

Temporal-Spatial Tubelet Embedding for Cloud-Robust MSI Reconstruction using MSI-SAR Fusion: A Multi-Head Self-Attention Video Vision Transformer Approach

Yiqun WANG¹, Lujun LI¹, Meiru YUE¹, Radu STATE¹

¹ University of Luxembourg, 1359 Kirchberg, Luxembourg
- (yiqun.wang, lujun.li, meiru.yue, radu.state)@uni.lu

Keywords: Cloud Removal, Video Vision Transformer, Temporal-Spatial Embedding, Time-series Image Reconstruction, MSI Images, SAR Images.

Abstract

Cloud cover in multispectral imagery (MSI) significantly hinders early-season crop mapping by corrupting spectral information. Existing Vision Transformer (ViT)-based time-series reconstruction methods, like SMTS-ViT, often employ coarse temporal embeddings that aggregate entire sequences, causing substantial information loss and reducing reconstruction accuracy. This study addresses the critical research question of how to maintain local temporal-spatial coherence during the reconstruction of cloud-contaminated MSI. We hypothesize that employing a localized tubelet embedding mechanism can mitigate the degradation caused by long-term sequence aggregation and better preserve inherent spectral-temporal patterns. To validate this hypothesis, a Video Vision Transformer (ViViT)-based framework with temporal-spatial fusion embedding for MSI reconstruction in cloud-covered regions is proposed in this study. Non-overlapping tubelets are extracted via 3D convolution with constrained temporal span ($t = 2$), ensuring local temporal coherence while reducing cross-day information degradation. Both MSI-only and SAR-MSI fusion scenarios are considered during the experiments. Comprehensive experiments on 2020 Trill County data demonstrate notable performance improvements: MTS-ViViT achieves a 2.23% reduction in MSE compared to the MTS-ViT baseline, while SMTS-ViViT achieves a 10.33% improvement with SAR integration over the SMTS-ViT baseline. The proposed framework effectively enhances spectral reconstruction quality for robust agricultural monitoring.

1. Introduction

Accurate early-season crop mapping is essential for agricultural monitoring, food security assessment, and sustainable resource management (Qader et al., 2021). Timely identification of crop types and phenological stages enables precision farming, optimized resource allocation, and early detection of crop stress conditions (Wang et al., 2023, Wang et al., 2024b, Wang et al., 2024a). Multispectral imagery (MSI) from satellites such as Sentinel-2 (Main-Knorn et al., 2017) provides rich spectral information spanning visible to shortwave infrared wavelengths, enabling precise crop classification and continuous phenological monitoring. However, MSI observations are inherently vulnerable to cloud cover, which frequently obscures critical observations during the early growing season. Cloud-induced data gaps severely limit the ability to capture temporal crop dynamics and make informed agricultural decisions, a problem particularly acute during early growth stages when phenological changes are most indicative of crop variety and health status.

Synthetic aperture radar (SAR) data from sensors such as Sentinel-1 (Torres et al., 2012) provides a complementary solution, as microwave signals penetrate cloud cover and deliver consistent all-weather, day-and-night observations. However, SAR fundamentally lacks the spectral richness and spatial detail of MSI, rendering it insufficient for precise discrimination between crop types at the spatial and spectral granularity required for operational agricultural applications. Existing approaches to address cloud cover rely on simplistic techniques such as linear temporal interpolation (Xia et al., 2012, Wang et al., 2024b, Wang et al., 2024a), closest spectral fit (Eckardt et al., 2013), or single-timeframe MSI-SAR fusion (Scarpa et al., 2018, Li et al., 2019, Jing et al., 2023, Wang et al., 2024c, Tu et al., 2025), which fail to capture the complex temporal-spectral

patterns necessary for accurate reconstruction under severe and prolonged cloud cover. A recent advanced multi-head self-attention Vision Transformer (ViT)-based model called SMTS-ViT (Li et al., 2025) for time-series MSI reconstruction leverages integrated MSI-SAR fusion to achieve superior results compared to single-time frameworks. However, it employs temporal embeddings that aggregate information uniformly across the entire sequence, which can lead to significant information loss due to temporal aliasing and ultimately degrade reconstruction quality.

In this paper, a Video Vision Transformer (Arnab et al., 2021) (ViViT)-based framework is proposed, which introduces temporal-spatial fusion embedding via 3D convolutional tubelet projection with constrained temporal span, rather than aggregating information uniformly across the entire time series. This thereby preserves local temporal coherence and reduces information loss from excessive temporal aggregation. Theoretically, the original convolutional patch projection from (Li et al., 2025) can be understood as a degenerate case of our time-spatial fusion embedding where temporal span equals the full sequence length. By limiting the temporal span, our approach reduces cross-day information degradation and more effectively captures transient spectral patterns essential for accurate MSI reconstruction in cloud-covered regions.

In conclusion, three main contributions are presented in this study: (1) A novel temporal-spatial fusion embedding method is introduced, enabling robust multispectral cloud removal and reconstruction through explicit integration of MSI and SAR data. (2) The masking strategy for SAR inputs is improved. Unlike the previous method (Li et al., 2025) that masks SAR whenever clouds are present in MSI, SAR data remain unmasked and available, so its all-weather sensing properties are

fully leveraged even in cloud-affected scenes. (3) The effectiveness of SAR-MSI fusion and the proposed temporal-spatial embedding framework, compared with MSI-only and standard ViT-based approaches, has been comprehensively validated by a carefully designed experimental setup and quantitative evaluation, clearly demonstrating superior performance under various cloud conditions.

2. Related Works

Cloud cover remains a pervasive obstacle in optical remote sensing for agricultural monitoring, often obscuring important ground information needed for accurate analysis. Traditional gap-filling strategies, such as temporal interpolation (Xia et al., 2012, Wang et al., 2024b, Wang et al., 2024a) or closest spectral fit (Eckardt et al., 2013), are limited in their ability to represent the underlying spectral and spatial variability crucial for monitoring crop development. Although advanced statistical inpainting methods with spatial regularization have been proposed, their effectiveness diminishes when large, contiguous cloud-covered regions occur during key phenological stages, leaving substantial gaps in usable observations.

Recent developments have shifted toward leveraging data fusion and deep learning to address these limitations. Notably, integrating SAR with optical MSI enables all-weather observation and supports robust cloud removal through multi-source data fusion, as demonstrated by deep residual neural networks and spatial-spectral network architectures. Approaches such as Convolutional Neural Network(CNN)-based networks (Scarpa et al., 2018, Li et al., 2019), denoising diffusion models (Jing et al., 2023), and multi-modal similarity attention mechanisms (Tu et al., 2025) have delivered significant improvements in Sentinel-2 imagery reconstruction by explicitly modeling the dependence between SAR and MSI inputs. While these multi-source fusion methods show clear advantages over single-modal and traditional approaches, these methods typically aggregate or process multi-modal data without explicitly leveraging time-series information, treating each image independently or focusing solely on spatial relationships.

Vision Transformers (ViTs) (Vaswani et al., 2017), with their powerful multi-head self-attention mechanism, have shown remarkable advantages in remote sensing image reconstruction. Unlike traditional CNNs that are limited by fixed local receptive fields, the multi-head self-attention in ViTs can efficiently capture both global and local dependencies across spatial and temporal dimensions, leading to richer feature representations and improved reconstruction quality. This global modeling ability allows ViTs to better leverage important temporal and spatial context, making them particularly effective for handling cloud-contaminated multispectral series.

Recent advanced work by (Zhao et al., 2023) proposes the Space-Eye network, which utilizes ViTs with single-head attention for cloud removal by processing time-series SAR and MSI data. This approach represents an early fusion of multi-modal and multi-temporal data, but the attention mechanism is limited in modeling complex dependencies due to the use of only a single attention head.

Nevertheless, (Li et al., 2025) introduces the SMTS-ViT architecture, which employs ViTs with multi-head self-attention to more effectively exploit temporal and spatial coherence in

MSI and SAR data for cloud-filling applications. By leveraging multi-head self-attention, SMTS-ViT can model richer and more diverse relationships across multiple dates and modalities, leading to improved reconstruction quality in cloud-covered regions. This marks a shift towards explicit time-series modeling and highlights the value of temporal aggregation and advanced attention mechanisms for robust agricultural monitoring under clouds.

However, SMTS-ViT aggregates temporal information across the entire sequence, which can dilute local spectral variations critical for precise cloud gap-filling. To address this, we propose SMTS-ViViT, an enhanced ViViT-based framework that leverages temporal-spatial fusion embedding with a controlled temporal span using 3D convolutional tubelet projection. Unlike previous approaches that over-aggregate temporally or rely on computationally intensive separate spatial and temporal embeddings, our method extracts nonoverlapping spatio-temporal tubelets with constrained temporal span ($t = 2$), thus preserving local temporal coherence and reducing information loss. Theoretically, standard convolutional patch projection from SMTS-ViT is a special case of our framework where the temporal span covers the full sequence, but limiting the span empirically improves reconstruction for both MSI-only and SAR-MSI fusion. Our comprehensive evaluation also demonstrates performance gains, establishing time-spatial fusion embedding as an effective solution for cloud-robust remote sensing image reconstruction.

3. Study Areas and Data Preparation

3.1 Study Areas

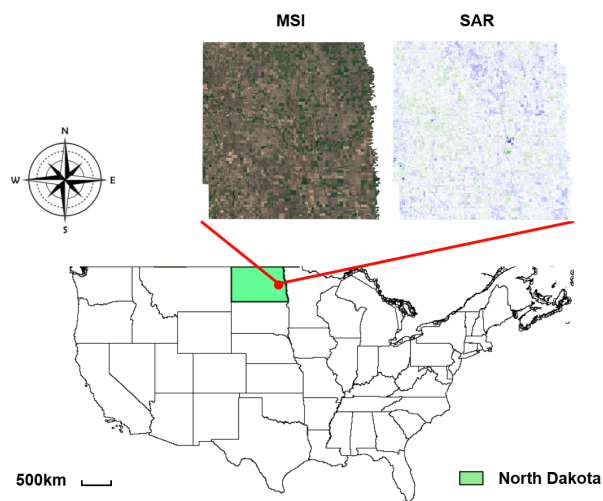


Figure 1. The Study Area: Traill County, located in North Dakota, the USA, 2020.

The study area selected for this research is Traill County, located in North Dakota, USA, as shown in Figure 1, which is an agriculturally significant region characterized by intensive crop production systems, making it an ideal testbed for evaluating time-series MSI image reconstruction and crop mapping applications. Traill County experiences pronounced seasonal variations in crop development, with early May marking the onset of critical phenological stages during corn and soybean emergence. The region's agricultural intensity and well-documented crop management practices provide a representative context

for testing reconstruction methodologies under realistic cloud cover conditions.

3.2 Data Preparation

For this research, two complementary optical and microwave remote sensing platforms are utilized: Sentinel-1 SAR and Sentinel-2 MSI. The Sentinel-1 mission provides a consistent all-weather monitoring capability with two operational satellites maintaining a 6-day nominal revisit cycle over mid-latitude regions. The SAR data includes two polarization bands: VV (vertical transmit, vertical receive) and VH (vertical transmit, horizontal receive), which capture distinct geometric and moisture-related surface scattering properties valuable for vegetation characterization and structural analysis.

The Sentinel-2 constellation comprises two identical satellites with a combined 5-day nominal revisit frequency, delivering high-resolution multispectral observations across 13 spectral bands spanning the visible, near-infrared, and shortwave infrared regions. For this study, 11 spectral bands are utilized. They are selected for their relevance to vegetation monitoring and crop phenology analysis: B1 (Coastal Aerosol, 60 m), B2 (Blue, 10 m), B3 (Green, 10 m), B4 (Red, 10 m), B5 (Red Edge 1, 20 m), B6 (Red Edge 2, 20 m), B7 (Red Edge 3, 20 m), B8 (Near Infrared, NIR, 10 m), B8A (Red Edge 4, 20 m), B11 (Shortwave Infrared 1, SWIR1, 20 m), and B12 (Shortwave Infrared 2, SWIR2, 20 m). Additionally, cloud detection is carried out using Sentinel Hub's tools (Skakun et al., 2022), which generate cloud masks by filtering out cloud-covered regions in the MSI data.

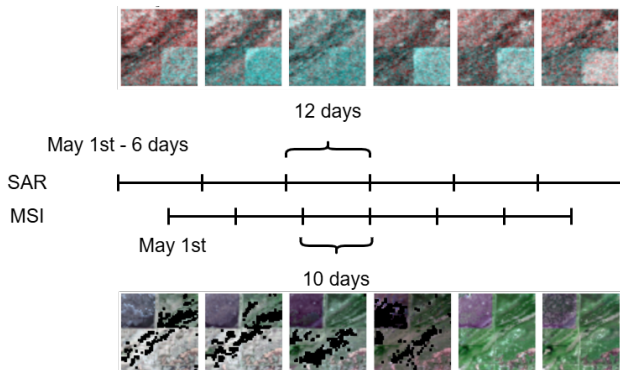


Figure 2. Multi-Temporal SAR and MSI Acquisition Scheme and Data Example. The black pixels in MSI present the cloud mask.

3.2.1 Temporal Coverage and Data Collection Period:

This paragraph details the overall temporal span and the scientific motivation for the data collection period used in the study. Our objective was to capture the critical phases of early-season crop development when phenological variability is highest and thus most informative for crop differentiation and monitoring. As shown in Figure 2, Sentinel-2 MSI composites were generated at 10-day intervals over a continuous 60-day window starting May 1st, covering phenomena like seed germination and early vegetative growth. For Sentinel-1 SAR, six observations were obtained by commencing acquisition six days earlier (April 25th) and extending the period to 72 days, with composited images produced every 12 days. This temporal design, shown in Figure 2, ensures that both MSI and SAR sequences are consistently paired and aligned with key agronomic events and realistic cloud cover variability.

3.2.2 Spatial Resolution Harmonization and Preprocessing:

To ensure consistent spatial resolution and geographic alignment across all data sources, we implemented comprehensive preprocessing and reprojection procedures. All Sentinel-1 SAR and Sentinel-2 MSI images were reprojected to a common Universal Transverse Mercator (UTM) projection and resampled to a uniform 30-meter ground pixel size using bilinear interpolation for spectral bands and nearest-neighbor resampling for categorical data.

3.2.3 Data Aggregation Strategy and Temporal Gap Management:

There is a practical challenge posed by missing or incomplete image coverage despite Sentinel-1 and Sentinel-2s nominal revisit cycles of 6 and 5 days, respectively. In real-world scenarios, acquisition gaps caused by satellite transmission, maintenance, processing delays, or scheduling priorities can disrupt temporal continuity. To guarantee complete and reliable time series for deep learning, we implemented a data aggregation approach: for each sensor, two successive revisit periods were composited, yielding aggregate SAR and MSI images at 12-day and 10-day intervals, respectively. This protocol ensures uninterrupted sequences, six MSI and six SAR frames as shown in Figure 2, across the study period. Pixelwise max compositing was employed to preserve spatial detail and minimize the effects of atmospheric interference or sensor artifacts. As a result, the final dataset maintains both sufficient temporal resolution to capture crop growth dynamics and full spatiotemporal coverage, ensuring its suitability for robust time-series modeling and fusion.

4. Methodology

4.1 Problem Definition

The time-series multispectral imagery affected by cloud occlusion can be represented as $X_{\text{MSI}} \in \mathbb{R}^{T \times C_{\text{MSI}} \times H \times W}$, where T denotes the temporal sequence length, C_{MSI} represents the total number of spectral channels in MSI data, and H, W denote the spatial height and width dimensions. The pixel value at position (x, y) in channel c at time t is denoted as $X_{\text{MSI}}(t, c, x, y)$. A binary cloud mask tensor $M_{\text{cloud}} \in \mathbb{R}^{T \times H \times W}$ identifies cloud-affected pixels in the MSI observations:

$$M_{\text{cloud}}(t, x, y) = \begin{cases} 0 & \text{if } X_{\text{MSI}}(t, :, x, y) \text{ is cloud-free} \\ 1 & \text{if } X_{\text{MSI}}(t, :, x, y) \text{ is cloud-occluded} \end{cases} \quad (1)$$

where T = temporal sequence length
 C_{MSI} = total number of MSI spectral channels
 H, W = spatial dimensions
 $X_{\text{MSI}}(t, c, x, y)$ = original MSI spectral value

Assuming uniform cloud coverage across all spectral channels at each spatio-temporal location, the cloud mask M_{cloud} is obtained from the Sentinel Hub's tools. Critically, the cloud mask is applied only to MSI data, as synthetic aperture radar (SAR) observations are unaffected by cloud cover and require no masking. Accordingly, SAR time-series data are represented as $X_{\text{SAR}} \in \mathbb{R}^{T \times C_{\text{SAR}} \times H \times W}$, where $C_{\text{SAR}} = 2$ (VV and VH polarizations for Sentinel-1), and all SAR observations are preserved without masking:

$$X_{\text{SAR}}(t, c, x, y) \text{ for all } t, c, x, y \text{ (no masking applied)} \quad (2)$$

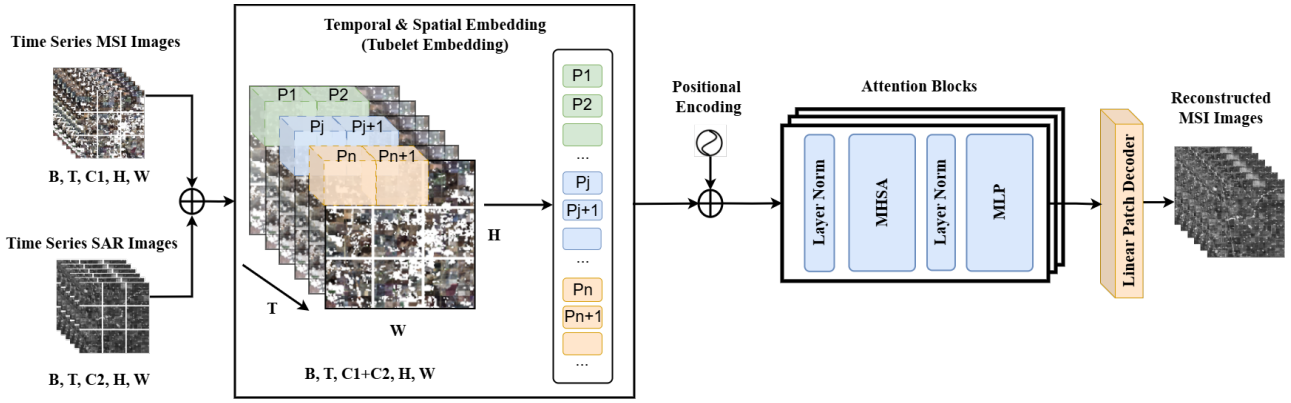


Figure 3. The Video Vision Transformer Structure with the Temporal-Spatial Tubelet Embedding for Cloud-Robust MSI Reconstruction using MSI-SAR Fusion.

For the MSI-only reconstruction task, the clouded MSI data $X_{\text{MSI}}^{\text{cloud}}$ is obtained by applying the cloud mask via element-wise multiplication:

$$X_{\text{MSI}}^{\text{cloud}}(t, c, x, y) = X_{\text{MSI}}(t, c, x, y) \cdot (1 - M_{\text{cloud}}(t, x, y)) \quad (3)$$

The ground-truth target sequence $Y \in \mathbb{R}^{T \times C_{\text{MSI}} \times H \times W}$ represents the cloud-free reference MSI imagery. For the MSI-only reconstruction scenario, the model prediction is defined as:

$$\hat{Y}_{\text{MSI}} = \text{Model}_{\text{MSI}}(X_{\text{MSI}}^{\text{cloud}}, M_{\text{cloud}}) \quad (4)$$

For the SAR-MSI fusion scenario, the combined input integrates both data modalities, where MSI is cloud-corrupted but SAR remains complete and unmasked. The fused model prediction becomes:

$$\hat{Y}_{\text{MSI}} = \text{Model}_{\text{Fusion}}(X_{\text{MSI}}^{\text{cloud}}, X_{\text{SAR}}, M_{\text{cloud}}) \quad (5)$$

where $X_{\text{MSI}}^{\text{cloud}}$ = cloud-corrupted MSI observations
 X_{SAR} = complete SAR observations
 M_{cloud} = MSI cloud mask
 \hat{Y}_{MSI} = reconstructed cloud-free MSI

The optimization objective for both scenarios is to minimize the reconstruction loss function $\mathcal{L}(\hat{Y}_{\text{MSI}}, Y)$, ensuring that predicted values closely approximate the true spectral distribution of cloud-free MSI observations while minimizing information loss across temporal and spectral dimensions. SAR serves as a complementary data source providing all-weather context without being subject to cloud masking, enabling improved reconstruction through multi-modal information fusion.

4.2 General Framework Architecture:

The proposed framework, as shown in Figure 3 consists of three principal components:

- (1) **Temporal-Spatial Fusion Embedding** via 3D convolutional tubelet projection,
- (2) a **Multi-Head Self-Attention (MHSA) Encoder**, and

- (3) a **Linear Patch Decoder**.

The framework exploits local temporal context preservation within MSI time-series to reconstruct spectral information in cloud-covered pixels, while integrating unmasked SAR observations that penetrate cloud cover as complementary all-weather features for multi-modal reconstruction.

4.2.1 Temporal-Spatial Fusion Embedding via 3D Convolutional Tubelet Projection: The key innovation is the introduction of a constrained temporal span in the tubelet extraction process. SMTS-ViT typically concatenates the entire time-series channels with spatial channels (i.e., temporal span $t = T$), creating a channel dimension $C' = T \times C$. This full-sequence aggregation approach leads to excessive cross-day information mixing and substantial information loss.

In contrast, our proposed temporal-spatial fusion embedding extracts non-overlapping spatio-temporal tubelets using 3D convolution with a short, fixed temporal span ($t = 2$). This design principle reduces information degradation compared to full-sequence aggregation: as $t \rightarrow T$, transient spectral variations critical for accurate cloud reconstruction become obscured within the aggregated channels.

The temporal-spatial fusion embedding operates on the input tensor directly via 3D convolutional filtering. Unlike 2D convolution applied independently to each timeframe, 3D convolution simultaneously processes spatial and temporal dimensions, enabling implicit modeling of local spatio-temporal patterns within the input volume.

For the MSI-only scenario, the input tensor $X_{\text{MSI}} \in \mathbb{R}^{T \times C_{\text{MSI}} \times H \times W}$ is transformed into a sequence of non-overlapping spatio-temporal tubelets. For the SAR-MSI fusion scenario, the concatenated input tensor $X \in \mathbb{R}^{T \times (C_{\text{MSI}} + C_{\text{SAR}}) \times H \times W}$ combines all spectral and polarization channels. The 3D convolution operates with kernel size (k_t, k_s, k_s) , where the temporal kernel dimension is $k_t = t$ (the constrained temporal span, fixed at $t = 2$ in our implementation), and the spatial kernel dimension is $k_s = 5$ (producing 5×5 pixel patches).

The 3D convolutional operation for extracting the p -th tubelet is defined as:

$$X'_{(p,t')} = \sum_{t''=0}^{k_t-1} \sum_{i=0}^{k_s-1} \sum_{j=0}^{k_s-1} \sum_{c=1}^C W_{(p,c,i,j,t'')} X_{(t'+t'',c,h+i,w+j)} + b_p \quad (6)$$

where $X'_{(p,t')}$ = embedding feature for the p -th tubelet at temporal position t'
 p = tubelet index
 t' = temporal anchor position of tubelet
 t'' = temporal offset within the convolution kernel
 i, j = spatial offsets within the convolution kernel
 $X_{(t'+t'',c,h+i,w+j)}$ = input pixel value at temporal position $(t'+t'')$
 $W_{(p,c,i,j,t'')}$ = 3D convolution kernel weight
 b_p = bias term for patch

This produces a sequence of patch tokens, each jointly encoding spatial structure and local temporal evolution. The linear embedding layer then projects each tubelet into the embedding space, following by the positional encoding, with dimension d_e :

$$T_{(p)} = \text{Linear}(X'_{(p,t')}) \in \mathbb{R}^{d_e} \quad (7)$$

where $T_{(p)}$ = p -th tubelet token embedding
 d_e = embedding dimension
 $\text{Linear}(\cdot)$ = learned linear transformation layer

4.2.2 Multi-Head Self-Attention Encoder with Joint Temporal-Spatial Attention: The MHSA mechanism operates directly on the sequence of spatial-temporal patch (tubelet) embeddings produced by 3D convolution, allowing the model to capture complex dependencies across both temporal and spatial dimensions in the data. Before entering the MHSA blocks, each tubelet embedding is summed with its corresponding positional encoding, ensuring the network is aware of the patch's location within both space and time. This joint encoding strategy enables the model to learn interactions between different moments and locations within the input sequence in a unified way.

In each self-attention block, for each attention head, query (Q), key (K), and value (V) matrices are computed from the input tubelet embeddings:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

where Q = query matrix, dimension: $N_{\text{patch}} \times d_k$
 K = key matrix, dimension: $N_{\text{patch}} \times d_k$
 V = value matrix, dimension: $N_{\text{patch}} \times d_v$
 d_k = dimension of key vectors

Multi-head attention employs h parallel attention heads to process the input, which are then concatenated and projected back to the embedding dimension:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (9)$$

where h = number of attention heads
 W^O = output projection matrix,
 dimension: $h \cdot d_v \times d_e$

In our framework, joint temporal-spatial attention is implemented by combining positional encoding with all tubelet (patch) embeddings and processing them through Multi-Head Self-Attention blocks. This structure enables the model to automatically discover and emphasize relevant relationships across both time and space, without requiring explicit masking. As a result, the attention mechanism flexibly aggregates long-range temporal and spatial features that are informative for reconstructing MSI data under cloud cover and synergistically integrating complementary SAR signals.

4.2.3 Linear Patch Decoder: The decoder consists of a linear projection layer that reshapes high-dimensional encoder features back into the original input time-series MSI structure. The encoder output feature maps are projected linearly to reconstruct patches corresponding to the input spatial-temporal resolution:

$$\hat{Y} \in \mathbb{R}^{T \times C_{\text{MSI}} \times H \times W} \quad (10)$$

where \hat{Y} = reconstructed MSI tensor,
 dimensions match the input tensor X_{MSI}

This linear reconstruction design is chosen for its simplicity and efficiency, enabling the model to learn direct mappings from learned representations to pixel-level reconstructions.

4.3 Multi-Scale Loss Function:

A multi-scale loss function is employed for the training process that combines two complementary objectives: Mean Squared Error (MSE) loss for pixel-level reconstruction accuracy, and Spectral Angle Mapper (SAM) loss for spectral consistency. The multi-scale formulation ensures reconstruction quality across multiple spatial resolutions:

$$\mathcal{L}_{\text{Multi-Scale}} = \sum_{s=1}^S w_{\text{MSE}}^{(s)} \cdot \mathcal{L}_{\text{MSE}}^{(s)} + w_{\text{SAM}}^{(s)} \cdot \mathcal{L}_{\text{SAM}}^{(s)} \quad (11)$$

where S = number of pyramid scales
 $w_{\text{MSE}}^{(s)}$ = weight for MSE loss at scale s
 $w_{\text{SAM}}^{(s)}$ = weight for SAM loss at scale s
 s = scale index

Images are rescaled using bilinear interpolation to generate the multi-scale pyramid.

MSE Loss measures global pixel-level fidelity:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N \cdot T \cdot C \cdot H \cdot W} \times \sum_{n=1}^N \sum_{t=1}^T \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W \left(y_{\text{true}}^{(n,t,c,h,w)} - y_{\text{pred}}^{(n,t,c,h,w)} \right)^2 \quad (12)$$

where $N = \text{batch size}$
 $y_{\text{true}}^{(n,t,c,h,w)}$ = true MSI image pixel spectral value
 $y_{\text{pred}}^{(n,t,c,h,w)}$ = predicted pixel spectral value

SAM Loss enforces spectral consistency by computing the angular distance between true and predicted spectral vectors at each pixel:

$$\mathcal{L}_{\text{SAM}} = \frac{1}{N \cdot T \cdot H \cdot W} \sum_{n=1}^N \sum_{t=1}^T \sum_{(h,w)=1}^{H \times W} \arccos \left(\frac{\langle \mathbf{y}_{\text{true},h,w}^{(n,t)}, \mathbf{y}_{\text{pred},h,w}^{(n,t)} \rangle}{\|\mathbf{y}_{\text{true},h,w}^{(n,t)}\|_2 \cdot \|\mathbf{y}_{\text{pred},h,w}^{(n,t)}\|_2 + \varepsilon} \right) \quad (13)$$

where $\mathbf{y}_{\text{true},h,w}^{(n,t)}$ = spectral vector at pixel (h, w) and time t
 $\mathbf{y}_{\text{pred},h,w}^{(n,t)}$ = predicted spectral vector at pixel (h, w) and time t
 $\langle \cdot, \cdot \rangle$ = dot product
 $\|\cdot\|_2$ = Euclidean norm
 ε = small constant to prevent numerical instability

5. Experiments

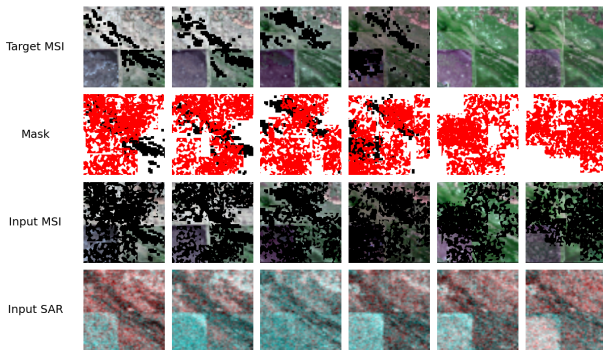


Figure 4. Overview of Multi-Modal Data Inputs and Cloud Masks. There are four rows: the first row shows target MSI reconstruction images, the second displays cloud masks (where black pixels indicate real cloud occlusion and red pixels indicate artificial cloud coverage), the third row contains input MSI data, and the fourth row presents input SAR data for the ViViT model. The x-axis indicates the timeline of data acquisition.

5.1 Experimental Setup

As shown in Figure 4, for the experimental setup, the MSI-only scenario uses a sequence of MSI as input, while the MSI and SAR fusion scenario uses both time-series MSI and co-registered SAR images as input. In both cases, a cloud mask is provided for each image, consisting of two parts: the original cloud mask and an additional artificial mask to simulate more cloud-covered areas. The output is the reconstructed time-series cloud-free MSI images. Four model variants were evaluated to assess the contribution of time-spatial fusion embedding:

- **MTS-ViT (Li et al., 2025)**: Baseline multi-temporal Vision Transformer with standard 2D convolutional patch projection using full-sequence temporal aggregation of MSI ($t = T = 6, C_{\text{total}} = C_{\text{MSI}} = 11$).
- **MTS-ViViT**: Proposed multi-temporal Video Vision Transformer with 3D convolutional tubelet extraction using constrained temporal span ($t = 2, C_{\text{total}} = C_{\text{MSI}} = 11$).
- **SMTS-ViT (Li et al., 2025)**: Baseline SAR-MSI fusion with standard temporal aggregation of both MSI and SAR ($t = T, C_{\text{total}} = C_{\text{MSI}} + C_{\text{SAR}} = 13$).
- **SMTS-ViViT**: Proposed SAR-MSI fusion with 3D tubelet projection ($t = 2, C_{\text{total}} = C_{\text{MSI}} + C_{\text{SAR}} = 13$).

5.2 Datasets and Settings

The dataset was partitioned into 80% for training and 20% for validation with spatial stratification. Training and validation data were subdivided into 60×60 pixel patches. To simulate realistic cloud occlusion, artificial cloud coverage was introduced during training using Gaussian smooth random noise with configurable cloud size (default `cloud_size = 0.3`, i.e., clouds initiate at 30% of image dimensions) and cloud count (default `#Clouds = 20` or `30`). It is important to clarify that the parameters `#Cloud = 20` and `#Cloud = 30` do not denote a total cloud cover percentage of 20% or 30% relative to the entire image area. Instead, `#Cloud` is defined as the specific count of individual artificial cloud patches randomly synthesized and distributed across the 60×60 pixel patches to simulate varying levels of occlusion severity. While the frequency of these occlusions is dictated by the `#Cloud` count, the spatial extent of each individual cloud is independently governed by a separate cloud size parameter, which is initialized at 30% of the image dimensions. Consequently, an increase from `#Cloud = 20` to `#Cloud = 30` represents a higher density of cloud artifacts and increasingly fragmented data gaps, rather than a linear increase in total pixel-wise cloud coverage. Cloud masks were applied via element-wise multiplication to corrupt MSI data while preserving SAR observations.

5.2.1 Hyperparameter Configuration: All models share identical hyperparameter settings: patch size 5×5 pixels, network depth 6 encoder layers, 8 multi-head self-attention heads, embedding dimension $d_e = 64$, and feed-forward hidden dimension $4 \times d_e = 256$. Training is conducted for 200 epochs with batch size 8 using Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) and learning rate $\eta_0 = 1 \times 10^{-3}$. Learning rate decay is applied at rate $\gamma = 0.95$ every 10 epochs. Random seed is fixed at 42 for reproducibility.

Loss function combines MSE and SAM with equal weights: $w_{\text{MSE}} = w_{\text{SAM}} = 0.5$. Multi-scale pyramid consists of $S = 3$ scales at resolutions $\{1.0, 0.5, 0.25\}$. The temporal span for baseline models is $t_{\text{ViT}} = T = 6$ time steps, while proposed models employ $t_{\text{ViViT}} = 2$ time steps. The 3D convolution kernel size is $(2, 5, 5)$ for temporal and spatial dimensions.

5.2.2 Robustness Assessment Model robustness is trained and evaluated across multiple cloud count configurations: `#Clouds = 20` and `30`. Each cloud configuration represents increasingly severe occlusion scenarios. For each configuration, a minimum of three independent runs are conducted using different random initializations, and results are averaged for statistical assessment.

Model	# Cloud	MSE ($\times 10^{-3}$) ↓	SAM ($\times 10^{-1}$) ↓	PSNR ↑	SSIM ↑
MTS-ViT	20	4.435	1.036	23.293	0.796
	30	7.131	1.402	22.256	0.727
MTS-ViViT	20	4.336	0.966	23.574	0.814
	30	5.002	1.131	22.925	0.772
SMTS-ViT	20	3.464	0.933	25.242	0.849
	30	3.966	1.071	24.052	0.783
SMTS-ViViT	20	3.106	0.857	25.649	0.867
	30	3.784	1.017	24.735	0.813
MTS-ViT (AVG)	-	5.783	1.219	22.775	0.762
MTS-ViViT (AVG)	-	4.669	1.049	23.250	0.793
SMTS-ViT (AVG)	-	<u>3.715</u>	<u>1.002</u>	<u>24.647</u>	<u>0.816</u>
SMTS-ViViT (AVG)	-	3.445	0.937	25.192	0.840

Table 1. Comprehensive evaluation results across all model configurations and cloud count scenarios. MTS-ViT and MTS-ViViT represent MSI-only reconstruction, while SMTS-ViT and SMTS-ViViT represent SAR-MSI fusion. Cloud counts (20, 30) represent increasing occlusion severity. Arrows indicate optimization direction: ↓ (lower better), ↑ (higher better). Best results highlighted in bold. The second-best results are underlined.

5.2.3 Evaluation Metrics Performance is quantified using four complementary metrics:

- (1) MSE measuring pixel-level reconstruction error,
- (2) SAM measuring angular deviation in spectral space,
- (3) Peak Signal-to-Noise Ratio (PSNR) evaluating signal quality in decibels, computed as

$$\text{PSNR}(X_{\text{MSI}}, \hat{Y}) = 10 \log_{10} \left(\frac{\text{MAX}_{X_{\text{MSI}}}^2}{\text{MSE}(X_{\text{MSI}}, \hat{Y})} \right) \quad (14)$$

where $\text{MAX}_{X_{\text{MSI}}} = 1$, image dynamic range for normalized data

- (4) Structural Similarity Index Measure (SSIM) assessing perceptual quality in range $[-1, 1]$:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (15)$$

where μ_x, μ_y = local means of x and y
 σ_x^2, σ_y^2 = local variances of x and y
 σ_{xy} = local covariance between x and y
 $C_1 = (K_1L)^2, C_2 = (K_2L)^2$,
 stabilization constants
 $L = 1$, image dynamic range for normalized data
 K_1, K_2 = small positive constants (typically $K_1 = 0.01, K_2 = 0.03$)

5.3 Results

Table 1 summarizes the quantitative performance of all models under varying cloud conditions, using four standard metrics: MSE ($\times 10^{-3}$), SAM ($\times 10^{-1}$), PSNR, and SSIM. Lower MSE and SAM scores, alongside higher PSNR and SSIM values, indicate better image reconstruction quality.

Across all groups, the proposed SMTS-ViViT consistently achieves the best average performance, with the lowest average MSE (3.445), lowest SAM (0.937), highest PSNR (25.192), and highest SSIM (0.840). SMTS-ViT ranks as the second-best approach, outperforming both baseline methods (MTS-ViT and MTS-ViViT) on all metrics. Under moderate cloud

levels (#Clouds = 20), SMTS-ViViT and SMTS-ViT demonstrate clear improvements over their standard counterparts, with SMTS-ViViT yielding the lowest errors and highest perceptual scores.

Importantly, when the cloud amount increases to 30, reconstruction quality decreases for all models, as expected. This demonstrates the robustness of the temporal-spatial embedding approach, especially when combined with the ViViT backbone, for reconstructing multispectral time series data under heavy cloud occlusions.

Comparing integration of SAR and MSI vs. MSI-only: Adding SAR auxiliary information substantially improves reconstruction performance. For #Clouds = 20:

Comparing MTS-ViT (MSI only) and SMTS-ViT (SAR+MSI), the inclusion of SAR reduces MSE by 21.98%, reduces SAM by 9.94%, increases PSNR by 8.37%, and increases SSIM by 6.66%.

Comparing MTS-ViViT and SMTS-ViViT, adding SAR decreases MSE by 28.37%, decreases SAM by 11.28%, increases PSNR by 8.09%, and increases SSIM by 6.51%.

These improvements are consistent with prior work showing that fusing SAR with optical data can significantly enhance cloud removal and image reconstruction performance. SAR provides robust complementary information when optical bands are corrupted by clouds, resulting in consistently better quantitative and perceptual results.

ViViT backbone benefits: For cloud=20, integrating the ViViT transformer backbone yields additional gains:

MTS-ViT vs. MTS-ViViT: ViViT reduces MSE by 2.23%.

SMTS-ViT vs. SMTS-ViViT: ViViT reduces MSE by 10.33%.

Similar improvements are observed under heavier cloud cover (#Clouds = 30), reinforcing the advantage of transformer-based architectures for time-series remote sensing tasks.

In summary, these results validate that fusing SAR data with MSI significantly boosts cloud removal accuracy over MSI-only baselines. The combination of our sparse-in-time modeling strategy with a transformer backbone (ViViT) yields state-of-the-art performance on all metrics and under all cloud conditions.



Figure 5. Cloud Removal and Reconstruction Results Across Models. Black pixels present cloud masks.

5.4 Visualization

To qualitatively assess model performance, a comprehensive visualization is presented in Figure 5. The figure displays ten rows: the input MSI image (with cloud mask), the reconstruction and error maps for SMTS-ViViT, SMTS-ViT, MTS-ViViT, and MTS-ViT, as well as the clear target MSI as a reference. The error maps are shown to highlight the deviations between each model's prediction and the ground truth.

From these visual comparisons, several trends are evident. SMTS-ViViT consistently produces reconstructions with the least visible residual error, closely matching the target MSI even in heavily clouded regions. As seen in the third row of the figure, the pink color of the error portion in the SMTS-ViViT result is lighter than the error hue in other results, indicating smaller errors. Furthermore, the errors in the time series images 1 and 3 also show significantly less error in the details.

Both SMTS-ViViT, SMTS-ViT recover more spatial detail and preserve spectral characteristics better than their counterparts using only MSI input. Error maps for SMTS-based models exhibit fewer and less intense artifacts.

A clearer distinction between the SMTS- and MTS-based reconstructions, as well as between the ViViT- and ViT-based reconstructions, is evident not only quantitatively but also visually. These qualitative differences further confirm the added value of fusing SAR information and highlight the effectiveness of the proposed temporal-spatial embedding approach for improved multispectral cloud removal and reconstruction.

6. Conclusions

This work proposes a ViViT-based model with temporal-spatial time embedding for cloud removal by fusing MSI and SAR, addressing the challenge of accurate time-series MSI reconstruction under different cloud conditions. Our main contribution is a temporal-spatial time embedding mechanism that allows the model to effectively leverage observations from multiple dates by explicitly encoding temporal information in both space and time, which leads to significantly better results than MSI-only or simple fusion approaches.

Our experiments show that combining SAR and MSI significantly boosts performance on all tested metrics, and the sparsity-aware temporal embedding further improves robustness under heavy clouds. The SMTS-ViViT model, which integrates this embedding with a transformer backbone, achieves the best results both quantitatively and visually.

However, the current version still aggregates MSI and SAR for each pixel as a mean feature across dates. For future work, we plan to feed all single-date MSI and SAR data separately and develop time embedding strategies that can flexibly accommodate different observation dates in different years or regions. Such methods will allow the network to understand and utilize information even when the set of available dates varies, opening new possibilities for robust cross-temporal, cross-regional analysis and applications in remote sensing.

References

- Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C., 2021. ViViT: A video vision transformer. *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.
- Eckardt, R., Berger, C., Thiel, C., Schmullius, C., 2013. Removal of optically thick clouds from multi-spectral satellite images using multi-frequency SAR data. *Remote Sensing*, 5(6), 2973–3006.
- Jing, R., Duan, F., Lu, F., Zhang, M., Zhao, W., 2023. Denoising diffusion probabilistic feature-based network for cloud removal in Sentinel-2 imagery. *Remote Sensing*, 15(9), 2217.
- Li, L., Wang, Y., State, R., 2025. Vision Transformer-Based Time-Series Image Reconstruction for Cloud-Filling Applications. *arXiv preprint arXiv:2506.19591*.
- Li, W., Li, Y., Chan, J. C.-W., 2019. Thick cloud removal with optical and SAR imagery via convolutional-mapping-deconvolutional network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4), 2865–2879.
- Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., Gascon, F., 2017. Sen2cor for sentinel-2. *Image and signal processing for remote sensing XXIII*, 10427, SPIE, 37–48.
- Qader, S. H., Dash, J., Alegana, V. A., Khwarahm, N. R., Tatem, A. J., Atkinson, P. M., 2021. The role of earth observation in achieving sustainable agricultural production in arid and semi-arid regions of the world. *Remote Sensing*, 13(17), 3382.
- Scarpa, G., Gargiulo, M., Mazza, A., Gaetano, R., 2018. A CNN-based fusion method for feature extraction from sentinel data. *Remote Sensing*, 10(2), 236.

Skakun, S., Wevers, J., Brockmann, C., Doxani, G., Aleksandrov, M., Batič, M., Frantz, D., Gascon, F., Gómez-Chova, L., Hagolle, O. et al., 2022. Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 274, 112990.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M. et al., 2012. GMES Sentinel-1 mission. *Remote sensing of environment*, 120, 9–24.

Tu, K., Yang, C., Qing, Y., Qi, K., Chen, N., Gong, J., 2025. Cloud removal with optical and SAR imagery via multimodal similarity attention. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226, 116–126.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y., Huang, H., State, R., 2023. Early crop mapping using dynamic ecoregion clustering: A USA-Wide study. *Remote Sensing*, 15(20), 4962.

Wang, Y., Huang, H., State, R., 2024a. Cross Domain Early Crop Mapping using CropSTGAN. *IEEE Access*.

Wang, Y., Huang, H., State, R., 2024b. Cross Domain Early Crop Mapping with Label Spaces Discrepancies using Multi-CropGAN. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 241–248.

Wang, Y., Zhang, B., Zhang, W., Hong, D., Zhao, B., Li, Z., 2024c. Cloud Removal With SAR-Optical Data Fusion Using a Unified SpatialSpectral Residual Network. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-20.

Xia, Q., Gao, X., Chu, W., Sorooshian, S., 2012. Estimation of daily cloud-free, snow-covered areas from MODIS based on variational interpolation. *Water Resources Research*, 48(9).

Zhao, M., Olsen, P., Chandra, R., 2023. Seeing through clouds in satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–16.