A cGAN-Based Approach for SAR-to-Optical Image Translation with Application to Cloud Removal

Ahmed Attia and Peter Hofmann *

Posted Date: 6 May 2026

doi: 10.20944/preprints202605.0243.v1

Keywords: synthetic aperture radar (SAR); generative adversarial networks (GANs); SAR-to-optical image translation; cloud removal; deep learning

 doi:10.20944/preprints202605.0243.v1

Article

A cGAN-Based Approach for SAR-to-Optical Image Translation with Application to Cloud Removal

Ahmed Attia and Peter Hofmann *

Deggendorf Institute of Technology
* Correspondence: peter.hofmann@th-deg.de

Abstract

Cloud cover remains a persistent challenge in optical remote sensing, limiting the usability of optical satellite imagery for continuous Earth observation. Synthetic Aperture Radar (SAR) data, in contrast, provides cloud-penetrating, all-weather imaging but lacks the spectral richness and is less visually interpretable compared to optical observations. Bridging these complementary modalities, this study investigates SAR-to-optical image translation using the Pix2Pix conditional generative adversarial network (cGAN). While existing research predominantly focuses on reconstructing only the visible (RGB) or near-infrared bands, this work employs the winter subset of the SEN12-MS dataset to address the full spectral range. The objectives are threefold: (i) to validate SAR-to-optical translation across all 13 Sentinel-2 spectral bands; (ii) to assess the reliability and reconstructability of each individual band; and (iii) to evaluate the performance of the translation model for cloud removal. Experimental results show that the model effectively learns the SAR-to-optical mapping and achieves high reconstruction quality across all spectral bands, though bandwise analysis reveals that reconstruction accuracy varies with spectral characteristics. When applied to the SEN12-MS-CR dataset, the model successfully generates cloud-free optical imagery that closely matches reference data, achieving performance competitive with state-of-the-art models such as DiffCR. Overall, the findings confirm the viability of SAR-to-optical translation for producing spectrally consistent, cloud-free optical imagery, thus enhancing the temporal continuity of Earth observation data. Two ablation studies further analyze the impact of different loss functions and the exclusion of 60 m bands.

Keywords: synthetic aperture radar (SAR); generative adversarial networks (GANs); SAR-to-optical image translation; cloud removal; deep learning

1. Introduction

Earth observation has become an indispensable tool for understanding and monitoring the planet's dynamic processes. Optical and radar remote sensing represent two complementary modalities at the core of modern monitoring systems. Optical sensors, such as the Multi-Spectral Instrument (MSI) onboard Sentinel-2, provide rich spectral and visual information essential for applications ranging from agricultural monitoring and land-cover classification to disaster response and forest monitoring [1]. However, these sensors are inherently constrained by atmospheric conditions, particularly cloud cover. Global estimates suggest that average cloud cover exceeds 66% [4–6], with about 55% over land surfaces [6], causing considerable data gaps in both spatial and temporal domains. For applications requiring consistent time series, e.g., agricultural monitoring, or where a specific scene must be observed at a given time, e.g., disaster monitoring, cloud cover represents a serious limitation [6]. The diversity of clouds —including thin and thick clouds as well as haze— together with the wide range of occlusion scenarios and their uneven distribution, poses an additional challenge for image reconstruction and the generalizability of cloud removal techniques [7]. In contrast, Synthetic Aperture Radar (SAR) sensors, operating in the microwave domain, such as the C-band instrument on Sentinel-1 mission offer all-weather, day-and-night imaging capabilities independent of solar illumination or

 doi:10.20944/preprints202605.0243.v1

cloud interference. However, the backscatter-based nature of SAR imagery introduces challenges related to speckle noise, geometric distortions, and a lack of intuitive spectral color information [2,3], necessitating advanced interpretation skills.

To bridge the gap between these two sensing modalities, SAR-to-optical image translation has emerged as a powerful generative approach. It aims to synthesize optical-like, cloud-free imagery from SAR data, combining the interpretability of optical observations with the reliability of radar acquisitions. The domain gap between SAR and optical imagery, however, poses significant challenges. For example, while two objects with identical structures may appear different in optical imagery due to their spectral responses, they can appear similar in SAR imagery, reflecting SAR's emphasis on structural rather than spectral properties [16]. Moreover, acquiring perfectly co-registered SAR–optical pairs is also non-trivial, as both spatial and temporal alignment must be ensured.

Recent advances in generative artificial intelligence (GenAI), particularly in conditional generative adversarial networks (cGANs) [8] and diffusion models [9], have made it possible to learn complex mappings between SAR and optical domains with visually plausible results [10]. These developments have opened new pathways for applications in land-cover classification, vegetation monitoring, disaster response, and particularly, cloud removal. Despite the rapid progress in this field, a significant gap remains in the literature. Most existing studies have focused primarily on reconstructing the visible RGB subset of optical imagery such as [2,10,11,13,23], with a few extending to the Near-Infrared (NIR) range such as [14,15]. Consequently, the full multispectral potential of missions such as Sentinel-2 remains largely underexplored in the context of generative reconstruction. Critical bands for vegetation analysis (Red-Edge) and soil moisture monitoring (SWIR) are rarely addressed.

Furthermore, while SAR-to-optical translation is often evaluated qualitatively, systematic assessments of its reliability across individual spectral bands remain scarce. Finally, although the approach shows promise for cloud removal, its effectiveness in this context has not yet been comprehensively validated across the full spectral range.

Accordingly, this study investigates the use of generative models for translating Sentinel-1 SAR imagery into the full multispectral stack of Sentinel-2 imagery. The objectives of this work are threefold:

1. To validate SAR-to-optical image translation across all 13 Sentinel-2 spectral bands, moving beyond standard RGB constraints.
2. To systematically assess the reliability and reconstructability of each individual spectral band.
3. To evaluate the performance of the translation model for the practical application of cloud removal.

By addressing these objectives, this work aims to provide a comprehensive and quantitative understanding of SAR-to-optical translation as a multimodal learning problem and to clarify its potential and limitations for enhancing the temporal continuity of Earth observation data.

2. Related Work

2.1. Cloud Removal Methods (Overview)

Cloud removal methods in optical remote sensing are commonly grouped into three categories: (i) single-image methods, (ii) multimodal-based methods, and (iii) multitemporal-based methods [7].

Single-image methods aim to reconstruct surface information using only the cloudy optical image and typically rely on statistical or physical priors (e.g., frequency filtering or atmospheric scattering models). Zhang et al. [33] proposed the Haze Optimized Transformation (HOT) for thin cloud and haze compensation, while He et al. [34] introduced the dark channel prior, later adapted for thin cloud removal in optical remote sensing. With deep learning, CNNs, U-Nets, and GAN-based architectures have also been applied in this setting; however, single-image methods remain fundamentally limited under dense cloud cover due to missing surface information [7].

Multimodal-based methods integrate auxiliary data from complementary sensors to improve restoration. In practice, the most significant progress has been achieved by fusing SAR and optical imagery, leveraging the cloud-penetrating capability of SAR. Meraner et al. [6] proposed DSen2-CR,

 doi:10.20944/preprints202605.0243.v1

combining Sentinel-1 and Sentinel-2 to enhance reconstruction under thick clouds and preserve spectral fidelity. Grohnfeldt et al. [5] demonstrated the effectiveness of conditional GANs for SAR–optical fusion, and Xu et al. [35] proposed GLF-CR using a global–local fusion strategy. This line of work motivates SAR-to-optical translation approaches, which either translate SAR features into optical-like imagery or fuse SAR with partially corrupted optical inputs, while still facing challenges related to data registration, modality differences, and SAR speckle.

Multitemporal-based methods exploit repeated observations to fill cloud-covered regions. Xu et al. [28] proposed multitemporal dictionary learning for reconstructing thin and thick cloud regions without explicit cloud masks. Ebel et al. [30] introduced UnCRtainTS, combining multispectral reconstruction with per-pixel uncertainty estimation. While highly effective under persistent cloud cover, multitemporal approaches can be sensitive to geometric misalignment and temporal land-cover changes [36], and they require consistent time series data; mono-temporal approaches avoid such requirements and reduce co-registration constraints [37]. Given the focus of this work on mono-temporal SAR–optical generation, the following subsections review SAR-to-optical translation methods and their use for cloud removal.

2.2. General SAR-to-Optical Image Translation

SAR-to-optical translation is defined as an image-to-image (I2I) task that synthesizes optical-like imagery from SAR data, despite the domain gap between backscatter intensity and spectral reflectance as well as SAR-specific artifacts such as speckle and geometric distortions. Before the emergence of generative AI, SAR-to-optical translation relied on heuristic or classical methods. Early approaches used pseudo-colorization of SAR channels or polarization composites to enhance interpretability [17], though they failed to reproduce true optical characteristics. Multisensor fusion methods, such as combining SAR with prior cloud-free optical data via intensity–hue–saturation (IHS) or wavelet-based transforms [18], provided partial solutions but depended on handcrafted features and complex preprocessing, limiting scalability and accuracy.

Recent deep learning approaches—most notably generative adversarial networks (GANs) [19] and diffusion models [9]—have enabled end-to-end learning of mappings from SAR to optical observations by modeling cross-modal feature relationships and pixel-level distributions. Most existing studies focus on reconstructing the visible (RGB) subset of Sentinel-2 (or extend to near-infrared), whereas comparatively fewer works address full multispectral reconstruction; correspondingly, the literature often distinguishes SAR-to-optical translation from SAR-to-multispectral (SAR-to-MS) translation. In this paper, the term "optical" denotes the full 13-band Sentinel-2 spectral range, and the proposed methodology extends SAR-to-optical translation to the complete multispectral domain.

The foundational GAN framework by Goodfellow et al. [19], later extended to Conditional GANs (cGANs) by Mirza and Osindero [8], enabled image-to-image translation tasks that made SAR-to-optical synthesis feasible. Fuentes Reyes et al. [20] optimized an unsupervised CycleGAN for SAR-to-optical translation, improving interpretability and reducing speckle through customized preprocessing and architectural refinements, although fine structural details remained difficult to preserve in urban scenes. Wang et al. [21] introduced a Supervised CycleGAN (S-CycleGAN) by incorporating a pixel-wise MSE loss, producing realistic optical images and showing strong potential for cloud removal. Gao et al. [10] extended this idea with a fusion-based GAN framework for high-resolution imagery. Instead of directly translating SAR to optical, they generated a simulated optical image from SAR data and fused it with both SAR and optical inputs to reconstruct more realistic results. An ablation study confirmed that this two-stage fusion strategy achieved superior performance among tested configurations.

Following the success of Vision Transformers (ViTs) [22], transformer-based architectures have also emerged. Zhao et al. [23] proposed the Hybrid Vision Transformer cGAN (HVT-cGAN), combining CNN and ViT branches to capture both local details and global semantics. A Convolutional Attention Fusion Module (CAFM) adaptively merged multiscale features, enhancing texture and color fidelity. Trained on the SEN1-2 dataset [24], HVT-cGAN achieved superior visual and quantitative results

 doi:10.20944/preprints202605.0243.v1

over previous GAN-based models. Park et al. [25] further improved this approach with a multiscale ViT-based cGAN architecture integrating perceptual loss and a two-phase transfer learning strategy to enhance realism and stability.

Diffusion models have recently entered this field. Bai et al. [11] introduced a conditional diffusion model for SAR-to-optical translation, which, despite limited experimentation due to computational constraints, demonstrated promising performance compared to GAN-based methods. Bai et al. [13] later extended the framework with color supervision to improve reconstruction fidelity. A more recent contribution, the Multi-Temporal Conditional GAN (MTcGAN) by Kwak and Park [26], was designed for early-stage crop monitoring and utilized SAR–optical pairs from reference and prediction dates to capture temporal dynamics, achieving superior spectral consistency compared to conventional methods.

These developments have motivated the use of SAR-to-optical synthesis as a data-driven strategy for cloud removal, discussed next.

2.3. SAR-to-Optical Image Translation for Cloud Removal

SAR-guided cloud removal exploits the cloud-penetrating capability of SAR to reconstruct missing optical observations in cloud-contaminated scenes. Early methods relied on traditional machine-learning-based signal processing: Huang et al. [27] introduced sparse representation-based cloud removal using SAR data, which Xu et al. [28] extended via multi-temporal dictionary learning. These approaches demonstrated the feasibility of SAR-assisted cloud reconstruction but struggled under heavy cloud cover or rapidly changing surface conditions.

With deep learning, multimodal fusion approaches became dominant. Enomoto et al. [14] applied a cGAN for cloud removal by fusing the RGB composite of a cloudy optical image with the cloud-free near-infrared (NIR) band to reconstruct a cloud-free RGB image. Although limited under dense cloud conditions, the approach was pivotal in demonstrating the potential of multimodal data fusion for cloud removal and laid the groundwork for subsequent studies. Building on this concept, Grohnfeldt et al. [5] introduced SAR-Opt-cGAN, a model designed to fuse Sentinel-1 SAR and Sentinel-2 optical data—marking the first use of SAR data within a cGAN framework for cloud mitigation. Their adaptation of the Pix2Pix architecture [29] allowed flexible multi-channel inputs and was trained on a subset of the SEN1-2 dataset [24]. The results confirmed the advantage of incorporating SAR data, validating the effectiveness of the approach for mitigating cloud cover.

Beyond direct cGAN fusion, specialized pipelines were proposed. Darbaghshahi et al. [2] introduced a dual-GAN framework with Dilated Residual Inception Blocks (DRIBs), translating SAR to optical in the first stage and fusing with cloudy optical inputs in the second; it showed strong qualitative results but moderate quantitative accuracy. Ebel et al. [30] advanced this direction with UnCRtainTS, which reconstructs multispectral imagery while predicting per-pixel uncertainty, improving performance and providing reliability indicators.

Recently, diffusion models have become highly competitive. Zou et al. [31] proposed DiffCR, a fast conditional diffusion framework and reported state-of-the-art on SEN12-MS-CR [32]. DiffCR employs a decoupled conditional architecture and a Time and Condition Fusion Block (TCFBlock), and predicts cloud-free images directly, enabling high-quality results in as few as 1–5 denoising steps with substantially reduced computational cost. In parallel, Meraner et al. [6] proposed DSen2-CR, a deep residual SAR–optical fusion network with a Cloud-Adaptive Regularized Loss (CARL) to preserve uncorrupted content while focusing reconstruction on clouded regions; evaluated on SEN12-MS-CR, it showed strong global generalization and high reconstruction accuracy across all 13 bands, with best performance on surface-related bands and higher errors on atmospheric bands.

Overall, the literature demonstrates a progression from classical SAR-guided reconstruction to deep multimodal fusion, uncertainty-aware prediction, and diffusion-based models, establishing SAR-to-optical translation as a strong paradigm for cloud removal. This work builds on that foundation by evaluating SAR-to-optical translation for cloud removal across the full 13-band Sentinel-2 spectrum and analyzing model behavior through bandwise evaluation and ablation studies.

 doi:10.20944/preprints202605.0243.v1

3. Materials and Methods

3.1. Problem Formulation

The task of SAR-to-optical translation is formulated as a supervised image-to-image translation problem. Let $\mathcal{X} \subset \mathbb{R}^{H\times W\times C_{in}}$ denote the domain of SAR inputs, where H and W represent the height and width of the image patches in pixels, and $C_{in} = 2$ corresponds to the dual-polarized Sentinel-1 channels (VV, VH) expressed in the decibel (dB) scale. Similarly, let $\mathcal{Y} \subset \mathbb{R}^{H\times W\times C_{out}}$ denote the domain of optical reference images, where $C_{out} = 13$ represents the full multispectral stack of Sentinel-2.

The objective is to learn a parametric mapping function $G_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by θ, such that for a given SAR input $x \in \mathcal{X}$, the generated optical image $\hat{y} = G_\theta(x)$ is indistinguishable from the ground truth optical image $y \in \mathcal{Y}$ in terms of spectral and spatial characteristics.

We assume access to a paired dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M}$ consisting of spatially co-registered SAR and optical patches acquired over the same geographical regions. The training objective is to find the optimal generator parameters θ^* that minimize a composite loss function $\mathcal{L}$ over the data distribution:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(G_\theta(x), y)] \tag{1}$$

where $\mathcal{L}$ is a weighted combination of reconstruction, adversarial, and perceptual losses designed to collectively encourage pixel-level accuracy, structural consistency, and spectral fidelity. Specific details on the loss components and optimization strategy are provided in Section 3.4.

3.2. Dataset

This study utilizes the *SEN12-MS* dataset [38], curated by Schmitt et al. SEN12-MS is a large-scale, globally distributed benchmark explicitly designed to advance research in multimodal Earth observation and deep learning. It comprises 180,662 georeferenced image triplets, each consisting of: (i) dual-polarized Sentinel-1 synthetic aperture radar (SAR) data in VV and VH polarization (σ^0 backscatter values in decibel scale); (ii) full Sentinel-2 multispectral imagery spanning all 13 bands; and (iii) MODIS land cover maps derived from the MCD12Q1 product and resampled to 10 m resolution. Each triplet is stored as a 256 × 256 pixel GeoTIFF at 10 m ground sampling distance, corresponding to a spatial coverage of approximately 2.56 × 2.56 km per patch. The dataset has a total size of 510 GB, reflecting its high complexity, diversity, and spatial resolution. For the purpose of SAR-to-optical translation, only the Sentinel-1 and Sentinel-2 modalities are utilized; the MODIS component is excluded. Sample data are depicted in Figure 1.

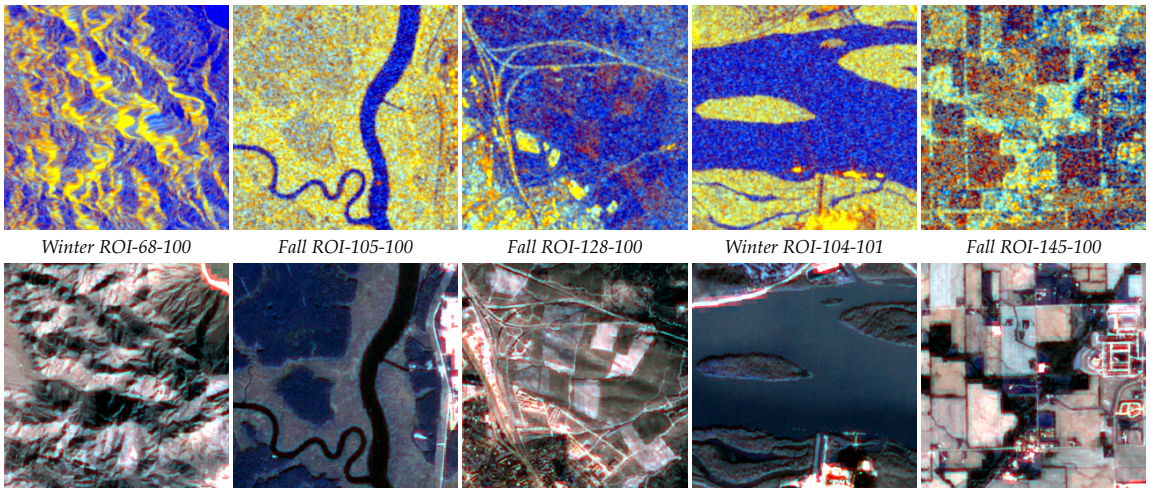


Figure 1. Sample pairs from the SEN12-MS dataset. Top row: Sentinel-1 SAR patches (R: VV, G: VH, B: VV/VH). Middle row: ROI IDs. Bottom row: corresponding Sentinel-2 multispectral patches (RGB visualization).

The Sentinel-1 component originates from ground-range-detected (GRD) products acquired in interferometric wide swath (IW) mode. The Sentinel-2 imagery was curated using a cloud-free mosaick-

 doi:10.20944/preprints202605.0243.v1

ing workflow on Google Earth Engine, ensuring that each region of interest (ROI) is represented by seasonally consistent, nearly cloud-free multispectral data. Importantly, all triplets underwent manual verification by remote sensing experts to ensure freedom from major artifacts, severe registration errors, or residual cloud contamination.

The ROIs were sampled globally across all inhabited continents and all four seasons of 2017 to maximize spatial and temporal diversity.

SEN12-MS is part of a broader family of datasets developed to foster multimodal remote sensing research. Its predecessor, SEN1-2 [24], provided SAR–optical pairs but lacked georeferencing, full spectral coverage, and dual-polarization SAR. SEN12-MS addressed these limitations, establishing a comprehensive multimodal benchmark. The family has since been extended by SEN12-MS-CR [32] (adding cloudy/cloud-free pairs) and SEN12-MS-CR-TS [39] (multimodal time series). This study focuses on SEN12-MS to leverage its global diversity for learning the core SAR-to-optical translation mapping.

3.2.1. Subset Selection and Preprocessing

The SEN12-MS dataset is divided into four seasonal subsets. Due to the computational demands of extensive experimentation and ablation studies, the *Winter* subset—being the smallest of the four and containing 31,825 paired images—was selected for training and evaluation.

A custom preprocessing pipeline was implemented to standardize the input domains. Following established best practices in the literature [6,40–43], Sentinel-1 backscatter values were clipped to fixed physical ranges of $[-25, 0]$ dB for VV polarization and $[-32.5, 0]$ dB for VH polarization to suppress radiometric outliers. Similarly, Sentinel-2 top-of-atmosphere (TOA) reflectance values were clipped to the range $[0, 10,000]$. After clipping, all values were linearly normalized to the interval $[-1, 1]$ to align with the Tanh activation function of the generator's output layer. The original image size of 256×256 pixels was preserved, and no data augmentation was applied given the inherent global diversity of the dataset.

3.3. Network Architecture

To address the SAR-to-optical translation task, we employ the Pix2Pix model, a cGAN originally introduced by Isola et al. [29]. The Pix2Pix framework has gained substantial recognition not only in general image-to-image translation [1], but also within the domain of SAR-to-optical translation. It is frequently employed as a baseline model as in [6,10,23,26,41,44], or further adapted and extended in various studies [5,44]. Owing to its demonstrated effectiveness and widespread adoption in related research, Pix2Pix was selected as the primary model for the experiments conducted in this work.

The framework consists of a generator G and a discriminator D trained in a minimax game, where G learns to map a dual-polarized SAR input x to a multispectral optical output y, while D attempts to distinguish between real pairs $\{x, y\}$ and synthesized pairs $\{x, G(x)\}$.

The generator is implemented as a U-Net encoder–decoder [45]. Unlike standard autoencoders, U-Net employs skip connections between mirror layers in the encoder and decoder stacks. These connections concatenate low-level feature maps from the downsampling path directly to the upsampling path, mitigating the information bottleneck and preserving structural details essential for SAR-to-optical alignment. The discriminator follows a PatchGAN architecture, which classifies local $N \times N$ image patches (specifically 70×70) as real or fake. This design emphasizes high-frequency texture accuracy and enforces local realism, while global structural coherence is implicitly guided by the reconstruction loss.

The training objective combines an adversarial loss $\mathcal{L}_{cGAN}$ to encourage realism with a pixel-wise reconstruction loss $\mathcal{L}_1$ to enforce data fidelity. The base objective is defined as:

$$G^* = \arg\min_G \max_D \; \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{\ell_1}(G) \tag{2}$$

 doi:10.20944/preprints202605.0243.v1

where λ controls the weight of the reconstruction term. Consistent with literature standards [5,29], we set $\lambda = 100$. To further enhance perceptual quality, our final model augments this base objective with supplementary structural (SSIM) and perceptual (LPIPS) loss terms, as detailed in the following section.

3.4. Loss Functions

While the standard Pix2Pix framework employs Binary Cross-Entropy (BCE) for the adversarial objective, we observed training instabilities consistent with vanishing gradients. To mitigate this, we replaced BCE with the Least Squares GAN (LSGAN) loss [46], which penalizes samples based on their distance to the decision boundary, yielding more stable gradients and smoother convergence.

To simultaneously ensure pixel-level fidelity, structural consistency, and perceptual realism, the final objective function combines four complementary terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LSGAN}} + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} \tag{3}$$

where:

- $\mathcal{L}_{\text{GAN}}$ denotes the adversarial loss (LSGAN), encouraging the generator to produce outputs indistinguishable from real images.
- $\mathcal{L}_{\text{L1}}$ represents the pixel-wise reconstruction loss (MAE), enforcing global structural similarity between the generated and target images.
- $\mathcal{L}_{\text{SSIM}}$ is the Structural Similarity Index Measure loss, promoting local structural consistency.
- $\mathcal{L}_{\text{LPIPS}}$ denotes the Learned Perceptual Image Patch Similarity loss, which captures high-level perceptual differences.
- $\lambda_{\text{L1}}, \lambda_{\text{SSIM}}$, and λ_{LPIPS} are weighting coefficients that control the relative contribution of each term.

The weighting coefficients were set to $\lambda_{\text{L1}} = 100$, following the original Pix2Pix configuration [29], and $\lambda_{\text{SSIM}} = \lambda_{\text{LPIPS}} = 50$. These values were selected to balance low-frequency accuracy (L1) with high-frequency perceptual details (SSIM/LPIPS). The efficacy of this composite loss over individual configurations is quantitatively and qualitatively analyzed in the ablation study (Section 4.4).

3.5. Experimental Setup and Training Strategy

All experiments were implemented in Python 3.11 and PyTorch 2.4.0 using the official Pix2Pix implementation (https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix) on an NVIDIA RTX A5000 GPU. The data consisted of the preprocessed SEN12-MS subset described in Section 3.2.1, where each sample paired a Sentinel-1 SAR patch (VV and VH channels) with a corresponding 13-band Sentinel-2 optical patch.

Training followed a staged adversarial optimization scheme. The generator was first warmed up independently for 20 epochs to mitigate early instability, encountered in the early experiments. Subsequently, both networks were jointly optimized for a total of 150 epochs with a 1:1 update ratio, where the discriminator minimized the least-squares error and the generator minimized the composite objective. We employed the Adam optimizer [47] with a batch size of 16, an initial learning rate (LR) of 1×10^{-4}, and momentum parameters $(\beta_1, \beta_2) = (0.5, 0.999)$. To accelerate convergence, the generator's LR was adjusted via a `ReduceLROnPlateau` scheduler, which halved the rate whenever the validation L1 loss stagnated for 10 epochs (patience). The discriminator's LR was kept constant throughout to preserve stable adversarial dynamics.

Model performance was monitored via quantitative metrics—specifically L1, SSIM, and LPIPS calculated on the validation set after each epoch—and qualitative visual inspections of generated samples every ten epochs. The checkpoint yielding the lowest validation L1 loss was retained as the best-performing model.

 doi:10.20944/preprints202605.0243.v1

3.6. Evaluation Metrics

The effectiveness of SAR-to-optical image translation depends not only on the choice of translation models but also on the methods employed for quality assessment. As discussed in [48], Image Quality Assessment (IQA) serves two key purposes: (i) to objectively evaluate the quality of results produced by different models, and (ii) to guide the optimization of network architectures and algorithms.

Moreover, evaluating SAR-to-optical translation requires assessing both pixel-level accuracy and perceptual quality. While some studies suggest metrics like SSIM, MSE, and LPIPS align well with human perception [48], our survey of 16 related studies (summarized in Table 1) indicates that SSIM, PSNR, and SAM remain the standard in current literature, whereas perceptual metrics like LPIPS appear less frequently. This trend aligns with the observations reported in the literature survey by [49]

Accordingly, we employed a hybrid evaluation strategy. We used PSNR and MAE/RMSE for pixel-wise fidelity, SSIM for structural integrity, and SAM for spectral consistency—crucial for multispectral data. To address perceptual realism often missed by pixel-based metrics, we also computed LPIPS.

Table 1. Frequency of common evaluation metrics in SAR-to-optical and cloud-removal studies.

Metric	References	Frequency
SSIM	[1,7,11,16,23,25,31,40,41,43,55,56]	12
PSNR	[1,2,7,23,25,31,40,41,43,55–57]	12
SAM	[1,2,5,16,23,40,41,43,55,58]	11
FID	[11,23,25,31,55,58]	6
RMSE	[2,5,7,16,40]	5
LPIPS	[1,7,31,56,58]	5
MAE	[2,40]	2
MSE	[41,57]	2

3.6.1. Metric Definitions

SSIM

The Structural Similarity Index (SSIM) [50] evaluates perceptual quality by comparing local luminance, contrast, and structure, thereby accounting for visual sensitivity to structural distortion:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{4}$$

where μ_x, μ_y are means, σ_x^2, σ_y^2 are variances, and σ_{xy} is the covariance of the pixel intensities computed over corresponding small image patches (typically using an 11×11 sliding window). Values close to 1 indicate strong structural similarity.

PSNR

The Peak Signal-to-Noise Ratio (PSNR) quantifies reconstruction quality by measuring the ratio of maximum signal power to the Mean Squared Error (MSE). For two images x and y, it is defined as

$$\mathrm{PSNR}(x,y) = 10 \cdot \log_{10}\left(\frac{MAX^2}{\mathrm{MSE}(x,y)}\right), \tag{5}$$

where the MSE is given by

$$\mathrm{MSE}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2. \tag{6}$$

Here, x_i and y_i denote the pixel values of the generated and reference images, N is the total number of pixels, and MAX represents the maximum pixel intensity (e.g., 255). While higher PSNR values indicate lower distortion, the metric is limited by its purely pixel-wise formulation and often correlates weakly with human visual perception [31].

 doi:10.20944/preprints202605.0243.v1

SAM

The Spectral Angle Mapper (SAM) [54] evaluates spectral fidelity by treating the spectrum of each pixel as an n-dimensional vector and measuring the angle between the generated vector x and the reference vector y:

$$\mathrm{SAM}(x, y) = \arccos\left(\frac{\langle x, y\rangle}{\|x\|_2 \cdot \|y\|_2}\right), \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product and $\| \cdot \|_2$ is the Euclidean norm. The result is typically expressed in degrees, with smaller values indicating higher spectral similarity. Because SAM considers only vector direction rather than magnitude, it is invariant to illumination changes, making it ideal for multispectral analysis [1].

LPIPS

The Learned Perceptual Image Patch Similarity (LPIPS) [52] measures distance in deep feature space using a pretrained network. It captures high-level semantic similarity closer to human visual judgment than traditional metrics:

$$\mathrm{LPIPS}(x, y) = \sum_l w_l \cdot \|f_l(x) - f_l(y)\|_2, \tag{8}$$

where $f_l(\cdot)$ represents the feature activations in layer l and w_l denotes learned weights. Lower values indicate higher perceptual realism. Since standard LPIPS models are trained on 3-channel inputs, we computed the metric using the representative RGB stack of the multispectral images.

MAE & RMSE

Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) provide direct quantitative assessments of pixel-wise error, with RMSE being more sensitive to outliers:

$$\mathrm{MAE} = \frac{1}{N}\sum |y_i - x_i|, \quad \mathrm{RMSE} = \sqrt{\frac{1}{N}\sum (y_i - x_i)^2}. \tag{9}$$

These metrics are calculated using the same pixel-wise notation (x_i, y_i, N) defined previously for PSNR/MSE.

A brief summary of the evaluated aspects, advantages, and limitations of each metric is provided in Table 2.

Table 2. Summary of evaluation metrics for SAR-to-multispectral translation.

Metric	Aspect Evaluated	Key Characteristic
SSIM	Structural similarity	Captures local texture/structure; intensity-based
PSNR	Pixel-level fidelity	Interpretable, but weakly correlated with perception
SAM	Spectral fidelity	Illumination invariant; ignores spatial context
LPIPS	Perceptual similarity	Deep feature-based; aligns with human vision
MAE/RMSE	Reconstruction accuracy	Direct error estimate; sensitive to outliers (RMSE)

4. Results

4.1. SAR-to-Optical Translation

We validate the trained model on both qualitative and quantitative metrics to assess its capability for direct SAR-to-optical image translation. Unlike most prior work that generates only 3–4 optical bands, the trained Pix2Pix model produces all 13 Sentinel-2 spectral bands.

Figure 2 presents some representative examples demonstrating the model's performance on diverse geographic and terrain conditions. In row (a), the model achieves strong color fidelity and accurately reconstructs spectral characteristics of vegetation and water bodies. Row (b) shows an urban scene where the model successfully delineates the city layout and captures the river traversing the area,

 doi:10.20944/preprints202605.0243.v1

preserving fine structural details. Row (c) demonstrates the model's ability to capture topographic features and surface relief variation from SAR backscatter intensity. Finally, row (d) depicts a coastal scene, where the model accurately reconstructs the land–water boundary and shoreline geometry while producing a plausible optical appearance. Across all examples, the generated optical images exhibit minimal artifacts and maintain structural consistency with ground truth.

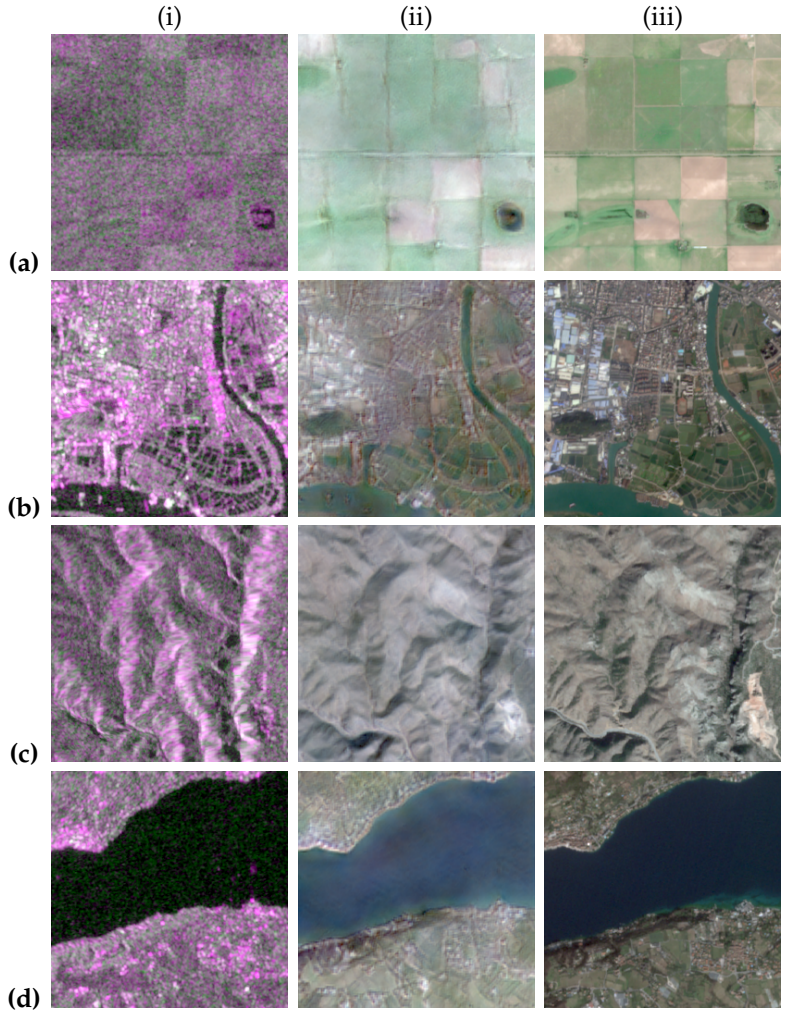


Figure 2. Qualitative results on SAR-to-Optical Translation. Columns: **(i)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(ii)** generated optical image, and **(iii)** ground-truth Sentinel-2 image. For visualization purposes, all optical images are depicted in RGB (B4, B3, B2) batch.

To quantitatively assess the translation quality, we evaluate the model using standard image fidelity and spectral metrics (Table 3). The model achieves SSIM of 0.888 and PSNR of 32.63 dB across all 13 spectral bands, with excellent spectral angle mapper (SAM) of 4.41°, indicating strong preservation of spectral characteristics. These metrics are within the range reported for recent SAR-to-optical translation methods. Detailed quantitative comparison with state-of-the-art cloud removal methods is provided in Section 4.3.

Table 3. Quantitative results on SAR-to-Optical Translation.

SSIM	PSNR (dB)	LPIPS	SAM (°)	MAE	RMSE
0.888	32.63	0.173	4.41	140.72	233.69

These results demonstrate that the model effectively reconstructs full-spectrum optical imagery from SAR inputs with high fidelity across structural, spectral, and radiometric dimensions.

 doi:10.20944/preprints202605.0243.v1

4.2. Results Across Individual Optical Bands

Another objective of this work was to assess the model's ability to reliably reconstruct each optical band individually and to evaluate the uniformity of its accuracy across the spectrum. For this purpose, the trained model was evaluated separately for all 13 Sentinel-2 bands. The results are summarized in Table 4.

When comparing reconstruction quality across bands, we focus on the Structural Similarity Index (SSIM). Unlike MAE or RMSE, which depend on the absolute magnitude and dynamic range of reflectance values (varying significantly between spectral bands), SSIM measures structural similarity based on local intensity patterns. While not entirely invariant to scale, SSIM provides a more robust and interpretable basis for cross-band comparison in this context.

Table 4. Per-band quantitative validation results. Each Sentinel-2 band's central wavelength, spectral designation, and native spatial resolution are listed for reference. Best results are highlighted in green and worst in red. Arrows (↑ / ↓) indicate whether higher or lower values denote better performance, respectively.

Band	PSNR (dB) ↑	SSIM ↑	Central Wavelength [nm]	Spectral / Resolution [m]
B1	36.53	**0.9758**	443	Aerosols / 60
B2	**37.49**	0.9506	490	Blue / 10
B3	35.67	0.9199	560	Green / 10
B4	32.84	0.8639	665	Red / 10
B5	33.68	0.9007	705	Red Edge / 20
B6	31.62	0.8536	740	Red Edge / 20
B7	30.37	0.8253	783	Red Edge / 20
B8	*29.62*	*0.7738*	842	NIR / 10
B8A	29.62	0.8071	865	Red Edge / 20
B9	33.99	0.9388	945	Water Vapour / 60
B10	32.12	0.9386	1375	Cirrus / 60
B11	29.92	0.8309	1610	SWIR / 20
B12	31.48	0.8586	2190	SWIR / 20

Notably, Band 8 (NIR), despite its native spatial resolution of 10 m, exhibits the lowest reconstruction performance among all spectral bands, including those with coarser resolutions, as illustrated in Figure 3. This indicates a weaker correlation between SAR backscatter and NIR reflectance compared to other spectral regions, likely due to their differing sensitivities to surface structure and vegetation properties. This limitation may also stem from the use of winter-only training samples, where vegetation-related information captured by the NIR band is comparatively scarce.

In contrast, the 60 m atmospheric correction bands—B1 (Aerosols), B9 (Water Vapour), and B10 (Cirrus)—are reconstructed with high structural similarity. B1 achieves the highest SSIM overall (0.9758). This high performance likely stems from the lower spatial complexity and smoother textural characteristics of these bands compared to land-cover bands. The model can more easily recover these spatially homogeneous signals, even after they are upsampled to 10 m.

These findings, however, contrast with those reported in [6], where the authors observed that 10 m and 20 m bands achieved the highest reconstruction quality, while 60 m bands performed worst. However, their study employed an optical-SAR fusion approach on the full SEN12-MS-CR dataset. In their case, the presence of input optical channels likely allowed the model to exploit the high spatial detail of the 10 m bands directly. In our SAR-only translation setting, the model must synthesize optical textures purely from SAR structure; thus, it performs best on bands with lower spatial complexity (atmospheric bands) and struggles where SAR-optical physical correlation is weak.

To visually complement this quantitative assessment, representative grayscale examples for each Sentinel-2 band are provided in Appendix A.

 doi:10.20944/preprints202605.0243.v1

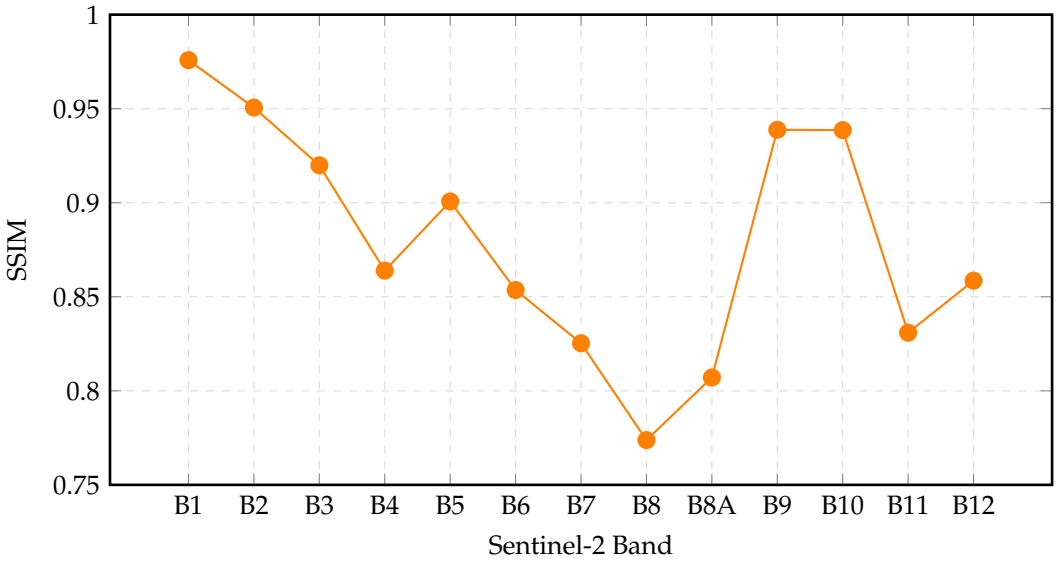


Figure 3. Per-band SSIM for the Pix2Pix model trained on the full winter subset.

4.3. Cloud Removal

To evaluate the practical utility of the learned SAR-to-optical translation, we assess the model on the cloud removal task using the SEN12-MS-CR benchmark (Section 3.2). This dataset provides co-registered triplets: Sentinel-1 SAR, cloud-contaminated Sentinel-2 optical observations, and ground-truth cloud-free Sentinel-2 references. Since the SAR modality is unaffected by atmospheric conditions, mapping SAR to optical effectively bypasses cloud occlusion.

For quantitative evaluation, we employ a subset of 2,000 samples from the winter portion of SEN12-MS-CR. To ensure consistency with the training phase, this evaluation data undergoes the identical preprocessing and clipping steps described in Section 3.2.1. Table 5 presents the quantitative performance. Despite not being explicitly trained for cloud removal or on the SEN12-MS-CR dataset specifically, the proposed approach achieves strong results. The Spectral Angle Mapper (SAM) score of 3.86° indicates high spectral fidelity in the reconstructed observations, while the best-in-class PSNR (33.65 dB) and LPIPS (0.152) reflect superior reconstruction quality and perceptual similarity.

Table 5. Quantitative evaluation on SEN12-MS-CR for cloud removal.

Model	SSIM ↑	PSNR (dB) ↑	LPIPS ↓	SAM (°) ↓
DSen2-CR [6]	0.878	–	–	8.07
FAT [40]	0.880	31.85	–	4.19
DiffCR [31]	**0.902**	31.77	–	5.82
Pix2Pix (This Study)	0.899	**33.65**	**0.152**	**3.86**

Notably, the Pix2Pix model outperforms several specialized approaches, including DiffCR [31], a current state-of-the-art model for this benchmark. This is a significant finding, as Pix2Pix is frequently treated as a lower-performing baseline in related literature. These results suggest that when trained on high-quality paired data, the direct SAR-to-optical translation approach is highly competitive for cloud removal tasks.

Qualitative results, presented in Figure 4, further illustrate the model's effectiveness. The network successfully reconstructs cloud-free optical imagery solely from SAR inputs, regardless of the occlusion level. Rows (a) and (b) demonstrate the restoration of regions affected by thin cloud cover, preserving underlying structural and textural details. In partially occluded scenarios (c) and (d), the outputs maintain clear boundaries and consistent color representation. Furthermore, even in fully cloud-covered scenes such as (e), the model generates realistic optical imagery, effectively reproducing urban and structural features that are completely invisible in the cloud-contaminated reference.

 doi:10.20944/preprints202605.0243.v1

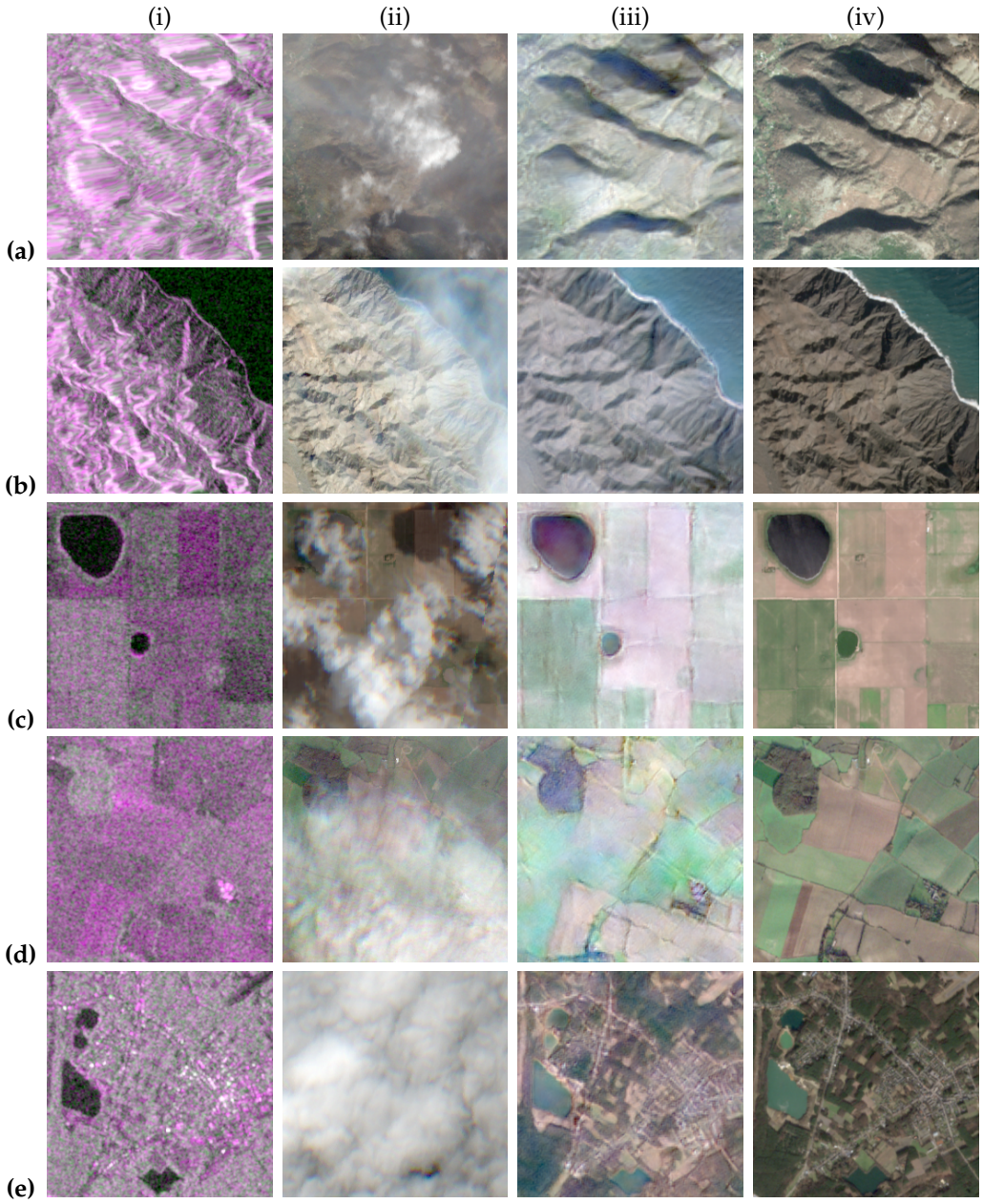


Figure 4. Qualitative results of the model trained on the full winter subset of the SEN12-MS dataset, evaluated on the complementary SEN12-MS-CR dataset for the **cloud removal** task. Columns: **(i)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(ii)** reference cloud-contaminated optical image, **(iii)** generated cloud-free optical image, and **(iv)** reference cloud-free optical image (ground truth).

In summary, both qualitative and quantitative metrics confirm the model's ability to generate high-quality, cloud-free optical images across all 13 spectral bands. Because the model relies exclusively on SAR input to generate the optical output, the density or presence of clouds in the target area does not influence the reconstruction process, emphasizing the robustness of SAR-to-optical translation in addressing cloud cover.

4.4. Ablation Studies

To assess the individual contributions of the loss functions and spectral bands to the model's reconstruction capability, we conducted two ablation studies. All models were trained using identical hyperparameters and datasets (20% of the winter subset) to ensure fair comparison.

 doi:10.20944/preprints202605.0243.v1

4.4.1. Impact of Loss Components

We evaluated four training configurations to isolate the effects of structural ($\mathcal{L}_{\text{SSIM}}$) and perceptual ($\mathcal{L}_{\text{LPIPS}}$) losses against a baseline adversarial setup:

1. $\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{L1}}$,
2. $\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{SSIM}}$,
3. $\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{LPIPS}}$, and
4. the full objective: $\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{LPIPS}}$.

Table 6 summarizes the quantitative results. The baseline configuration ($\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{L1}}$) yields the lowest performance across most metrics. This deficiency is also visually apparent, as the generated images lack textural detail (see Figure 5b).

Integrating $\mathcal{L}_{\text{SSIM}}$ yields the highest quantitative scores for SSIM and PSNR. However, qualitative inspection reveals that this configuration suffers from perceptual inconsistencies and reduced realism (Figure 5c). This aligns with prior findings that SSIM may overemphasize low-level intensity matching at the expense of high-frequency texture fidelity [59]. Conversely, incorporating $\mathcal{L}_{\text{LPIPS}}$ significantly improves edge retention and textural realism (Figure 5d), reflected in the best LPIPS score of 0.213, though at the cost of slight color deviations compared to the SSIM-based model.

The full combination of losses achieves the optimal trade-off. It maintains competitive pixel-level accuracy (MAE/RMSE) while ensuring structural coherence and perceptual realism, as evidenced in Figure 5(e). However, it introduces slight smoothing compared to the LPIPS-only configuration, a tendency likely driven by $\mathcal{L}_{\text{SSIM}}$. This confirms the necessity of a composite loss function that balances the sharpening effect of LPIPS with the structural regularization of SSIM for effective SAR-to-optical translation, consistent with Similar results reported in [2,25].

Table 6. Quantitative ablation results across loss configurations. Best values are in **bold**.

Configuration	SSIM ↑	PSNR ↑	LPIPS ↓	SAM (°) ↓	MAE ↓	RMSE ↓
$\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{L1}}$	0.820	26.38	0.287	7.88	229	441
+ $\mathcal{L}_{\text{SSIM}}$	**0.862**	**27.67**	0.399	**6.67**	198	**380**
+ $\mathcal{L}_{\text{LPIPS}}$	0.842	27.58	**0.213**	7.05	201	385
Full Combination	0.859	27.65	0.224	6.71	**195**	382

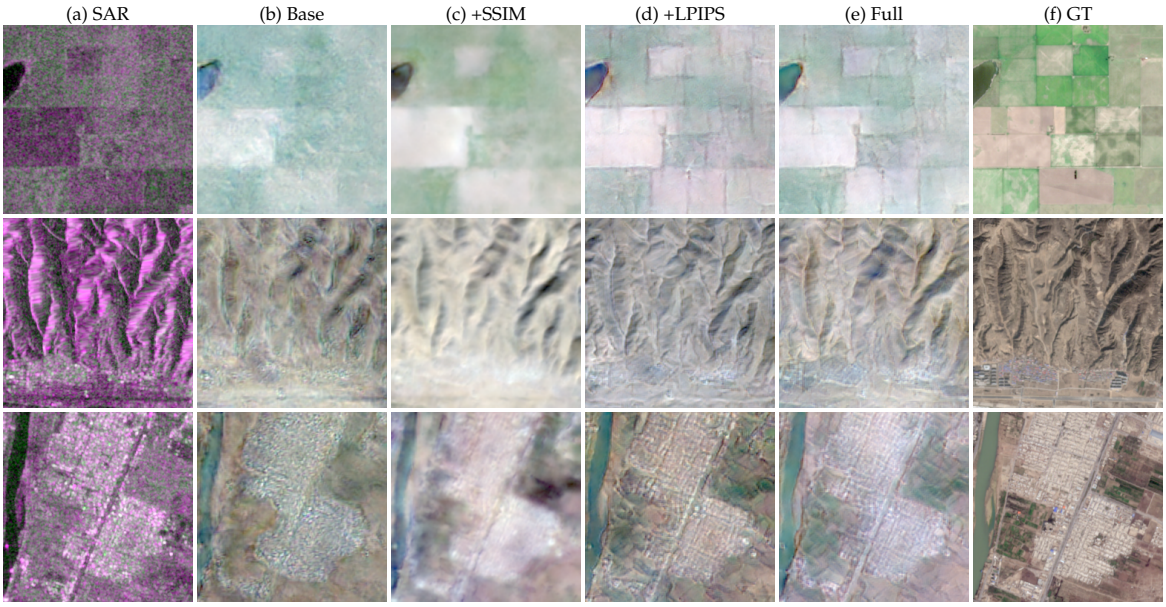


Figure 5. Visual comparison of loss configurations. (a) Input SAR, (b) Baseline ($\mathcal{L}_{\text{GAN}}+\mathcal{L}_{\text{L1}}$), (c) Baseline w/ SSIM, (d) Baseline w/ LPIPS, (e) Proposed full loss, (f) Ground Truth.

 doi:10.20944/preprints202605.0243.v1

4.4.2. Influence of 60m Spectral Bands

Sentinel-2 includes three 60 m resolution bands (B1, B9, B10) primarily used for atmospheric correction. We hypothesized that including these lower-resolution bands might introduce noise or training instability due to upsampling artifacts, since they were resampled to 10 m. To test this, we trained a model strictly excluding these bands.

Contrary to this hypothesis, excluding the 60 m bands did not improve performance. Instead, radiometric fidelity decreased, with both SSIM and PSNR dropping compared to the full 13-band model (Table 7). Interestingly, the Spectral Angle Mapper (SAM) score improved slightly when excluding these bands (6.33° vs. 6.71°). This indicates a trade-off: the model trained on fewer bands learned tighter spectral relationships among the remaining subsets (improving SAM) but lacks the enhanced image quality derived from the atmospheric correction information in the 60 m bands. It is worth noting, however, that these differences are marginal. Similarly, qualitative inspection in Figure 6 reveals only subtle differences in favour of including the full spectrum.

Table 7. Performance comparison: All 13 bands vs. excluding 60 m bands (B1, B9, B10).

Setting	SSIM ↑	PSNR (dB) ↑	SAM (°) ↓
All 13 bands	**0.859**	**27.65**	6.71
Excluding 60m	0.826	26.47	**6.33**

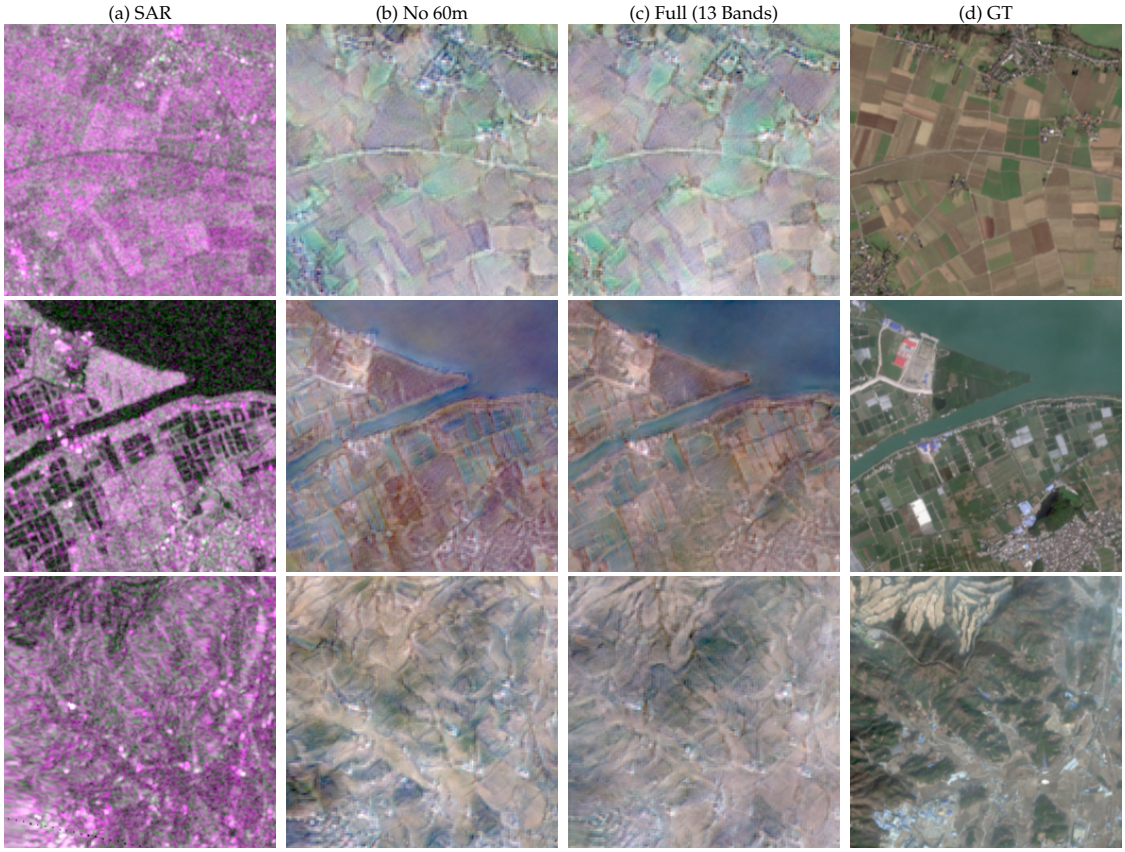


Figure 6. Qualitative results comparing models trained with and without the 60 m bands. (a) SAR input, (b) 10-band model output, (c) 13-band model output, (d) Ground Truth.

This trend is further confirmed by the per-band analysis in Table 8. While SSIM remains relatively stable across configurations, PSNR drops for nearly all 10 m and 20 m bands when the atmospheric bands are removed, although the absolute differences are small.

 doi:10.20944/preprints202605.0243.v1

Table 8. Per-band performance comparison. **No 60m**: Model trained without B1, B9, B10. **13**: Full model.

Band	SSIM (No 60m)	SSIM (13)	PSNR (No 60m)	PSNR (13)
B1	—	0.9634	—	29.749
B2	**0.9431**	0.9430	**34.15**	34.06
B3	**0.9064**	0.9060	31.69	**31.72**
B4	**0.8348**	0.8340	28.04	**28.15**
B5	0.8707	**0.8729**	28.10	**28.26**
B6	0.8199	**0.8231**	26.74	**26.93**
B7	0.7890	**0.7925**	25.69	**25.88**
B8	0.7375	**0.7416**	25.53	**25.74**
B8A	0.7687	**0.7722**	25.10	**25.31**
B9	—	0.8666	—	24.617
B10	—	0.8826	—	24.571
B11	0.7774	**0.7809**	23.40	**23.73**
B12	0.8034	**0.8073**	24.96	**25.26**

5. Discussion and Future Work

Despite the remarkable performance achieved by the proposed framework, specific limitations remain. Since the translation relies solely on SAR data, which inherently contains speckle noise, the trained model struggles to generate realistic optical images when the SAR inputs lack distinct structural information. In such cases, the model appears unable to discern meaningful spatial patterns and instead interprets some parts of the scene as noise, resulting in noise-like optical outputs, as illustrated in Figure 7.

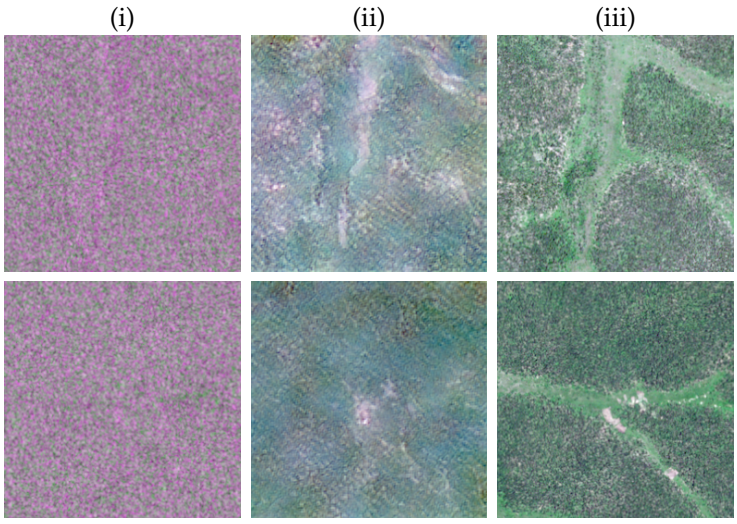


Figure 7. Qualitative examples illustrating the limitation of the SAR-to-optical translation model when the SAR input lacks clear structural information. Columns: **(i)** SAR input, **(ii)** model-generated optical image, and **(iii)** reference cloud-free Sentinel-2 image.

To address these generative artifacts, future research should consider alternative architectures such as Diffusion Models. Unlike GANs, which can suffer from training instability, diffusion models explicitly learn to reverse noise processes to synthesize data. Recent state-of-the-art performance on the SEN12-MS-CR dataset by methods like *DiffCR* [31] suggests that diffusion-based approaches [11,13,42,61] may offer superior fidelity and stability for SAR-to-optical translation. Additionally, performance could be enhanced by integrating attention mechanisms. Approaches such as the Spatial Attention GAN (SpA-GAN) [57] demonstrate that adaptively weighting features allows the network to prioritize relevant spatial details while suppressing noise or occlusions. Incorporating similar attention modules into SAR-to-optical architectures would likely improve feature consistency and spectral accuracy in complex scenes.

 doi:10.20944/preprints202605.0243.v1

Beyond that, reducing noise in both image domains as much as possible before training and application could enhance the results as well. Since the imaging instruments are based on completely different technologies, different types of noise occur in principle. However, the different signal characteristics generate different noise through interactions with the Earth's surface and the atmosphere: while SAR noise is mostly induced by backscattering effects, optical noise is rather caused by the atmosphere's water content and other aerosols. The incompatibility of noise between both domains remains a challenge in AI-based SAR-to-optical translation.

The results demonstrate the potential of genAI-based methods to fill data gaps in a given image domain by data from another image domain. Since the methods are not restricted to SAR and optical images, they could also be applied for other purposes, such as the reconstruction of historic images, or translating between image data of different spectral characteristics, such as multi-spectral, hyperspectral, RGB or B/W.

6. Conclusions

This study explored the potential of SAR-to-optical image translation as a generative framework for synthesizing multispectral optical imagery from radar data, specifically aiming to mitigate the limitations of cloud-contaminated optical remote sensing. Using co-registered Sentinel-1 and Sentinel-2 data from the SEN12-MS dataset, we trained a Pix2Pix conditional generative adversarial network (cGAN) to translate dual-polarized SAR inputs into full-spectrum optical outputs across all 13 Sentinel-2 bands.

Our results confirm that meaningful spectral and spatial correlations exist between the SAR and optical domains, enabling the reliable reconstruction of high-fidelity optical imagery. Reconstruction quality varied across spectral bands, reflecting inherent differences in wavelength sensitivity and signal characteristics. Notably, the model's performance in cloud removal demonstrated that SAR-to-optical translation can effectively generate cloud-free imagery without explicit training for this task, achieving results comparable to or surpassing several state-of-the-art approaches. Furthermore, ablation experiments indicated that combining SSIM and LPIPS losses with the standard GAN objective significantly enhances reconstruction consistency and perceptual realism.

While these findings validate the feasibility of SAR-to-optical translation, certain challenges remain. Model performance tends to decrease in textureless or spectrally complex regions, and temporal generalization is currently limited by training exclusively on winter data. Overall, this work establishes that GAN-based architectures can reconstruct full-spectrum, cloud-free optical imagery from SAR data with competitive accuracy, reinforcing the potential of generative approaches to enhance the temporal continuity of optical remote sensing.

Author Contributions: Conceptualization, A.A. and P.H.; methodology, A.A.; software, A.A.; validation, A.A. and P.H.; formal analysis, P.H.; investigation, A.A. and P.H.; data curation, A.A.; writing—original draft preparation, A.A.; writing—review and editing, P.H. and A.A.; visualization, A.A.; supervision, P.H. All authors have read and agreed to the published version of the manuscript.

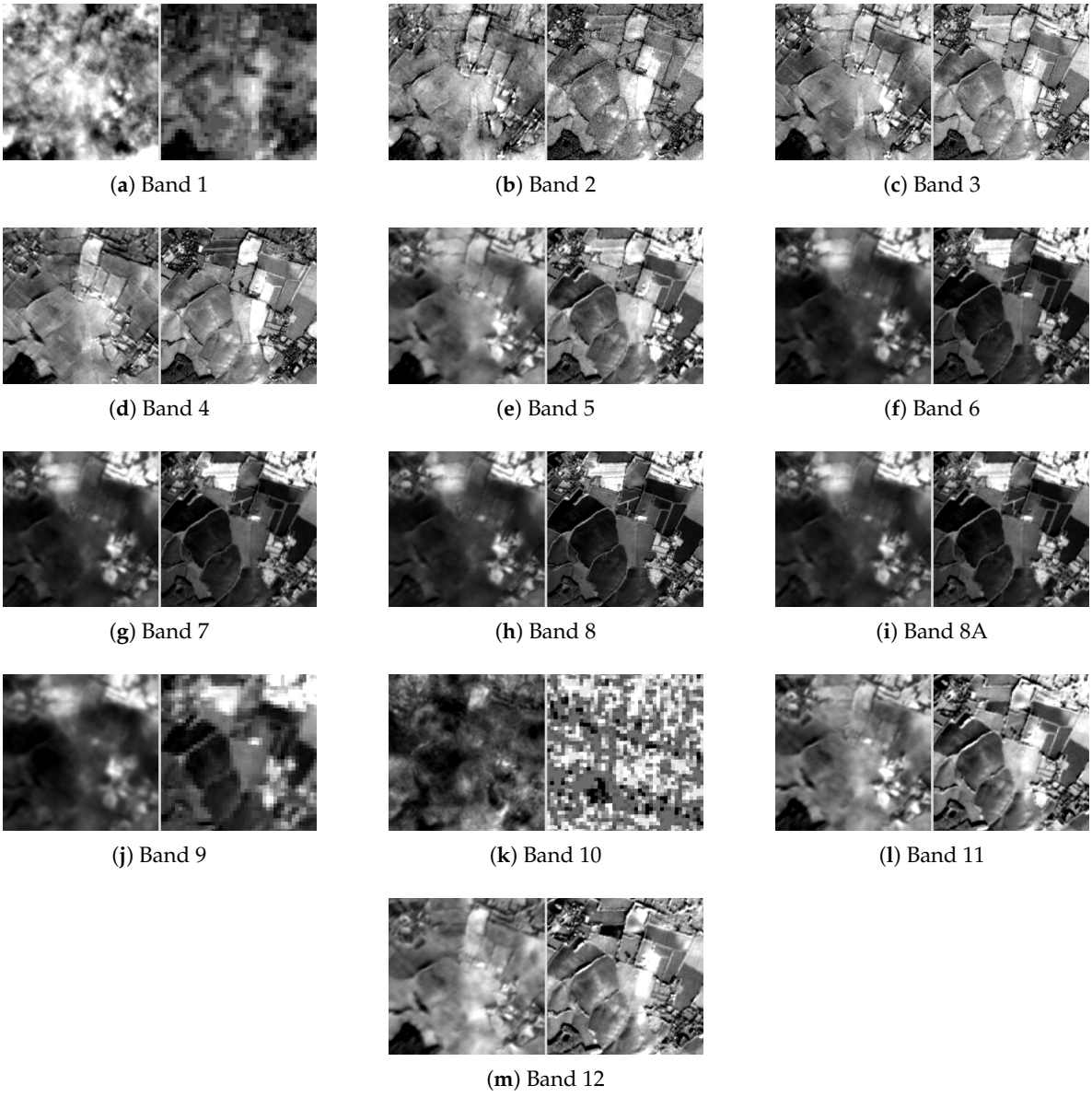
Funding: This research received no external funding

Data Availability Statement: The data presented in this study are available in the library of the Technical University of Munich (TUM) at https://doi.org/10.14459/2019mp1474000, reference number [38].

Conflicts of Interest: The authors declare no conflicts of interest

 doi:10.20944/preprints202605.0243.v1

Appendix A. Bandwise Grayscale Reconstructions



(**a**) Band 1 (**b**) Band 2 (**c**) Band 3

(**d**) Band 4 (**e**) Band 5 (**f**) Band 6

(**g**) Band 7 (**h**) Band 8 (**i**) Band 8A

(**j**) Band 9 (**k**) Band 10 (**l**) Band 11

(**m**) Band 12

Figure A1. Bandwise grayscale reconstructions for all Sentinel-2 bands. For all bands: generated left, ground truth right.

References

1. Liu, Y.; Han, Q.; Yang, H.; Hu, H. High-Resolution SAR-to-Multispectral Image Translation Based on S2MS-GAN. *Remote Sens.* **2024**, *16*, 4045.
2. Darbaghshahi, F.N.; Mohammadi, M.R.; Soryani, M. Cloud Removal in Remote Sensing Images Using Generative Adversarial Networks and SAR-to-Optical Image Translation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–9.
3. Shen, K.; Vivone, G.; Yang, X.; Lolli, S.; Schmitt, M. A benchmarking protocol for SAR colorization: From regression to deep learning approaches. *Neural Netw.* **2024**, *169*, 698–712.
4. Wang, Z.; Zhao, L.; Meng, J.; Han, Y.; Li, X.; Jiang, R.; Chen, J.; Li, H. Deep Learning-Based Cloud Detection for Optical Remote Sensing Images: A Survey. *Remote Sens.* **2024**, *16*, 4583.
5. Grohnfeldt, C.; Schmitt, M.; Zhu, X. A Conditional Generative Adversarial Network to Fuse Sar And Multispectral Optical Data For Cloud Removal From Sentinel-2 Images. In Proceedings of the IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1726–1729.

 doi:10.20944/preprints202605.0243.v1

6. Meraner, A.; Ebel, P.; Zhu, X.X.; Schmitt, M. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 333–346.
7. Ning, J.; Xie, L.; Yin, J.; Liu, Y. Cloud Removal Advances: A Comprehensive Review and Analysis for Optical Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 15914–15930.
8. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
9. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239.
10. Gao, J.; Yuan, Q.; Li, J.; Zhang, H.; Su, X. Cloud Removal with Fusion of High Resolution Optical and SAR Images Using Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 191.
11. Bai, X.; Pu, X.; Xu, F. Conditional Diffusion for SAR to Optical Image Translation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5.
12. Zhao, W.; Jiang, N.; Liao, X.; Zhu, J. HVT-cGAN: Hybrid Vision Transformer cGAN for SAR-to-Optical Image Translation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–17.
13. Bai, X.; Xu, F. SAR to Optical Image Translation with Color Supervised Diffusion Model. In Proceedings of the IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 963–966.
14. Enomoto, K.; Sakurada, K.; Wang, W.; Fukui, H.; Matsuoka, M.; Nakamura, R.; Kawaguchi, N. Filmy Cloud Removal on Satellite Imagery with Multispectral Conditional Generative Adversarial Nets. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1533–1541.
15. He, W.; Yokoya, N. Multi-Temporal Sentinel-1 and -2 Data Fusion for Optical Image Simulation. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 389.
16. Liu, P.; Li, J.; Wang, L.; He, G. Remote Sensing Data Fusion With Generative Adversarial Networks: State-of-the-art methods and future research directions. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 295–328.
17. Schmitt, M.; Hughes, L.H.; Körner, M.; Zhu, X.X. Colorizing Sentinel-1 SAR images using a variational autoencoder conditioned on Sentinel-2 imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *XLII-2*, 1045–1051.
18. Zhang, W.; Xu, M. Translate SAR Data into Optical Image Using IHS and Wavelet Transform Integrated Fusion. *J. Indian Soc. Remote Sens.* **2019**, *47*, 125–137.
19. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661.
20. Fuentes Reyes, M.; Auer, S.; Merkle, N.; Henry, C.; Schmitt, M. SAR-to-Optical Image Translation Based on Conditional Generative Adversarial Networks—Optimization, Opportunities and Limits. *Remote Sens.* **2019**, *11*, 2067.
21. Wang, L.; Xu, X.; Yu, Y.; Yang, R.; Gui, R.; Xu, Z.; Pu, F. SAR-to-Optical Image Translation Using Supervised Cycle-Consistent Adversarial Networks. *IEEE Access* **2019**, *7*, 129136–129149.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
23. Zhao, W.; Jiang, N.; Liao, X.; Zhu, J. HVT-cGAN: Hybrid Vision Transformer cGAN for SAR-to-Optical Image Translation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–17.
24. Schmitt, M.; Hughes, L.H.; Zhu, X.X. The SEN1-2 dataset for deep learning in SAR-optical data fusion. In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*; 2018; Volume IV-1, pp. 141–146.
25. Park, S.; Lee, H.; Lee, S. SAR-to-Optical Image Translation Using Vision Transformer-Based CGAN. *IEEE Sensors J.* **2025**, *25*, 18503–18514.
26. Kwak, G.-H.; Park, N.-W. Assessing the Potential of Multi-Temporal Conditional Generative Adversarial Networks in SAR-to-Optical Image Translation for Early-Stage Crop Monitoring. *Remote Sens.* **2024**, *16*, 1199.
27. Huang, B.; Li, Y.; Han, X.; Cui, Y.; Li, W.; Li, R. Cloud Removal From Optical Satellite Imagery With SAR Imagery Using Sparse Representation. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1046–1050.
28. Xu, M.; Jia, X.; Pickering, M.; Plaza, A.J. Cloud Removal Based on Sparse Representation via Multitemporal Dictionary Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2998–3006.
29. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv* **2018**, arXiv:1611.07004.
30. Ebel, P.; Sainte Fare Garnot, V.; Schmitt, M.; Wegner, J.D.; Zhu, X.X. UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series. *arXiv* **2023**, arXiv:2304.05464.

 doi:10.20944/preprints202605.0243.v1

31. Zou, X.; Li, K.; Xing, J.; Zhang, Y.; Wang, S.; Jin, L.; Tao, P. DiffCR: A Fast Conditional Diffusion Framework for Cloud Removal From Optical Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14.
32. Ebel, P.; Meraner, A.; Schmitt, M.; Zhu, X.X. Multisensor Data Fusion for Cloud Removal in Global and All-Season Sentinel-2 Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5866–5878.
33. Zhang, Y.; Guindon, B.; Cihlar, J. An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images. *Remote Sens. Environ.* **2002**, *82*, 173–187.
34. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 1956–1963.
35. Xu, F.; Shi, Y.; Ebel, P.; Yu, L.; Xia, G.-S.; Yang, W.; Zhu, X.X. GLF-CR: SAR-Enhanced Cloud Removal with Global-Local Fusion. *arXiv* **2022**, arXiv:2206.02850.
36. Mvogo, J.N.; Noumsi, W.A.V.; Wirba, P.B. Exploration of machine learning techniques for cloud removal and gap filling on Sentinel-2 time series images for better exploitation in far North Cameroon. *Discover Appl. Sci.* **2025**, *7*, 843.
37. Hofmann, P.; Trofanisin, N.; Wöllmann, S. Automatic Delineation of Burned Forest Areas from Satellite Imagery to Analyze and Manage Wildfires. In Proceedings of the 2024 14th International Conference on Advanced Computer Information Technologies (ACIT), Ceske Budejovice, Czech Republic, 16–18 September 2024; pp. 766–771.
38. Schmitt, M.; Hughes, L.H.; Qiu, C.; Zhu, X.X. 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*; 2019; Volume IV-2/W7, pp. 153–160.
39. Ebel, P.; Xu, Y.; Schmitt, M.; Zhu, X.X. SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14.
40. Xiang, X.; Tan, Y.; Yan, L. Cloud-Guided Fusion With SAR-to-Optical Translation for Thick Cloud Removal. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15.
41. Li, C.; Liu, X.; Li, S. Transformer Meets GAN: Cloud-Free Multispectral Image Reconstruction via Multisensor Data Fusion in Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13.
42. Aydin, K.; Hanna, J.; Borth, D. SAR-to-RGB Translation with Latent Diffusion for Earth Observation. *arXiv* **2025**, arXiv:2504.11154.
43. Liu, R.; Meng, S.; Peng, Y.; Tian, X. TransFusion-CR: Two-Phase SAR-to-Optical Translation and Deep Feature Fusion for Cloud Removal. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11.
44. Fu, X.; Kouyama, T.; Seki, S.; Nakamura, R.; Yoshikawa, I. Advanced SAR-To-Optical Image Translation Techniques using Jaxa's High-Resolution Land-Use and Land-Cover Map. In Proceedings of the IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 7367–7370.
45. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
46. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.K.; Wang, Z.; Smolley, S.P. Least Squares Generative Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2813–2821.
47. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
48. Zhang, J.; Zhou, J.; Li, M.; Zhou, H.; Yu, T. Quality Assessment of SAR-to-Optical Image Translation. *Remote Sens.* **2020**, *12*, 3472.
49. Xiong, Q.; Li, G.; Yao, X.; Zhang, X. SAR-to-Optical Image Translation and Cloud Removal Based on Conditional Generative Adversarial Networks: Literature Survey, Taxonomy, Evaluation Indicators, Limits and Future Directions. *Remote Sens.* **2023**, *15*, 1137.
50. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612.
51. Tanchenko, A. Visual-PSNR measure of image quality. *J. Vis. Commun. Image Represent.* **2014**, *25*, 874–878.
52. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
53. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv* **2018**, arXiv:1706.08500.

 doi:10.20944/preprints202605.0243.v1

54. Kruse, F.A.; Lefkoff, A.B.; Boardman, J.W.; Heidebrecht, K.B.; Shapiro, A.T.; Barloon, P.J.; Goetz, A.F.H. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote Sens. Environ.* **1993**, *44*, 145–163.
55. Chen, Y.; Zhu, Z.; Huang, Y.; Wang, P.; Huang, B.; Mura, M.D. MSF: A Multi-Scale Fusion Generative Adversarial Network for SAR-to-Optical Image Translation. In Proceedings of the IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 9058–9061.
56. Guo, Z.; Liu, J.; Cai, Q.; Zhang, Z.; Mei, S. Learning SAR-to-Optical Image Translation via Diffusion Models With Color Memory. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 14454–14470.
57. Pan, H. Cloud Removal for Remote Sensing Imagery via Spatial Attention Generative Adversarial Network. *arXiv* **2020**, arXiv:2009.13015.
58. Kim, S.-H.; Chung, D. Conditional Brownian Bridge Diffusion Model for VHR SAR to Optical Image Translation. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5.
59. Nilsson, J.; Akenine-Möller, T. Understanding SSIM. *arXiv* **2020**, arXiv:2006.13846.
60. Lin, D.; Xu, G.; Wang, X.; Wang, Y.; Sun, X.; Fu, K. A Remote Sensing Image Dataset for Cloud Removal. *arXiv* **2019**, arXiv:1901.00600.
61. Wang, M.; Hu, S.; Song, Y.; Shi, Y. SAR-DeCR: Latent Diffusion for SAR-Fused Thick Cloud Removal. *Remote Sens.* **2025**, *17*, 2241.