Transformer-Based SAR-Optical Fusion Network for Cloud Removal in Satellite Imagery

Thanh-Thoai Le[1], Trong-An Bui[2,*], Hai Tran[1], Yang-Lang Chang[2], Pei-Jun Lee[3]

***Abstract*—Cloud cover significantly limits the usability of optical satellite imagery in various remote sensing applications. This paper presents a Transformer-based fusion framework for cloud removal in optical satellite imagery, leveraging synthetic aperture radar (SAR) data. The proposed approach adopts a Swin Transformer-based self-attention mechanism to capture spatial dependencies and an Attention Difference Mechanism to refine SAR-optical feature integration. By selectively enhancing cloud-occluded regions, the method ensures high-fidelity image reconstruction while preserving both spectral and structural details. Experimental results show that the proposed approach achieves a PSNR of 31.32 dB and SSIM of 0.93 for partially cloud-covered images, and a PSNR of 29.83 dB and SSIM of 0.86 for fully cloud-covered images.**

***Index Terms*—Cloud Removal, SAR-Optical Fusion, Swin Transformer, Attention Mechanism.**

I. Introduction

Optical satellite imagery plays a crucial role in environmental monitoring, agriculture, and urban planning [1]. However, cloud cover often obscures critical information, significantly reducing image quality and usability [2].

Recent advancements in deep learning, particularly transformer architectures, have demonstrated promising results in cloud removal. With self-attention mechanisms, transformers effectively capture spatial and temporal dependencies, making them well-suited for processing satellite imagery [3].

This study proposes SOTF-Net (SAR-Optical Transformer Fusion Network), a transformer-based fusion framework that integrates synthetic aperture radar (SAR) data to mitigate cloud cover in optical satellite imagery. By enhancing image reconstruction quality, this approach contributes to advancements in remote sensing and environmental monitoring.

II. Proposed Methodology

The proposed **SAR-Optical Transformer Fusion Network (SOTF-Net)** reconstructs cloud-free optical images utilizing SAR data and Transformer-based fusion, as illustrated in Figure 1. The methodology consists of three key components: Feature extraction, Transformer-based Fusion, and Cloud Removal.

Feature extraction is performed separately for SAR and optical images using *CNN-based encoders*. The *SAR encoder* captures structural details from radar backscatter, while the *optical encoder* preserves spectral information. To enhance feature representation, a *channel-spatial transform* (1) is applied, adjusting the spatial resolution and channel depth.

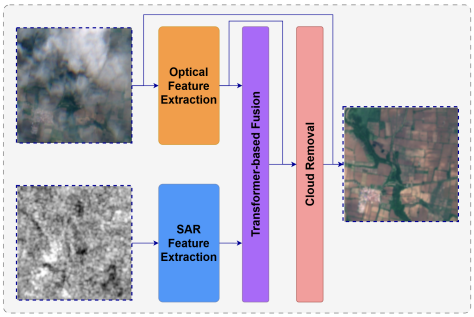

Fig. 1: Proposed SOTF-Net Architecture

$$G_m = \Psi\big(F_m, (C_m, H, W)\big) \tag{1}$$

$$\rightarrow \big(C_m^* = C_m \cdot k^2,\ H^* = \frac{H}{k},\ W^* = \frac{W}{k}\big) \tag{2}$$

where: F_m is the input SAR or optical image; C_m, H, W denote the original number of channels, height, and width; C_m^* increases to $C_m \cdot k^2$, expanding the channel space for *multi-scale feature extraction*; H^* and W^* are reduced by k, downsampling spatial dimensions while preserving important details.

The CNN-based encoders extract (3) refined features from the transformed SAR Z_{sar} and optical Z_{opt} feature maps.

$$Z_{\text{opt}} = \mathcal{E}_{\text{opt}}(G_{\text{opt}}), \quad Z_{\text{sar}} = \mathcal{E}_{\text{sar}}(G_{\text{sar}}) \tag{3}$$

where:

$$\mathcal{E}_m(X) = \rho\Big(V_m^{(2)} * \rho\big(V_m^{(1)} * X + d_m^{(1)}\big) + d_m^{(2)}\Big) \tag{4}$$

$\mathcal{E}_m$ represents the independent SAR and optical encoders; $V_m^{(1)}, V_m^{(2)}$ are convolutional filters that capture hierarchical features; $d_m^{(1)}, d_m^{(2)}$ are bias terms; $\rho(\cdot)$ is the ReLU activation function; and $*$ denotes the convolution operation.

The extracted features are then passed to the **Transformer-based Fusion module**. To effectively integrate SAR and optical features, the Transformer-based Fusion Module employs a Swin Transformer [4], which captures both local and global dependencies, following the flow illustrated in Figure 2. First, the optical and SAR feature maps are divided into non-overlapping windows of size $M \times M$ using a partition function $\mathcal{P}(\cdot)$ (5)

$$Z_{\text{opt}}^{(w)} = \mathcal{P}(Z_{\text{opt}}, M), \quad Z_{\text{sar}}^{(w)} = \mathcal{P}(Z_{\text{sar}}, M) \tag{5}$$

This work was supported by the National Science and Technology Council, Taiwan, under Grant NSTC-113-2222-E-027-011.

*Corresponding author: Trong-An Bui, trongan93@ntut.edu.tw.

[1]Ho Chi Minh University of Education, HCMC, Vietnam.

[2]National Taipei University of Technology, Taipei, Taiwan (R.O.C).

[3]National Taiwan University of Science and Technology, Taipei, Taiwan (R.O.C).

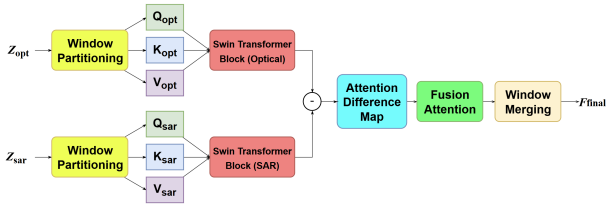

Fig. 2: Transformer-based Fusion Flowchart

Self-attention is computed independently for optical and SAR features to capture distinct modality-specific representations (6) and (7).

$$\mathcal{A}_{\text{self, opt}}(Z_{\text{opt}}^{(w)}) = \sigma\left(\frac{Q_{\text{opt}}^{(w)} K_{\text{opt}}^{(w)T}}{\sqrt{d}}\right) V_{\text{opt}}^{(w)} \tag{6}$$

$$\mathcal{A}_{\text{self, sar}}(Z_{\text{sar}}^{(w)}) = \sigma\left(\frac{Q_{\text{sar}}^{(w)} K_{\text{sar}}^{(w)T}}{\sqrt{d}}\right) V_{\text{sar}}^{(w)} \tag{7}$$

where: Q, K, V are the feature representations that are projected into query (Q), key (K), and value (V) spaces using learned weight matrices; $\sigma(\cdot)$ is softmax normalization; $\sqrt{d}$ is ormalizes values based on feature dimensionality d to maintain numerical stability.

The *Attention Difference Map* $\Delta_{\text{att}}^{(w)}$ is then calculated to highlight the discrepancies between the SAR and the optical attention features (8).

$$\Delta_{\text{att}}^{(w)} = \mathcal{A}_{\text{self, sar}}(Z_{\text{sar}}^{(w)}) - \mathcal{A}_{\text{self, opt}}(Z_{\text{opt}}^{(w)}) \tag{8}$$

This difference map is used to guide the Fusion Attention Mechanism, allowing SAR to contribute to cloud-occluded regions (9).

$$F_{\text{fusion}} = \lambda \cdot \sigma\left(\frac{Q_{\text{opt}}^{(w)} (\Delta_{\text{att}}^{(w)} K_{\text{sar}}^{(w)T})}{\sqrt{d}}\right) V_{\text{sar}}^{(w)} + (1-\lambda) \cdot \mathcal{A}_{\text{self, opt}}(Z_{\text{opt}}^{(w)}) \tag{9}$$

where λ is fusion factor, balances the contribution of SAR-based attention and original optical self-attention, allowing adaptive feature integration.

Finally, windowed representations (10) are merged back into the full-resolution feature space.

$$F_{\text{fused}} = \mathcal{M}(F_{\text{fusion}}) \tag{10}$$

The fused feature representation is then passed to the **Cloud Removal Module**, which reconstructs a high-quality, cloud-free optical image by merging fused features with the pre-extracted optical features (11).

$$F_{\text{rec}} = F_{\text{fused}} + Z_{\text{opt}} \tag{11}$$

A convolutional network further refines spatial details (12).

$$F_{\text{restored}} = \sigma\left(\text{Conv}_{3\times3}\left(\text{Conv}_{3\times3}\left(F_{\text{rec}}\right)\right)\right) \tag{12}$$

where: $\text{Conv}_{3\times3}$is a standard convolution operation applied twice to progressively refine spatial details and remove noise; $\sigma(\cdot)$ is ReLU activation function.

To ensure *spectral consistency*, a residual connection is applied (13).

$$X_{\text{final}} = F_{\text{restored}} + X_{\text{opt,cloudy}} \tag{13}$$

where: $X_{\text{opt,cloudy}}$ is original cloudy optical image.

This process *effectively removes cloud cover* while preserving *structural and spectral fidelity*, ensuring an accurate and visually appealing reconstructed optical image.

III. Experiments

The proposed method is evaluated on the SEN12MS-CR dataset [5], which provides co-registered SAR and optical image pairs with cloud contamination. Figure 3 presents visual comparisons of cloud removal results under both partial and full cloud cover conditions. The method effectively reconstructs fine spatial details while preserving spectral consistency across varying levels of occlusion.

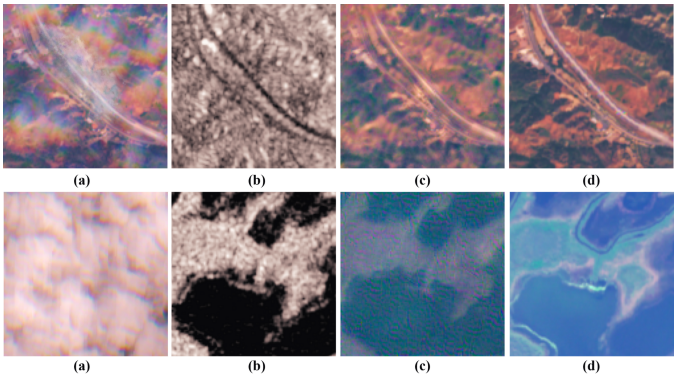

Fig. 3: Visual comparison of cloud removal results under different cloud coverage levels. Top row: Partial Cloud Cover (≈50%); Bottom row: Full Cloud Cover (100%). Each row shows (a) Cloudy Input, (b) SAR Data, (c) Predicted Output, and (d) Ground Truth. **PSNR↑/ SSIM↑** (Partial): **31.32/ 0.93**; (Full): **29.83/ 0.86**.

IV. Conclusion

This paper presents SOTF-Net, a Transformer-based SAR-optical fusion framework for cloud removal in satellite imagery. By combining Swin Transformer blocks with an Attention Difference Mechanism, the model effectively restores cloud-free images while preserving spectral and structural features. Future work will explore multi-temporal fusion strategies and improve generalization across diverse geographic regions.

References

[1] T.-A. Bui, P.-J. Lee, J. Liobe, V. Barzdenas, and D. Udris, "Region of interest-focused dynamic enhancement (roide) for satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025.

[2] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020.

[3] R. Singh, M. Biswas, and M. Pal, "Enhanced cloud detection in Sentinel-2 imagery using K-means clustering embedded transformer-inspired models," *Journal of Applied Remote Sensing*, vol. 18, no. 3, p. 034516, 2024.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[5] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5866–5878, 2021.