Beyond the Clouds: Reliable and Cloud-Aware Spatiotemporal Fusion via Adversarial Regression Wavelets

Sichen Lu[1,2], **Mingfei Li**[1,2], **Juanjuan Jing**[1,2], **Junhua Yu**[1,3], **Lei Yang**[1,2], **Boyang Nie**[1] and **Jinsong Zhou**[1,2,*]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2]School of Optoelectronics, University of Chinese Academy of Sciences, Beijing 100049, China
[3]State Key Laboratory of Internet of Things for Smart City and Department of Ocean Science and Technology, University of Macau, Macau, 999078, China.
{lusichen23, limingfei24}@mails.ucas.ac.cn, {jingjj, yangl, nieby, zhoujs}@aircas.ac.cn, yc37969@um.edu.mo *

Abstract

Spatiotemporal fusion (STF) bridges the gap between temporal and spatial resolutions in satellite imagery, enabling effective monitoring of Earth's surface dynamics. However, existing methods rely on cloud-free reference images, a constraint that fails in realistic, cloud-prone scenarios. To overcome this, we propose the **Cl**oud-**A**ware **W**avelet **G**enerative **A**dversarial **N**etwork (CLAW-GAN), a novel framework for high-fidelity reconstruction under cloud-contaminated conditions. CLAW-GAN introduces Regression Wavelet Analysis (RWA) to decouple spectral backgrounds from structural details. While a change-aware gated mechanism accounts for land-cover changes, the Frequency-Separated Fusion (FSF) module then independently integrates these components. To ensure visual realism, a multi-scale discriminator operates in the wavelet domain, enforcing consistency across high-frequency subbands to minimize artifacts. Evaluated on the newly introduced **G**lobal **C**loud-shrouded **A**gricultural **R**egions (GCAR) benchmark and the simulated Daxing dataset, CLAW-GAN achieves state-of-the-art performance and demonstrates superior robustness across varying cloud coverage.

1 Introduction

Acquiring satellite imagery with both high spatial resolution and high temporal frequency is critical for monitoring dynamic Earth surface processes [Chaminé *et al.*, 2021]. Unfortunately, due to hardware constraints, existing sensors invariably suffer from a trade-off between spatial detail and revisit cycles [Xiao *et al.*, 2023]. Spatiotemporal fusion (STF) mitigates this limitation by blending frequent coarse-resolution data with sparse high-resolution observations [Li *et al.*, 2025]. This technique provides synthetic time series that bridge the resolution gap, enabling precise tracking of rapid environmental changes such as crop phenology and urbanization [Guo and Shi, 2023].

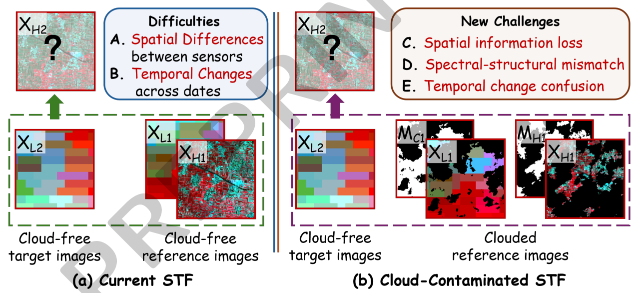

Figure 1: Challenges in (a) idealized cloud-free STF and (b) the cloud-constrained STF task addressed in this study.

As illustrated in Figure 1 (a), existing STF methods typically rely on an idealized clear-sky assumption and face two fundamental difficulties: spatial differences between heterogeneous sensors and temporal changes across dates. However, when reference images are contaminated by clouds, these models lack the necessary spatial guidance, which often leads to severe artifacts or blurred outputs. This problem is further intensified in the common but unexplored scenario where both high- and low-resolution references are simultaneously obscured.

Establishing a robust fusion framework under these compromised conditions is complicated by three further challenges (Figure 1 (b)). First, thick clouds and shadows cause significant **spatial information loss** by creating extensive data gaps that hinder the recovery of coherent textures. Second, **spectral–structural mismatch** arises when the inherent correlation between high- and low-resolution observations is disturbed by cloud pollution, hindering the unified spectral–spatial mapping. Third, **temporal change confusion** occurs when separating genuine land-cover variations from cloud artifacts, leading to potential semantic errors during feature propagation.

To bridge this critical gap, we present CLAW-GAN, a novel **Cl**oud-**A**ware **W**avelet **G**enerative **A**dversarial **N**etwork designed for robust STF under atmospheric interfer-

*Corresponding author.

ence. The model first decouples spectral backgrounds from structural details via Regression Wavelet Analysis (RWA) [Amrani *et al.*, 2016], establishing a unique pipeline where high-frequency (HF) and low-frequency (LF) components are processed independently. Guided by cloud masks and a reliability map from the Temporal Consistency Evaluator (TCE), the Frequency-Separated Fusion (FSF) model adaptively integrates decoupled feature streams by prioritizing stable land-cover information. Furthermore, a multi-scale discriminator with the stationary wavelet transform (SWT)[Nason and Silverman, 1995] is used to enforce textural and spectral fidelity across scales. We also introduce the **G**lobal **C**loud-shrouded **A**gricultural **R**egions (GCAR) benchmark to provide a large-scale evaluation platform with authentic cloud conditions. By combining wavelet-domain processing with adversarial training, CLAW-GAN delivers fusion images that are both temporally consistent and geographically faithful. The primary contributions of this work are summarized as follows:

- We formalize the cloud-constrained STF task and propose the cloud-reliable CLAW-GAN framework. To support this research, we establish GCAR, a large-scale benchmark built with real cloud-contaminated imagery from worldwide agricultural zones.
- A Wavelet-Pyramid Generator is designed to decouple LF backgrounds and HF details via RWA. This approach, integrated with the FSF module, effectively minimizes cross-frequency interference and recovers land-cover features with high fidelity, even under significant atmospheric interference.
- We develop a multi-scale discriminator that operates in the wavelet domain using SWT. It enforces consistency across directional HF subbands, thereby reducing synthetic artifacts and promoting geographically realistic outputs.
- Comprehensive experiments on the proposed GCAR benchmark and the cloud-simulated Daxing dataset demonstrate that CLAW-GAN achieves state-of-the-art (SOTA) performance and exhibits strong robustness across varying cloud coverage levels.

2 Related Work

2.1 Spatiotemporal Fusion for Remote Sensing

STF methods can be broadly categorized into five classes: weight function-based, unmixing-based, Bayesian-based, hybrid, and deep learning-based [Zhu *et al.*, 2018].

Weight-based methods like STARFM [Gao *et al.*, 2006] and ESTARFM [Zhu *et al.*, 2010] propagate reflectance changes by calculating spectral-temporal similarity weights. Alternatively, unmixing approaches such as MMT [Zhukov *et al.*, 1999] and OBSUM [Guo *et al.*, 2024] utilize linear decomposition, which can cause spectral distortions in heterogeneous regions. While Bayesian frameworks offer a probabilistic foundation for fusion, their high computational cost limits large-scale deployment [Li *et al.*, 2013]. Hybrid strategies, notably FSDAF [Zhu *et al.*, 2016], combine multiple techniques but often rely on empirical parameters that restrict their generalizability.

Recent advances in STF have been significantly driven by deep learning. CNN-based models like EDCSTFN [Tan *et al.*, 2019] and ECPW [Zhang *et al.*, 2024] effectively extract hierarchical spatial features for detail enhancement. Transformer-based frameworks such as SwinSTFM [Chen *et al.*, 2022] utilize self-attention mechanisms to model long-range dependencies and capture global contextual information. Diffusion-based frameworks reconstruct target images through an iterative denoising process that progressively transforms random noise into structured outputs [Huang *et al.*, 2024]. GAN-based methods, such as GAN-STFM [Tan *et al.*, 2022] and OPGAN [Song *et al.*, 2022], employ adversarial training to synthesize realistic textures and enhance spatial details.

Despite these advances, a key limitation of current STF methods is their reliance on the idealized cloud-free reference assumption. When input images are cloud-corrupted, traditional approaches lack reliable spatial guidance, which leads to inaccurate change propagation. While cloud-removal can be employed as preprocessing, such two-stage pipelines often introduce error accumulation and spectral shifts, ultimately degrading final fusion quality. Therefore, developing a unified framework that jointly handles cloud corruption and spatiotemporal reconstruction remains a significant research gap.

2.2 Wavelet Transforms

Wavelet Transform (WT) is a powerful tool for decomposing signals into localized spatial and frequency components. In image processing, the Discrete Wavelet Transform (DWT) [Mallat, 1989] separates data into frequency subbands through decimated filtering. In contrast, the SWT [Nason and Silverman, 1995] achieves translation invariance by omitting the down-sampling step, thereby better preserving edge information and mitigating artifacts. Furthermore, RWA [Amrani *et al.*, 2016] integrates wavelet decomposition with regression modeling to enhance feature separation and textural clarity. These multi-scale capabilities are particularly suited for isolating and synthesizing features across heterogeneous resolutions [Ji *et al.*, 2023], making wavelet theory a robust foundation for spatiotemporal fusion.

Wavelet-based approaches have demonstrated strong utility across multiple remote sensing tasks. WTSO [Yang *et al.*, 2019] effectively captures multi-resolution features in hyperspectral image classification, while IRPruneDet [Zhang *et al.*, 2025] enhances infrared target detection by improve robustness in identifying subtle features. DTWSR [Du *et al.*, 2025] efficiently process spectral-spatial information for image super-resolution while preserving high-frequency details. Cloud removal [Zi *et al.*, 2023] utilizes wavelet-based frameworks to effectively separate cloud artifacts from surface features across frequency bands. In STF, ECPW [Zhang *et al.*, 2024] introduced a CNN-based network with DWT and employed a wavelet-aware loss to recover details.

3 Method

3.1 Problem Formulation

Conventional STF assumes cloud-free high-resolution (HR) and low-resolution (LR) reference images, denoted as $\mathbf{X}_{\mathrm{H1}}$

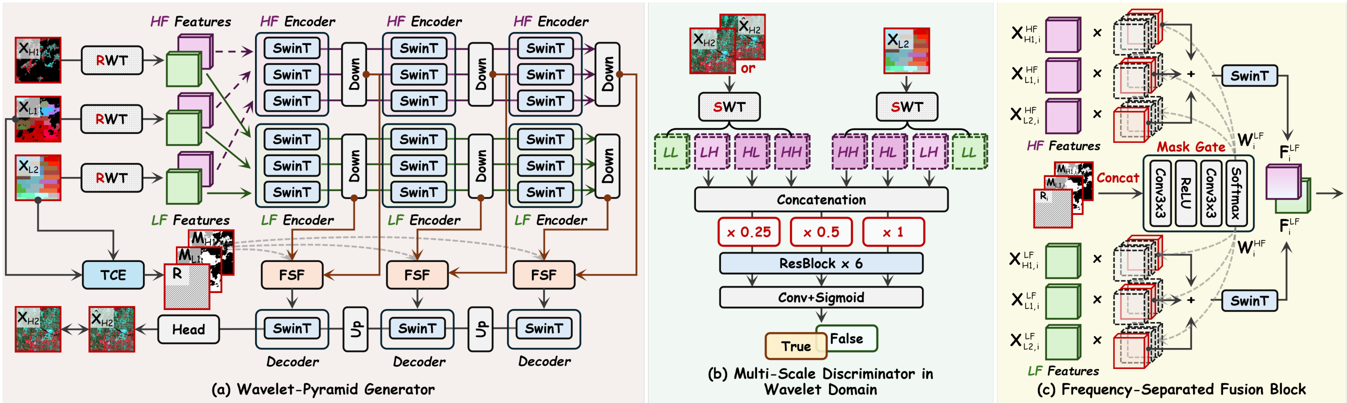

Figure 2: Flowchart of the proposed CLAW-GAN. Illustration of the (a) Wavelet-Pyramid Generator, (b) Multi-Scale Discriminator in Wavelet Domain and (c) Frequency-Separated Fusion Block.

and $\mathbf{X}_{\mathrm{L1}}$, respectively. As shown in Figure 1(a), the goal under this idealized setting is to generate a cloud-free HR image $\mathbf{X}_{\mathrm{H2}}$ at the target time by fusing spatial details from the references with temporal changes from the LR observation $\mathbf{X}_{\mathrm{L2}}$, where all images share the same dimensions $\mathbb{R}^{C\times H\times W}$.

In practice, however, clouds and other atmospheric contaminants frequently corrupt reference imagery. This work addresses the more realistic and challenging scenario of cloud-constrained STF, where both $\mathbf{X}_{\mathrm{H1}}$ and $\mathbf{X}_{\mathrm{L1}}$ are simultaneously contaminated (Figure 1(b)). We denote binary cloud masks $\mathbf{M}_{\mathrm{H1}}$ and $\mathbf{M}_{\mathrm{L1}}$, where values of 0 and 1 indicate cloudy and clear pixels, respectively. The reconstruction task is then formulated as learning a mapping function $\mathcal{F}$ that estimates the clean target from the corrupted observations:

$$\hat{\mathbf{X}}_{\mathrm{H2}} = \mathcal{F}\left(\{\mathbf{X}_{\mathrm{H1}}, \mathbf{M}_{\mathrm{H1}}\}, \{\mathbf{X}_{\mathrm{L1}}, \mathbf{M}_{\mathrm{L1}}\}, \mathbf{X}_{\mathrm{L2}}\right). \tag{1}$$

3.2 Overall Framework of CLAW-GAN

CLAW-GAN follows a hierarchical pipeline built upon wavelet decoupling. As shown in Figure 2 (a), The Wavelet-Pyramid Generator $\mathcal{G}$ first employs RWA to decompose each input source into LF spectral backgrounds and HF structural details. These components are then processed separately through parallel Swin Transformer (SwinT) encoders to capture multi-scale spatial contexts. Subsequently, separate gated units adaptively aggregate these streams, utilizing cloud masks and reliability maps to fill information gaps. Finally, a SwinT-based pyramidal decoder synthesizes the fused representations into a cloud-free estimate $\hat{\mathbf{X}}_{\mathrm{H2}}$.

To ensure textural realism, a Multi-scale Discriminator $\mathcal{D}$ evaluates the results in the wavelet domain. Using the SWT, the discriminator extracts directional HF subbands from $\hat{\mathbf{X}}_{\mathrm{H2}}, \mathbf{X}_{\mathrm{H2}}$ and $\mathbf{X}_{\mathrm{L2}}$, respectively. By evaluating these concatenated features across multiple scales, the discriminator identifies spectral artifacts and directional inconsistencies that are often overlooked in the image domain. This frequency-aware supervision guides the model toward generating geographically plausible content, ensuring high-fidelity reconstruction even under significant atmospheric interference.

3.3 Wavelet-Pyramid Generator

Wavelet-based Spectral Decomposition

Atmospheric interference distorts both spectral irradiance and spatial details, inducing artifacts during direct pixel-domain fusion. To mitigate this, we map multi-source inputs into the wavelet domain using RWA. Unlike conventional transforms, RWA exploits inter-band correlations via regression-based prediction to achieve a more compact and physically consistent representation. For a spectral vector $\mathbf{v}$ at any pixel, RWA applies a lifting operation that splits it into odd ($\mathbf{v}_o$) and even ($\mathbf{v}_e$) samples:

$$\mathbf{h} = \mathbf{v}_o - \mathbf{v}_e, \quad \mathbf{l} = \mathbf{v}_e + \lfloor \mathbf{h}/2 \rfloor, \tag{2}$$

where $\mathbf{h}$ captures spectral differences and $\mathbf{l}$ represents the local average. The operator $\lfloor\cdot\rfloor$ denotes the floor function. A regression model then estimates the HF component from its LF counterpart to yield a prediction $\hat{\mathbf{h}}$. The final decomposed features for each source $s \in \{H1, L1, L2\}$ are:

$$\mathbf{F}_s^{\mathrm{LF}} = \mathbf{l}, \quad \mathbf{F}_s^{\mathrm{HF}} = \mathbf{h} - \mathrm{round}(\hat{\mathbf{h}}). \tag{3}$$

This mapping effectively isolates global spectral trends ($\mathbf{F}_s^{\mathrm{LF}}$) from fine textures ($\mathbf{F}_s^{\mathrm{HF}}$), providing a stable foundation for robust feature extraction under varying atmospheric conditions.

Hierarchical Spatial Modeling via SwinT Blocks

CLAW-GAN adopts a hierarchical SwinT backbone to capture non-local spatial dependencies efficiently. To handle high-resolution fusion with manageable computation, the SwinT block divides the feature maps into non-overlapping local windows of size $M \times M$. Inside each window, multi-head self-attention (W-MSA) is computed as:

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{4}$$

where $Q, K, V \in \mathbb{R}^{M^2\times d}$ denote the query, key, and value matrices, d is the query dimension, and B is a learnable relative position bias. To enable information exchange across neighboring regions, a shifted windowing (SW-MSA) mechanism is utilized in alternating layers.

Through successive stages of resolution adjustment and channel expansion, this hierarchical design enables the generator to model geographic semantics across scales, effectively recovering missing content in cloud-occluded regions.

Temporal Consistency Evaluator

To manage land-cover variations between acquisition dates, the TCE processes the stacked low-resolution pair $[\mathbf{X}_{\text{L1}}, \mathbf{X}_{\text{L2}}]$ through two 3×3 convolutional layers. With intermediate ReLU and final Sigmoid activations, the module outputs a single-channel reliability map $\mathbf{R} \in [0, 1]^{H \times W}$, which quantifies surface stability at the original resolution. $\mathbf{R}$ highlights temporally consistent regions, thereby preventing the transfer of outdated or irrelevant features from the reference to the target reconstruction.

Frequency-Separated Fusion

The linear relationships required for a standard inverse RWA transformation are disrupted during multi-source fusion, making direct frequency-to-image recovery no longer feasible. We therefore introduce the FSF module to aggregate frequency components independently, preventing cross-frequency interference (Figure 2 (c)).

At each pyramid scale i, a mask-gated network $\mathcal{M}$ generates weights $\mathbf{W}_i$ from the stacked cloud masks and reliability map:

$$\mathbf{W}_i = \text{Softmax}\big(\mathcal{M}([\mathbf{M}_{\text{H1},i}, \mathbf{M}_{\text{L1},i}, \mathbf{R}_i])\big). \tag{5}$$

For each frequency $\phi \in \{\text{LF}, \text{HF}\}$, the fused features are calculated as a weighted sum of the three input streams:

$$\mathbf{F}_i^{\phi} = \mathbf{W}_{\text{H1},i}^{\phi}\mathbf{F}_{\text{H1},i}^{\phi} + \mathbf{W}_{\text{L1},i}^{\phi}\mathbf{F}_{\text{L1},i}^{\phi} + \mathbf{W}_{\text{L2},i}^{\phi}\mathbf{F}_{\text{L2},i}^{\phi}. \tag{6}$$

This design dynamically rebalances information sources based on cloud coverage and temporal consistency. It prioritizes high-resolution textures from $\mathbf{X}_{\text{H1}}$ in clear, stable regions while relying more on the temporal dynamics from $\mathbf{X}_{\text{L2}}$ when reference pixels are occluded or altered. This selective aggregation ensures that the fused features maintain both structural sharpness and spectral accuracy.

3.4 Multi-scale Discriminator in Wavelet Domain

To ensure the fidelity of synthesized high-frequency details, we implement a multi-scale discriminator $\mathcal{D}$ that operates directly in the wavelet domain, as illustrated in Figure 2(b). For each multispectral channel, the input image is decomposed via the SWT into one LF approximation subband (LL) and three directional HF detail subbands (LH, HL, and HH). By maintaining the original spatial resolution, SWT ensures that the HF subbands (carrying horizontal, vertical, and diagonal textures) remain precisely aligned with its corresponding spatial location. This design is motivated by the observation that fine-grained textures and potential artifacts are most discriminative in the HF domain, whereas the spectral background in the LL subband is already well-constrained by the generator's reconstruction losses.

$\mathcal{D}$ adopts a hierarchical architecture that processes these wavelet subbands at three spatial scales: the original resolution, $1/2\times$, and $1/4\times$. Each sub-discriminator consists of residual blocks with spectral normalization to stabilize adversarial training. By verifying the consistency of directional HF subbands across multiple scales, $\mathcal{D}$ guides $\mathcal{G}$ to produce outputs with sharp edges and realistic spatial semantics.

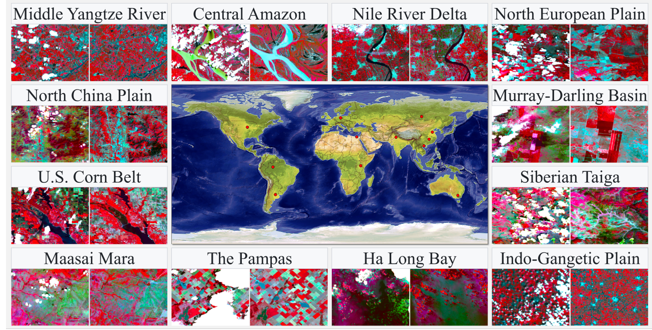

Figure 3: Distribution and representative imagery of the 12 GCAR study sites. Red markers indicate collection locations across diverse geographical regions. Each image set displays a high-resolution pair, with the reference on the left and the target on the right.

3.5 Loss Functions

To achieve stable reconstruction under atmospheric interference, the CLAW-GAN framework is optimized within the Least Squares GAN (LSGAN) [Mao *et al.*, 2017] paradigm, which provides stable gradient flow and mitigates the vanishing gradient problem. While $\mathcal{D}$ follows the standard LSGAN update, $\mathcal{G}$ is trained by minimizing a composite objective function that balances adversarial realism with reconstruction accuracy:

$$\mathcal{L}_G = \mathcal{L}_{MSE} + \alpha\mathcal{L}_{SSIM} + \beta\mathcal{L}_{cos} + \gamma\mathcal{L}_{adv}, \tag{7}$$

where $\mathcal{L}_{MSE}$ is the mean squared error that reduces overall pixel-wise discrepancies; $\mathcal{L}_{SSIM}$ is a multi-scale structural similarity loss that preserves perceptual details; $\mathcal{L}_{cos}$ enforces spectral alignment across all channels via cosine similarity; and $\mathcal{L}_{adv}$ denotes the adversarial loss from the LSGAN formulation. By jointly optimizing pixel-level accuracy, structural consistency, and spectral alignment, CLAW-GAN effectively corrects cloud-induced artifacts while maintaining high spectral fidelity.

4 Experiment

4.1 Datasets

GCAR

As depicted in Figure 3, the GCAR dataset facilitates STF resilience assessment across 12 diverse biomes spanning six continents, such as the Amazon rainforest, Nile Delta, Siberian Taiga, and U.S. Corn Belt. It utilizes Landsat-8/9 OLI Surface Reflectance (30 m) as HR data and MODIS MOD09A1 (500 m) as LR data. Covering February 2022 to March 2025, the dataset features authentic reference-time cloud cover between 5% and 45%. In our experiments, the Pampas (Argentina) and Indo-Gangetic Plain (India) are reserved for testing, while the remaining regions are used for training. Further details on the dataset construction pipeline are provided in Appendix A.

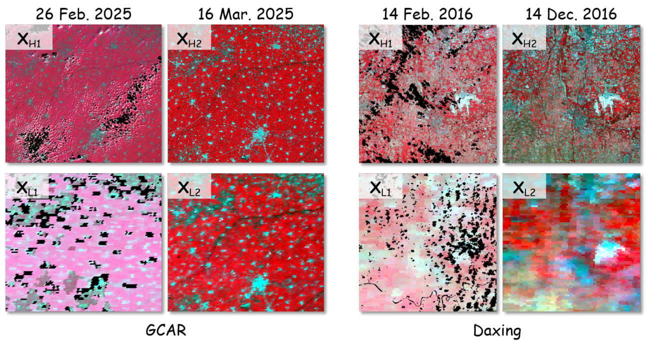

Figure 4: The pseudo-colored images of the test pairs on the GCAR (left) and 15% cloud-contaminated Daxing (right) datasets.

Daxing

The Daxing dataset [Li *et al.*, 2020] is a well-established benchmark containing 29 MODIS-Landsat pairs (2013-2019) across urban, cropland, and mountainous landscapes in Beijing, China. To simulate cloud-constrained scenarios, we superimposed 30 real cloud masks (5-45% coverage) extracted from contemporary Landsat-9 imagery onto the original cloud-free references. For each task, the reference pair was selected based on temporal proximity to the target date. Five image pairs from February 2016 to November 2017 are used for testing, with the rest for training, ensuring model evaluation under realistic atmospheric interference.

4.2 Settings

Implementation Details

All experiments used PyTorch with an NVIDIA RTX 3090 GPU. Images were processed as 256×256 patches with a stride of 128. Training used the AdamW optimizer (initial learning rate 2×10^{-4}) for 200 epochs with a batch size of 10. For the loss function, $\alpha = \beta = 1$ and $\gamma = 0.01$.

Evaluation Metric

Fusion quality is assessed using five metrics, including Root Mean Square Error (RMSE), Structural Similarity Index (SSIM) [Wang *et al.*, 2004], Correlation Coefficient (CC) [Kaneko *et al.*, 2003], Spectral Angle Mapper (SAM) [Yuhas *et al.*, 1992], and Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) [Khan *et al.*, 2009]. Lower RMSE, SAM, and ERGAS values and higher SSIM and CC values indicate better performance.

Comparisons methods

We compare CLAW-GAN with seven representative STF methods, including traditional approaches (STARFM [Gao *et al.*, 2006], OBSUM [Guo *et al.*, 2024], FSDAF [Zhu *et al.*, 2016]), a wavelet-integrated CNN (ECPW [Zhang *et al.*, 2024], GAN-based architectures (OPGAN [Song *et al.*, 2022], GANSTFM [Tan *et al.*, 2022]) and a Transformer-based method (SwinSTFM [Chen *et al.*, 2022]). All methods were trained and tested under identical settings for a fair comparison.

4.3 Qualitative Comparisons

Visual performance is evaluated on the GCAR benchmark and the Daxing dataset with 15% simulated cloud coverage, the test image pairs are illustrated in Figure 4. Figure 5 provides detailed comparisons including full-scene fused images, enlarged local patches, and their corresponding Average Absolute Difference (AAD) maps. For a more comprehensive assessment, additional visualizations, full-image AAD maps, and scatter plots are provided in the Appendix B.

Traditional STF methods struggle to reconstruct discernible ground features, typically resulting in severe blurring or spectral distortions in regions affected by clouds. While deep learning approaches improve sharpness, they frequently introduce inconsistent textures and artifacts along cloud boundaries. In contrast, CLAW-GAN generates coherent imagery with high fidelity, effectively preserving distinct field boundaries and authentic urban textures. Substantially lower AAD errors further validate the robustness of our model, showing that it maintains accurate inference of realistic land-cover semantics even under heavy cloud cover.

4.4 Quantitative Comparisons

Our method achieves SOTA on both the GCAR and Daxing datasets across all five evaluation metrics (Table 1). The detailed band-wise analysis is available in Appendix B. Under cloud-constrained conditions, the performance of traditional STF methods declines significantly. This limitation is primarily attributed to their reliance on linear mapping or weighted averaging assumptions, which are insufficient to address the complex nonlinear spectral-spatial distortions and extensive data gaps caused by cloud cover. In contrast, CLAW-GAN demonstrates superior robustness in both radiometric reconstruction and structural preservation despite substantial information loss.

4.5 Ablation Study

Ablation studies are conducted on the GCAR and Daxing datasets to comprehensively examine the impact of key components within the CLAW-GAN framework, focusing on frequency-domain representation and architectural design.

Effectiveness of Wavelet-Domain Processing

We first evaluate the necessity of frequency-domain processing by comparing CLAW-GAN against several spatial and alternative wavelet baselines. In the generator, the RWA-based decomposition is replaced by a direct spatial-domain baseline and the standard DWT. This comparison isolates the advantage of RWA in decoupling spectral and spatial information. Similarly, the contribution of the SWT in the discriminator is assessed by substituting it with DWT or standard spatial inputs. The omission or replacement of these components results in a consistent performance decline across all metrics (Table 2). These results confirm that regression-based decorrelation in RWA and shift-invariant supervision from SWT are both critical for maintaining reconstruction fidelity under cloud-constrained condition.

Contribution of Architectural Components

To assess the impact of individual architectural modules, we perform a series of ablation studys on the core components of

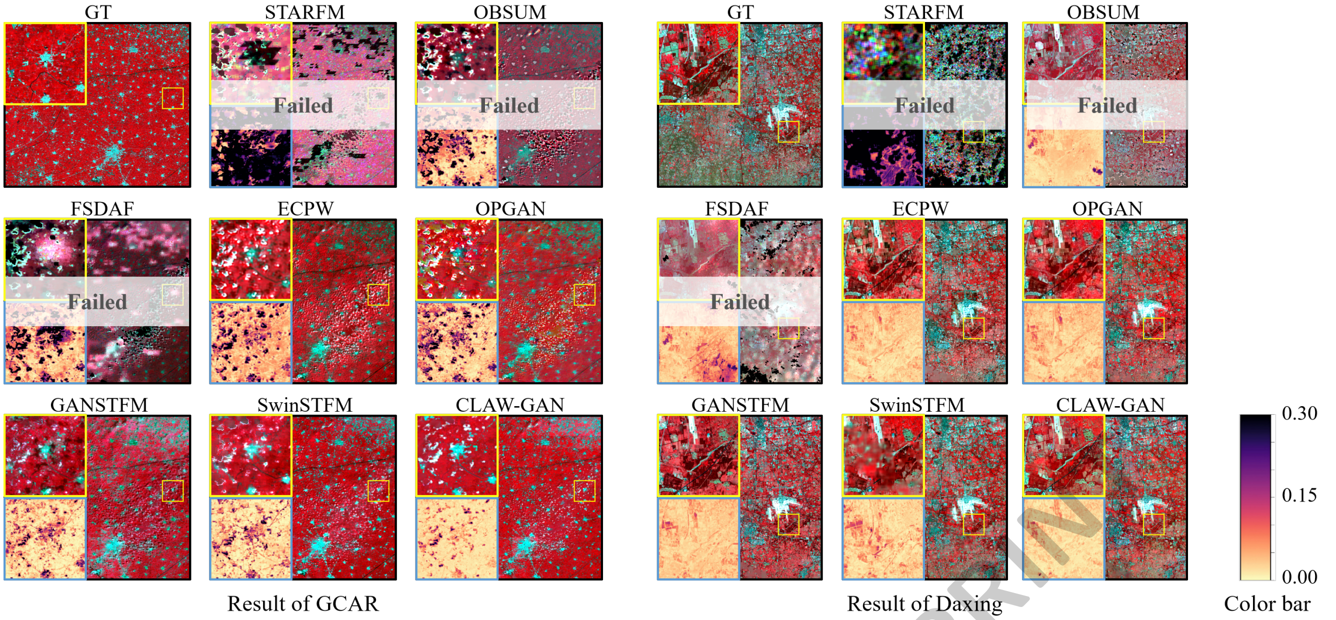

Figure 5: Visual comparison of different fusion methods on the GCAR benchmark (left) and 15% cloud-contaminated Daxing dataset (right). Each panel presents the ground truth (GT), fused results from various models, zoomed-in local patches, and their corresponding AAD maps.

Dataset	Index	STARFM	OBSUM	FSDAF	ECPW	OPGAN	GANSTFM	SwinSTFM	CLAW-GAN
Cloud	**RMSE**	0.1755	0.0323	0.0436	0.0231	0.0221	0.0199	0.0190	**0.0171**
	SSIM	0.2800	0.7615	0.6468	0.8028	0.8143	0.8425	0.8421	**0.8749**
	CC	0.0961	0.4068	0.2548	0.4882	0.5228	0.6229	0.6994	**0.7664**
	SAM	36.9926	30.1781	37.7981	22.8199	22.2913	19.4206	17.7108	**15.5415**
	ERGAS	15.9126	4.5907	4.9659	3.0484	2.8966	2.4919	2.4947	**2.0655**
Daxing	**RMSE**	0.2547	0.0233	0.0417	0.0173	0.0184	0.0196	0.0158	**0.0156**
	SSIM	0.0963	0.8285	0.6769	0.8681	0.8644	0.8376	0.8812	**0.8855**
	CC	0.1694	0.6556	0.3979	0.7939	0.8084	0.7966	0.8193	**0.8252**
	SAM	43.4291	16.8555	28.1614	12.1648	11.8672	14.2050	11.2382	**11.2176**
	ERGAS	20.7836	1.9610	3.2693	1.5854	1.5953	2.0411	1.4709	**1.4566**

Table 1: Quantitative performance evaluation on the GCAR benchmark and 15% cloud-contaminated Daxing dataset in terms of mean RMSE, SSIM, CC, SAM and ERGAS. The **bold** and the underline indicate the best and the second-best results respectively.

the framework. Specifically, we remove TCE to examine its role in handling land-cover variations. FSF is replaced with a direct feature concatenation operation to evaluate its effectiveness in frequency-separated processing. Furthermore, the pyramidal hierarchy is tested by restricting fusion to a uniform resolution, while the multi-scale design of the discriminator is compared against a fixed-scale variant.

Removing or simplifying any of these components leads to a consistent decline across all evaluation metrics (Table 2). These findings confirm that the combined design, which integrates change awareness from the TCE, decoupled processing in FSF, and a hierarchical multi scale architecture, is essential for preserving spectral and structural integrity when reference data is obscured.

4.6 Case Study

To evaluate the performance boundaries of CLAW-GAN, we conducted a case study on the Daxing dataset with simulated cloud levels of 15%, 30%, and 45%. The corresponding cloud masks are shown in Figure 6. Quantitative results for deep learning methods are summarized in Figure 7, with bar charts for RMSE and line plots for SSIM. As expected, all models degrade with increasing occlusion, yet CLAW-GAN maintains the highest stability, with its metrics declining at a markedly slower rate than competing approaches.

Band-wise analysis (detailed in Appendix C) indicates that shorter wavelengths are more susceptible under heavy cloud cover. This sensitivity likely arises from the higher spatial complexity and subtler spectral variations in these bands, which complicate the recovery of fully obscured pixels. CLAW-GAN delivers consistently high accuracy across

Dataset	Index	Wavelet-Domain Processing					Architectural Design (w/o)				Full Set
		-/-	-/SWT	DWT/SWT	RWA/-	RWA/DWT	TCE	FSF	Pyramid	MS-Disc	
GCAR	**RMSE**	0.0188	0.0186	0.0182	0.0185	0.0179	0.0176	0.0182	0.0179	0.0175	**0.0171**
	SSIM	0.8565	0.8592	0.8654	0.8618	0.8692	0.8685	0.8602	0.8634	0.8712	**0.8749**
	CC	0.7284	0.7356	0.7482	0.7512	0.7554	0.7521	0.7388	0.7445	0.7592	**0.7664**
	SAM	16.754	16.982	16.445	16.712	16.124	15.985	16.842	16.512	15.824	**15.5415**
	ERGAS	2.3521	2.3842	2.2954	2.3412	2.1845	2.1456	2.3124	2.2541	2.1124	**2.0655**
Daxing	**RMSE**	0.0178	0.0174	0.0168	0.0171	0.0163	0.0160	0.0172	0.0169	0.0159	**0.0156**
	SSIM	0.8612	0.8665	0.8732	0.8694	0.8785	0.8792	0.8684	0.8712	0.8821	**0.8855**
	CC	0.7924	0.7985	0.8092	0.8045	0.8164	0.8185	0.8012	0.8054	0.8212	**0.8252**
	SAM	12.542	12.312	11.954	12.124	11.645	11.584	12.421	12.185	11.412	**11.2176**
	ERGAS	1.7245	1.6842	1.6124	1.6452	1.5421	1.5245	1.6954	1.6241	1.4985	**1.4566**

Table 2: Quantitative results of ablation experiments for CLAW-GAN. In the WT ablation columns, the notation X/Y specifies the transform types used in the generator and discriminator, respectively, where ”-” denotes that no wavelet transform was applied. ”w/o” represents results obtained without a specific component. **Bold** values denote the best performance.

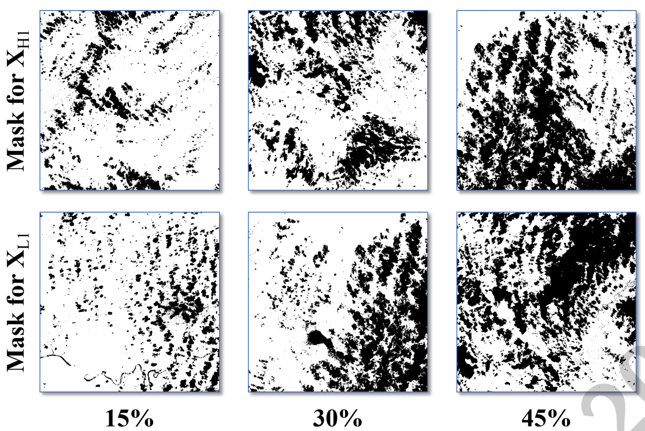

Figure 6: Simulated cloud masks at varied coverage levels. The top and bottom rows display the masks for $\mathbf{X}_{H1}$ and $\mathbf{X}_{L1}$, respectively. From left to right, the cloud cover fraction is 15%, 30% and 45%.

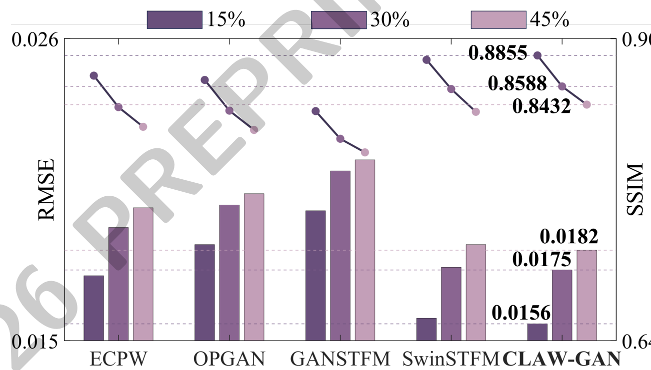

Figure 7: Performance trends of deep learning methods across varied cloud cover levels. Bar charts and line plots represent the mean RMSE and SSIM results, respectively, as the cloud coverage increases from 15% to 45%.

the spectrum and outperforms other models in overall averages, despite comparable results with SwinSTFM in a few isolated parameters. Visual AAD maps and scatter plots further confirm these findings, demonstrating that CLAW-GAN yields the smallest deviations from ground truth across diverse land-cover types. Together, these results underscore the reliable of our framework for continuous time-series monitoring, ensuring high-fidelity data reconstruction even under extreme atmospheric conditions.

5 Conclusion

This paper introduces CLAW-GAN, a wavelet-based adversarial framework designed for robust STF under cloud-contaminated conditions. The Wavelet-Pyramid Generator utilizes RWA to separate global spectral trends from fine-scale textures. The resulting components are independently fused by the FSF module, which is guided by a TCE to adaptively handle land-cover changes. A multi-scale discriminator operating in the SWT domain further enforces the geographical authenticity of reconstructed high-frequency details. Comprehensive evaluations on the proposed GCAR benchmark and the simulated Daxing dataset confirm that CLAW-GAN achieves SOTA accuracy and structural preservation, demonstrating strong robustness even under 45% cloud coverage. Future work will investigate mask-free and uncertainty-aware modeling to generalize across diverse atmospheric scenarios and application domains. This work establishes a reliable foundation for continuous Earth observation in suboptimal imaging conditions.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by the National natural science foundation of China under Grant 42427801.

References

[Amrani *et al.*, 2016] Naoufal Amrani, Joan Serra-Sagristà, Valero Laparra, Michael W. Marcellin, and Jesus Malo. Regression wavelet analysis for lossless coding of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 54(9):5616–5627, 2016.

[Chaminé *et al.*, 2021] Helder I. Chaminé, Alcides J. S. C. Pereira, Ana C. Teodoro, and José Teixeira. Remote sensing and GIS applications in earth and environmental systems sciences. *SN Applied Sciences*, 3(12):870, November 2021.

[Chen *et al.*, 2022] Guanyu Chen, Peng Jiao, Qing Hu, Linjie Xiao, and Zijian Ye. SwinSTFM: Remote Sensing Spatiotemporal Fusion Using Swin Transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022.

[Du *et al.*, 2025] Peng Du, Hui Li, Han Xu, Paul Barom Jeon, Dongwook Lee, Daehyun Ji, Ran Yang, and Feng Zhu. Diffusion transformer meets multi-level wavelet spectrum for single image super-resolution, 2025.

[Gao *et al.*, 2006] Feng Gao, J. Masek, M. Schwaller, and F. Hall. On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8):2207–2218, August 2006.

[Guo and Shi, 2023] Dizhou Guo and Wenzhong Shi. Object-Level Hybrid Spatiotemporal Fusion: Reaching a Better Tradeoff Among Spectral Accuracy, Spatial Accuracy, and Efficiency. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:8007–8021, 2023.

[Guo *et al.*, 2024] Houcai Guo, Dingqi Ye, Hanzeyu Xu, and Lorenzo Bruzzone. OBSUM: An object-based spatial unmixing model for spatiotemporal fusion of remote sensing images. *Remote Sensing of Environment*, 304:114046, April 2024.

[Huang *et al.*, 2024] He Huang, Wei He, Hongyan Zhang, Yu Xia, and Liangpei Zhang. STFDiff: Remote sensing image spatiotemporal fusion with diffusion models. *Information Fusion*, 111:102505, 2024.

[Ji *et al.*, 2023] Deyi Ji, Feng Zhao, and Hongtao Lu. Guided patch-grouping wavelet transformer with spatial congruence for ultra-high resolution segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 920–928, 2023.

[Kaneko *et al.*, 2003] Shun'ichi Kaneko, Yutaka Satoh, and Satoru Igarashi. Using selective correlation coefficient for robust image registration. *Pattern Recognition*, 36(5):1165–1173, May 2003.

[Khan *et al.*, 2009] Muhammad Murtaza Khan, Luciano Alparone, and Jocelyn Chanussot. Pansharpening quality assessment using the modulation transfer functions of instruments. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3880–3891, January 2009.

[Li *et al.*, 2013] Aihua Li, Yanchen Bo, Yuxin Zhu, Peng Guo, Jian Bi, and Yaqian He. Blending multi-resolution satellite sea surface temperature (SST) products using Bayesian maximum entropy method. *Remote Sensing of Environment*, 135:52–63, August 2013.

[Li *et al.*, 2020] Jun Li, Yunfei Li, Lin He, Jin Chen, and Antonio Plaza. Spatio-temporal fusion for remote sensing data: an overview and new benchmark. *Science China Information Sciences*, 63(4):140301, mar 2020.

[Li *et al.*, 2025] Kangning Li, Zilin Xie, Xiaojun Qiao, Jinzhong Yang, and Jinbao Jiang. An Improved Spatiotemporal Fusion Framework for Land-Cover Temporal Harmonization of High-Resolution Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:11731–11750, 2025.

[Mallat, 1989] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[Mao *et al.*, 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[Nason and Silverman, 1995] G. P. Nason and B. W. Silverman. *The Stationary Wavelet Transform and some Statistical Applications*, pages 281–299. Springer New York, New York, NY, 1995.

[Song *et al.*, 2022] Yiyao Song, Hongyan Zhang, He Huang, and Liangpei Zhang. Remote sensing image spatiotemporal fusion via a generative adversarial network with one prior image pair. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.

[Tan *et al.*, 2019] Zhenyu Tan, Liping Di, Mingda Zhang, Liying Guo, and Meiling Gao. An Enhanced Deep Convolutional Model for Spatiotemporal Image Fusion. *Remote Sensing*, 11(24):2898, January 2019.

[Tan *et al.*, 2022] Zhenyu Tan, Meiling Gao, Xinghua Li, and Liangcun Jiang. A Flexible Reference-Insensitive Spatiotemporal Fusion Model for Remote Sensing Images Using Conditional Generative Adversarial Network. *IEEE Transactions on Geoscience and Remote Sensing*, page 1–13, January 2022.

[Wang *et al.*, 2004] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[Xiao *et al.*, 2023] Juan Xiao, Ashwani Kumar Aggarwal, Nguyen Hong Duc, Abhinandan Arya, Uday Kiran Rage, and Ram Avtar. A review of remote sensing image spatiotemporal fusion: Challenges, applications and recent trends. *Remote Sensing Applications: Society and Environment*, 32:101005, November 2023.

[Yang *et al.*, 2019] Lina Yang, Hailong Su, Cheng Zhong, Zuqiang Meng, Huiwu Luo, Xichun Li, Yuan Yan Tang, and Yang Lu. Hyperspectral image classification using wavelet transform-based smooth ordering. *International Journal of Wavelets, Multiresolution and Information Processing*, 17(06):1950050, 2019.

[Yuhas *et al.*, 1992] RobertaH. Yuhas, Alexander Goetz, and J. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop*, 1:147–149, June 1992.

[Zhang *et al.*, 2024] Xingjian Zhang, Shuang Li, Zhenyu Tan, and Xinghua Li. Enhanced wavelet based spatiotemporal fusion networks using cross-paired remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211:281–297, May 2024.

[Zhang *et al.*, 2025] Mingjin Zhang, Jin Feng, Handi Yang, Jie Guo, Yunsong Li, and Xinbo Gao. Irprunedext: Efficient infrared small target detection via musical wavelet-regularized channel pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(12):20229–20242, 2025.

[Zhu *et al.*, 2010] Xiaolin Zhu, Jin Chen, Feng Gao, Xuehong Chen, and Jeffrey G Masek. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sensing of Environment*, 114(11):2610–2623, November 2010.

[Zhu *et al.*, 2016] Xiaolin Zhu, Eileen H. Helmer, Feng Gao, Desheng Liu, Jin Chen, and Michael A. Lefsky. A flexible spatiotemporal method for fusing satellite images with different resolutions. *Remote Sensing of Environment*, 172:165–177, 2016.

[Zhu *et al.*, 2018] Xiaolin Zhu, Fangyi Cai, Jiaqi Tian, and Trecia Williams. Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions. *Remote Sensing*, 10(4):527, March 2018.

[Zhukov *et al.*, 1999] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhackel. Unmixing-based multisensor multiresolution image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3):1212–1226, May 1999.

[Zi *et al.*, 2023] Yue Zi, Haidong Ding, Fengying Xie, Zhiguo Jiang, and Xuedong Song. Wavelet integrated convolutional neural network for thin cloud removal in remote sensing images. *Remote Sensing*, 15(3), 2023.