

Multiscale Restoration of Missing Data in Optical Time-Series Images With Masked Spatial–Temporal Attention Network

Zaiyan Zhang^{ID}, *Student Member, IEEE*, Jining Yan^{ID}, *Senior Member, IEEE*, Yuanqi Liang^{ID}, Jiaxin Feng, Haixu He^{ID}, and Li Cao

Abstract—Remote sensing images often suffer from substantial data loss due to factors such as thick cloud cover and sensor limitations. Existing methods for imputing missing values in remote sensing images fail to fully exploit spatiotemporal auxiliary information, which restricts the accuracy of their reconstructions. To address this issue, this article proposes a novel deep learning-based approach called multiscale masked spatial–temporal attention network (MS²TAN) for reconstructing time-series remote sensing images. First, we introduce an efficient spatiotemporal feature extractor based on masked spatial–temporal attention (MSTA) to capture high-quality representations of spatiotemporal neighborhood features surrounding missing regions while significantly reducing the computational complexity of the attention mechanism. Second, a multiscale restoration network composed of MSTA-based feature extractors (MFEs) is designed to progressively refine missing values by exploring spatiotemporal neighborhood features at different scales. Third, we propose a “pixel-structure-perception” multiobjective joint optimization method to enhance the visual quality of the reconstructed results from multiple perspectives and to preserve more texture structures. Finally, quantitative experimental results under multitemporal inputs on two public datasets demonstrate that the proposed method outperforms competitive approaches, achieving a 9.76%/9.30% reduction in mean absolute error (MAE) and a 0.56/0.62 dB increase in peak signal-to-noise ratio (PSNR), along with stronger texture and structural consistency. Ablation experiments further validate the contribution of the core innovations to imputation accuracy.

Index Terms—Masked spatial–temporal attention (MSTA), missing data restoration, multiobjective joint optimization, multiscale restoration, time-series remote sensing images.

I. INTRODUCTION

OVER the past few decades, remote sensing data has been extensively used in various industries. Among them,

high spatial resolution remote sensing imagery is particularly beneficial for applications such as vegetation monitoring, land cover mapping, and land cover change detection. However, fine spatial resolution images suffer from inevitable information loss caused by internal factors (e.g., sensor malfunction) and external factors (e.g., atmospheric conditions), which restrict their applications in different domains [1].

This article addresses the issue of missing data in remote sensing images involving multiple spectra. Common tasks include resolving the Landsat-7 ETM+ sensor scan line corrector off (SLC-off) problem and removing thick clouds. The key challenge is to estimate the missing regions and fill the gaps with predicted pixels, ensuring visual and semantic consistency with the surrounding pixels to enhance data usability.

Researchers have proposed various methods to recover missing data in remote sensing images. Early approaches to missing value restoration can be broadly categorized into three types [1]: spatial-based, temporal-based, and spatiotemporal-based data recovery. These methods have shown promising results in specific scenarios with low resolution and low missing rates. However, most of them rely on linear models and struggle to handle complex and detailed scenes. Moreover, due to limited reference information, the generated images often exhibit blurriness and lack continuous textures.

In recent years, the rapid advancement of deep learning theory and computer hardware [2] has led to significant progress in remote sensing image restoration using deep learning-based methods. These methods can be broadly categorized into two types: convolutional neural network (CNN)-based [3] and vision Transformer (ViT)-based [4] approaches. Compared to traditional statistical models, CNNs exhibit strong nonlinear expressive power, allowing for efficient extraction of spatial features from remote sensing data and significantly improving the accuracy of image restoration. However, these structures struggle to fully exploit time-series information, resulting in a bottleneck in restoration accuracy. ViTs, based on self-attention mechanism [5], possess a global receptive field, enabling comprehensive and efficient utilization of both temporal sequences and images to enhance reconstruction results [6]. However, due to the high resolution and long time-series of remote sensing images, token sequences become excessively long, leading to extremely high complexity in attention computations.

To efficiently mine spatiotemporal information in remote sensing images, we apply a self-attention mechanism sepa-

Received 18 November 2024; revised 20 April 2025; accepted 26 May 2025. Date of publication 29 May 2025; date of current version 11 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42471505, in part by Guangxi Key Research and Development Program under Grant (Guike) AB25069111, and in part by the National Key Research and Development Program of China under Grant 2022YFC3800700. (Corresponding author: Jining Yan.)

Zaiyan Zhang, Jining Yan, Yuanqi Liang, Jiaxin Feng, and Haixu He are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China, and also with the Engineering Research Center of Natural Resource Information Management and Digital Twin Engineering Software, Ministry of Education, Wuhan 430074, China (e-mail: zzaiyan@cug.edu.cn; yanjn@cug.edu.cn; naji@cug.edu.cn; fjxpen1234@cug.edu.cn; 20161001925@cug.edu.cn).

Li Cao is with the Second Surveying and Mapping Institute of Hunan Province, Changsha 410029, China, and also with the Key Laboratory of Natural Resources Monitoring and Supervision in Southern Hilly Region, Ministry of Natural Resources, Changsha 410029, China (e-mail: cl@img.net). Digital Object Identifier 10.1109/TGRS.2025.3574799

rately in the temporal and spatial dimensions and alternate between the two [7], greatly reducing computational complexity. To address the problem of significant distributional differences between missing and nonmissing values in remote sensing images, we apply missing values mask and diagonal mask [8] to the attention matrix, proposing masked spatial-temporal attention (MSTA) to enhance the expressive power of spatiotemporal attention and optimize the spectral discrepancy at the transition regions.

For full exploitation of the spatiotemporal neighborhood features at different scales [9], we further propose a multiscale restoration network. The network consists of MSTA-based feature extractors (MFEs) with different embedding scales, which progressively refine the reconstruction of missing information from coarse to fine granularity levels, achieving higher restoration accuracy [10]. For model training, we propose a “pixel-structure-perception” multiobjective joint optimization method, using pixel-wise loss as the basic loss for the restoration task and incorporating structural loss [11] and perceptual loss [12] to optimize the model’s results from the perspectives of structure, texture, shape, and spatial relations, thereby achieving high-quality image inpainting.

Finally, we performed both quantitative and qualitative experiments on two datasets, comparing our approach with several mainstream methods. In addition, we carried out ablation studies on the key innovations and analyzed the tradeoffs between effectiveness and efficiency across models of different sizes.

In summary, our main contributions are as follows.

- 1) We propose a deep learning-based method multiscale masked spatial-temporal attention network (MS²TAN) for reconstructing missing data in time-series remote sensing images. Our method utilizes a multiscale restoration network to learn an end-to-end mapping between incomplete and complete image sequences. MS²TAN achieved higher restoration accuracy than mainstream methods in quantitative experiments, and showed better visual effects in eliminating gaps in real data.
- 2) To address the challenges of high resolution and long time-series in remote sensing data, we introduce the MSTA mechanism. MSTA effectively extracts spatiotemporal features, improves the utilization of spatiotemporal context information, reduces color transition artifacts at the boundaries of missing value areas, and significantly reduces the computational complexity of self-attention.
- 3) For model training, we propose a “pixel-structure-perception” multiobjective joint optimization method. This method considers pixel-wise reconstruction error, structural reconstruction error, and perceptual error, resulting in restoration results with enhanced visual quality and preserved texture and structural details.

The rest of this article is organized as follows. In Section II, we review the existing methods for reconstructing missing information in remote sensing images. The network architecture and methodology details of our proposed model are presented in Section III. In Section IV, we showcase the

results of missing data reconstruction in both quantitative and qualitative experiments, compare them with mainstream methods, and validate the effectiveness of each component through validation studies. Finally, our conclusion and future prospects are summarized in Section V.

II. RELATED WORKS

A. Traditional Methods

Early research on the restoration of missing data in remote sensing images can be roughly divided into three categories: spatial-based, temporal-based, and spatiotemporal hybrid methods.

1) *Spatial-Based Methods*: Spatial-based methods rely solely on the valid information within the image itself to predict the missing data. The most commonly used approach is spatial interpolation methods [13]. In addition, methods based on partial differential equations (PDEs) [14] and variational methods [15], [16], [17] have also been utilized for reconstructing missing values. Furthermore, patch-based methods have found extensive application [18], [19], [20]. In general, spatial-based methods are suitable for reconstructing small missing areas or regions with regular textures. However, the accuracy of the reconstruction cannot be guaranteed, particularly for large regions or complex textures.

2) *Temporal-Based Methods*: Temporal-based methods utilize observations of the same location at different times from satellites to restore missing data. These methods include histogram matching-based approaches [21], temporal interpolation-based methods [22], replacement-based methods [23], and regression-based methods [24], [25], [26], among others. However, the differences between different time phases restrict the application of these methods.

3) *Spatiotemporal Hybrid Methods*: To overcome these limitations, spatiotemporal hybrid methods integrate the spatial and temporal correlations to reconstruct missing data under various conditions. For instance, improved nearest neighbor pixel interpolation methods [27], methods based on spatiotemporal Markov random field model [28], spatiotemporal weighted regression model [29], methods based on group sparse representation [30], methods based on the cascade of spatiotemporal reconstruction [31], and methods based on low-rank tensor decomposition [32], [33] have been proposed. However, most of these methods rely on linear models and struggle to handle complex and intricate scenes.

B. Deep Learning-Based Methods

In recent years, deep learning methods based on CNN and Transformer have been widely applied.

1) *CNN-Based Methods*: CNNs have shown high efficiency in extracting spatial features from remote sensing data, leading to significant improvements in the accuracy of remote sensing image restoration. Malek et al. [34] applied a context encoder [35] to reconstruct thick clouds in remote sensing images. CNNs combined with GAN structures [36], [37] were used to generate realistic reconstructed images. Zhang et al. [38] proposed a CNN-based spatiotemporal spectra (STS-CNN) framework, which was further developed into

a progressive spatiotemporal patch grouping framework [39]. CNNs that incorporate temporal inputs through channel-wise concatenation [40] merge feature maps of target images and temporal images, introducing auxiliary information for missing data restoration. Stucker et al. [41] used temporal self-attention for CNN's feature map sequences in conjunction with U-Net to repair temporal images. However, CNNs lack a true understanding of time-series, which hinders the efficient integration of temporal and spatial information in these methods and limits the utilization of long temporal sequences as auxiliary information, ultimately affecting the restoration accuracy.

2) *ViT-Based Methods*: ViT [4] has shown excellent performance in many tasks in the vision domain. Xu et al. [42] applied spatial self-attention to feature maps to capture the distribution of cloud thickness. Christopoulos et al. [43] utilized axial attention to remove thick clouds in remote sensing images. Recently, Liu et al. [44] used spatial attention and channel attention to remove cloud cover in the images, and Zhang et al. [45] explored the application of cascaded temporal and spatial attention in image restoration. However, the self-attention used in ViT requires computing the correlations between all pairs of patches (including all times). In the context of processing time-series remote sensing images, the images often have high resolutions and long time-series, resulting in a large number of tokens and extremely high computational complexity.

Bertasius et al. [7] conducted a detailed comparison of various forms of spatial-temporal attention and proposed separated spatial-temporal attention that achieves a balance between efficiency and performance. We further improved the separated spatial-temporal attention by introducing missing value masks and diagonal masks [8], resulting in MSTA, which exhibits superior performance in the task of missing value restoration. Compared to CNN-based methods and original ViT methods, MSTA enables efficient processing of spatiotemporal information and leverages long-term temporal information to assist in missing value reconstruction, leading to more precise reconstruction results.

III. METHODOLOGY

A. Problem Definition and Overall Framework

1) *Problem Definition*: The original time-series remote sensing image with missing values can be represented as $X \in \mathbb{R}^{T \times C \times H \times W}$, where T denotes the length of the time-series, C represents the number of channels, and H and W denote the spatial dimensions of the region. $X_{(t,c,i,j)}$ denotes the value of channel c at position (i, j) at time t . To account for the missing values, we introduce the missing value hint tensor $M \in \mathbb{R}^{T \times C \times H \times W}$, which is defined as shown in (1). In practical applications, the missing value hint tensor M can be obtained from the QA band or cloud detection algorithms such as Fmask [46] and S2Cloudless [47]

$$M_{(t,c,i,j)} = \begin{cases} 0 & \text{if } X_{(t,c,i,j)} \text{ is missing} \\ 1 & \text{if } X_{(t,c,i,j)} \text{ is observed.} \end{cases} \quad (1)$$

The target sequence Y in $\mathbb{R}^{T \times C \times H \times W}$ represents the actual data, while \tilde{Y} in $\mathbb{R}^{T \times C \times H \times W}$ represents the data repaired by the

model. Therefore, the definition of the multitemporal remote sensing image reconstruction model is given by (2). Here, the output \tilde{Y} of the model is the initial reconstruction result, and \odot denotes the Hadamard product. By replacing the observed real values in X with \tilde{Y} , we obtain the final reconstruction result \tilde{Y}_{out} as shown in (3)

$$\tilde{Y} = \text{Model}(X, M) \quad (2)$$

$$\tilde{Y}_{\text{out}} = \tilde{Y} \odot (1 - M) + X \odot M. \quad (3)$$

Our objective is to make \tilde{Y} closely match the data distribution of Y in order to achieve high-quality reconstructed images.

2) *Overall Framework*: The proposed framework for time-series image recovery consists of a multiscale restoration network (called MS²TAN) with MFEs and a “pixel-structure-perception” multiobjective joint optimization method, as depicted in Fig. 1.

The framework utilizes the MS²TAN to learn a complex nonlinear mapping from incomplete to complete information. It employs the MFE at different scales to extract temporal and spatial features for reconstruction. Finally, the network parameters are optimized using the “pixel-structure-perception” multiobjective joint optimization method, and the trained parameters are used for inference. The details of these components will be discussed in Sections III-B–III-D.

B. MSTA-Based Feature Extractor

The structure of the MFE module is illustrated in Fig. 2. It consists of position encoding and a cascade of L MSTA units. The input to this module is a high-dimensional feature vector $E \in \mathbb{R}^{TN \times d_{\text{emb}}}$, which is obtained by embedding the time-series remote sensing image X and the missing information mask M into blocks (as described in detail in Section III-C1). Here, T represents the time-series length, N represents the number of patches in a single remote sensing image, and d_{emb} represents the dimension of the token vector corresponding to each patch. Subsequently, the position encoding introduces spatial-temporal positional semantics, followed by L layers of MSTA units, and ultimately produces a sequence with incorporated spatiotemporal correlation features.

1) *Masked Self-Attention*: The Transformer model utilizes a self-attention mechanism to model sequences. It maps the input to query vector Q , key vector K , and value vector V . The attention score between Q and K is computed using scaled dot-product, followed by the application of the Softmax function to obtain attention weights A . The final output is the attention-weighted V , as shown in the following equation:

$$H = AV = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V$$

$$\text{where } [Q, K, V] = X \cdot W_{\text{QKV}}. \quad (4)$$

To address the challenge of disparate distribution of missing values in remote sensing data and enhance feature extraction capability, we introduce missing value masks and diagonal masks into self-attention.

1) The missing value mask sets the attention scores from patches with too high a missing rate to other patches

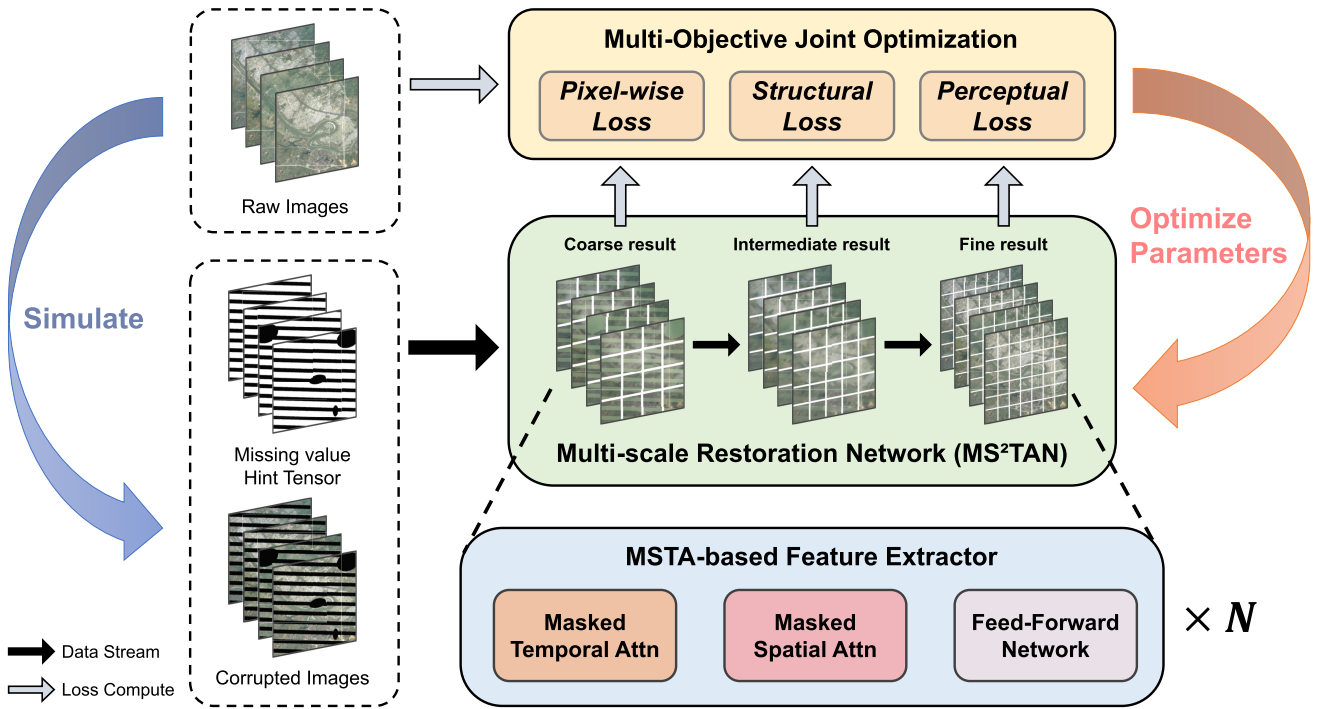


Fig. 1. Overall flowchart of the proposed method consists of two main components: a multiscale restoration network (named MS²TAN) with MFEs, and a “pixel-structure-perception” multiobjective joint optimization method.

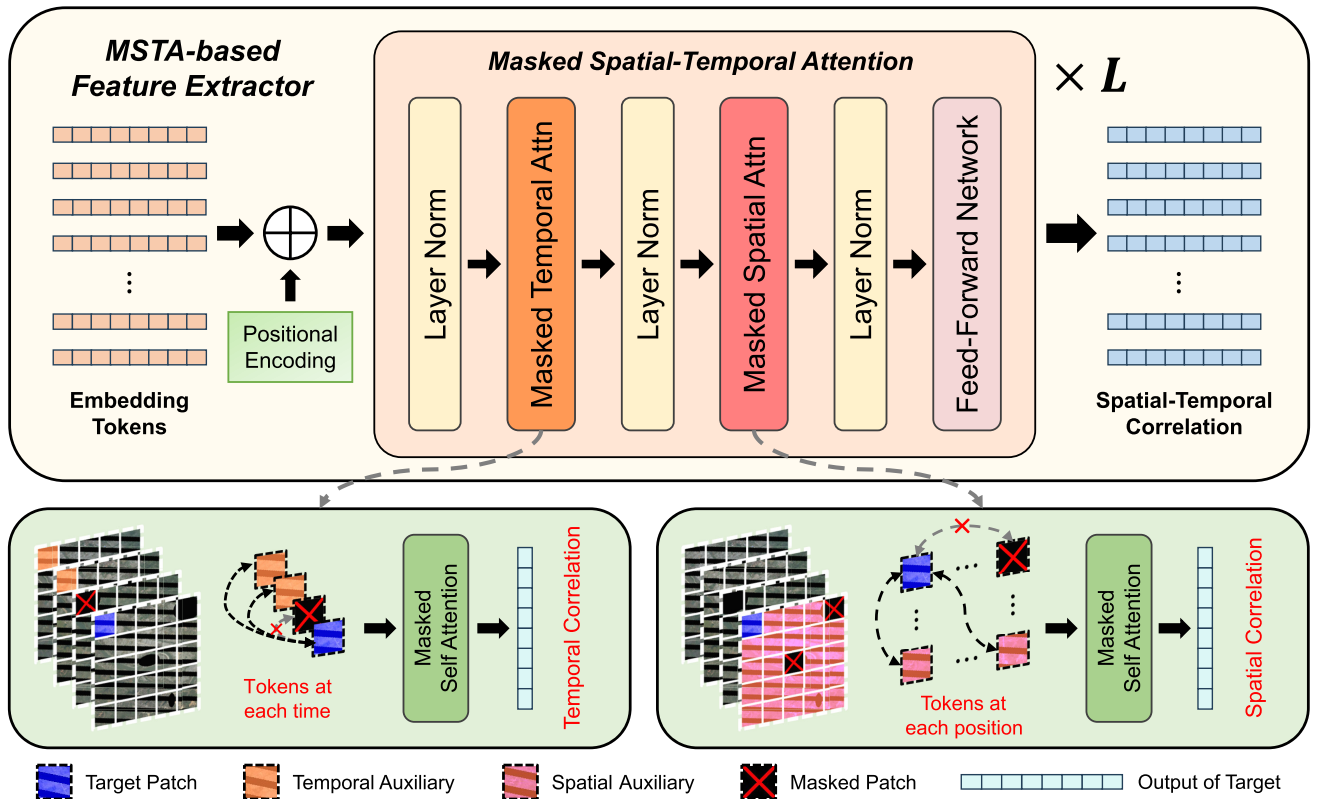


Fig. 2. Illustration of the structure of the MFE. The input sequence is augmented with positional encoding (PE) to incorporate positional information. It then undergoes L layers of MSTA operations to obtain the output sequence, which combines temporal and spatial contextual features. The masked temporal/spatial attention, introduced below the image, respectively, employs masked self-attention to capture the data distribution features in the temporal/spatial neighborhood.

as $-\infty$, effectively masking the influence of missing values. This enables the model to focus on extracting

useful information from nonmissing data to fill in the gaps, without being affected by the missing data.

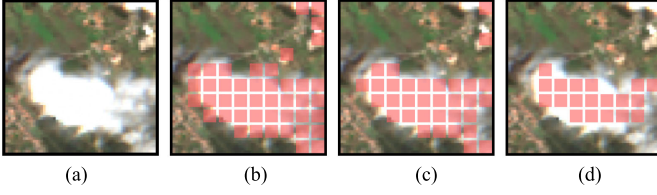


Fig. 3. Missing value masks obtained from different C_{\max} Values. The red squares in the figure indicate the masked patches. The scale of the mask is determined by the patch size, which is consistent with the embedding. This example uses a resolution of 80×80 and a patch size of 8. (a) Cloudy image. (b) $C_{\max} = 0.25$. (c) $C_{\max} = 0.5$. (d) $C_{\max} = 0.75$.

- 2) The diagonal mask sets the diagonal of the attention matrix as $-\infty$, preventing each step from attending to itself and forcing it to rely on the other $TN - 1$ steps for inference. This helps capture the spatiotemporal feature correlation in high-dimensional space.

More specifically, for the attention score matrix $Sc \in \mathbb{R}^{TN \times TN}$, the attention mask operation ApplyMask is defined as shown in the following equation:

$$\text{ApplyMask}(Sc)_{(i,j)} = \begin{cases} -\infty & \text{Mask}(i, j) = 1 \\ Sc_{(i,j)} & \text{Mask}(i, j) \neq 1. \end{cases}$$

$$\text{Mask}(i, j) = 1 \text{ only if } (\text{MR}(E(i)) > C_{\max} \text{ or } i = j). \quad (5)$$

Here, $\text{Mask}()$ checks whether the mask condition is satisfied, $\text{MR}(e)$ represents the missing rate of the corresponding patch e (which can be calculated from the hint tensor M), and C_{\max} is a hyperparameter that controls the maximum allowable missing rate. Fig. 3 shows the missing value masks obtained from different C_{\max} as an example of a cloudy image.

Applying (5) to the attention scores in (4), we obtain the expression for a single masked attention head, as shown in (6). Finally, by linearly projecting and combining multiple attention heads with the residual connection, we obtain the output of the masked self-attention, as shown in (7)

$$\begin{aligned} H^{(i)} &= A'V \\ &= \text{Softmax} \left(\text{ApplyMask} \left(\frac{Q \cdot K^T}{\sqrt{d_{\text{qkv}}}} \right) \right) \cdot V \end{aligned} \quad (6)$$

$$\text{MaskedAttn}(X, M) = \text{Proj}([H^{(1)}, H^{(2)}, \dots, H^{(h)}]) + X \quad (7)$$

where A' denotes the attention weights with mask applied, $\text{MaskedAttn}()$ represents the masked self-attention operation, $\text{Proj}()$ stands for projection head, and h denotes the number of attention heads.

2) *Masked Spatial-Temporal Attention*: MSTA is used to capture the spatiotemporal correlations within the input sequence, and consists of masked temporal attention (MTA), masked spatial attention (MSA), layer normalization (LN), and feed-forward network (FFN). These components are elaborated upon below.

For an input sequence $E \in \mathbb{R}^{TN \times d_{\text{emb}}}$, MTA first reshapes E to place the time dimension and the feature dimension at the end, obtaining a sequence e_t of length T and d_{emb} dimensions.

Subsequently, for each position in space, the MaskedAttn operation is performed along the temporal direction to obtain the output of MTA, as shown in the following equation:

$$\begin{aligned} e_t &= \text{Reshape}(E) \in \mathbb{R}^{N \times T \times d_{\text{emb}}} \\ \text{MTA}(E, M) &= \text{Reshape}(\text{MaskedAttn}(e_t, M)). \end{aligned} \quad (8)$$

Similarly, after the corresponding dimension transformation of the input sequence, the MaskedAttn operation is conducted along the spatial direction for each temporal image, yielding the output of MSA, as shown in the following equation:

$$\begin{aligned} e_s &= \text{Reshape}(E) \in \mathbb{R}^{T \times N \times d_{\text{emb}}} \\ \text{MSA}(E, M) &= \text{Reshape}(\text{MaskedAttn}(e_s, M)). \end{aligned} \quad (9)$$

FFN consists of two linear layers separated by a rectified linear unit (ReLU) activation function, with a residual connection established between the input and output. The expression for FFN is presented in the following equation:

$$\text{FFN}(X) = \text{Linear}(\text{ReLU}(\text{Linear}(X))) + X. \quad (10)$$

MSTA sequentially applies MTA and MSA to the input token sequence, exploring the spatiotemporal correlations among patches. Subsequently, FFN is used to fuse spatiotemporal features and introduce nonlinear transformations, as shown in the following equation:

$$\begin{aligned} U &= \text{MTA}(\text{LN}(E), M) \\ V &= \text{MSA}(\text{LN}(U), M) \\ \text{MSTA}(E, M) &= \text{FFN}(\text{LN}(V)). \end{aligned} \quad (11)$$

3) *Positional Encoding*: In the Transformer, PE is added to the input sequences to introduce positional semantics, as depicted in (12). Here, $\text{pos} = t \times N + n \in [0, TN - 1]$ represents the spatiotemporal joint coordinate derived from the temporal index $t \in [0, T - 1]$ and spatial index $n \in [0, N - 1]$, where N denotes the total number of patches in a single image. The function $\text{PosEnc}(\text{pos}, \text{dim})$ indicates the value of the dimth dimension of the PE for the posth patch in the image sequence

$$\begin{aligned} \text{PosEnc}(\text{pos}, 2i) &= \sin(\text{pos} \cdot 10000^{-2i/d_{\text{emb}}}) \\ \text{PosEnc}(\text{pos}, 2i + 1) &= \cos(\text{pos} \cdot 10000^{-2i/d_{\text{emb}}}). \end{aligned} \quad (12)$$

After PE, the input sequence will undergo L layers of MSTA to explore deeper spatiotemporal features. Finally, the overall expression of the MFE module is shown in (13), where the symbol L denotes stacking L layers, and $\text{MFE}()$ represents the MSTA-based feature extractor

$$\text{MFE}(E, M) = \{\text{MSTA}(E + \text{PosEnc}, M)\}^L. \quad (13)$$

C. Multiscale Restoration Network

The MS²TAN is a multiscale restoration network composed of S residual-connected restoration modules and observed value replacement, as illustrated in Fig. 4. Each module consists of patch embedding, MFE, and patch unembedding. These restoration modules can capture the spatiotemporal correlations of pixels at different scales to predict missing values and ultimately replace observed values as outputs.

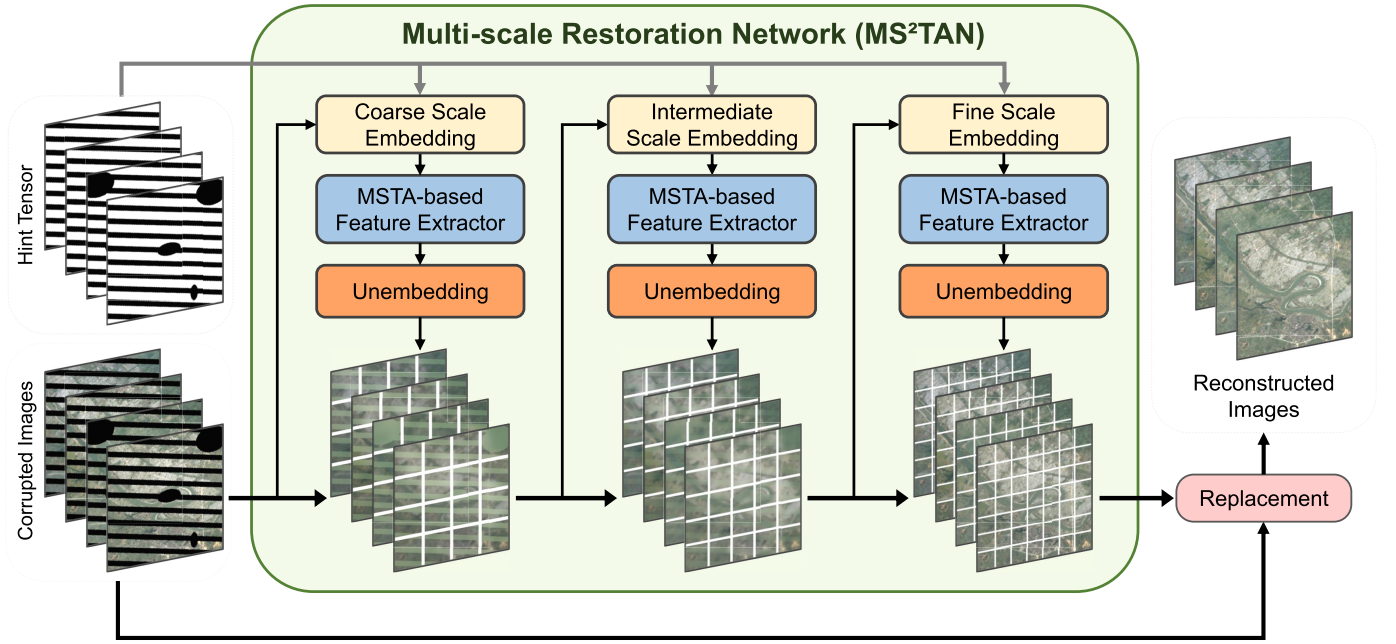


Fig. 4. Network structure diagram of proposed multiscale restoration network (MS²TAN), which contains several residual-connected restoration modules consisting of embedding, MFE, and unembedding.

1) *Embedding and Unembedding*: Following the approach of ViT, each image is partitioned into a series of patches, with the size of each patch determined by the hyperparameter P . For each temporal image, $N = HW/P^2$ patches are obtained. These patches are then aggregated and reorganized to form a patch sequence of length TN , with each patch represented as a CP^2 -dimensional vector. Equation (14) delineates the processing of the t th temporal image X_{in}^t

$$A^t = [p_1^t, p_2^t, \dots, p_N^t] = \text{Patchify}(x_{in}^t, P) \quad (14)$$

where $x_{in}^t \in \mathbb{R}^{C \times H \times W}$ and $A^t \in \mathbb{R}^{N \times CP^2}$.

Next, a linear layer is utilized to project the vectors corresponding to each patch into a high-dimensional representation space. The images from different time steps are then aggregated to obtain the complete token sequence E

$$E^t = \text{Linear}(A^t) \in \mathbb{R}^{N \times d_{\text{emb}}} \quad (15)$$

$$E = [E^1, E^2, \dots, E^T] \in \mathbb{R}^{TN \times d_{\text{emb}}}.$$

The process of transforming X_{in} step by step to obtain E as described above is referred to as embedding, which can be represented by (16). Conversely, if this operation is reversed, i.e., E is restored to a patch sequence and then rearranged into an image, it is referred to as unembedding, as shown in (17). Fig. 5 illustrates the processes of embedding and unembedding on images

$$E = \text{Embedding}(X_{in}, P) \quad (16)$$

$$X_{out} = \text{Unembedding}(E, P). \quad (17)$$

2) *MS²TAN*: MS²TAN consists of components such as embedding, MFE, and unembedding, where the patch size P is an important hyperparameter, and the embedding scale of the i th layer is denoted as $P^{(i)}$. The process of obtaining the embedding vector α from the input is described in (18).

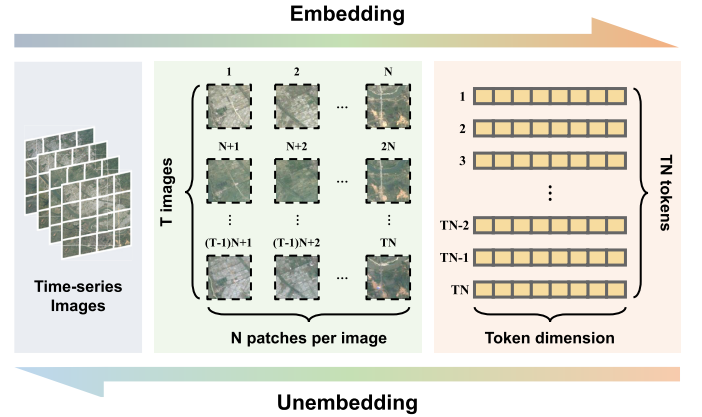


Fig. 5. Illustration of embedding and unembedding processes. Embedding represents a series of transformations from left to right, and unembedding represents the reverse.

Subsequently, MFE is utilized to mine deep spatiotemporal correlation features β , as shown in (19). Finally, the features are decoded, unembedded back into image form, and added to the input, resulting in an intermediate result as depicted in (20), where $\tilde{Y}^{(i)}$ represents the i th intermediate result

$$\alpha = \text{Embedding}(\tilde{Y}^{(i-1)}, P^{(i)}) \quad (18)$$

$$\beta = \text{MFE}(\alpha, M) \quad (19)$$

$$\tilde{Y}^{(i)} = \text{Unembedding}(\beta, P^{(i)}) + \tilde{Y}^{(i-1)}. \quad (20)$$

Let $\tilde{Y}^{(0)} = X_{in}$, and sequentially obtain S intermediate results of the reconstruction. Eventually, the S th result is taken as the output of MS²TAN, as shown in the following equation:

$$\tilde{Y} = \text{MS}^2\text{TAN}(X_{in}, M) = \tilde{Y}^{(S)}. \quad (21)$$

3) *Observed Value Replacement*: For the original reconstruction results of this network, denoted as \tilde{Y} , the observed

values are replaced with their nonmissing parts to obtain the final result \tilde{Y}_{out} , as shown in (3).

4) *Multiscale Restoration Strategy*: Common ViT-based models use a fixed patch size, which has been shown to be inefficient in dense prediction tasks like image segmentation and restoration [48]. In such tasks, smaller patch sizes often achieve better performance, but also come with greater computational costs. In order to balance the effect and efficiency, we apply a multiscale restoration strategy, applying a larger patch size at a scale close to the input to obtain a coarse result, and applying a smaller patch size at a scale close to the output to obtain a fine result.

D. Multiobjective Joint Optimization

To fully exploit the reconstruction capability of the model, this article proposes a multiobjective joint optimization method to train the network parameters. This method utilizes the “pixel-structure-perception” multiobjective loss function to optimize the results generated by the model from the perspectives of structure, color, texture, shape, and spatial relationships, achieving high-quality image inpainting.

1) *Pixel-Wise Loss*: Pixel-wise loss disregards the overall integrity of the image, treating the image as a collection of pixels, and comparing pixel-by-pixel to generate the image against the target image. It serves as the foundational loss for image reconstruction tasks. Equation (22) illustrates its calculation process. The pixel-wise loss calculated here includes both missing and observed parts. The reduction of the loss in the observed part does not directly improve the performance of the model. However, it does smooth the model output, making the model training process more stable, and hence is also included

$$\mathcal{L}_{\text{pixel-wise}} = \frac{1}{\text{TCHW}} \|\eta - y\|_2^2. \quad (22)$$

2) *Structural Loss*: Pixel-wise loss is commonly employed in various time-series and visual tasks but fails to consider the correlation between pixels. Structural loss uses structural similarity [49] to measure the difference from the target image to optimize the visual consistency of the reconstruction results in terms of structure, contrast, and luminance, and has been shown to have better performance in image reconstruction tasks [50]. Equation (23) specifies the computation process of structural loss, where μ_η and μ_y denote the means of η and y , respectively, σ_η and σ_y denote the variances of η and y , respectively, $\sigma_{\eta y}$ denotes the covariance between η and y , and C_1 and C_2 are constants

$$\begin{aligned} \mathcal{L}_{\text{structural}} &= 1 - \text{SSIM}(\eta, y) \\ &= 1 - \frac{(2\mu_\eta\mu_y + C_1)(2\sigma_{\eta y} + C_2)}{(\mu_\eta^2 + \mu_y^2 + C_1)(\sigma_\eta^2 + \sigma_y^2 + C_2)}. \end{aligned} \quad (23)$$

3) *Perceptual Loss*: Perceptual loss [12] is widely used in tasks like super-resolution and image generation, where pretrained CNNs are used as the feature network to extract perceptual features such as texture details and image style. Specifically, we use a VGG-16 [51] with the fully connected layer removed as the feature network and the L2 loss of the

feature map differences as the final loss. The calculation of perceptual loss is shown in (24), where $\psi()$ denotes the feature network and d_f denotes the feature dimension

$$\mathcal{L}_{\text{perceptual}} = \frac{1}{d_f} \|\psi(\eta) - \psi(y)\|_2^2. \quad (24)$$

4) *Multiobjective Loss Function*: For each scale of MS²TAN, the expression of the loss function $\mathcal{L}^{(i)}$ corresponding to its intermediate output $\tilde{Y}^{(i)}$ is shown in (25), where λ_1 , λ_2 , and λ_3 denote the weights of the pixel-wise loss, structural loss, and perceptual loss, respectively. The multiobjective loss function \mathcal{L} considers the intermediate results of each scale, using their mean value as the overall loss, as shown in (26)

$$\mathcal{L}^{(i)}(\eta, y) = \lambda_1 \mathcal{L}_{\text{pixel-wise}} + \lambda_2 \mathcal{L}_{\text{struct}} + \lambda_3 \mathcal{L}_{\text{perceptual}} \quad (25)$$

$$\mathcal{L} = \frac{1}{S} \sum_{i=1}^S \mathcal{L}^{(i)}(\tilde{Y}^{(i)}, Y). \quad (26)$$

IV. EXPERIMENTS

A. Settings

To assess the performance of the proposed model in different scenarios, we conducted quantitative and qualitative experiments on two datasets and compared it with some mainstream methods. In addition, we conduct ablation experiments on the key innovations and discuss the balance between effectiveness and efficiency for models with different sizes.

1) *Compared Algorithms*: Based on the type of inputs and outputs, common image restoration methods can be categorized into three main types: single-input single-output (SISO), dual-input single-output (DISO), and multi-input multi-output (MIMO). The SISO algorithms are the simplest type, but they struggle to achieve excellent performance, so we do not compare this category with ours. The DISO algorithm takes the target image and a temporal reference image as input and synthesizes a single reconstructed image. For these methods, we use the less missing temporal neighbor of the target image as the reference. The DISO methods we compare include simple replacement (replace missing values by copying data from reference image), LLHM [21], WLR [24], and STS-CNN [38]. The MIMO method utilizes a multitemporal image containing missing data as input and outputs a reconstructed multitemporal image, allowing the model to fully take into account temporal semantics. Compared methods of this type include last padding (last), nearest neighbor padding (nearest), and linear interpolation (linear), which are commonly used in cartography [52], as well as the UTILIZE [41].

2) *Evaluation Metrics*: Through the sliding windows approach, we reconstructed and assessed data gaps across all time steps to facilitate a more comprehensive analysis. The quantitative experiments utilized mean absolute error (MAE), spectral angle mapper (SAM) [53], mean peak signal-to-noise ratio (PSNR), and mean structural similarity (SSIM) [49] as evaluation metrics. For real data experiments, due to the lack of spatiotemporally synchronized ground truth, traditional full-reference image quality evaluation metrics (e.g., PSNR and SSIM) cannot be directly applied to the quantitative assessment, so we combine the visual effect and no-reference

TABLE I

QUANTITATIVE EVALUATION RESULTS UNDER **DUAL-TEMPORAL** INPUTS COMPARED TO THE **DISO** METHODS, USING THE LANDSAT AND EarthNet2021 DATASETS. THE RESULTS FOR THE BEST-PERFORMING METHOD FOR EACH METRIC ARE HIGHLIGHTED IN **BOLD**

Method	Type	Landsat				EarthNet2021			
		↓ MAE	↓ SAM	↑ PSNR	↑ SSIM	↓ MAE	↓ SAM	↑ PSNR	↑ SSIM
Replace	DISO	0.0210	2.81	32.67	0.8913	0.0146	3.10	34.04	0.9452
LLHM [21]		0.0161	1.95	34.64	0.9104	0.0121	2.59	35.54	0.9566
WLR [24]		0.0110	1.33	36.88	0.9312	0.0113	2.41	36.08	0.9605
STS-CNN [38]		0.0091	1.12	37.79	0.9423	0.0102	2.18	37.07	0.9647
MS²TAN (Ours)	DIDO	0.0081	1.02	38.54	0.9503	0.0095	2.04	37.60	0.9671

image evaluation metrics including blind/referenceless image spatial quality evaluator (BRISQUE) [54] and natural image quality evaluator (NIQE) [55] to perform the evaluation.

3) *Implement Details*: For MS²TAN, we utilized the Adam optimizer with parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for training, and set the batch size to 8. The initial learning rate was set to 4×10^{-4} , decayed by a factor of 0.5 every 100 epochs, and an early stopping strategy of 30 epochs on the validation set to prevent overfitting. For the other models we compared, we conducted our experiments while following the original authors' training settings and environment. The proposed MS²TAN was implemented with PyTorch 1.12 framework and trained with an NVIDIA GeForce RTX 3090 24-GB GPU on a Ubuntu 20.04 environment. In addition, the test code of MS²TAN is openly available on GitHub at <https://github.com/CUG-BEODL/MS2TAN>

B. Datasets

We conducted experiments on two datasets, Landsat and EarthNet2021 [56]. Their details are described in the Appendix.

1) *Landsat*: The Landsat dataset contains a large number of top-of-atmosphere (TOA) time-series captured by Landsat-5/7/8 satellites, with approximately uniform time intervals. The original images are cropped into small patches of size 120×120 pixels with a spatial resolution of 30 meters. Each image includes six bands (blue, green, red, NIR, SWIR1, and SWIR2), as well as the SLC-off or cloud mask provided in the QA band. In terms of spatial extent, we focus on sequences from multiple study regions in China and the United States.

2) *EarthNet2021*: The EarthNet2021 [56] dataset comprises over 28 000 Sentinel-2 TOA time-series observations from the European continent, with uniformly spaced time intervals. Each image consists of four bands (blue, green, red, and NIR), at a size of 128×128 pixels, downsampled to a spatial resolution of 20 m, accompanied by cloud probability maps. We utilized 80% of sequences from the training split for training purposes, with the remaining set aside for validation. For testing, we used sequences from the *iid* test split.

3) *Preprocessing*: We follow the preprocessing protocol of EarthNet2021 to crop the values of the optical images to the range [0, 10 000] and then normalize them to the unit range [0, 1]. To match the multiscale patch division strategy of MS²TAN, we resize the input images to a common multiple of the patch size of the MSTA at different scales. In our

experiments, we use three scales (patch size is 8, 10, and 12, respectively) which have a least common multiple of 120. We therefore centrally cropped the images in the EarthNet2021 dataset to 120×120 pixels, in line with the Landsat dataset.

4) *Data Gap Synthesis*: Since it is difficult for models trained in simulated cloud mask datasets to exhibit generalization capability in real clouds [57], we synthesize the dataset from cloud masks of real cloud coverage samples. Specifically, we use cloud detection algorithms to obtain binary cloud masks from real thick cloud images and apply them randomly to cloud-free images to create cloudy-cloudless image sequence pairs for training and validation, hence the dataset has a cloud distribution consistent with reality. For the SLC-off problem, we instead obtain missing masks from the QA band and synthesize them in a similar manner.

C. Results

To quantitatively compare restoration results, we evaluated our proposed approach as well as mainstream methods on synthetic data gaps. Specifically, we performed comparison experiments under two different conditions: dual-temporal inputs (versus the DISO algorithms), and multitemporal inputs (versus the MIMO algorithms). The quantitative evaluation results on the Landsat and EarthNet2021 datasets are shown in Tables I and II, and visual examples are illustrated in Figs. 6 and 7. Furthermore, Figs. 8 and 9 show the restoration results in real data gaps to validate the generalization performance of MS²TAN.

1) *Comparison Under Dual-Temporal Inputs*: We conducted comparison experiments with the DISO algorithms using synthetic data gaps in the Landsat and EarthNet2021 datasets. As illustrated in Fig. 6, (a) and (b) represent the two images with data gaps used as input, while (c)–(g) sequentially display the recovery results obtained from five methods, with (h) showing the ground truth of the target image. As depicted in (c)–(f), all compared methods to some extent produced discontinuous fine features. This is because the time-assisted images were unable to completely cover the missing regions, leading the LLHM, and WLR to rely on Laplacian prior regularization method (LRPM) [24] to fill in the remaining gaps, resulting in very blurry noise bands in the reconstruction results. Although the end-to-end strategy-based STS-CNN model can repair these blanks, it fails to accurately restore the original features. In contrast, the proposed MS²TAN, integrating a “pixel-structure-perception”

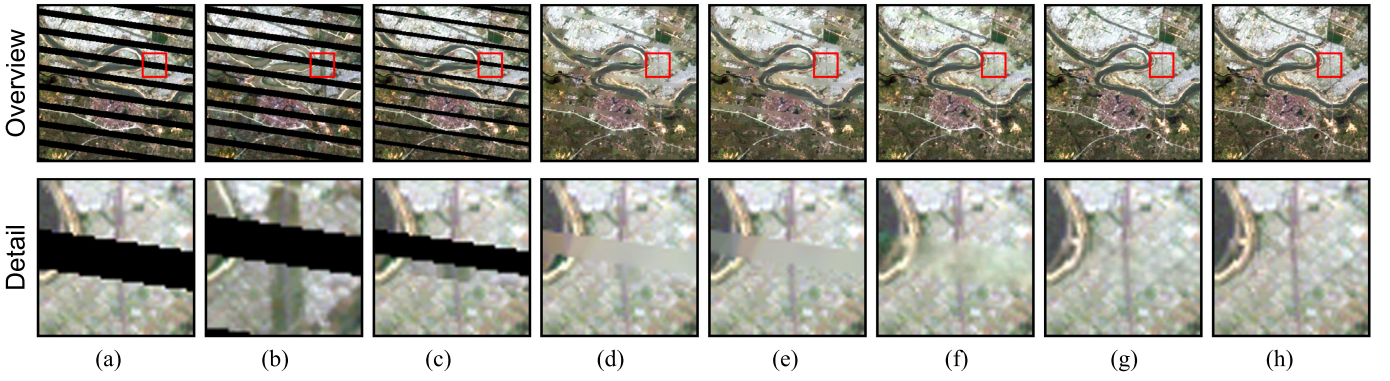


Fig. 6. Visual comparison results between MS²TAN and the DISO algorithms on the Landsat dataset, with dead pixels shown in black. Row 2 details the highlighted area in row 1. From left to right: (a) Input 1 (target) and (b) Input 2 (reference). (c)–(g) Restoration results of each method. (h) Ground truth.

TABLE II

QUANTITATIVE EVALUATION RESULTS UNDER MULTITEMPORAL INPUTS COMPARED TO THE MIMO METHODS, USING THE LANDSAT AND EarthNet2021 DATASETS. THE RESULTS FOR THE BEST-PERFORMING METHOD FOR EACH METRIC ARE HIGHLIGHTED IN **BOLD**

Method	Type	Landsat				EarthNet2021			
		↓ MAE	↓ SAM	↑ PSNR	↑ SSIM	↓ MAE	↓ SAM	↑ PSNR	↑ SSIM
Last	MIMO	0.0223	3.14	32.11	0.8862	0.0148	3.17	33.68	0.9439
Nearest		0.0198	2.54	33.28	0.8957	0.0128	2.74	35.04	0.9534
Linear		0.0103	1.26	37.20	0.9348	0.0110	2.35	36.48	0.9620
U-TILISE [41]		0.0082	1.05	38.45	0.9491	0.0086	1.87	38.29	0.9702
MS²TAN (Ours)		0.0074	0.96	39.01	0.9552	0.0078	1.71	38.91	0.9728

multiobjective joint optimization approach, achieves superior texture and structural detail recovery, as shown in (g). In the quantitative evaluation shown in Table I, MS²TAN outperforms other comparative methods across all evaluation metrics. Compared to the previous SOTA method STS-CNN, MS²TAN exhibits a 10.99%/6.86% (in two datasets, respectively) reduction in MAE and a 0.75/0.53 dB increase in PSNR, showcasing its robust capability in utilizing spatiotemporal data.

2) *Comparison Under Multitemporal Inputs:* Similar to the dual-temporal input experiment, we conducted comparative experiments with the MIMO algorithms under multitemporal input conditions. Visual comparative results are illustrated in Fig. 7, displaying, from top to bottom: the input image sequence, the restoration outcomes of five methods, and the ground truth. Since traditional linear methods rely on time-series information, they cannot restore regions in the input sequence where all time steps are missing. In contrast, UTILIZE can recover data gaps of any shape, but still exhibits significant color shifts in irregularly shaped gaps, reducing the quality of the results, as shown in the first three columns of Fig. 7. Benefiting from the MSTA mechanism, MS²TAN can effectively handle the edges of missing regions, achieving more harmonious seams. Notably, the multiobjective optimization makes MS²TAN perform better in restoring structural and textural details in the missing regions. In some extreme cases (e.g., the upper-left region of the first eight columns), due to a lack of sufficient information, MS²TAN may not accurately reproduce the actual changes but can still output coherent images with natural color transitions. As shown in the quantitative evaluation results in Table II,

compared to the UTILIZE, MS²TAN exhibits a 9.76%/9.30% reduction in MAE, a 0.56/0.62 dB increase in PSNR, and achieves higher visual consistency.

3) *Comparison Between the Two Strategies:* The dual-temporal experiments can be seen as obtained by restricting the multitemporal basis (described in detail in Section IV-A1), so that for long time sequences, both methods can be used. According to our experimental results (see Tables I and II), the results of multitemporal inputs tend to be better due to the fact that more information can be acquired at the same time, which was the initial motivation for our study of long time-series image restoration. Therefore, for long time-series images, the MIMO approach is the first choice. When the image cannot constitute a long time-series, dual-temporal, or single-temporal methods can be used as a supplement.

4) *Restoration of the Real Data Gaps:* To evaluate the generalization performance of the proposed MS²TAN, we applied pretrained model weights to real ETM+ SLC-off sequences and thick cloud-covered sequences. Figs. 8 and 9 present restoration results of two sequences with real data gaps and the corresponding no-reference assessment metrics. In the restored ETM+ SLC-off images, the roads and rivers, which are divided by striped missing areas, are rationally connected in the restoration results with the original structure maintained. Obvious ground surface changes brought by long time intervals are also taken into account. MS²TAN utilizes MSTA to learn the distribution of missing data, thereby capturing temporal evolution patterns from neighboring frames and fitting spatial textures, achieving high-precision data restoration. For the thick cloud-covered images, the method accurately restored

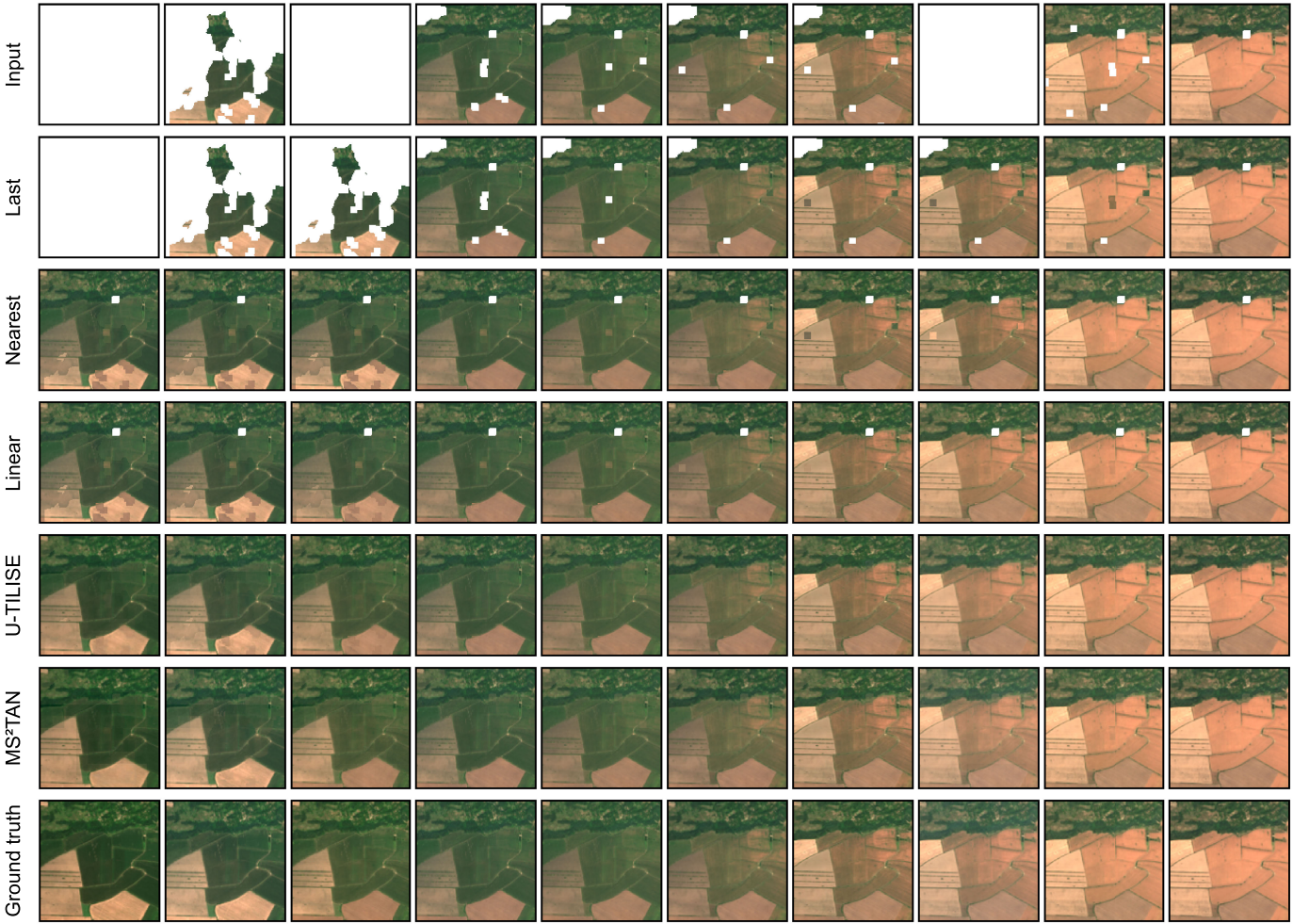


Fig. 7. Visual comparison results between MS²TAN and the MIMO algorithms on the EarthNet2021 dataset, with thick cloud shown in white. Each row shows a time-ordered image sequence. From top to bottom: input sequence (row 1), restoration results of each method (rows 2–6), and ground truth (row 7).

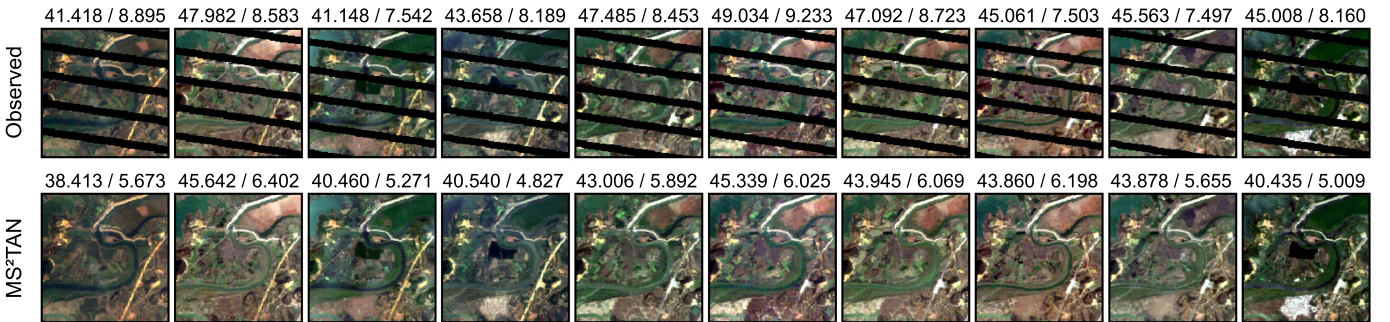


Fig. 8. Visualization of the results of the restoration of real ETM + SLC-off images from the Landsat dataset. The dead pixels are shown in black, and the values above the image are the corresponding no-reference image quality assessments (BRISQUE and NIQE, the lower the better). The first and second rows show the observed sequence and the restoration results of the MS²TAN, respectively.

all missing data, including both large-scale and fragmented gaps. Based on the assistance of long time-series, continuous large-area data gaps can also be repaired reasonably well and keep the original evolutionary trend. Irregularly shaped data gaps (e.g., the images in the sixth column in Fig. 9) are repaired to blend in with the spatial domain, preserving high textural and structural coherence. It is also evident from the BRISQUE and NIQE metrics that the quality of the restored images is significantly higher than before, maintaining the

same level as the nondegraded images. By applying MSTA and multiobjective joint optimization, MS²TAN exhibits excellent performance in handling complex and large-scale data gaps, which often present challenges in traditional data recovery methods.

D. Validation Studies

To validate the effectiveness of the current architecture, we experimented with some variants of MS²TAN and obtained

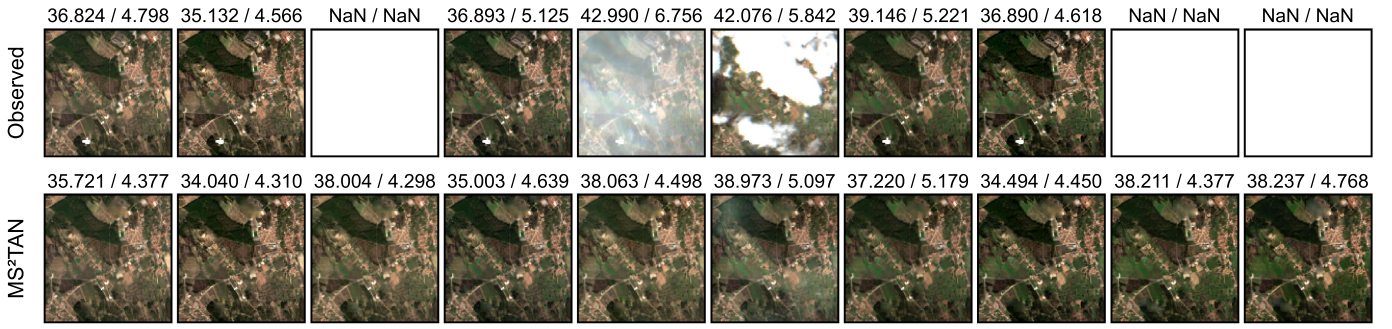


Fig. 9. Visualization of the results of the restoration of real thick cloud cover images from the EarthNet2021 dataset. The clouds are shown in white, and the values above the image are the corresponding no-reference image quality assessments (BRISQUE and NIQE, the lower the better). The first and second rows show the observed sequence and the restoration results of the MS²TAN, respectively.

TABLE III

QUANTITATIVE EVALUATION OF DIFFERENT MS²TAN VARIANTS THAT DERIVED FROM WHETHER SEPARATED SPATIAL–TEMPORAL ATTENTION (SEP. ATTN), ATTENTION MASK MECHANISM (ATTN MASK), AND MULTISCALE RESTORATION STRATEGY (MULTISCALE) ARE ENABLED. STATISTICS INFORMATION AND EVALUATION METRICS ARE PROVIDED ON THE RIGHT. ALL EXPERIMENTS ARE PERFORMED UNDER MULTITEMPORAL INPUTS

Features			Statistics		Landsat				EarthNet2021			
Sep. Attn	Attn Mask	Multi-Scale	Params (M)	FLOPs (G)	↓ MAE	↓ SAM	↑ PSNR	↑ SSIM	↓ MAE	↓ SAM	↑ PSNR	↑ SSIM
✓	✓	✓	4.33	8.756	0.0074	0.96	39.01	0.9552	0.0078	1.71	38.91	0.9728
	✓	✓	4.33	14.626	0.0073	0.95	39.08	0.9560	0.0083	1.81	38.53	0.9714
✓		✓	4.33	8.756	0.0096	1.18	37.47	0.9384	0.0112	2.38	36.37	0.9609
✓	✓		6.61	10.130	0.0086	1.07	38.09	0.9459	0.0095	2.05	37.60	0.9672
			6.61	16.473	0.0101	1.25	37.24	0.9356	0.0127	2.73	35.19	0.9538

quantitative results as shown in Table III. We discuss each of the mechanisms used below.

1) *Separated Spatial–Temporal Attention*: The original ViT, after embedding images into token sequences, feeds the entire sequence into a multihead self-attention module. However, the computational complexity of self-attention is proportional to the square of the sequence length. For a sequence of remote sensing images (t images, n patches per image), the computational complexity is as high as $O(t^2 \times n^2)$. Whereas separating temporal attention from spatial attention and computing them sequentially reduces the complexity to $O(t^2 + n^2)$. Although the perceptual field of the separated spatial–temporal attention is reduced in a single operation, we alternate multiple temporal and spatial attention modules to spread the perceptual field to the whole spatiotemporal range. Meanwhile, the extension of the perceptual field along the temporal axis and the spatial plane matches the characteristics of the time-series remote sensing images, which have small displacement in time and large spatial correlation, and can fully integrate the spatial and temporal information. The first two rows of Table III reveal that separated spatial–temporal attention can decrease computational costs by 40.1% with comparable performance. While the two forms of spatial–temporal attention exhibit minor performance differences, there is no statistically significant difference. Moreover, separated spatial–temporal attention achieves better performance–efficiency tradeoff, delivering superior results under equivalent computational loads.

2) *Attention Mask in MSTA*: To verify the effectiveness of the attention mask in MSTA, we visualized the distribution of attention. Fig. 10 shows the distribution of temporal attention

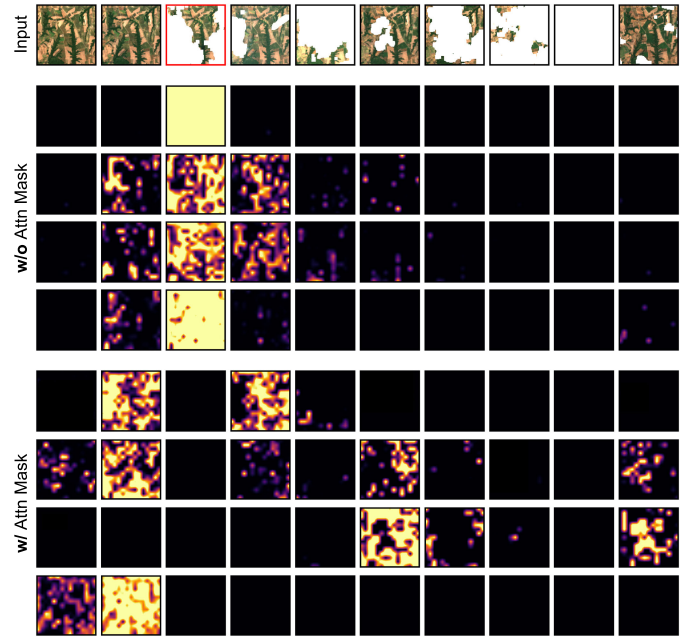


Fig. 10. Visualization of temporal attention distribution in the first MSTA unit. We present the attention scores for repairing the third frame (highlighted with a red box in row 1) of the sequence. It is observed that, after incorporating the mask, the temporal attention is no longer restricted to the vicinity of the target time but instead focuses on a broader range of valid pixels. From top to bottom: Input (row 1), w/o Mask (rows 2–5), and w/ Mask (rows 6–9).

with and without the attention mask, while Fig. 11 illustrates the spatial attention distribution. Without the mask, certain attention heads focus excessively on patches with missing data,

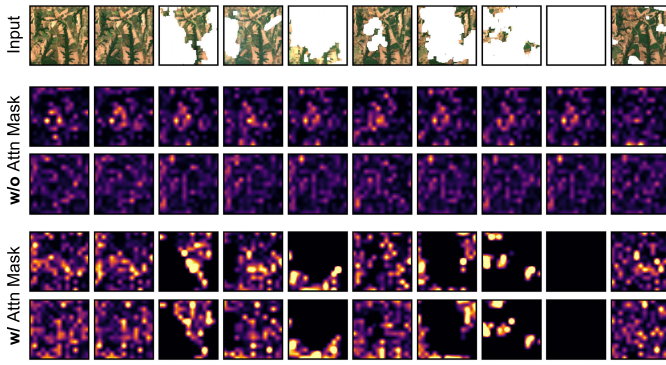


Fig. 11. Visualization of spatial attention distribution in the first MSTA unit. We show the average attention scores for imputing each patch. With the mask, spatial attention focuses on the edges of the valid pixels, extracting textural and structural features, rather than spreading around as it would be without the mask. From top to bottom: Input (row 1), **w/o** Mask (rows 2 and 3), and **w/** Mask (rows 4 and 5).

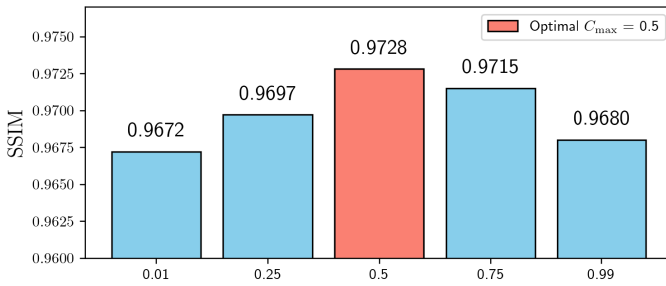


Fig. 12. Ablation experiment results for optimal C_{\max} value in MSTA. The horizontal axis represents different C_{\max} values, and the vertical axis represents the SSIM in the EarthNet2021 dataset.

while others focus excessively on their own patch, providing no additional information to impute missing values and leading to unnecessary computational costs. With the mask mechanism in place, the attention heads no longer focus on missing-data patches or themselves, but instead extract information from spatiotemporal domains, obtaining a larger effective receptive field to achieve higher Imputation accuracy.

Within the MSTA framework, the maximum allowable missing rate, C_{\max} , is a crucial hyperparameter. Fig. 3 shows the attention masks obtained from different C_{\max} as an example of a cloud image. It can be observed that lower C_{\max} values tend to discard an excessive number of valid patches leading to loss of details, whereas higher values result in the retention of superfluous thick cloud patches which distort token distribution. Furthermore, ablation experiments on the C_{\max} selection (see Fig. 12) reveal that moderate C_{\max} values achieve optimal reconstruction accuracy. Based on this empirical evidence, the present study adopts $C_{\max} = 0.5$ for subsequent experiments.

3) *Multiscale Restoration*: MS²TAN employs a multiscale restoration strategy to refine the restoration results layer by layer. To verify its effectiveness, we take the SLC-off reconstruction task as an example and compare the intermediate outputs of each layer in Fig. 13. With the gradual deepening of the restoration stages, the restoration results are gradually refined to better capture the details and structural features in the image. Specifically, the coarse repair stage addresses global textures and structural issues, while the fine repair stage further

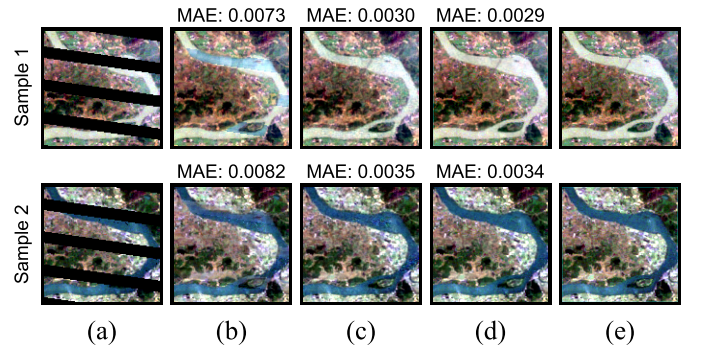


Fig. 13. Visualization of intermediate results for multiscale restoration. The values above indicate the MAE compared to the ground truth. (a) Input. (b)–(d) Intermediate outputs at each scale. (e) Ground truth.

TABLE IV

ABLATION ANALYSIS FOR **MULTIOBJECTIVE JOINT OPTIMIZATION**, WHERE λ_1 , λ_2 , AND λ_3 DENOTE THE WEIGHTS OF THE PIXEL-WISE LOSS, STRUCTURAL LOSS, AND PERCEPTUAL LOSS, RESPECTIVELY

Loss weights			EarthNet2021			
λ_1	λ_2	λ_3	↓ MAE	↓ SAM	↑ PSNR	↑ SSIM
1	-	-	0.0120	2.55	35.76	0.9572
0.5	0.5	-	0.0099	2.13	37.29	0.9657
0.9	0.1	-	0.0091	1.97	37.94	0.9686
0.5	-	0.5	0.0111	2.38	36.37	0.9609
0.9	-	0.1	0.0102	2.20	37.09	0.9645
-	0.5	0.5	0.0210	4.43	31.72	0.9241
0.9	0.05	0.05	0.0078	1.71	38.91	0.9728

optimizes local details and spectral consistency. However, despite the effectiveness of the multiscale approach, introducing too many scale levels leads to diminishing returns, with further improvements becoming negligible or even introducing overfitting and increased computational costs. Therefore, selecting an appropriate number of scale levels is crucial.

4) *Multiojective Joint Optimization*: In multiojective joint optimization, the weights assigned to each loss function are critical. To investigate this, we conducted an ablation study on the allocation of loss weights. Table IV presents different weight configurations for pixel-wise loss (λ_1), structural loss (λ_2), and perceptual loss (λ_3), along with their corresponding results on the EarthNet2021 dataset. The findings show that combining pixel-wise loss with either structural or perceptual loss leads to significant performance improvements (rows 2 and 4). The performance gains reach their maximum when pixel-wise loss dominates the weighting scheme (rows 3 and 5). On this basis, combining both simultaneously achieves the best results (the last row). However, when only structural and perceptual losses are applied, the model struggles to converge, resulting in poor performance (row 6). Therefore, pixel-wise loss plays a dominant role in multiojective optimization, and its combination with structural and perceptual losses is essential for achieving high restoration accuracy and visual consistency.

5) *Balance Between Effectiveness and Efficiency*: Scalability is a key strength of ViT-based models, and MS²TAN

TABLE V

HYPERPARAMETER SETTINGS AND EVALUATION OF MS²TAN WITH DIFFERENT SIZES. THE HYPERPARAMETERS INCLUDE THE RESTORATION SCALES NUMBER (S), PATCH SIZE (P), EMBEDDING DIMENSIONS (d_{EMB}), ATTENTION HEADS NUMBER (h), DIMENSIONS OF VECTORS Q , K , AND V IN SELF-ATTENTION (d_{QKV}), AND THE MSTA UNITS NUMBER (L), WHERE BRACKETED LISTS INDICATE THE HYPERPARAMETERS FOR EACH SCALE

Model	Hyperparameters						Statistics		EarthNet2021			
	S	P	d_{emb}	h	d_{qkv}	L	Params (M)	FLOPs (G)	↑ MAE	↑ SAM	↑ PSNR	↑ SSIM
MS ² TAN	3	(12, 10, 8)	(256, 192, 128)	(8, 6, 4)	(32, 32, 32)	(2, 2, 2)	4.33	8.756	0.0078	1.71	38.91	0.9728
MS ² TAN-L	3	(12, 10, 8)	(384, 256, 192)	(8, 6, 4)	(48, 48, 48)	(4, 4, 4)	17.04	34.363	0.0076	1.67	39.03	0.9733
MS ² TAN-S	2	(12, 10)	(192, 128)	(8, 6)	(24, 24)	(2, 2)	2.07	3.567	0.0089	1.92	38.07	0.9692

TABLE VI
ABLATION ANALYSIS FOR PE

Positional Encoding	EarthNet2021			
	↓ MAE	↓ SAM	↑ PSNR	↑ SSIM
✓	0.0078	1.71	38.91	0.9728
✗	0.0092	1.98	37.84	0.9683

is no exception. To strike a balance between effectiveness and efficiency, we adjusted several critical hyperparameters to obtain two MS²TAN variants and evaluated them on the EarthNet2021 dataset (see Table V). The experimental results reveal that when the model is too small (MS²TAN-S), performance significantly degrades; conversely, when the model is too large (MS²TAN-L), there is only a slight performance gain, while the computational cost increases drastically, making the approach less efficient. Overall, the current configuration of parameters and computational load represents an optimal “sweet spot” where good performance is maintained without excessive computational overhead.

6) *Positional Encoding*: In Transformer-based models, PE is shown to be effective and necessary [5]. We follow the base setting of the previous paper and add PE to MS²TAN. The spatial-temporal feature extraction of MS²TAN has been performed alternatively. After PE, in temporal attention, the input T tokens have relative position relations among them; in spatial attention, the input $(H/P) \times (W/P)$ tokens also have relative position relations. Thus, spatiotemporal position semantics can be introduced. Table VI shows the results of the ablation experiments of PE, and it can be found that the model performance is comprehensively leading after adding PE, which verifies its effectiveness.

V. CONCLUSION

This article presents a novel approach to reconstruct time-series remote sensing images using the multiscale MSTA network (MS²TAN). MS²TAN takes an image sequence with arbitrary data gaps as input and generates a clear and complete time-series. Compared with the existing methods, the proposed MSTA mechanism improves the efficiency of spatiotemporal information utilization, reduces the high computational cost of the original ViT, and obtains higher restoration accuracy. In addition, the application of the multiscale restoration strategy and multiobjective joint optimization further enhances the texture and structural consistency of the reconstructed images.

TABLE VII

DETAILS OF THE LANDSAT AND EarthNet2021 DATASETS

	Landsat	EarthNet2021
Data source	Landsat-5/7/8	Sentinel-2
Restoration task	SLC-off & Thick Cloud	Thick Cloud
Spatial scope	China & USA	Europe
Temporal scope	2001 - 2011	2016 - 2020
Spectral bands	6 bands	4 bands
Image size	120 × 120	128 × 128
Spatial Resolution	30-m	20-m
Temporal length	10	30
Temporal interval	average 30 days	constant 5 days
Average missing rate	35.8%	40.5%
Number of samples	11913	28123
– Train	7625	18955
– Valid	1905	4949
– Test	2383	4219

MS²TAN achieves higher imputation accuracy than the mainstream methods in quantitative experiments and exhibits better visual effects in eliminating real data gaps. Ablation experiments also confirm the contributions of primary innovations to the results.

Despite the effectiveness of the proposed method in addressing ETM+ SLC-off and thick cloud cover, there are some unavoidable limitations. For instance, the model needs to simulate various shapes of data gaps in the dataset to adequately learn to extract the features of the missing value distribution in order to achieve better restoration results, which leads to the model needing more time for training. Therefore, future research could incorporate network components pretrained on large-scale datasets to improve model convergence speed and generalization performance.

APPENDIX

Table VII details the statistics of the dataset we used, including key information such as data source, resolution, and number of sequences.

REFERENCES

- [1] H. Shen et al., “Missing information reconstruction of remote sensing data: A technical review,” *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 61–85, Sep. 2015.

- [2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [3] Y. LeCun et al., "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, Jan. 1989, pp. 396–404.
- [4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2020, pp. 1–21. [Online]. Available: <https://openreview.net/pdf?id=YicbFdNTTy>
- [5] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.
- [6] H. F. Shen, X. C. Meng, and L. P. Zhang, "An integrated framework for the spatio-temporal-spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Sep. 2016.
- [7] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, p. 4.
- [8] W. Du, D. Côté, and Y. Liu, "SAITS: Self-attention-based imputation for time series," *Expert Syst. Appl.*, vol. 219, Jun. 2023, Art. no. 119619.
- [9] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6721–6729.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [11] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
- [12] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 694–711.
- [13] C. Zhang, W. Li, and D. Travis, "Gaps-fill of SLC-off Landsat ETM+ satellite image using a geostatistical approach," *Int. J. Remote Sens.*, vol. 28, no. 22, pp. 5103–5122, Nov. 2007.
- [14] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, 2000, pp. 417–424.
- [15] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1621–1633, Dec. 1997.
- [16] Q. Yuan, L. Zhang, and H. Shen, "Hyperspectral image denoising with a spatial-spectral view fusion strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2314–2325, Jul. 2014.
- [17] Q. Cheng, H. Shen, L. Zhang, and P. Li, "Inpainting for remotely sensed images with a multichannel nonlocal total variation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 175–187, Jan. 2014.
- [18] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.
- [19] K. He and J. Sun, "Image completion approaches using the statistics of similar patches," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2423–2435, Dec. 2014.
- [20] Q. Cheng, H. Shen, L. Zhang, and Z. Peng, "Missing information reconstruction for single remote sensing images using structure-preserving global optimization," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1163–1167, Aug. 2017.
- [21] P. Scaramuzza and J. Barsi, "Landsat 7 scan line corrector-off gap-filled product development," *Proc. Pecora*, vol. 16, pp. 23–27, 2005.
- [22] J. Chen, X. Zhu, J. E. Vogelmann, F. Gao, and S. Jin, "A simple and effective method for filling gaps in Landsat ETM+ SLC-off images," *Remote Sens. Environ.*, vol. 115, no. 4, pp. 1053–1064, Apr. 2011.
- [23] G. Gao and Y. Gu, "Multitemporal Landsat missing data recovery based on tempo-spectral angle model," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3656–3668, Jul. 2017.
- [24] C. Zeng, H. Shen, and L. Zhang, "Recovering missing pixels for Landsat ETM+ SLC-off imagery using multi-temporal regression analysis and a regularization method," *Remote Sens. Environ.*, vol. 131, pp. 182–194, Apr. 2013.
- [25] X. Li, H. Shen, L. Zhang, H. Zhang, Q. Yuan, and G. Yang, "Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7086–7098, Nov. 2014.
- [26] J. Zhang, M. K. Clayton, and P. A. Townsend, "Missing data and regression models for spatial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1574–1582, Mar. 2015.
- [27] X. Zhu, F. Gao, D. Liu, and J. Chen, "A modified neighborhood similar pixel interpolator approach for removing thick clouds in Landsat images," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 521–525, May 2012.
- [28] Q. Cheng, H. Shen, L. Zhang, Q. Yuan, and C. Zeng, "Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model," *ISPRS J. Photogramm. Remote Sens.*, vol. 92, pp. 54–68, Jun. 2014.
- [29] B. Chen, B. Huang, L. Chen, and B. Xu, "Spatially and temporally weighted regression: A novel method to produce continuous cloud-free Landsat imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 27–37, Jan. 2017.
- [30] X. Li, H. Shen, H. Li, and L. Zhang, "Patch matching-based multitemporal group sparse representation for the missing information reconstruction of remote-sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3629–3641, Aug. 2016.
- [31] J. Wu, T. Li, L. Lin, and C. Zeng, "Progressive gap-filling in optical remote sensing imagery through a cascade of temporal and spatial reconstruction models," *Remote Sens. Environ.*, vol. 311, Sep. 2024, Art. no. 114245.
- [32] M. K. Ng, Q. Yuan, L. Yan, and J. Sun, "An adaptive weighted tensor completion method for the recovery of remote sensing images with missing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3367–3381, Jun. 2017.
- [33] H. He, J. Yan, L. Wang, D. Liang, J. Peng, and C. Li, "Bayesian temporal tensor factorization-based interpolation for time-series remote sensing data with large-area missing observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 506.
- [34] S. Malek, F. Melgani, Y. Bazi, and N. Alajlan, "Reconstructing cloud-contaminated multispectral images with contextualized autoencoder neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2270–2282, Apr. 2018.
- [35] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
- [36] L. Sun, Y. Zhang, X. Chang, Y. Wang, and J. Xu, "Cloud-aware generative network: Removing cloud from optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 691–695, Apr. 2020.
- [37] M. Shao, C. Wang, W. Zuo, and D. Meng, "Efficient pyramidal GAN for versatile missing data reconstruction in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626014.
- [38] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4274–4288, Aug. 2018.
- [39] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang, "Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 148–160, Apr. 2020.
- [40] Y. Chen, L. Tang, X. Yang, R. Fan, M. Bilal, and Q. Li, "Thick clouds removal from multitemporal ZY-3 satellite images using deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 143–153, 2020.
- [41] C. Stucker, V. S. F. Garnot, and K. Schindler, "U-TILISE: A sequence-to-sequence model for cloud removal in optical satellite time series," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 1007.
- [42] M. Xu, F. Deng, S. Jia, X. Jia, and A. J. Plaza, "Attention mechanism-based generative adversarial networks for cloud removal in Landsat images," *Remote Sens. Environ.*, vol. 271, Mar. 2022, Art. no. 112902.
- [43] D. Christopoulos, V. Ntouskos, and K. Karantzalos, "CloudTran: Cloud removal from multitemporal satellite images using axial transformer networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 1125–1132, May 2022.
- [44] H. Liu, B. Huang, and J. Cai, "Thick cloud removal under land cover changes using multisource satellite imagery and a spatiotemporal attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 1440.
- [45] C. Zhang, S. Wang, L. Zhong, Q. Chen, and C. Rao, "Cascaded temporal and spatial attention network for solar adaptive optics image restoration," *Astron. Astrophys.*, vol. 674, p. A126, Jun. 2023.

- [46] Z. Zhu and C. E. Woodcock, "Object-based cloud and cloud shadow detection in Landsat imagery," *Remote Sens. Environ.*, vol. 118, pp. 83–94, Mar. 2012.
- [47] L. Baetens, C. Desjardins, and O. Hagolle, "Validation of Copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure," *Remote Sens.*, vol. 11, no. 4, p. 433, Feb. 2019.
- [48] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.
- [49] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.
- [50] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for neural networks for image processing," 2015, *arXiv:1511.08861*.
- [51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Jan. 2015, pp. 1–14. [Online]. Available: <https://arxiv.org/pdf/1409.1556v6>
- [52] J. Inglada et al., "Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery," *Remote Sens.*, vol. 7, no. 9, pp. 12356–12379, Sep. 2015.
- [53] F. A. Kruse et al., "The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, nos. 2–3, pp. 145–163, 1993.
- [54] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *Proc. 45th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2011, pp. 723–727.
- [55] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Apr. 2012.
- [56] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler, "EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1132–1142.
- [57] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.



Zaiyan Zhang (Student Member, IEEE) is currently pursuing the B.E. degree in data science and big data technology with China University of Geosciences, Wuhan, China.

His research interests include remote sensing image processing, multimodal data fusion, low-level vision, land-cover change detection, and computer vision.



Jining Yan (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is currently an Associate Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. His research interests include remote sensing data management and time-series analysis.



Yuanqi Liang is currently pursuing the B.E. degree in data science and big data technology with China University of Geosciences, Wuhan, China.

Her research interests include time-series analysis, change detection, and spatiotemporal data processing.



Jiaxin Feng received the bachelor's degree from Shanghai University of Electric Power, Shanghai, China, in 2022. He is currently pursuing the master's degree with the School of Computer Science, China University of Geosciences, Wuhan, China.

His research interests include remote sensing data processing, time-series analysis, and soil erosion.



Haixu He received the bachelor's and master's degrees from China University of Geosciences, Wuhan, China, in 2020 and 2023, respectively, where he is currently pursuing the Ph.D. degree with the School of Computer Science.

His research interests include remote sensing data processing, time-series analysis, and change detection.



Li Cao received the M.S. degree in resource and environment from China University of Geosciences, Wuhan, China, in 2011.

He is currently a Senior Engineer with the Second Surveying and Mapping Institute of Hunan Province, Changsha, China. His research interests include nature resource monitoring with remote sensing.