Zhao, Y., Vatsavai, R.R., 2022. Deep residual network with multi-image attention for imputing under clouds in satellite imagery. In: 2022 26th International Conference on Pattern Recognition. ICPR, IEEE, pp. 643–649.

Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z., 2023. Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5505–5514.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2019. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4471–4480.

Zeng, Y., Fu, J., Chao, H., 2020. Learning joint spatial-temporal transformations for video inpainting. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. Springer, pp. 528–543.

Zeng, Y., Lin, Z., Lu, H., Patel, V.M., 2021. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14164–14173.

Zhang, K., Fu, J., Liu, D., 2022. Flow-guided transformer for video inpainting. In: European Conference on Computer Vision. Springer, pp. 74–90.

Zhang, J., Jia, X., Hu, J., Tan, K., 2021. Moving vehicle detection for remote sensing video surveillance with nonstationary satellite platform. IEEE Trans. Pattern Anal. Mach. Intell. 44 (9), 5185–5198.

Zhang, H., Mai, L., Xu, N., Wang, Z., Collomosse, J., Jin, H., 2019. An internal learning approach to video inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2720–2729.

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 J. Photogramm. Remote Sens. 162, 148–160.

Zhang, Q., Yuan, Q., Li, Z., Sun, F., Zhang, L., 2021. Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images. ISPRS J. Photogramm. Remote Sens. 177, 161–173.

Zhang, Q., Yuan, Q., Zeng, C., Li, X., Wei, Y., 2018. Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network. IEEE Trans. Geosci. Remote Sens. 56 (8), 4274–4288.

Zhao, M., Olsen, P., Chandra, R., 2023. Seeing through clouds in satellite images. IEEE Trans. Geosci. Remote Sens..

Zheng, C., Cham, T.-J., Cai, J., Phung, D., 2022. Bridging global context interactions for high-fidelity image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11512–11522.

Zheng, W.-J., Zhao, X.-L., Zheng, Y.-B., Lin, J., Zhuang, L., Huang, T.-Z., 2023. Spatial-spectral-temporal connective tensor network decomposition for thick cloud removal. ISPRS J. Photogramm. Remote Sens. 199, 182–194.

Zou, X., Yang, L., Liu, D., Lee, Y.J., 2021. Progressive temporal feature alignment network for video inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16448–16457.