STGAN-CR: A Swin Transformer-Enhanced GAN Framework for Effective Cloud Removal in Satellite Imagery

1st Hongming Zhu
School of Software Engineering
Tongji University
Shanghai, China
zhu_hongming@tongji.edu.cn

2nd Zeju Wang
School of Software Engineering
Tongji University
Shanghai, China
2231550@tongji.edu.cn

3rd Manxin Xu
School of Software Engineering
Tongji University
Shanghai, China
2333095@tongji.edu.cn

4th Jinfeng Jiang
School of Software Engineering
Tongji University
Shanghai, China
2231513@tongji.edu.cn

5th Jipeng Zhang
School of Software Engineering
Tongji University
Shanghai, China
2333081@tongji.edu.cn

6th Hongfei Fan
School of Software Engineering
Tongji University
Shanghai, China
fanhongfei@tongji.edu.cn

7th Qin Liu
School of Software Engineering
Tongji University
Shanghai, China
qin.liu@tongji.edu.cn

* Bowen Du
School of Software Engineering
Tongji University
Shanghai, China
dubowen91@163.com

***Abstract*—Advancements in satellite remote sensing have enhanced Earth observation, enabling the acquisition of images crucial for various remote sensing applications. However, cloud cover degrades image quality and obstructs surface information. Traditional cloud removal techniques struggle to restore both low-level and high-level features, especially under complex conditions. To address these challenges, we propose STGAN-CR, a novel framework integrating Swin Transformer with Generative Adversarial Networks (GANs) to optimize image detail recovery. Leveraging the Swin Transformer's global modeling capabilities, our method enhances feature extraction and restoration, overcoming limitations of conventional models. We introduce a new evaluation metric focused on scene classification accuracy post-de-clouding to better assess practical utility. Extensive experiments and ablation studies show that STGAN-CR outperforms existing models in visual quality and classification performance. These advancements offer an effective solution for enhancing the quality and utility of remote sensing images, balancing the restoration of both low-level and high-level features, and providing more meaningful de-clouded images for downstream applications.**

***Index Terms*—Cloud removal; Transformer; Generative Adversarial Network; Deep learning; Remote sensing**

I. Introduction

Advancements in satellite remote sensing technology have significantly expanded the potential for Earth observation, enabling the acquisition of large-scale remote sensing images. These images play a crucial role in various applications such as environmental monitoring, urban planning, and disaster assessment. However, the presence of cloud cover severely degrades the visual quality of these images and obstructs crucial surface information, thereby impeding their practical utility [1]. Effective cloud removal techniques are therefore essential to restore lost data and enhance the reliability and accuracy of remote sensing applications [2].

Despite significant advancements in cloud removal methods, the task remains formidable due to the complex and variable nature of cloud cover [2]. Deep learning methods, particularly those utilizing Synthetic Aperture Radar (SAR), have shown remarkable abilities to handle these complexities [3] [4]. Models such as DSen2-CR [5] and GLF-CR [6] leverage multimodal data to restore low-level features like pixel values and brightness, as well as high-level features such as geographic contours and semantic details. However, these models often fail to balance the restoration of these two feature levels, focusing more on pixel accuracy while neglecting the integration of semantic and geographic integrity essential for practical applications of remote sensing data.

Furthermore, integrating SAR and multispectral data using Generative Adversarial Networks (GANs) has improved de-clouding results. Models like McGANs [7], CycleGAN [8], SAR-Opt-cGAN [9], and SpA-GAN [10] maintain high-level image features but are limited by the use of traditional Convolutional Neural Networks (CNNs), which are less efficient

* corrresponding author Bowen Du

DOI nunber 10.18293/SEKE2024-059

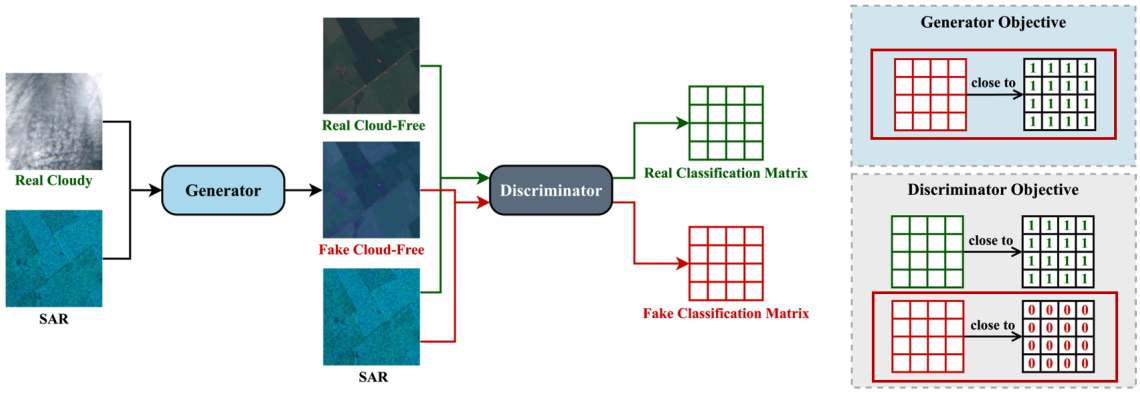


Fig. 1. Overview of the proposed STGAN-CR algorithm.

in capturing long-range pixel associations and global features. There remains a gap in research exploring the combination of GANs and Transformer models, which could synergize GANs' high-level feature reconstruction with Transformers' low-level feature processing to overcome the limitations of CNN-based frameworks and enhance detailed and contextually accurate image restoration.

Current evaluation practices for remote sensing image de-clouding primarily focus on low-level feature recovery using metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), which assess pixel-level accuracy but often overlook the restoration of higher-level features. The analysis of the SEN12MS-CR dataset by Gawlikowski et al. [11] highlights the significant impact of cloud cover on downstream tasks like scene classification. This underscores the need for robust evaluation metrics that assess both low-level and high-level feature restoration, aligning more closely with the operational utility of de-clouded images.

To address these challenges, we propose a novel cloud removal approach named STGAN-CR (Swin Transformer-enhanced GAN-based Cloud Removal). This framework leverages the Swin Transformer to enhance feature extraction and restoration, optimizing the recovery of both low-level and high-level features. By incorporating conditional adversarial training, STGAN-CR ensures comprehensive feature restoration in remote sensing images. The Swin Transformer-based architecture for both the generator and discriminator significantly improves the extraction of global high-level features, overcoming the limitations of traditional convolution-based models.

Furthermore, we introduce a new evaluation metric focused on the scene classification accuracy of de-clouded images. This metric aims to more accurately assess the practical utility of de-clouded images by evaluating the restoration success of high-level features essential for downstream remote sensing tasks [12]. Extensive experiments and ablation studies demonstrate that STGAN-CR significantly outperforms existing methods in terms of both visual quality and classification performance.

In summary, our key contributions are as follows:

- **Advanced De-clouding Framework**: A novel cloud removal framework that integrates generative models and considers both low-level and high-level features within the optimization objectives, ensuring comprehensive restoration across varying levels of detail.
- **Swin Transformer-Based Architecture**: Utilizing Swin Transformer for both the generator and discriminator to enhance the extraction of global high-level features, providing superior performance in capturing complex spatial dependencies.
- **New Metric for Scene Classification**: Development of an innovative evaluation metric focused on scene classification accuracy post-de-clouding, measuring the restoration success of high-level features and offering a relevant assessment of image quality for practical applications.

II. Methodology

A. STGAN-CR Framework Overview

In this study, we propose the STGAN-CR framework, which leverages the Swin Transformer to enhance cloud removal from remote sensing images. The framework employs a conditional adversarial training strategy involving two primary components: the generator G and the discriminator D. The generator G takes as input a cloud-covered multispectral image O and a corresponding Synthetic Aperture Radar (SAR) image S, generating a de-clouded pseudo multispectral image F:

$$F = G(O, S)$$

The discriminator D evaluates the authenticity and quality of the generated images. It processes both real cloud-free multispectral images R with their corresponding SAR images S and the pseudo cloud-free multispectral images F with the same SAR images. The discriminator outputs a classification matrix M, which assesses the authenticity of the images in local regions:

$$M_r = D(R, S) \quad \text{and} \quad M_f = D(F, S)$$

By including SAR images, which provide crucial spatial structural information about the terrain beneath the clouds, the

discriminator can more accurately determine the authenticity of the multispectral images, especially in densely or extensively cloud-covered areas. This adversarial training strategy effectively restores terrain structures and semantic information, enhancing the overall quality and utility of the generated cloud-free images.

B. Optimization Strategy and Loss Functions

To guide the adversarial learning process, we design specific optimization objectives and loss functions for both the generator and the discriminator.

1) Adversarial Loss Function for the Generator: The primary goal of the generator G is to produce images that closely resemble real cloud-free images. This is achieved through an adversarial loss function:

$$L_{\text{GAN}} = \frac{1}{H_M W_M} \sum_{i,j} \left((M_{f_{i,j}} - 1)^2 \right)$$

where H_M and W_M are the height and width of the classification matrix, respectively, and $M_{f_{i,j}}$ is the value at position (i, j) in the pseudo classification matrix.

2) Discriminator's Loss Function: The discriminator D aims to differentiate between real and generated pseudo images. Its loss function L_D is designed to improve discrimination accuracy:

$$L_D = \frac{1}{H_M W_M} \sum_{i,j} \left((M_{r_{i,j}} - 1)^2 + (M_{f_{i,j}} - 0)^2 \right)$$

This formulation allows the discriminator to independently judge the authenticity of images in each local region, enhancing its overall discrimination capability.

3) Charbonnier Loss: To further refine the generator's performance, we incorporate the Charbonnier loss, a modified version of the L1 loss that is particularly effective in smoothing the gradient of the loss function:

$$L_{Charbonnier} = \frac{1}{H_I W_I C_I} \sum_{i,j,k} \sqrt{(R_{i,j,k} - F_{i,j,k})^2 + \epsilon^2}$$

where ϵ is a small constant ensuring numerical stability, and $R_{i,j,k}$ and $F_{i,j,k}$ are the pixel values of the real and generated images, respectively.

4) Final Loss Function for the Generator: The final loss function for the generator combines the adversarial loss and the Charbonnier loss to guide the comprehensive restoration of both high-level and low-level features:

$$L_G = L_{GAN} + \lambda_1 L_{Charbonnier}$$

where λ_1 is a weight coefficient that tunes the influence of the Charbonnier loss based on specific application needs.

C. Generator and Discriminator Design

1) Generator Design: The generator in the STGAN-CR framework leverages the Swin Transformer to enhance global high-level feature extraction. The architecture involves several key components:

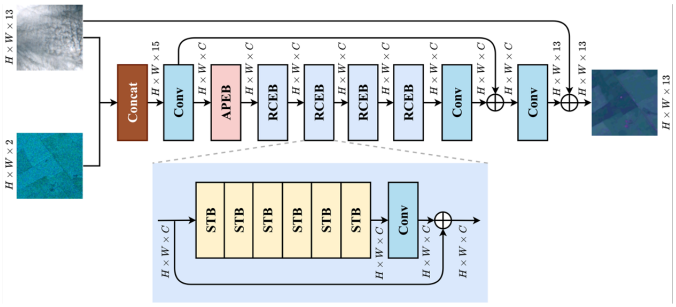


Fig. 2. Architecture of Generator

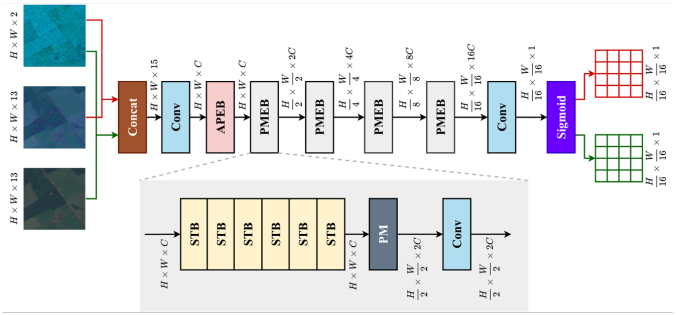


Fig. 3. Architecture of Discriminator

a) Initial Convolution Operation: Before extracting features with the Transformer, an initial 3 × 3 convolutional kernel is applied for preliminary feature extraction, resulting in F_{init}. This operation maps the image to a higher-dimensional feature space, preparing it for subsequent Transformer-based feature extraction.

b) Absolute Position Encoding Block (APEB): To introduce positional information, an Absolute Position Encoding Block (APEB) is applied, yielding F_{pos}. This block compensates for the Transformer's lack of inherent positional encoding.

c) Residual Connection Encoder Blocks (RCEB): The features F_{pos} undergo deep feature extraction through multiple layers of Residual Connection Encoder Blocks (RCEB), producing F_{deep}. The initial and deep features are then combined using residual connections to form the final features F_{final}.

d) Final Convolution and Output: The features are downscaled back to the dimensions of the multispectral images through another 3 × 3 convolutional operation, resulting in I_{res}. This result is added to the cloud-covered multispectral image O to form the final output:

$$I_{final} = I_{res} + O$$

2) Discriminator Design: The discriminator uses a PatchGAN architecture, which outputs a classification matrix rather than a single scalar. This matrix assesses the authenticity of the input image in corresponding subregions. Key components include:

TABLE I
COMPARATIVE ANALYSIS OF CLOUD REMOVAL METHODS ON SEN12MS-CR DATASET

Method	PSNR	SSIM	MAE	SAM	Scene Class. Acc.
SFGAN	27.22	0.8533	0.0358	10.445	42.66%
DSen2-CR	28.00	0.8718	0.0310	9.469	53.09%
GLF-CR	28.23	0.8632	0.0320	8.887	50.93%
STGAN-CR (Ours)	**28.56**	**0.8859**	**0.0295**	**8.630**	**57.40%**

a) Patch Merging Encoder Blocks (PMEB): The discriminator employs Patch Merging Encoder Blocks (PMEB), which progressively reduce the feature dimensions while increasing vector dimensions. This compression enhances feature expression capabilities.

b) Final Convolution and Classification: After the PMEB sequence, a final 3 × 3 convolution reduces the feature dimensions to one, producing the final classification matrices M_r or M_f. This design ensures effective evaluation of the authenticity of local areas within the overall image.

D. Scene Classification Accuracy

Traditional metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) primarily focus on low-level feature recovery. However, they often overlook the restoration of high-level features crucial for practical remote sensing applications. To address this, we introduce scene classification accuracy as a novel metric.

1) Calculation and Implementation: Scene classification accuracy evaluates the effectiveness of cloud removal methods in restoring high-level features essential for accurate scene classification. It is calculated as follows:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$$

where n is the number of samples, y_i is the true scene category of the ith sample, $\hat{y}_i$ is the predicted category, and I is an indicator function that equals 1 when $y_i = \hat{y}_i$ and 0 otherwise.

The scene classifier employed in this research is trained on the SEN12MS dataset, which includes scene classification labels. By evaluating the restored high-level features through scene classification accuracy, we provide a more direct assessment of the practical utility of de-clouded images in real-world applications.

III. EXPERIMENT AND RESULTS

A. Dataset and Implementation Details

The experiments in this study use the SEN12MS-CR dataset, built on the SEN12MS dataset, which includes observations for 175 non-overlapping Regions Of Interest (ROI) across the globe during all four seasons. Each ROI contains cloud-free multispectral images, cloud-covered multispectral images, and SAR images, with an average cloud cover of about 48%, reflecting real-world scenarios [12]. Each ROI spans approximately 52 × 40 km, subdivided into multiple image segments of 256 × 256 pixels, with a 50% overlap, resulting in about 700 segments per ROI. The dataset comprises 108,941 triplets for training, 6,535 for validation, and 6,742 for testing, with multispectral images normalized to [0, 1].

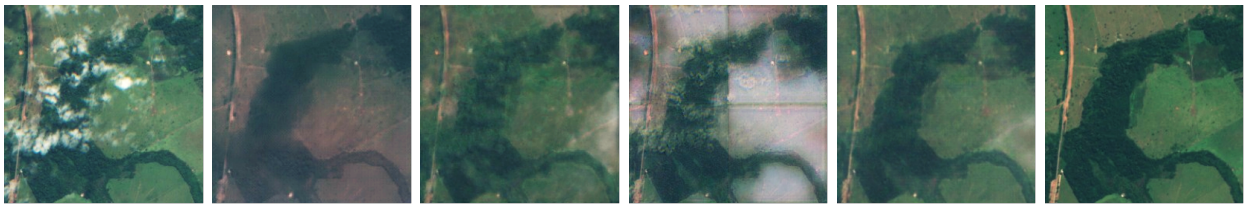
The generator architecture features an initial feature vector length of 60 and 4 Residual Connection Encoder Blocks (RCEBs) with 4 Swin Transformer Blocks (STBs) each, while the discriminator uses a PatchGAN architecture with 4 Patch Merging Encoder Blocks (PMEBs). Both models are trained using the Adam optimizer with a batch size of 18, employing a cosine annealing schedule for learning rates, initially set at 3×10^{-4} for the generator and 3×10^{-6} for the discriminator, over 50 epochs. Data augmentation techniques such as random cropping, flipping, and rotation are applied to enhance model robustness. The source code and pre-trained models for STGAN-CR are available at https://github.com/Major-333/cloud-removal, providing comprehensive instructions for replicating our experiments.

B. Evaluation Metrics

Traditional metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) have been foundational in evaluating the fidelity of cloud-removed images, focusing primarily on low-level features like pixel values, brightness, and contrast. However, these metrics often fail to address higher-level features such as object contours and semantic information, which are crucial for the practical utility of remote sensing data in applications like environmental monitoring and urban planning. To provide a more comprehensive evaluation, we introduce scene classification accuracy, which assesses the effectiveness of cloud removal methods in restoring high-level features essential for accurate scene classification.

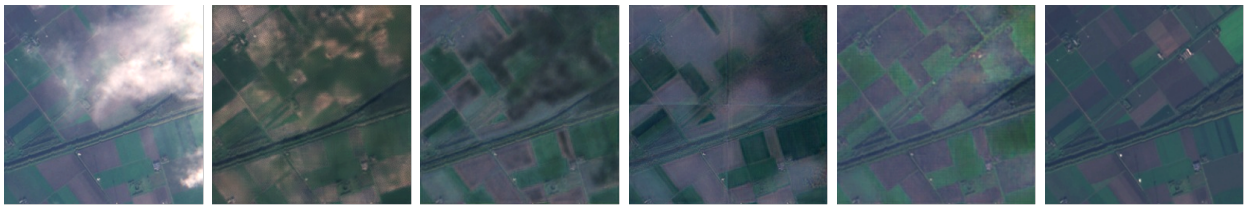
C. Quantitative Evaluation

We compare STGAN-CR with three leading de-clouding methods: Simulation-Fusion GAN (SFGAN), DSen2-CR, and GLF-CR. These benchmarks are chosen for their demonstrated effectiveness in handling cloud-covered remote sensing images. SFGAN is renowned for its generative model-based approach, serving as a standard reference in numerous studies due to its innovative use of GANs. DSen2-CR leverages SAR data with convolutional neural networks to enhance data reliability and image clarity. GLF-CR, employing Transformer technologies, achieves notable improvements in key image quality metrics such as PSNR, showcasing the effectiveness of Transformers in capturing complex spatial relationships in remote sensing data.



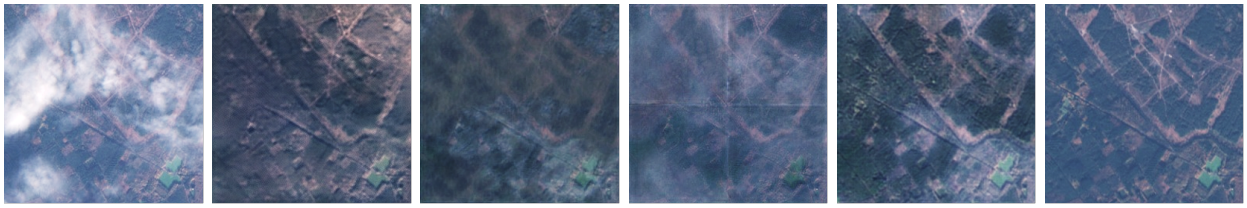
(a) Cloudy MS (b) DSen2_CR (c) SFGAN (d) GLF-CR (e) Ours (f) Ground Truth

Fig. 4. Visual Comparison of Cloud Removal Methods Across Sparse Grasslands



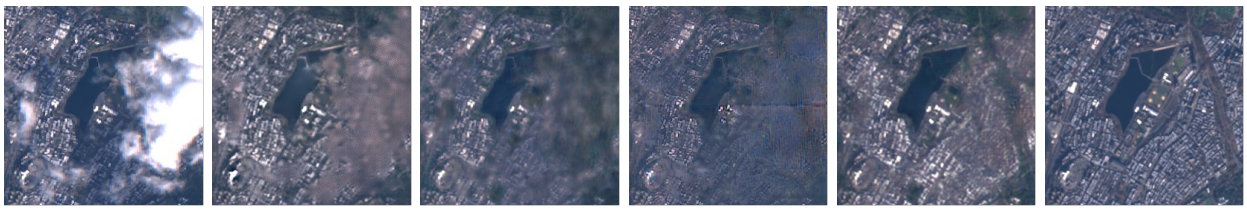
(a) Cloudy MS (b) DSen2_CR (c) SFGAN (d) GLF-CR (e) Ours (f) Ground Truth

Fig. 5. Visual Comparison of Cloud Removal Methods Across Croplands



(a) Cloudy MS (b) DSen2_CR (c) SFGAN (d) GLF-CR (e) Ours (f) Ground Truth

Fig. 6. Visual Comparison of Cloud Removal Methods Across Forest



(a) Cloudy MS (b) DSen2_CR (c) SFGAN (d) GLF-CR (e) Ours (f) Ground Truth

Fig. 7. Visual Comparison of Cloud Removal Methods Across Urban

The models are evaluated using PSNR, SSIM, Mean Absolute Error (MAE), Spectral Angle Mapper (SAM), and scene classification accuracy. PSNR and SSIM measure the fidelity and structural similarity to the original images, MAE evaluates the average magnitude of errors between the predicted and true values, SAM assesses spectral fidelity, and scene classification accuracy evaluates the restoration of high-level features essential for accurate scene classification.

Table I shows the results. STGAN-CR outperforms other methods across all metrics, demonstrating superior performance in restoring both low-level and high-level features.

D. Subjective Visual Comparison

We selected four representative regions for visual comparison: sparse grasslands, farmland, forest, and urban areas. These regions were chosen due to their diverse geographical characteristics and the varying complexity they present for cloud removal. Figures 4 to 7 illustrate the de-clouding effects produced by STGAN-CR and three other methods (SFGAN, DSen2-CR, and GLF-CR). In the sparse grasslands region, STGAN-CR closely matches the true cloud-free image in overall color and successfully identifies and restores parts of trees obscured by clouds, presenting a realistic visual effect. In farmland areas, STGAN-CR restores the farmland patterns more neatly, with clearer demarcations between different patches of farmland compared to other methods, demonstrating distinctly superior visual effects.

In the dense forest region, where cloud cover obscures several patches of forest, STGAN-CR not only delineates the forest contours sharply but also recovers intricate tree patterns, delivering the highest realism. Urban areas pose one of the most significant challenges due to the detailed structures of buildings and roads. STGAN-CR effectively restores detailed

architectural patterns and urban semantics, significantly outperforming other methods, which often produce unclear and vague outlines. These results demonstrate that STGAN-CR achieves the best de-clouding visual effects across various regions, maintaining clearer contours and more realistic visual effects compared to other methods. This highlights STGAN-CR's superior capability in analyzing deep semantic information and restoring corresponding terrain features.

E. Ablation Study

TABLE II
ABLATION ANALYSIS OF STGAN-CR METHODS ON SEN12MS-CR DATASET

Method	PSNR	SSIM	MAE	SAM	Scene Class. Acc.
w/o L_{GAN}	28.82	0.8886	0.0288	8.436	51.65%
STGAN-CR	28.56	0.8859	0.0295	**8.630**	**57.40%**

To demonstrate the effectiveness of the conditional adversarial training framework in STGAN-CR, we conducted an ablation study focusing on the loss function, which includes both the Charbonnier loss ($L_{Charbonnier}$) and the GAN loss (L_{GAN}). The study evaluated the model's performance with and without the L_{GAN} component. The results, as shown in Table II, indicate that excluding the L_{GAN} component leads to slight improvements in low-level feature metrics like PSNR, SSIM, and MAE, as these metrics favor pixel-level accuracy which deep regression models naturally optimize.

However, the inclusion of the L_{GAN} component significantly enhances the model's ability to recover high-level features, as evidenced by a substantial increase in scene classification accuracy. This suggests that while models without L_{GAN} excel in low-level feature restoration, they lag in high-level feature recovery essential for practical applications. The L_{GAN} component sacrifices a small portion of low-level accuracy to greatly improve high-level feature restoration, demonstrating its critical role in balancing the overall performance of the de-clouding model.

Our findings highlight that combining both $L_{Charbonnier}$ and L_{GAN} provides the most comprehensive restoration of remote sensing images, balancing low-level and high-level feature recovery. This dual-loss approach ensures that the generated cloud-free images are not only visually accurate but also semantically meaningful, thus better supporting downstream tasks such as environmental monitoring and urban planning.

IV. CONCLUSION

This study introduces STGAN-CR, a novel approach for cloud removal in remote sensing images by integrating Swin Transformer with Generative Adversarial Networks. Our method effectively bridges the gap between restoring low-level details and high-level semantic features, achieving superior performance across various metrics on the SEN12MS-CR dataset. The incorporation of a new performance metric, scene classification accuracy, provides a nuanced evaluation that aligns with practical demands in Earth observation. Future research could extend the applicability of STGAN-CR to multi-temporal datasets and other image enhancement tasks, leveraging its robust capabilities. The integration of advanced cloud detection technologies is anticipated to further refine the precision of cloud removal processes, enhancing the usability and functionality of remote sensing data for more accurate and reliable analyses.

ACKNOWLEDGMENT

This research was sponsored by the Science and Technology Commission of Shanghai Municipality (No. 23511103100), the National Key R&D Program of China (No. 2023YFC3805305), the Natural Science Foundation of Shanghai (No. 21ZR1465100), the Fundamental Research Funds for the Central Universities (No. 22120220658), and the National Natural Science Foundation of China (No. 61702374). We sincerely thank all these funding sources for their support.

REFERENCES

[1] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 7, pp. 3826–3852, 2013.

[2] R. Cao, Y. Chen, J. Chen, X. Zhu, and M. Shen, "Thick cloud removal in landsat images based on autoregression of landsat time-series data," *Remote Sensing of Environment*, vol. 249, p. 112001, 2020.

[3] Q. Zhang, Q. Yuan, Z. Li, F. Sun, and L. Zhang, "Combined deep prior with low-rank tensor svd for thick cloud removal in multitemporal images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 177, pp. 161–173, 2021.

[4] C. Long, X. Li, Y. Jing, H. Shen *et al.*, "Bishift networks for thick cloud removal with multitemporal remote sensing images," *International Journal of Intelligent Systems*, vol. 2023, 2023.

[5] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620301398

[6] F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu, "Glf-cr: Sar-enhanced cloud removal with global–local fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 192, pp. 268–278, 2022.

[7] K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and N. Kawaguchi, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 48–56.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[9] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse sar and multispectral optical data for cloud removal from sentinel-2 images," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 1726–1729.

[10] H. Pan, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," *arXiv preprint arXiv:2009.13015*, 2020.

[11] J. Gawlikowski, P. Ebel, M. Schmitt, and X. X. Zhu, "Explaining the effects of clouds on remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9976–9986, 2022.

[12] 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," *arXiv preprint arXiv:1906.07789*, 2019.