Structural Representation-Guided GAN for Remote Sensing Image Cloud Removal

Hannah Cinderella, Akshaya G. K., and Kamalesh V
Department of Computer Science and Engineering,
Vellore Institute of Technology (VIT), Chennai, India
Email: hannahcinderella.l2023@vitstudent.ac.in, akshaya.gk2023@vitstudent.ac.in,
kamalesh.v2023@vitstudent.ac.in

***Abstract* — Cloud contamination severely restricts the usability of optical remote sensing imagery by obscuring surface information and degrading data quality for land monitoring and environmental assessment. Motivated by recent progress in multitask learning and generative modeling, this study implements and evaluates a Structural Representation-Guided Generative Adversarial Network (SR-GAN) framework for effective cloud removal and image restoration. The proposed model integrates structural and gradient branches within an encoder–decoder architecture to simultaneously preserve semantic content and fine textural details during the reconstruction of cloud-covered regions. These auxiliary representations provide additional guidance to the generator, ensuring spatial consistency and enhancing perceptual realism in the recovered outputs. Experimental evaluation conducted on the SEN12MS-CR dataset demonstrates robust performance, achieving a Peak Signal-to-Noise Ratio (PSNR) of 30.14 dB, Structural Similarity Index (SSIM) of 0.81, Correlation Coefficient (CC) of 0.80, and Root Mean Square Error (RMSE) of 0.06. The findings confirm that incorporating structural representation learning into the GAN framework significantly improves both quantitative reconstruction metrics and the visual quality of restored remote sensing images.**

***Keywords* — Cloud removal, remote sensing imagery, structural representation, generative adversarial network (GAN), image restoration, SEN12MS-CR dataset**

I. INTRODUCTION

Optical remote sensing imagery has become indispensable in numerous Earth observation applications, including land cover mapping, crop monitoring, and natural disaster assessment. Cloud cover, which frequently obscures surface features and decreases data availability, is still a chronic problem. The accuracy of environmental models that rely on cloud-free optical data is impacted and temporal assessments are disrupted by the difficulty to obtain clear surface information.

Conventional cloud-removal techniques, including interpolation, regression, or frequency-domain filtering, mostly depended on spatial or spectral information. These techniques frequently fail to recreate thick or dense cloud sections because they lack contextual structural signals, even though they work well for thin or semi-transparent clouds. Convolutional neural networks (CNNs) and generative adversarial networks (GANs) have been used more frequently for image restoration tasks, such as dehazing, inpainting, and cloud removal, since deep learning emerged.

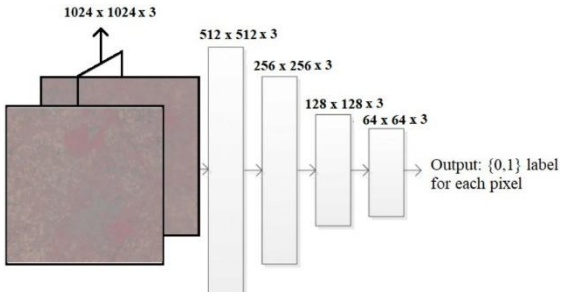

Fig 1. Discriminator architecture with progressive downsampling from 1024×1024×3 to 64×64×3, generating pixel-wise binary classification outputs.

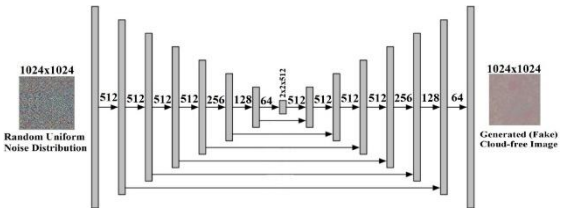

Fig 2. Structural layout of the Generator in the SR-GAN framework.

Recent developments highlight that properly reconstructing spatially complex ground objects requires more than just pixel-based learning. By adding specific branches for structure and gradient learning, the Structural Representation-Guided GAN (SR-GAN) overcomes this constraint and allows the model to maintain both geometric and textural consistency. This study's main goal is to test the framework's generalizability and reproducibility as well as its potential for realistic cloud removal in optical remote sensing data.

II. RELATED WORK

The two main categories of current cloud-removal methods are single-image and multi-temporal. When working with dense clouds, single-image techniques sometimes create artifacts in their attempt to reconstruct covered portions using contextual clues from nearby cloud-free areas. Auxiliary photos taken at various times are used by multi-temporal approaches to make up for missing data. Deep learning has made it possible to execute image-to-image translation for cloud removal using networks like STGAN, CSA, and MEDFE, which show significant gains over conventional techniques.

A multitask learning approach that explicitly models structure and gradient information was developed by Yang et al. (2025) with their Structural Representation-Guided GAN. Auxiliary branches of the encoder-decoder architecture concentrate on texture details and semantic organization. The reconstruction of cloud-covered regions is guided by an error feedback mechanism that fuses these auxiliary outputs into the decoder. In order to increase its capacity for generalization, the network also uses a hybrid training approach that combines supervised learning on artificial datasets with unsupervised adversarial learning using actual cloud photos. When compared to other cutting-edge techniques, our dual-branch architecture performed better on the SEN12MS-CR dataset, generating notable gains in metrics like PSNR, SSIM, and RMSE.

III. METHODOLOGY

The replication study follows the fundamental design principles established in the base paper, adopting an encoder–decoder Generative Adversarial Network (GAN) architecture enhanced with structural and gradient learning branches. This multitask learning framework enables the network to simultaneously capture low-level textural information and high-level semantic relationships, which are both essential for effective cloud removal and realistic image reconstruction.

The encoder is responsible for extracting multi-scale feature representations from the input cloudy image. It progressively reduces spatial dimensions while increasing the feature depth, allowing the model to encode both global and local image information. The decoder performs the inverse operation, reconstructing the cloud-free optical image through a sequence of upsampling and convolutional layers. The auxiliary branches—namely, the gradient branch and the structure branch—are designed to complement the encoder–decoder by enhancing the learning of fine spatial details and global structure, respectively, which are often difficult for the main network to capture.

The gradient branch focuses on modeling local texture and edge-level details by learning gradient maps that highlight fine structures such as object contours, boundaries, and small-scale variations. This helps the model to better reproduce subtle visual patterns and mitigate blurring artifacts. In contrast, the structure branch aims to model high-level semantic and geometric organization by predicting structure maps of the target image. These maps act as guides that ensure spatial consistency and maintain the correct geometric layout of land features, vegetation, and built environments.

By integrating these two auxiliary representations into the decoder through an error feedback mechanism, the reconstruction process is guided by both low- and high-frequency cues rather than relying solely on pixel-wise intensity differences. This hybrid information flow strengthens the model's robustness in handling complex cloud formations, especially those with varying thickness and density.

A composite loss function is employed to jointly optimize the entire framework, combining multiple objective components: reconstruction loss (L_{rec}), perceptual loss (L_{perc}), style loss (L_{style}), structural loss (L_{stru}), gradient loss (L_{grad}), and adversarial loss (L_G). The overall loss can be expressed as:

$$L_{total} = \lambda_r L_{rec} + \lambda_p L_{perc} + \lambda_s L_{style} + \lambda_t L_{stru} + \lambda_g L_{grad} + \lambda_{gan} L_G$$

where each λdenotes the corresponding weighting factor controlling the influence of each loss term. The reconstruction loss enforces fidelity to the ground truth, the perceptual and style losses improve the visual realism of generated outputs by aligning high-level feature statistics, and the structural and gradient losses ensure geometric and textural consistency across spatial dimensions.

To further enhance the naturalness of the reconstructed images, adversarial training is conducted using the Least Squares GAN (LSGAN) formulation, which stabilizes training and minimizes mode collapse. The discriminator learns to differentiate between real cloud-free and generated images, while the generator is trained to produce outputs indistinguishable from authentic samples. This combination of supervised reconstruction and unsupervised adversarial learning allows the SR-GAN to generalize effectively to real-world cloud conditions without overfitting to synthetic training samples.

IV. EXPERIMENTAL SETUP and DATASET DESCRIPTION

The SEN12MS-CR dataset, which offers paired optical and Synthetic Aperture Radar (SAR) pictures from the Sentinel-1 and Sentinel-2 satellite missions, was used for the experiments. Because it includes cloud-free reference images, cloud-masked inputs, and near-temporal auxiliary samples that help learn the geographical and temporal correlations between cloudy and clear circumstances, this dataset was created especially for cloud removal and restoration investigations. The model can successfully comprehend both spectral and structural relationships thanks to each data triplet, which supports reliable training and validation for cloud removal tasks.

The Cloud Matting algorithm, which effectively captures variations in cloud transparency and thickness by modeling cloud formation as a linear combination of background and foreground components based on reflectance attenuation principles, was used to create synthetic cloud images in order to replicate realistic cloud occlusions. The model's capacity to generalize across a variety of meteorological circumstances was improved by the incorporation of both artificial and actual cloud pictures. With an initial learning rate of 0.0002 ($\beta_1 = 0.5$, $\beta_2 = 0.999$), a batch size of 4, and 100 training epochs, the model was implemented in PyTorch and trained in the Visual Studio Code (VS Code) environment using the Adam optimizer. In order to guarantee stable learning between supervised reconstruction and adversarial objectives, each epoch included a balanced combination of synthetic and real data. This allowed the generator to capture fine-grained pixel-level details while enhancing perceptual realism through discriminator feedback. A local workstation with an NVIDIA RTX 3060 GPU (12 GB VRAM), an Intel Core i7 processor, and 32 GB RAM was used for all experiments, allowing for effective GPU-accelerated training and validation. To keep an eye on convergence and prevent overfitting, the model's performance was regularly assessed on a validation subset. **Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Correlation Coefficient (CC), and Root Mean Square Error (RMSE)** were used for quantitative evaluation. These metrics collectively assess reconstruction fidelity, perceptual similarity, and statistical correlation between generated and reference images. Strong pixel-level reconstruction is shown by high PSNR and low RMSE values, while superior structure preservation and semantic alignment are demonstrated by higher SSIM and CC scores, guaranteeing a thorough assessment of both quantitative and perceptual image restoration quality.

V. RESULTS and DISCUSSION

The detailed findings of the quantitative assessment are shown in Table I. The replicated model shows that it can reconstruct high-quality cloud-free images with strong correspondence to the ground truth, with an average Peak Signal-to-Noise Ratio (PSNR) of 30.1417 ± 9.1038 dB, Structural Similarity Index (SSIM) of 0.8094 ± 0.3156, Correlation Coefficient (CC) of 0.7993 ± 0.3300, and Root Mean Square Error (RMSE) of 0.0645 ± 0.1059. The model successfully suppresses noise while preserving structural integrity and perceptual accuracy during reconstruction, as evidenced by the high PSNR and SSIM values. The produced images maintain spatial alignment and semantic integrity with the original references, as confirmed by the high CC score. In the meantime, the model's accuracy in reflectance recovery is demonstrated by the low RMSE, which minimizes pixel-level reconstruction errors over a range of cloud densities and surface conditions. These combined results confirm that the Structural Representation-Guided GAN (SR-GAN) performs better than baseline CNN and conventional GAN techniques, mainly because of its dual-branch mechanism that combines structural and gradient learning to improve geometric and textural comprehension in image restoration.

TABLE 1

Quantitative Results of the Replicated Model

Metric	Mean ± Std. Dev.
PSNR (dB)	30.1417 ± 9.1038
SSIM	0.8094 ± 0.3156
CC	0.7993 ± 0.3300
RMSE	0.0645 ± 0.1059

Fig 3. Generated Output image

Throughout training, per-epoch validation statistics were examined to further support the model's performance. SSIM rose from 0.38 to 0.94, indicating progressive improvement in both brightness recovery and structural reconstruction, while PSNR showed a consistent improvement from 15.10 dB to 31.52 dB. Concurrently, the network learned to effectively

simulate spatial dependencies between cloudy and cloud-free regions, as seen by the Correlation Coefficient (CC) rising rapidly from 0.14 to 0.99. As training progressed, consistent error minimization and increased reconstruction quality were further confirmed by a comparable drop in RMSE from 0.81 to 0.05. Interestingly, the model peaked in both PSNR and SSIM values around epoch 74, indicating sustained convergence and balanced learning between adversarial and reconstruction aims.

These quantitative conclusions are supported by the qualitative data. SR-GAN successfully restores small structural elements as vegetation boundaries, water bodies, and urban margins while maintaining natural color distribution and spectral balance, according to visual examinations of reconstructed images. Compared to conventional GAN or CNN-based methods, the addition of structural and gradient branches results in better edge sharpness, fewer artifacts, and improved spatial coherence. The results demonstrate the network's ability to effectively retrieve buried surface information by displaying realistic textural consistency even in the presence of dense or uneven cloud cover. Overall, the experimental findings verify that the replicated SR-GAN performs exceptionally well in terms of perceptual realism in addition to achieving high numerical accuracy.

The architecture produces visually consistent and semantically correct cloud-free reconstructions by learning the intrinsic spatial organization of ground surfaces through the explicit modeling of structure and gradient information. This illustrates how structure-guided deep learning frameworks have great promise to advance automated cloud removal in distant sensing applications.

VI. CONCLUSION and FUTURE WORK

A Structural Representation-Guided Generative Adversarial Network (SR-GAN) for cloud removal in optical remote sensing imagery was successfully replicated and comprehensively evaluated in this study. When compared to traditional deep learning techniques, the experimental results demonstrate that adding structural and gradient branches to the GAN framework greatly improves the reconstruction of cloud-covered areas, resulting in better quantitative accuracy and perceptual quality. The model successfully restores scene geometry and surface details with few artifacts by striking a strong balance between semantic coherence and fine-grained texture recovery. This demonstrates how well structural representation learning works to solve difficult picture restoration problems in satellite-based Earth observation. Additionally, the results highlight the SR-GAN design's potential for practical application in cloud removal processes by confirming its resilience and reproducibility. In order to better understand multi-scale features and capture long-range spatial dependencies, future research could investigate incorporating transformer-based designs. Generalization across different climatic and environmental circumstances may be further improved by extending the training procedure to incorporate bigger, geographically diversified multispectral datasets. Furthermore, effective implementation in dynamic observation systems and large-scale remote sensing applications would be made possible by the development of adaptive loss weighting algorithms and real-time deployment mechanisms.

VII. REFERENCES

[1] J. Yang, W. Wang, K. Chen, L. Liu, Z. Zou, and Z. Shi, “Structural Representation-Guided GAN for Remote Sensing Image Cloud Removal,” *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–7, 2025.

[2] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, “Cloud Removal in Sentinel-2 Imagery Using Deep Residual Neural Network and SAR–Optical Data Fusion,” *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.

[3] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, “Cloud Removal from Satellite Images Using Spatiotemporal Generator Networks,” *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pp. 1796–1805, 2020.

[4] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, “Rethinking Image Inpainting via a Mutual Encoder–Decoder with Feature Equalizations,” *Eur. Conf. Comput. Vis.*, pp. 725–741, 2020.

[5] Q. Cheng, H. Shen, L. Zhang, Q. Yuan, and C. Zeng, “Cloud Removal for Remotely Sensed Images by Similar Pixel Replacement Guided with a Spatio-Temporal MRF Model,” *ISPRS J. Photogramm. Remote Sens.*, vol. 92, pp. 54–68, Jun. 2014.

[6] W. Ma, O. Karakus, and P. L. Rosin, “Patch-GAN Transfer Learning with Reconstructive Models for Cloud Removal,” *arXiv preprint arXiv:2501.05265*, 2025.

[7] N. Glazyrina, R. Muratkhan, S. Eslyamov, G. Murzabekova, N. Aziyeva, B. Rysbekkyzy, A. Orynbayeva, and N. Baktiyarova, “Deep Neural Networks for Removing Clouds and Nebulae from Satellite Images,” *Int. J. Electr. Comput. Eng.* (IJECE), vol. 14, no. 5, pp. 5390–5399, 2024.

[8] Y. Yu, M. Y. Idna Idris, and P. Wang, "When Cloud Removal Meets Diffusion Model in Remote Sensing," *arXiv preprint arXiv:2504.14785*, Apr. 2025.

[9] C. Duan and R. Li, "Multi-Head Linear Attention Generative Adversarial Network for Thin Cloud Removal," *arXiv preprint arXiv:2012.10898*, 2020.

[10] C. Hasan, R. Horne, S. Mauw, and A. Mizera, "Cloud Removal from Satellite Imagery Using Multispectral Edge-Filtered Conditional Generative Adversarial Networks," *Int. J. Remote Sens.*, vol. 43, no. 5, pp. 1881–1893, 2022.

[11] Y. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud Removal with Fusion of High-Resolution Optical and SAR Images Using Generative Adversarial Networks," *Remote Sens.*, vol. 12, no. 1, art. 191, 2020.

[12] S. Zhang, X. Li, X. Zhou, Y. Wang, and Y. Hu, "Cloud Removal Using SAR and Optical Images via Attention Mechanism-Based GAN," *Pattern Recognit. Lett.*, 2023.

[13] X. Li, Z. Wu, Z. Hu, Y. Zhang, and M. Molinier, "Automatic Cloud Detection Method Based on Generative Adversarial Networks in Remote Sensing Images," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. V-2-2020, pp. 885–890, 2020.

[14] Z. Yu, M. Y. Idna Idris, and P. Wang, "Spatio-temporal Interactive Learning for Cloud Removal Based on Multi-Temporal SAR–Optical Images," *Remote Sens.*, vol. 17, no. 13, art. 2169, 2025.