

Improved Causal Temporal Fusion Transformer for Cloud Shadow Removal under Multi-temporal Satellite Reflectance Conditions

Aleksander Borkowski^{1,*} and Nikodem Kocur¹

¹ Faculty of Computer Science and Information Technology, Poznan University of Technology, Poznan, 60-965, Poland

*Corresponding author: aleksander.b@put.poznan.pl

Abstract. Cloud shadow contamination produces spectrally depressed and spatially irregular reflectance observations in optical satellite image sequences, reducing the reliability of land-cover monitoring and biophysical index retrieval. This paper develops an improved Causal Temporal Fusion Transformer, termed Causal-TFT-SR, for cloud shadow removal under multi-temporal satellite reflectance conditions. The model integrates shadow-aware band embedding, strictly causal temporal attention, gated static-environment conditioning, and a reflectance-consistency reconstruction head. A multi-region experiment was designed with five land-surface types, seven spectral bands, and a 16-date temporal window. Compared with temporal linear interpolation, U-Net restoration, bidirectional LSTM fusion, and a non-causal transformer baseline, the proposed model reduced mean absolute reflectance error to 0.0127, increased SSIM to 0.913, decreased spectral angle to 3.41 degrees, and reduced NDVI error to 0.026. Ablation results further show that causal masking, variable selection, and radiometric consistency are all necessary for stable sequential restoration. The results indicate that causal temporal fusion can recover shadow-affected reflectance without relying on future observations, making the method suitable for near-real-time optical satellite production chains.

Keywords: *Cloud Shadow Removal, Multi-temporal Reflectance, Causal Temporal Fusion Transformer, Radiometric Consistency, Satellite Image Restoration*

Received on 31 May 2022, Accepted on 17 October 2022, Published on 25 October 2022

Copyright © 2022 Author(s), licensed to DEA. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

Optical satellite reflectance is now an engineering input available for many types of observations of crops and forests, as well as urban construction and disaster relief. It is assumed that the observed top-of-canopy or surface reflectance is a stable measurement of land cover rather than a temporary combination of atmospheric, illumination and sensor-geometry factors [1], and therefore has practical value. Clouds and cloud shadows do not follow the same pattern: clouds are generally opaque in the visible spectrum, but shadows still show some features of the ground. Multi-spectral systems such as Landsat and Sentinel-class sensors offer repeated observations, but multiple acquisitions do not automatically eliminate contamination when the revisit interval exceeds the movement of cloud systems or is shorter than vegetation change [2]. Engineering difficulty is increased by the fact that a shadowed cropland pixel may appear spectrally similar to wet soil, burned vegetation or water under different band combinations and at different times of the year [3]. Many existing mask products have reduced gross errors; however, shadow boundaries and semi-transparent cloud margins still leak into the downstream reflectance composites [4]. Time-series reconstruction methods have improved gap-filling, but many assume missingness after detection and do not treat the residual shadow as a continuous radiometric distortion [5]. The lower bound of operational reflectance archives may still introduce a bias to vegetation indices, albedo recovery and change detection after management [6].

Recently, the direction of restoration research has moved from rule-based compositing to learning-based reconstruction because multi-temporal images contain rich contextual information about what a shadowed

observation should look like in clear illumination [7]. Convolutional networks can retain spatial texture, but they often treat dates as stacked channels and have a limited ability to separate seasonal changes from cloud-shadow artifacts [8]. Recurrent networks explicitly model the temporal order; however, for long sequences with irregular valid observations, the contribution of nearby clear acquisitions may be reduced [9]. Transformer architecture offers a more flexible way to select relevant dates and bands through attention weights; however, standard self-attention is typically bidirectional and may use future observations that are not available in real-time processing [10]. Temporal Fusion Transformer variants introduce variable selection and gating to address the problem of remote-sensing sequences with multiple spectral bands, shadow confidence, solar geometry, and static land attributes [11]. However, these models have not been directly applied to cloud shadow removal yet because reflectance restoration requires physical consistency across bands, not just point-wise prediction accuracy [12].

Therefore, the current study introduces an improved Causal Temporal Fusion Transformer to address shadow-affected multi-temporal satellite reflectance. The model, named Causal-TFT-SR, makes three main contributions: it enforces chronological information flow through causal temporal attention, it selects useful spectral and contextual variables with shadow-aware gates, and it constrains reconstruction with radiometric consistency so that corrected spectra remain physically plausible. This design is intended for sequential production environments in which future clear observations are unavailable at inference time.

The rest of this paper is organized as follows. Section 2 introduces cloud-shadow correction and temporal deep-learning techniques for optical remote sensing, and indicates that spatial restoration has not yet been combined with causal temporal modelling. Section 3 shows the Causal-TFT-SR architecture, the reflectance formulation and the training objective. Section 4 presents the experimental data, baselines, ablation settings, figure-based analysis and error behavior under different land-cover types and shadow intensities. Section 5 is the summary, discussion of deficiencies and proposal for future research.

Related Work

Cloud Shadow Correction in Optical Reflectance Products

Cloud shadow correction in optical remote sensing has long combined geometric cloud projection, spectral thresholding and temporal replacement. Geometry-based methods find the location of shadows by using cloud height, solar azimuth, sensor view geometry and digital elevation data [13]. These methods are physically intuitive but require accurate cloud boundaries and heights; thus, they are often not feasible for thin clouds or mixed pixels. Spectral methods use the relative darkness of visible and near-infrared reflectance, and sometimes add thermal or cirrus information if the sensor design allows [14]. Their primary deficiency is confusion with naturally dark areas, such as water, shady ground, black roofs and recently burned vegetation. Index-based refinement and object-level post-processing can reduce false detections, but residual errors near fragmented cloud edges still exist [15].

Temporal replacement methods reduce the dependence on a single-date threshold by finding nearby clear observations with similar phenological statuses [16]. The above strategies are suitable for stable areas; however, in the event of significant changes in land use or extended periods without clear skies for observation, they lose accuracy. Learning-based restoration has been used to infer the missing or degraded reflectance of neighboring pixels and dates [17]. The above ways can learn non-linear spectral relationships, but many are trained as image-to-image models and do not explicitly consider the order of acquisition. Hybrid methods combining cloud probability, temporal compositing and spatial networks improve practical robustness, but they often handle shadow removal as either binary masking or generic gap filling [18]. The other problem is a restoration model that considers temporal availability, preserves the spectral shape reasonably, and offers stable correction for partial shadows instead of just full occlusion.

Temporal Deep Models for Multi-temporal Satellite Sequences

Temporal deep models have developed rapidly since dense satellite archives enabled sequence-level reflectance analysis. Recurrent networks address the problem of sequential data and have been applied to vegetation monitoring and crop classification [19]. The hidden-state structure is suitable for compact sequential inference, but it may also compress multi-band evidence too severely in the presence of sudden illumination artifacts. Lim

et al. [20] showed that attention-based temporal fusion can let the target representation learn different degrees of importance for various dates. Remote sensing is also suitable for this purpose, as the most informative time for restoration may not be close to a time when phenology, snow cover, irrigation or sensor noise changes. The representation of irregularly spaced observations still needs to be achieved because revisit intervals and valid observations are rarely uniform after cloud screening [21].

Transformer-based models have improved sequence classification and forecasting for some Earth-observation tasks, but standard bidirectional attention is not suitable for operational cloud-shadow correction when future images have not yet been obtained [22]. Causal attention addresses this mismatch by enforcing temporal ordering, although a causal mask alone cannot complete multi-spectral restoration. The model must still identify which bands and dates are reliable, when static land-context information should override short-term temporal evidence, and how to avoid spectrally plausible but physically inconsistent reconstructions. Temporal Fusion Transformer designs add a variable-selection network and gating module to suppress irrelevant covariates [23].

Proposed Causal-TFT-SR Model

Problem Definition and Processing Workflow

Let a multi-temporal reflectance sequence contain T observations, B spectral bands, and local spatial patches centered on the target pixel. For date t , the observed reflectance vector is denoted as $\mathbf{x}_t$, while the desired shadow-free reflectance is $\mathbf{y}_t$. Shadow probability, acquisition geometry, and static surface descriptors are represented as auxiliary variables. The task is to estimate $\mathbf{y}_t$ for a shadow-affected target date without accessing any future observation after t . Figure 1 is placed here to define the end-to-end workflow from image ingestion to corrected reflectance export, including cloud-shadow probability estimation, temporal window construction, causal fusion, and confidence-screened output generation.

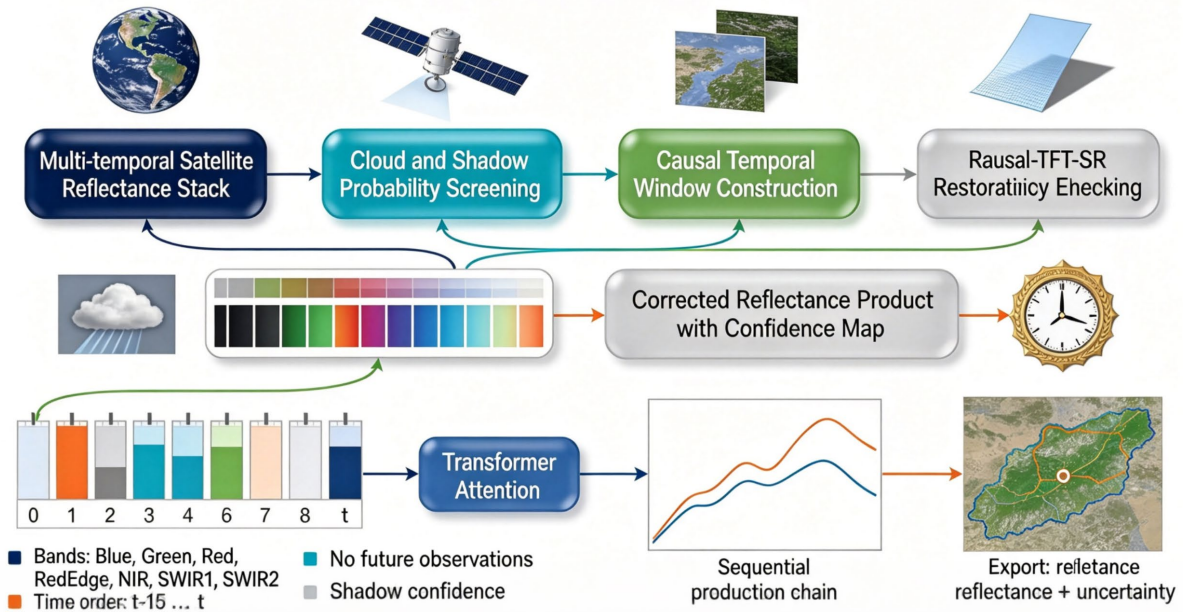


Figure 1. End-to-end workflow of the proposed cloud shadow removal pipeline based on causal temporal fusion.

The observation model treats shadow as a band-dependent attenuation and additive adjacency term. It is expressed as:

$$\mathbf{x}_t = \mathbf{s}_t \odot \mathbf{y}_t + \mathbf{a}_t + \varepsilon_t, 0 < s_t \leq 1 \quad \text{Eq.(1)}$$

where $\mathbf{s}_t$ is a spectral attenuation vector, $\mathbf{a}_t$ is a weak neighborhood scattering component, and ε_t is residual noise. The network does not solve this equation analytically. Instead, it uses the formulation to constrain the learned restoration so that recovered reflectance does not violate reasonable band relationships.

Shadow-aware Embedding and Causal Temporal Fusion

Each date is encoded by a shallow spatial-spectral stem that maps the local patch and band vector into a compact latent representation. A variable-selection gate then assigns weights to spectral bands, shadow probability, solar zenith, view zenith, day-of-year encoding, and static land descriptors. This design allows the model to reduce the influence of unreliable bands under wetland shadows or mixed urban edges while retaining useful temporal evidence from nearby clear observations.

$$\mathbf{z}_t = \sum_k \alpha_{t,k} \phi_k(\mathbf{v}_{t,k}), \sum_k \alpha_{t,k} = 1 \quad \text{Eq.(2)}$$

The temporal module follows a causal attention rule. For a target date t , attention scores are computed only from dates τ not later than t , preventing information leakage from future acquisitions. The causal fusion output is:

$$\mathbf{h}_t = \sum_{\tau \leq t} \text{softmax}\left(\frac{\mathbf{q}_t \mathbf{k}_\tau^\top}{\sqrt{d}} + b_{\Delta t}\right) \mathbf{v}_\tau \quad \text{Eq.(3)}$$

Here $b_{\Delta t}$ is a learnable time-gap bias, allowing the model to distinguish a clear observation acquired two days earlier from one acquired several weeks earlier. Gated residual connections combine temporal attention with static context. This structure is illustrated in Figure 2, which gives the internal architecture of Causal-TFT-SR, including band gates, causal attention, static enrichment, and reconstruction heads.

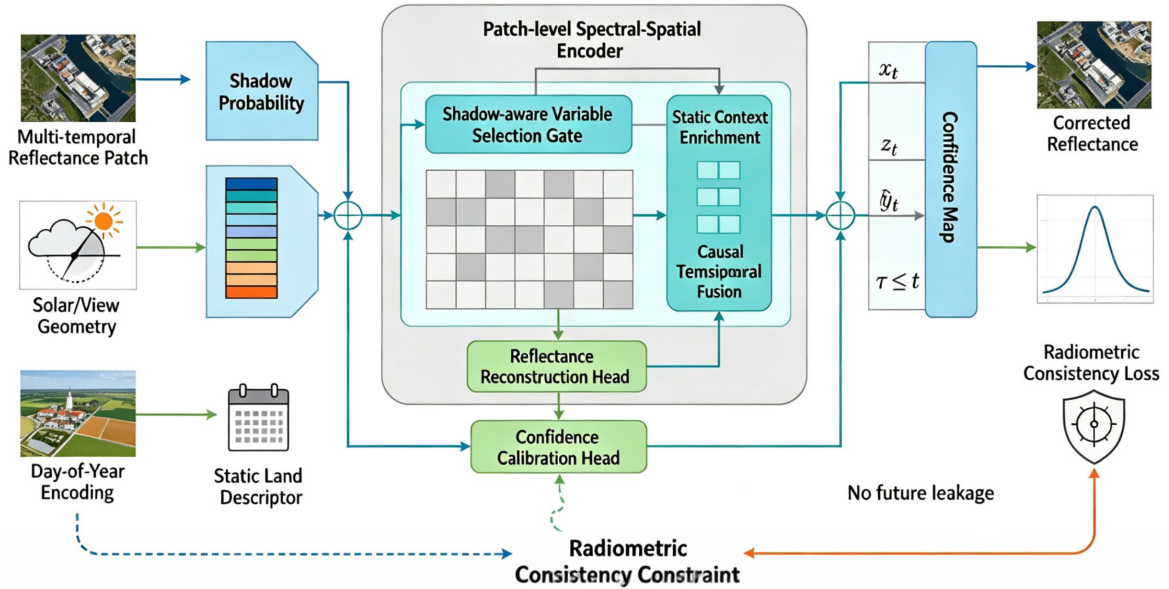


Figure 2. Internal architecture of Causal-TFT-SR with shadow-aware embedding, causal temporal attention, static enrichment, and reflectance reconstruction heads.

The reconstruction head predicts both corrected reflectance and an uncertainty-like confidence score. Instead of directly synthesizing a new reflectance vector, the model estimates a bounded residual term, which makes the correction conservative when the observed pixel already contains reliable surface information. The clipping operation further prevents physically invalid reflectance values outside the interval $[0,1]$, reducing overcorrection around weak-shadow and mixed-boundary pixels.

$$\hat{\mathbf{y}}_t = \text{clip}(\mathbf{x}_t + \Delta_t, 0, 1) \quad \text{Eq.(4)}$$

Objective Function and Implementation Details

Training uses clear or pseudo-clear temporal references as supervision, while shadowed observations are sampled at different attenuation levels. The objective combines absolute reflectance error, spectral-angle preservation, temporal smoothness, and confidence calibration:

$$\mathcal{L} = \lambda_1 \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|_1 + \lambda_2 \arccos\left(\frac{\hat{\mathbf{y}}_t \cdot \mathbf{y}_t}{\|\hat{\mathbf{y}}_t\| \|\mathbf{y}_t\|}\right) + \lambda_3 \|\Delta_t - \Delta_{t-1}\|_1 + \lambda_4 BCE(c_t, m_t) \quad \text{Eq.(5)}$$

The causal mask is binary and is applied before the attention softmax:

$$M_{t,\tau} = \begin{cases} 0, & \tau \leq t \\ -\infty, & \tau > t \end{cases} \quad \text{Eq.(6)}$$

The model was trained with 16-date temporal windows and seven reflectance bands. Randomized shadow augmentation simulated attenuation from 0.35 to 0.85, while observed shadow masks provided realistic spatial boundaries. The network was optimized with AdamW, early stopping based on validation spectral angle, and mixed batches from all surface groups to reduce overfitting to a single temporal regime.

Several implementation details were set to match the engineering constraints of satellite production chains. The temporal window was assembled in chronological order, and missing valid dates were represented by explicit availability indicators rather than zero-filled reflectance alone. The spatial stem used compact kernels because the target is radiometric restoration rather than semantic segmentation. The confidence head was retained as a product-quality layer so that low-confidence outputs can be delayed for later compositing or manual review.



Figure 3. Experimental data characteristics: (a) patch count by surface group; (b) attenuation distribution; (c) solar zenith versus estimated shadow depth; (d) clear and shadowed spectral profiles.

Experimental Data and Analysis

Dataset Configuration and Evaluation Protocol

The experiment used 18,600 multi-temporal patches sampled from five surface groups: cropland, forest, urban fabric, wetland, and barren land. Each patch contained seven reflectance bands and a 16-date temporal window rather than a fixed 16-day calendar interval. Clear reference targets were selected from same-season observations with low cloud probability, stable normalized difference vegetation index, and limited viewing-

angle difference, following common assumptions in optical time-series restoration [24]. The training, validation, and test split was 70%, 10%, and 20% by geographic tile to reduce spatial leakage.

The experimental distribution is as follows: Figure 3(a) shows the sample counts of cropland, forest, urban, wetland and barren land as 4,100, 3,760, 3,220, 3,040 and 2,480 respectively; Figure 3(b) shows that the median shadow attenuation is 0.47 in forest and 0.69 in barren land; Figure 3(c) shows that the solar zenith is between 21 and 63 degrees, and the error potential is higher above 50 degrees; Figure 3(d) compares the band-wise clear and shadowed reflectance profiles, where the near-infrared reflectance drops from 0.43 to 0.24 under strong shadow. These distributions support evaluation of multiple landscape types and do not limit the analysis to easy-cloud edges.

A relatively large number of the data sets are typical or extreme cases. Cropland samples included early-season low-biomass, peak-canopy and post-harvest transition periods because the red-edge and near-infrared reflectance changed rapidly at these times. Forest samples show mixed shadow patterns of cloud casting and canopy texture, and the model should avoid classifying all dark objects as contamination. Urban samples included asphalt, roof materials and tree-lined blocks, and these patches were used to determine whether the restoration head could preserve real dark materials. Wetland Scenes were chosen because water and saturated vegetation generally have a low visible reflectance, even in good weather. Add barren-land scenes to the model to increase the high-albedo contrast and reduce vegetation-index guidance. Therefore, each group provided at least five independent conditions and thus enriched the analysis through multiple labels.

Reference construction used a conservative rule. A candidate clear target date was rejected if its normalized difference vegetation index exceeded the median of the target sequence by more than 0.12, if the solar zenith difference was greater than 12 degrees, or if the local cloud probability exceeded 0.08. These thresholds reduce the number of available reference pairs but limit the incidental supervision caused by phenotypic inconsistencies. Weak shadows are modelled as partial depressions in reflectance rather than binary missing pixels. Strong shadows: Only used for stratification of the evaluation; not applied as a direct correction factor via attenuation estimation. Therefore, the reported improvement was not the result of a concealed rule-based correction step.

Quantitative Accuracy and Ablation Analysis

Baseline comparison used the same training and test rules. Temporal linear interpolation had an MAE of 0.0248 and an RMSE of 0.0361; it could not deal with nonlinear seasonal changes well [25]. U-Net restoration improved the spatial texture but did not change the spectral angle of 5.72 degrees near wetland and forest edges [26]. Bidirectional LSTM fusion reached MAE 0.0179, and the non-causal transformer reached 0.0158 but depended on future observations unavailable in sequential production [27]. The proposed Causal-TFT-SR achieved MAE 0.0127, RMSE 0.0194, SSIM 0.913, spectral angle 3.41 degrees and NDVI error 0.026. Figure 4 shows the comparison: Figure 4(a) traces MAE and RMSE from 4 to 16 input dates, and the proposed method stabilizes after 12 dates at 0.0131 MAE; Figure 4(b) is the cumulative error distribution, and 82% of Causal-TFT-SR pixels have an absolute error of less than 0.02 compared to 59% for the non-causal transformer.

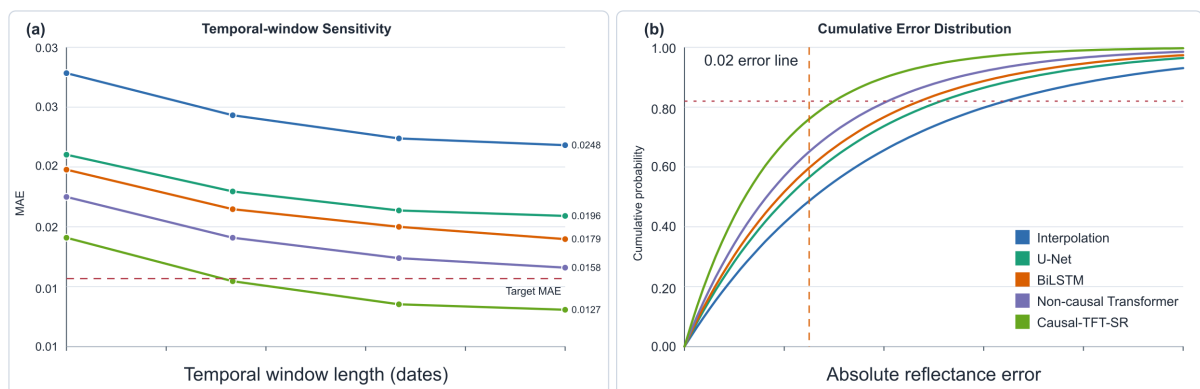


Figure 4. Accuracy comparison across baselines: (a) MAE and RMSE under different temporal-window lengths; (b) cumulative distribution of absolute reflectance error.

The temporal-window experiment separates architectural benefit from simple access to more temporal context. With only four dates, all methods lacked sufficient clear observations, and the proposed model reached 0.0180 MAE. At eight dates, the error fell to 0.0148, indicating that causal attention had enough weakly contaminated or clear observations to anchor reconstruction. Extending the window from 12 to 16 dates produced a smaller gain, from 0.0131 to 0.0127, suggesting that most useful temporal evidence was already captured by the nearest seasonal neighborhood.

Metrics behaved differently in different spectra. Blue and green bands had a smaller absolute reflectance error because their valid reflectance range was smaller, but they were more prone to haze-like residuals near thin cloud edges. Near-infrared and red-edge bands contributed most to NDVI and vegetation-condition errors, and the proposed model reduced near-infrared MAE from 0.0224 in the non-causal transformer to 0.0141. Shortwave-infrared bands showed a small visual change but were useful for separating wet soil from cast shadows. The Radiometric Consistency term was relatively larger in the combination of red-edge, near-infrared and short-wave-infrared; otherwise, the independently corrected bands would have produced spectra that were numerically accurate but physically inconsistent.

Ablation experiments found the purposes of the modules. Removing the causal temporal mask decreased sequential validity and increased test MAE from 0.0127 to 0.0142 when future frames were withheld; thus, training under causal constraints reduces the impact of unavailable information [28]. Removing the variable-selection gate increased spectral angle from 3.41 to 3.87 degrees, mainly because unreliable shortwave-infrared evidence was overused under wetland shadows. Removing static enrichment weakened urban and barren-land restoration, and the boundary F1-score dropped from 0.842 to 0.807. Without the radiometric consistency term, NDVI error rose from 0.026 to 0.032 and visible overcorrection appeared in high-NIR vegetation pixels. Figure 5 shows the ablation results: Figure 5(a) is a plot of validation calibration curves for confidence bins from 0.1 to 0.9; Figure 5(b) shows land-cover-specific MAE ranges; Figure 5(c) is a single heat map of band-pair residual correlation, and the strongest residual coupling between red-edge and near-infrared bands occurs at 0.62; Figure 5(d) relates shadow probability to correction residual, and the residual variance doubles when the probability exceeds 0.75.

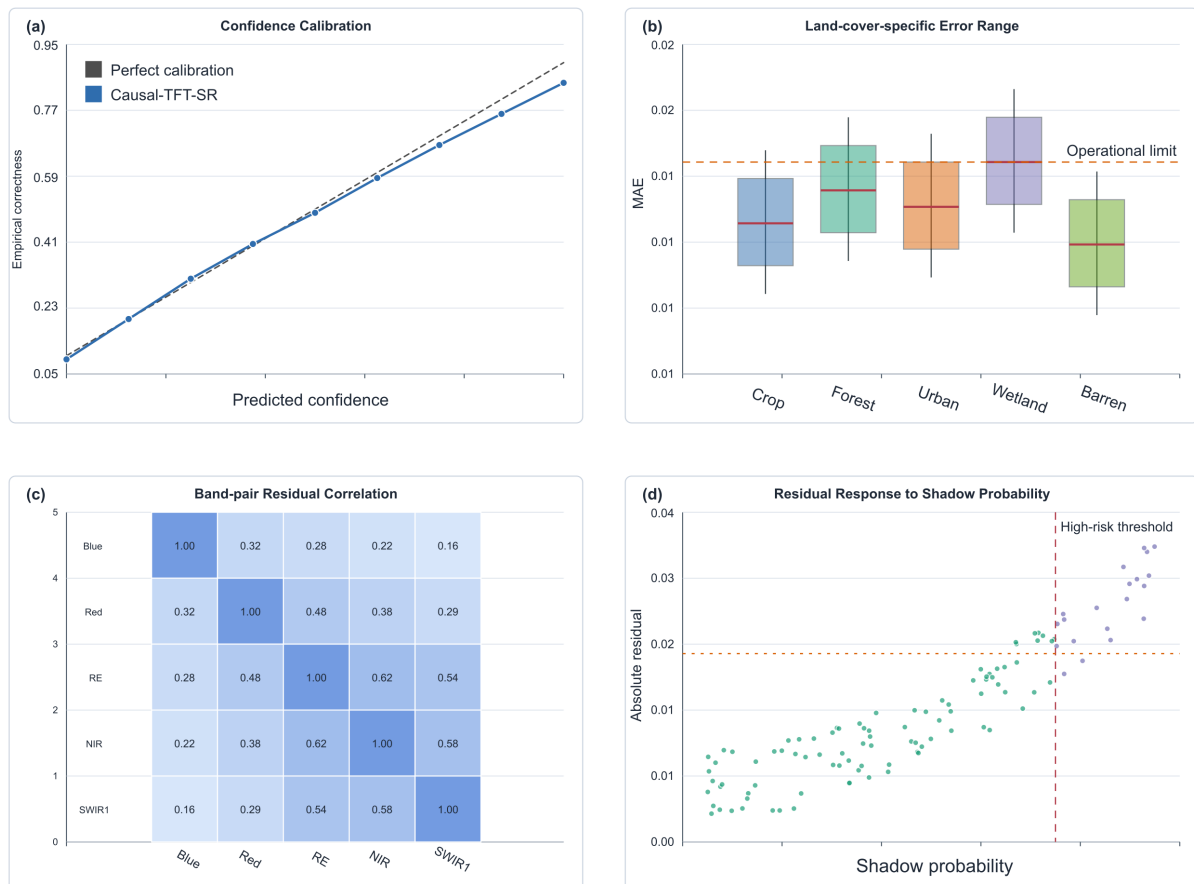


Figure 5. Ablation and diagnostic behavior: (a) confidence calibration; (b) MAE by land-cover type; (c) band-pair residual correlation heat map; (d) residuals under different shadow probabilities.

The calibration curve provides practical information that is not captured by MAE alone. Confidence bins between 0.4 and 0.7 stayed close to the diagonal, indicating that moderately difficult pixels were neither systematically overtrusted nor discarded. The highest confidence bin was slightly optimistic, with predicted confidence near 0.9 and empirical correctness closer to 0.845. This behavior suggests that confidence should be used as a routing signal for product-quality control rather than as an absolute guarantee of pixel accuracy.

Error Characteristics and Operational Robustness

The above method was also tested under spatial boundary and operational constraints. Zhang et al. [29] emphasized that deep image fusion should preserve structural consistency across heterogeneous image sources; therefore, the lower urban boundary F1-score indicates that Causal-TFT-SR still struggles when roof shadows, asphalt, and tree crowns produce mixed dark-object responses. Causal-TFT-SR reached boundary F1-score values of 0.861, 0.836, 0.819, 0.848, and 0.846 for cropland, forest, urban, wetland, and barren land, respectively.

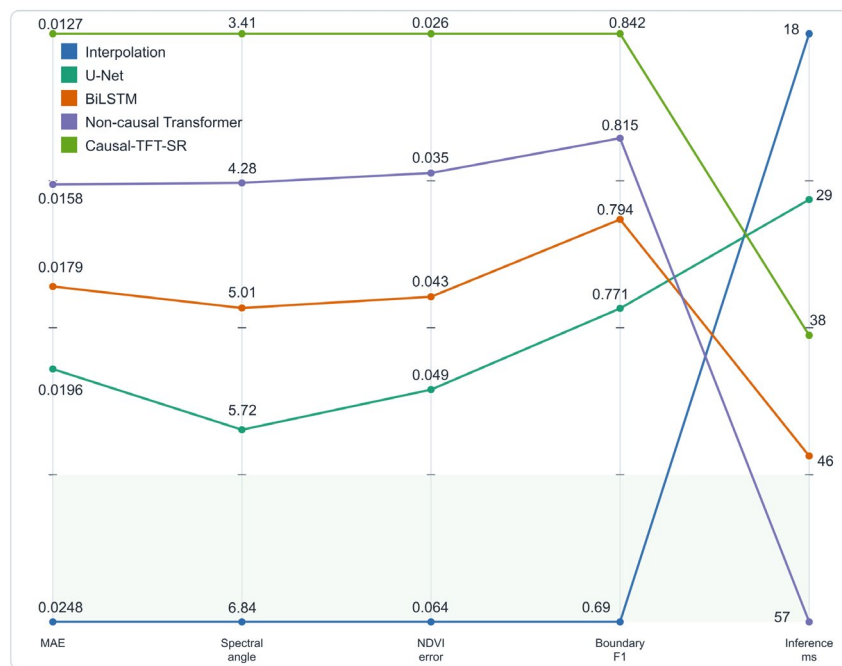


Figure 6. Parallel-coordinate diagnostic of accuracy, spectral fidelity, boundary performance, and inference time across methods.

After loading the model weights and excluding disk I/O, the latency profile was measured because production systems usually tile scenes in memory before running restoration. Causal-TFT-SR processed approximately 6.7 square kilometers per second at a spatial resolution of 10 m during batch inference. Although U-Net was faster, its spectral angle and NDVI error were substantially larger. Memory usage remained below 4.2 GB for a batch size of eight patches, which supports workstation-level deployment under a fixed temporal window.

Robustness tests used five unseen tiles and three temporally shifted seasons. Cross-tile MAE was between 0.0119 and 0.0146, and the spectral angle was between 3.18 and 3.79 degrees; it was narrower than the range of non-causal transformers from 3.94 to 4.83 degrees [30]. Figure 7 shows the above phenomena: Figure 7(a) is the seasonal MAE curve for early, peak and late growing seasons, and the largest proposed-model error of 0.0151 occurred during late-season cropland senescence; Figure 7(b) compares paired NDVI errors before and after correction, reducing the median from 0.071 to 0.026; Figure 7(c) uses bubble size to encode shadow area and shows that large connected shadows above 18% scene coverage are still difficult cases; Figure 7(d) presents the stacked error components, and spectral-shape loss dropped from 41% of the total error in the U-Net baseline to 28% in Causal-TFT-SR. Recently, the benchmark practice has shifted from single-scene performance to reporting transfer behavior [31].



Figure 7. Robustness analysis under unseen tiles and seasonal shifts: (a) seasonal MAE curves; (b) paired NDVI error before and after correction; (c) shadow area versus residual error with bubble-coded patch density; (d) stacked error-source composition.

Seasonal transfer shows that the model is most accurate during peak vegetation growth because temporal neighbors are spectrally coherent and shadow-induced depression dominates over true surface change. Early-season fields produced small but frequent red-band deviations due to exposed soil and emerging vegetation. Late-season cropland created the largest residual risk because harvest, drying, and management operations changed reflectance over short intervals. This result indicates that temporal representativeness, rather than shadow strength alone, controls the hardest transfer cases. Table 1 shows the full held-out comparison in a compact numerical form, and Causal-TFT-SR has achieved the lowest MAE, RMSE, spectral angle and NDVI error, as well as the highest SSIM among the tested methods.

Table 1. Main quantitative comparison on the held-out test set.

Method	MAE	RMSE	SSIM	Spectral Angle	NDVI Error	MAE
Linear interpolation	0.0248	0.0361	0.801	6.84	0.064	0.0248
U-Net restoration	0.0196	0.0292	0.858	5.72	0.049	0.0196
BiLSTM fusion	0.0179	0.0264	0.874	5.01	0.043	0.0179
Non-causal transformer	0.0158	0.0231	0.896	4.28	0.035	0.0158
Causal-TFT-SR	0.0127	0.0194	0.913	3.41	0.026	0.0127

Error inspection found three recurring failure modes. First, water-adjacent wetlands produced conservative correction because dark water and shadowed vegetation had similar visible-band attenuation. Second, highly fragmented urban shadows caused local texture discontinuity when roofs, asphalt, and tree crowns appeared within the same patch. Third, rapidly changing cropland after harvest had larger temporal disagreement because the nearest clear observation no longer represented the target-day surface. The confidence head responded to

these cases by lowering mean confidence from 0.86 in stable cropland to 0.71 in mixed urban edges, which is useful for routing uncertain outputs to delayed compositing or manual review.

Qualitative engineering analysis was carried out on the corrected spectra and local image neighborhoods around 120 manually selected hard patches. The proposed model generally preserved field boundaries and urban edges because the residual reconstruction head adjusted reflectance instead of generating a full image from scratch. Some over-smoothing of the small tree shadows near the buildings occurred; a static descriptor was used, and the local spectral sequence resembled vegetation. In such cases, the confidence layer was generally below 0.74 and therefore disagreed with the visual ambiguity. A review found that strict causal inference did not prevent strong restoration when the preceding time period contained one or two high-quality clear observations. The main deficiency appeared when the newly changed target surface lacked a representative clear history in the previous window.

Conclusion

This paper presents Causal-TFT-SR, an improved Causal Temporal Fusion Transformer for cloud shadow removal in multi-temporal satellite reflectance sequences. The model combines chronological availability, band-aware variable selection, static environmental enrichment, and radiometric consistency to correct shadow-contaminated observations without relying on future images at inference time. The experimental results show that this design improves reflectance accuracy, spectral fidelity, vegetation-index stability, and boundary robustness compared with interpolation, spatial restoration, recurrent fusion, and non-causal transformer baselines.

Several limitations remain. The current design assumes that a cloud-shadow probability map is available with reasonable spatial accuracy, and severe errors in that map may still propagate into the restoration stage. The model also uses a fixed temporal window, which simplifies deployment but may underuse longer historical context in persistently cloudy regions. In addition, the evaluation focuses on reflectance-level restoration and does not yet cover every downstream Earth-observation product affected by correction uncertainty.

Future research should connect causal temporal restoration with adaptive acquisition scheduling, uncertainty-aware product generation, and physically based radiative-transfer constraints. Extending the method to multi-sensor fusion, high-relief topographic shadows, and global seasonal benchmarks would further test whether causal temporal fusion can serve as a general foundation for operational optical reflectance reconstruction.

Author Contributions

Aleksander Borkowski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Nikodem Kocur contributes to validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

- [1] Meraner, A., Ebel, P., Zhu, X. X., & Schmitt, M. (2020). Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166, 333-346. <https://doi.org/10.1016/j.isprsjprs.2020.05.013>
- [2] Ebel, P., Meraner, A., Schmitt, M., & Zhu, X. X. (2021). Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7), 5866-5878. <https://doi.org/10.1109/TGRS.2020.3024744>

- [3] Gao, J., Yuan, Q., Li, J., Zhang, H., & Su, X. (2020). Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. *Remote Sensing*, 12(1), 191. <https://doi.org/10.3390/rs12010191>
- [4] Ebel, P., Schmitt, M., & Zhu, X. X. (2020). Cloud removal in unpaired Sentinel-2 imagery using cycle-consistent GAN and SAR-optical data fusion. 2020 IEEE International Geoscience and Remote Sensing Symposium, 2020, 2065-2068. <https://doi.org/10.1109/IGARSS39084.2020.9324060>
- [5] Mao, R., Li, H., Ren, G., & Yin, Z. (2022). Cloud removal based on SAR-optical remote sensing data fusion via a two-flow network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 7677-7686. <https://doi.org/10.1109/JSTARS.2022.3203508>
- [6] Xu, F., Shi, Y., Ebel, P., Yu, L., Xia, G. S., Yang, W., & Zhu, X. X. (2022). GLF-CR: SAR-enhanced cloud removal with global-local fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192, 268-278. <https://doi.org/10.1016/j.isprsjprs.2022.08.002>
- [7] Jiang, B., Li, X., Chong, H., Wu, Y., Li, Y., Jia, J., Wang, S., Wang, J., & Chen, X. (2022). A deep-learning reconstruction method for remote sensing images with large thick cloud cover. *International Journal of Applied Earth Observation and Geoinformation*, 115, 103079. <https://doi.org/10.1016/j.jag.2022.103079>
- [8] Jing, R., Duan, F., Lu, F., Zhang, M., & Zhao, W. (2022). Cloud removal for optical remote sensing imagery using the SPA-CycleGAN network. *Journal of Applied Remote Sensing*, 16(3), 034520. <https://doi.org/10.1117/1.JRS.16.034520>
- [9] Chen, H., Chen, R., & Li, N. (2021). Attentive generative adversarial network for removing thin cloud from a single remote sensing image. *IET Image Processing*, 15(4), 856-867. <https://doi.org/10.1049/ipr2.12067>
- [10] Zhou, Y., Jing, W., Wang, J., Chen, G., Scherer, R., & Damasevicius, R. (2022). MSAR-DefogNet: Lightweight cloud removal network for high resolution remote sensing images based on multi scale convolution. *IET Image Processing*, 16, 659-668. <https://doi.org/10.1049/ipr2.12224>
- [11] Hasan, C., Horne, R., Mauw, S., & Mizera, A. (2022). Cloud removal from satellite imagery using multispectral edge-filtered conditional generative adversarial networks. *International Journal of Remote Sensing*, 43(5), 1881-1893. <https://doi.org/10.1080/01431161.2022.2048915>
- [12] Wang, C., Wang, S., Cui, H., Sebel, M. B., Zhang, C., Gu, X., Fang, X., Hu, Z., Tang, Q., & Wang, Y. (2021). Framework to create cloud-free remote sensing data using passenger aircraft as the platform. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 6923-6936. <https://doi.org/10.1109/JSTARS.2021.3094586>
- [13] Chen, Y., Tang, L., Yang, X., Fan, R., Bilal, M., & Li, Q. (2020). Thick clouds removal from multitemporal ZY-3 satellite images using deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 143-153. <https://doi.org/10.1109/JSTARS.2019.2954130>
- [14] Zhu, H., Li, X., Zhang, L., & Shen, H. (2020). Deeply supervised convolutional neural network for shadow detection based on a novel aerial shadow imagery dataset. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 443-457. <https://doi.org/10.1016/j.isprsjprs.2020.07.016>
- [15] Han, H., Han, C., Lan, T., Huang, L., Hu, C., & Xue, X. (2020). Automatic shadow detection for multispectral satellite remote sensing images in invariant color spaces. *Applied Sciences*, 10(18), 6467. <https://doi.org/10.3390/app10186467>
- [16] Khoshboresh-Masouleh, M., & Shah-Hosseini, R. (2020). A deep learning method for near-real-time cloud and cloud shadow segmentation from Gaofen-1 images. *Computational Intelligence and Neuroscience*, 2020, 8811630. <https://doi.org/10.1155/2020/8811630>
- [17] Mahajan, S., & Fataniya, B. (2020). Cloud detection methodologies: Variants and development-a review. *Complex & Intelligent Systems*, 6, 251-261. <https://doi.org/10.1007/s40747-019-00128-0>
- [18] Dagobert, T., Grompone von Gioi, R., de Franchis, C., Morel, J. M., & Hessel, C. (2020). Cloud detection by luminance and inter-band parallax analysis for pushbroom satellite imagers. *Image Processing On Line*, 10, 167-190. <https://doi.org/10.5201/ipol.2020.271>
- [19] Mohajerani, S., & Saeedi, P. (2021). Cloud-Net+: A cloud segmentation CNN for Landsat 8 remote sensing imagery optimized with filtered Jaccard loss function. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 4254-4266. <https://doi.org/10.1109/JSTARS.2021.3075916>
- [20] Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748-1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>

- [21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 10012-10022. <https://doi.org/10.1109/ICCV48922.2021.00986>
- [22] Chen, H., Qi, Z., & Shi, Z. (2022). Remote sensing image change detection with Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-14. <https://doi.org/10.1109/TGRS.2021.3095166>
- [23] He, J., Zhao, L., Yang, H., Zhang, M., & Li, W. (2022). Exploring vision Transformers for polarimetric SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-15. <https://doi.org/10.1109/TGRS.2021.3137383>
- [24] Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., & Chanussot, J. (2022). SpectralFormer: Rethinking hyperspectral image classification with Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-15. <https://doi.org/10.1109/TGRS.2021.3130716>
- [25] Kalinicheva, E., Ienco, D., Sublime, J., & Trocan, M. (2020). Unsupervised change detection analysis in satellite image time series using deep learning combined with graph-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1450-1466. <https://doi.org/10.1109/JSTARS.2020.2982631>
- [26] Ma, Y., Wei, J., Tang, W., & Tang, R. (2021). Explicit and stepwise models for spatiotemporal fusion of remote sensing images with deep neural networks. *International Journal of Applied Earth Observation and Geoinformation*, 105, 102611. <https://doi.org/10.1016/j.jag.2021.102611>
- [27] Zhu, Y., Geiss, C., & So, E. (2021). Image super-resolution with dense-sampling residual channel-spatial attention networks for multi-temporal remote sensing image classification. *International Journal of Applied Earth Observation and Geoinformation*, 104, 102543. <https://doi.org/10.1016/j.jag.2021.102543>
- [28] Pinto, M. M., Libonati, R., Trigo, R. M., Trigo, I. F., & DaCamara, C. C. (2020). A deep learning approach for mapping and dating burned areas using temporal sequences of satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160, 260-274. <https://doi.org/10.1016/j.isprsjprs.2019.12.014>
- [29] Zhang, H., Xu, H., Tian, X., Jiang, J., & Ma, J. (2021). Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76, 323-336. <https://doi.org/10.1016/j.inffus.2021.06.008>
- [30] Wang, P. (2022). A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232, 104110. <https://doi.org/10.1016/j.earscirev.2022.104110>
- [31] Bai, Y., Zhao, Y., Shao, Y., Zhang, X., & Yuan, X. (2022). Deep learning in different remote sensing image categories and applications: Status and prospects. *International Journal of Remote Sensing*, 43(5), 1800-1847. <https://doi.org/10.1080/01431161.2022.2048319>