Title

Zineddine Bettouche, Ahmed Attia
Deggendorf Institute of Technology
Dieter-Görlitz-Platz 1, 94469 Deggendorf
{zineddine.bettouche, ahmed.attia}@th-deg.de

***Abstract*—**
***Index Terms*—**

I. Introduction

Continuous monitoring of the Earth's surface is important for applications such as crop monitoring, land cover classification, and environmental management. Optical satellite imagery provides rich spectral information, but cloud contamination limits its usability. Synthetic Aperture Radar (SAR) can penetrate clouds and provide structural information under all weather conditions. However, SAR and optical images capture different physical properties, making the reconstruction of cloud-free optical imagery from SAR a nontrivial problem.

Earlier approaches used signal processing techniques such as sparse representation and multi-temporal dictionary learning. These methods are limited under dense clouds or highly dynamic surfaces. Deep learning approaches, including conditional GANs, diffusion models, and transformer-based networks, improve reconstruction quality but may still struggle to accurately recover fine spatial details across all regions.

A central challenge is preserving local spatial structure while generating consistent optical reflectance. Pixel- or patch-based models can maintain fine details but may produce artifacts across larger areas, whereas global models may smooth out important spatial features. The heterogeneity between SAR and optical signals further complicates the task, requiring architectures capable of modeling complex cross-modal relationships.

In this work, we propose the Bidirectional Mamba Bridged UNet (BiMBU), an image-to-image translation architecture for SAR-to-optical cloud removal. BiMBU combines a U-Net encoder-decoder with a Bidirectional Mamba bottleneck, which processes spatial features in both directions to enhance feature representation while preserving high-resolution spatial details.

II. Related Work

Cloud contamination in optical remote sensing imagery hinders continuous Earth observation, limiting applications such as crop monitoring and land cover classification. Synthetic Aperture Radar (SAR) systems can penetrate clouds, enabling data acquisition in all weather conditions. This capability makes SAR data valuable for filling gaps in optical time series, driving research into SAR–optical fusion and SAR-to-optical image translation for cloud removal.

Early methods leveraged traditional signal processing techniques. Huang et al. [1] introduced sparse representation-based cloud removal using SAR data, which Xu et al. [2] extended via multi-temporal dictionary learning. These approaches, however, struggled under heavy cloud cover or highly dynamic surface changes.

Deep learning transformed the field. The foundational GAN framework was introduced by Goodfellow et al. [3], and Mirza & Osindero [4] extended it to conditional GANs (cGANs), ideal for image-to-image tasks like SAR-to-optical translation. Enomoto et al. [5] applied cGANs for cloud removal using NIR input, though dense clouds remained problematic. To address this, Grohnfeldt et al. [6] proposed SAR-Opt-cGAN to fuse Sentinel-1 SAR and Sentinel-2 optical data; Bermudez et al. [7] explored cGAN-based SAR-to-optical synthesis for crop classification. The SEN1-2 [8] and SEN12MS [9] datasets were pivotal for training deep models.

Advancements continued with Fuentes Reyes et al. [10] on cGAN optimization, Wang et al. [11] with supervised CycleGANs for translation, Meraner et al. [12] introducing DSen2-CR networks, Abady et al. [13] leveraging ProGANs, and Pan [14] with spatial-attention models. Gao et al. [15] developed fusion-based GAN approaches for high-resolution images.

Recent models show increasing complexity: Naderi Darbaghshahi et al. [16] proposed a two-GAN model with DRIBs, Ebel et al. [17] introduced UnCRtainTS with uncertainty prediction, Kwak & Park [18] proposed MTcGANs, and Liu et al. [19] developed S2MS-GAN.

Diffusion models have found their way into this field: Bai et al. [20] proposed a conditional diffusion model; Bai et al. [21] extended it with color supervision; Zou et al. [22] introduced the efficient DiffCR framework.

Vision Transformers (ViT) also made inroads: Dosovitskiy et al. [23] introduced the original ViT, while Park et al. [24] integrated multiscale ViT blocks into a cGAN for SAR-optical tasks. However, ViTs are computationally intensive.

Alternatives emerged via SSMs: Gu & Dao [25] introduced Mamba—an efficient, linear-complexity sequence model. U-Mamba [26] adapted this into a U-Net architecture, and Swin-UMamba [27] enhanced it further using ImageNet pretraining. Swin-UNet [28] remains a benchmark transformer-based segmentation model.

III. METHODOLOGY

A. Problem Formulation

Our objective is to generate Sentinel-2 optical images from Sentinel-1 SAR observations. This task can be formulated as an image-to-image translation problem. Let $S \in \mathbb{R}^{H\times W\times 2}$ denote a Sentinel-1 image of spatial resolution $H \times W$ with two polarization channels (VV and VH), and let $O \in \mathbb{R}^{H\times W\times 12}$ denote the corresponding Sentinel-2 image with 12 spectral bands. Given S, the goal is to predict O such that the generated image preserves the spatial structure from the SAR input while faithfully reconstructing the spectral information of the optical counterpart.

Formally, we aim to learn a mapping function

$$f_\theta : \mathbb{R}^{H\times W\times 2} \to \mathbb{R}^{H\times W\times 12},$$

parameterized by θ, that minimizes the reconstruction error between the predicted optical image $\hat{O} = f_\theta(S)$ and the ground truth O. The optimization objective is expressed as

$$\min_\theta \; L(f_\theta(S), O),$$

where L is the loss function. In this work, we adopt the Mean Absolute Error (MAE) loss:

$$L(\hat{O}, O) = \frac{1}{H \cdot W \cdot 12} \sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{12} \left|\hat{O}_{i,j,c} - O_{i,j,c}\right|.$$

While Sentinel-1 captures structural and backscatter information that is robust to weather and lighting conditions, Sentinel-2 provides rich spectral information sensitive to surface reflectance properties. The main challenge of this task is the heterogeneity between SAR and optical data, as they capture different physical properties of the Earth's surface. The model must therefore learn a consistent mapping from radar backscatter to optical reflectance.

B. Proposed Architecture: Bidirectional Mamba Bridged UNet (BiMBU)

We propose the Bidirectional Mamba Bridged UNet (BiMBU), an image-to-image translation architecture that integrates a convolutional U-Net structure with a Bidirectional Mamba (BiM) bottleneck for enhanced spatial feature extraction (cf. 1). Given an input tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times C_{in}}$, where spatial dimensions $H, W \in \{64, 128, 256\}$ and the number of input channels $C_{in} \in \{1, 2\}$, the model predicts an output tensor $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times W\times C_{out}}$ with $C_{out} = 12$. The model is composed of three main parts: a convolutional encoder, a Mamba-based bottleneck, and a convolutional decoder.

1) Convolutional Encoder: The encoder follows a standard U-Net downsampling path, consisting of four stages. Each stage comprises a convolutional block followed by a 2×2 max-pooling operation for downsampling. The convolutional block contains two 2D convolution layers, each followed by a Group Normalization and a GELU activation function. At each downsampling stage, the number of feature channels is doubled, allowing the model to learn increasingly abstract and high-level features while reducing spatial dimensions. High-resolution feature maps from each stage are preserved for later use in the decoder via skip connections.

2) Bidirectional Mamba Bottleneck: The bottleneck of the network serves to process the high-level features from the encoder. The input feature map $\mathbf{X}_{\text{enc}} \in \mathbb{R}^{B\times C\times H'\times W'}$ is flattened into a sequence of spatial tokens $\mathbf{X}_{\text{flat}} \in \mathbb{R}^{B\times (H'W')\times C}$. This sequence is then processed by the Bidirectional Mamba block, which contains two Mamba SSMs:

- A **forward Mamba** processes the sequence as is.
- A **backward Mamba** processes a reversed version of the sequence.

The outputs from both Mamba models are concatenated and fused using a linear layer, allowing the model to capture long-range spatial dependencies from two directions. The Mamba SSM's ability to selectively retain and propagate information is particularly effective for modeling complex spatial relationships. The resulting fused features are then reshaped back to their original spatial dimensions, $\mathbb{R}^{B\times C\times H'\times W'}$.

3) Convolutional Decoder: The decoder reconstructs the high-resolution output image from the bottleneck features. It consists of three upsampling stages. In each stage, a 2×2 transposed convolution doubles the spatial dimensions of the feature map. The upsampled features are then concatenated with the corresponding high-resolution feature maps from the encoder path via skip connections. This concatenation is processed by a convolutional block identical to those in the encoder. This process allows the network to combine high-level, abstract features with fine-grained spatial details to produce a precise output.

4) Output Head: Finally, a 1×1 convolutional layer is applied to the output of the final decoder stage. This layer maps the feature channels to the desired number of output channels, $C_{out} = 12$, producing the final output tensor $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times W\times 12}$.

C. Baseline Models

We compare our proposed model BiMBU against the following baselines:

- **Pix2Pix GAN** [29]: A conditional adversarial network for paired image-to-image translation, employing a U-Net generator and a PatchGAN discriminator. It combines reconstruction loss (L1) with adversarial loss to encourage both fidelity and realism in generated images.
- **Swin-UNet** [28]: A purely Transformer-based U-shaped encoder–decoder architecture using hierarchical Swin Transformer blocks with shifted windows and skip-connections, enabling multi-scale context aggregation across spatial dimensions.
- **U-Mamba** [26]: A hybrid encoder–decoder network integrating convolutional residual blocks with Mamba-based state-space modeling to capture both local features and long-range dependencies. Originally developed for biomedical image segmentation, U-Mamba's architecture is adapted here for cross-modal remote-sensing translation.

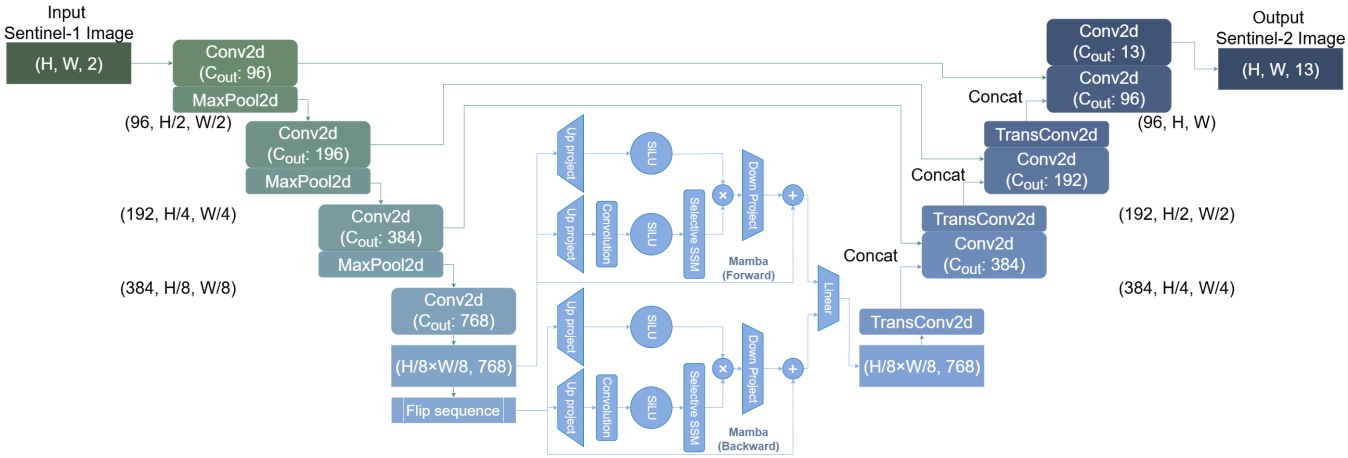

Figure 1. Bidirectional Mamba Bridged UNet (BiMBU) Architecture

IV. Datasets & Experimental Setup

A. Copernicus Sentinels: 1 and 2

We use paired observations from the Copernicus Sentinel-1 and Sentinel-2 missions. Sentinel-1 provides C-band SAR data in Interferometric Wide Swath (IW) mode with dual polarisation (VV+VH), while Sentinel-2 delivers multispectral optical imagery with 13 bands at spatial resolutions of 10–60 m. These complementary modalities form the basis for SAR-to-optical translation.

The dataset was collected from [*source*] over [*region*] during [*timeframe*]. Preprocessing included radiometric calibration, speckle filtering, and coregistration of Sentinel-1 with Sentinel-2, followed by resampling to a common resolution, cloud masking, and band-wise normalisation. In total, [*number*] paired samples were extracted and split into [*train/val/test split*].

B. Implementation and Training Configuration

We implement the BiMBU model in PyTorch, leveraging GPU-optimized operations for its `Mamba` and `Conv2D` modules to ensure computational efficiency. For reproducibility, we release the complete source code for our implementation in a public repository [**?**]. For training and inference, we use an AI-server with a single NVIDIA A100 80GB GPU with 64 CPU cores and 512 GB RAM, using CUDA 12.4 and PyTorch 2.6.0+cu124. BiMBU and baseline models were trained with a batch size of 128 for up to 150 epochs, using early stopping with patience of 15 and saving the best model based on validation loss. We use Adam optimizer with a learning rate of 10^{-4} and `ReduceLROnPlateau` scheduler (patience: 7, factor: 0.5).

C. Evaluation Metrics

We employ a set of metrics to evaluate reconstruction accuracy from multiple perspectives.

Pixel-level fidelity is assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which quantify absolute and squared deviations, respectively. Peak Signal-to-Noise Ratio (PSNR) measures the ratio between signal and reconstruction error, while the Structural Similarity Index (SSIM) captures perceived structural similarity in spatial patterns.

To evaluate spectral consistency, Spectral Angle Mapper (SAM) is used to measure the angular deviation between predicted and reference spectral signatures, and Spectral Information Divergence (SID) to compare their probabilistic distributions.

We assess downstream utility by computing errors on remote sensing indices derived from reconstructed bands: the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Water Index (NDWI), and the Normalized Burn Ratio (NBR).

Per-band metrics (e.g., SSIM, MAE, RMSE per Sentinel-2 band) are reported to highlight reconstruction quality at the individual spectral channel level. All metrics are computed after reversing the normalization to the original physical scale.

V. Results and Analysis

A. Quantitative Results

As shown in Table I, BiMBU consistently achieves the best performance across all crop sizes and input-channel configurations. For instance, at 256×256 with VV+VH input, BiMBU reaches an MAE of 49.61, RMSE of 210.80, PSNR of 39.78 dB, and SSIM of 0.9651, surpassing all baselines. At smaller crops, the gains are even more pronounced: with 64×64 inputs, BiMBU improves SSIM to 0.9812 and PSNR to 44.14 dB, while keeping MAE and RMSE lower than competing models. These results indicate that BiMBU preserves both pixel-level accuracy and structural similarity better than the baselines across all tested scales. Figure 2 shows 3 example visualizations of model outputs. From left to right: target Sentinel-2 image, BiMBU, GAN, SwinUnet, and UMamba.

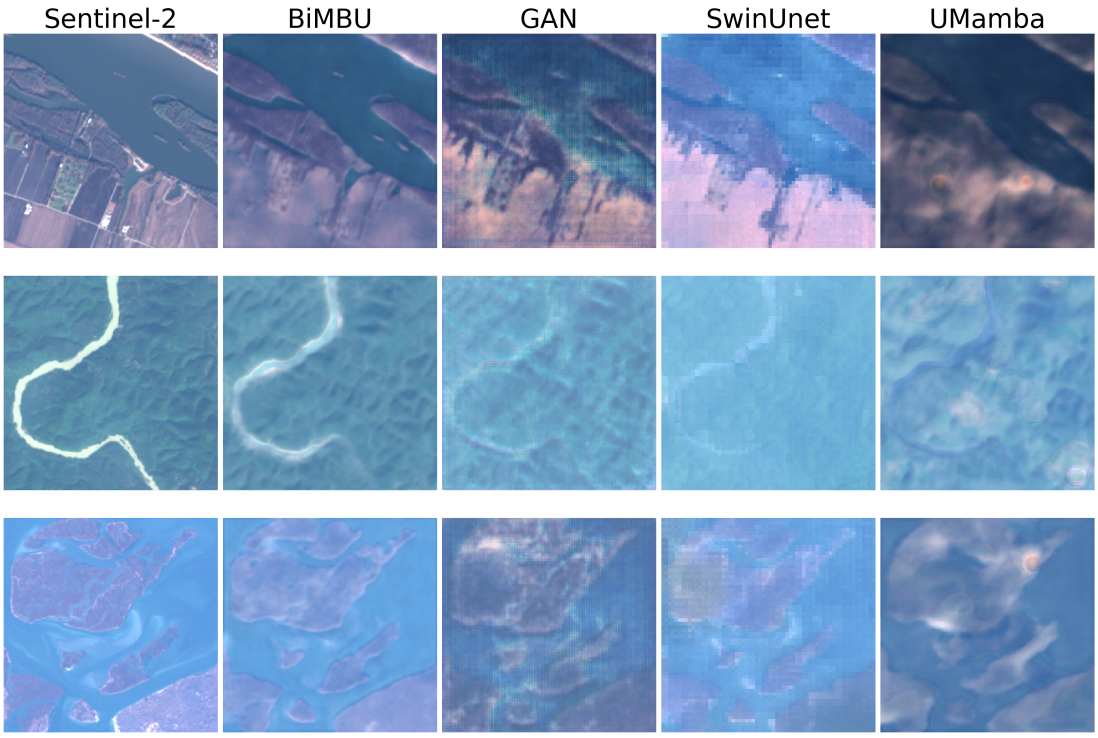

Figure 2. Example visualizations of model outputs. From left to right: target Sentinel-2 image, BiMBU, GAN, SwinUnet, and UMamba.

B. Spectral Shape and Index-based Evaluation

Table II reports spectral-shape metrics and index-based errors. BiMBU obtains the lowest SAM (4.938) and SID (0.0169), showing closer agreement to the reference spectral signatures. For vegetation and water indices, BiMBU also yields the lowest errors: NDVI MAE of 0.0624 and RMSE of 0.0911, NDWI MAE of 0.0565 and RMSE of 0.0844, and NBR MAE of 0.0875 and RMSE of 0.1212. These values confirm that the reconstructed bands are suitable for downstream applications such as vegetation monitoring and water detection.

C. Per-Band Metrics

Figure 3 presents per-band SSIM values. BiMBU consistently outperforms the baselines across most Sentinel-2 bands. The largest margins appear in the near-infrared (B8) and shortwave infrared (B11, B12) bands, where BiMBU maintains SSIM above 0.96 while others fall below 0.95. These results highlight BiMBU's ability to reconstruct bands with higher structural fidelity.

D. Computational Complexity

Table III summarizes model size and computational cost. BiMBU has 26.2M parameters and an inference time of 9.5 ms, larger than UMamba (10.1M, 8.3 ms) but smaller in multiply–accumulate operations (94.7 G vs. 123.3 G). Memory usage is higher (397.9 MB), reflecting the model's expanded capacity. These numbers show that BiMBU achieves accuracy improvements while maintaining tractable computational requirements.

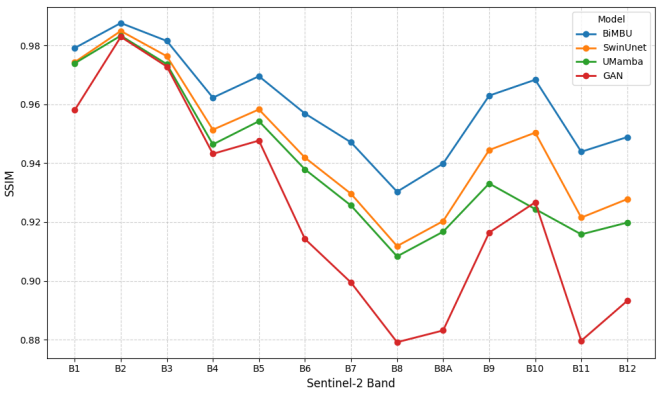

Figure 3. Per-band performance comparison of the BiMBU, SwinUnet, UMamba, and GAN models using the Structural Similarity Index (SSIM). The plot illustrates how the SSIM score for each model varies across the 13 different Sentinel-2 spectral bands.

VI. Conclusion

References

[1] B. Huang, Y. Li, X. Han, Y. Cui, W. Li, and R. Li, "Cloud removal from optical satellite imagery with sar imagery using sparse representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1046–1050, 2015.

[2] M. Xu, X. Jia, M. Pickering, and A. J. Plaza, "Cloud removal based on sparse representation via multitemporal dictionary learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2998–3006, 2016.

[3] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

Table I
QUANTITATIVE RESULTS UNDER DIFFERENT CROPPING SIZES AND INPUT CHANNEL SELECTIONS.

Model	MAE ↓	RMSE ↓	PSNR (dB) ↑	SSIM ↑
Crop: 256×256, Channels: VV+VH				
GAN	75.37	324.72	35.82	0.9338
SwinUnet	60.09	253.32	37.97	0.9529
UMamba	67.03	288.88	36.65	0.9484
BiMBU	**49.61**	**210.80**	**39.78**	**0.9651**
Crop: 256×256, Channels: VV only				
GAN	79.32	337.56	35.66	0.9336
SwinUnet	93.65	393.64	34.23	0.9281
UMamba	71.59	308.51	36.35	0.9466
BiMBU	**59.99**	**247.50**	**38.26**	**0.9591**
Crop: 256×256, Channels: VH only				
GAN	78.56	339.31	35.33	0.9328
SwinUnet	82.72	348.25	35.05	0.9345
UMamba	76.55	328.45	35.60	0.9400
BiMBU	**44.42**	**191.26**	**40.50**	**0.9673**
Crop: 128×128, Channels: VV+VH				
GAN	69.57	295.12	37.12	0.9481
SwinUnet	45.90	197.50	40.36	0.9652
UMamba	35.47	159.47	42.61	0.9749
BiMBU	**31.05**	**137.48**	**43.85**	**0.9791**
Crop: 128×128, Channels: VV only				
GAN	70.44	297.78	37.15	0.9481
SwinUnet	90.35	379.52	34.60	0.9312
UMamba	44.37	198.13	40.82	0.9688
BiMBU	**39.66**	**178.07**	**41.63**	**0.9720**
Crop: 128×128, Channels: VH only				
GAN	70.24	298.88	37.23	0.9498
SwinUnet	63.62	272.05	37.41	0.9498
UMamba	43.88	193.92	40.89	0.9684
BiMBU	**33.96**	**151.92**	**42.96**	**0.9757**
Crop: 64×64, Channels: VV+VH				
GAN	68.06	291.62	37.54	0.9545
SwinUnet	41.58	181.62	41.11	0.9687
UMamba	35.14	158.47	42.67	0.9757
BiMBU	**31.41**	**139.95**	**43.75**	**0.9797**
Crop: 64×64, Channels: VV only				
GAN	70.48	300.88	37.10	0.9490
SwinUnet	47.65	203.04	40.15	0.9648
UMamba	33.60	149.58	42.96	0.9749
BiMBU	**30.09**	**134.05**	**44.14**	**0.9812**
Crop: 64×64, Channels: VH only				
GAN	69.78	298.66	37.25	0.9491
SwinUnet	45.00	194.12	40.54	0.9662
UMamba	38.83	175.52	41.92	0.9739
BiMBU	**32.12**	**143.12**	**43.58**	**0.9787**

[4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[5] R. Enomoto *et al.*, "Image restoration from satellite cloud contamination using multispectral conditional gan," *Remote Sensing (likely)*, 2017.

[6] C. Grohnfeldt, M. Schmitt, and X. X. Zhu, "A conditional generative adversarial network to fuse sar and multispectral optical data for cloud removal from sentinel-2 images," in *ISPRS TC III Mid-term Symposium*, 2018.

[7] J. Bermudez, P. Happ, A. Boulch *et al.*, "Synthesis of multispectral optical images from sar/optical multitemporal data using conditional gans," in *IGARSS*, 2018.

[8] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The sen1-2 dataset for deep learning in sar-optical data fusion," in *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-1, 2018, pp. 141–146.

[9] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," in *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, 2019, pp. 153–160.

[10] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "Sar-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits," *Remote Sensing*, vol. 11, no. 17, p. 2067, 2019.

[11] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu, "Sar-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129 136–129 149, 2019.

[12] 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.

[13] L. Abady *et al.*, "Gan generation of synthetic multispectral satellite images," 2020, online; ResearchGate.

[14] B. Pan *et al.*, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," 2020.

[15] F. Gao *et al.*, "Cloud removal with fusion of high-resolution optical and sar images using generative adversarial networks," *Remote Sensing*, vol. 12, no. 1, p. 191, 2020.

[16] F. Naderi Darbaghshahi and M. R. Mohammadi, "Cloud removal in remote sensing images using generative adversarial networks with dilated residual inception blocks," in *IGARSS*, 2021.

[17] P. Ebel *et al.*, "Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series," *arXiv preprint*, 2022.

[18] H. Kwak and S. Park, "Assessing the potential of multi-temporal conditional gans in sar-to-optical image translation for early-stage crop monitoring," *Remote Sensing*, vol. 16, no. 7, p. 1199, 2024.

[19] J. Liu *et al.*, "High-resolution sar-to-multispectral image translation based on s2ms-gan," *Remote Sensing*, vol. 16, no. 21, p. 4045, 2024.

[20] W. Bai *et al.*, "Conditional diffusion for sar to optical image translation," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[21] ——, "Sar to optical image translation with color supervised diffusion model," 2024.

[22] H. Zou *et al.*, "Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images," 2023.

[23] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint*, 2020.

[24] S. Park *et al.*, "Sar-to-optical image translation using vision transformer-based cgan," in *IGARSS*, 2025.

[25] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint*, 2023.

[26] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint*, 2024.

[27] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, Y. Yu, Y. Liang, G. Shi, S. Zhang, H. Zheng, and S. Wang, "Swin-umamba: Mamba-based unet with imagenet-based pretraining," *arXiv preprint*, 2024.

[28] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: *Unet*-like pure transformer for medical image segmentation," pp. 205–218, 2023.

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

Table II
SPECTRAL SHAPE CONSISTENCY (SAM, SID) AND INDEX-BASED EVALUATION (MAE AND RMSE FOR NDVI, NDWI, AND NBR) ACROSS MODELS. ↓ INDICATES LOWER IS BETTER.

Model	**SAM ↓**	**SID ↓**	**NDVI MAE ↓**	**NDVI RMSE ↓**	**NDWI MAE ↓**	**NDWI RMSE ↓**	**NBR MAE ↓**	**NBR RMSE ↓**
GAN	8.274	0.0464	0.1090	0.1694	0.0971	0.1666	0.1147	0.1553
SwinUnet	6.048	0.0222	0.0731	0.1071	0.0652	0.1004	0.0961	0.1351
UMamba	6.993	0.0297	0.0804	0.1134	0.0735	0.1074	0.1101	0.1534
BiMBU	**4.938**	**0.0169**	**0.0609**	**0.0911**	**0.0542**	**0.0844**	**0.0821**	**0.1212**

Table III
MODEL COMPLEXITY ANALYSIS

Model	**Params (M)**	**Size (MB)**	**Inference (ms)**	**GPU Mem (MB)**	**MACs (G)**
GAN	54.43	217.73	2.10	346.16	19.48
SwinUnet	27.15	108.61	8.93	177.44	7.78
UMamba	10.13	40.53	7.29	186.05	123.27
BiMBU	26.19	104.77	9.51	397.87	94.72