Technische Hochschule Deggendorf
Fakultät Angewandte Informatik

Bachelor Künstliche Intelligenz

Erzeugung optischer Fernerkundungsdaten (Sentinel-2) auf Basis von Radar-Fernerkundungsdaten (Sentinel-1) mittels generativer KI

Generation of optical remote sensing data (Sentinel-2) based on radar remote sensing data (Sentinel-1) using generative AI

Bachelorarbeit zur Erlangung des akademischen Grades:

Bachelor of Science (B.Sc.)

an der Technischen Hochschule Deggendorf

Vorgelegt von:
Ahmed Attia
Matrikelnummer: 00815907

Prüfungsleitung:
Dr. Peter Hofmann

Am: XX. Monat 20XX

Erklärung

Name des Studierenden: Ahmed Attia

Name des Betreuenden: Dr. Peter Hofmann

Thema der Abschlussarbeit:

Erzeugung optischer Fernerkundungsdaten (Sentinel-2) auf Basis von Radar-Fernerkundungsdaten (Sentinel-1) mittels generativer KI

. . . .

1. Ich erkläre hiermit, dass ich die Abschlussarbeit gemäß § 35 Abs. 7 RaPO (Rahmenprüfungsordnung für die Fachhochschulen in Bayern, BayRS 2210-4-1-4-1-WFK) selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

   Deggendorf, ...
   Datum Unterschrift des Studierenden

2. Ich bin damit einverstanden, dass die von mir angefertigte Abschlussarbeit über die Bibliothek der Hochschule einer breiteren Öffentlichkeit zugänglich gemacht wird:

   ◯ Nein

   ◯ Ja, nach Abschluss des Prüfungsverfahrens

   ◯ Ja, nach Ablauf einer Sperrfrist von . . . Jahren.

   Deggendorf, ...
   Datum Unterschrift des Studierenden

Bei Einverständnis des Verfassenden vom Betreuenden auszufüllen:

Eine Aufnahme eines Exemplars der Abschlussarbeit in den Bestand der Bibliothek und die Ausleihe des Exemplars wird:

◯ Befürwortet

◯ Nicht befürwortet

Deggendorf, ...
Datum Unterschrift des Betreuenden

Contents

Abstract 1

1. Introduction 2

2. Background & Motivation 4

2.1. Remote Sensing . . . 4
2.2. Copernicus: Europe's eyes on Earth . . . 6
2.2.1. Sentinel-1 . . . 7
2.2.2. Sentinel-2 . . . 8
2.3. Cloud Removal . . . 10
2.4. Generative Artificial Intelligence . . . 13
2.5. SAR-to-optical image translation . . . 15
2.6. SAR-to-Optical Image Translation for Cloud Removal . . . 17
2.7. Application and Relevance to KIWA Project . . . 19

3. Methodology & Data Basis 21

3.1. Problem Formulation . . . 21
3.2. Datasets . . . 22
3.2.1. SEN12-MS . . . 22
3.2.2. SEN12 datasets Family . . . 23
3.2.3. Subset Selection & Preprocessing . . . 24
3.3. Pix2Pix Model . . . 25
3.4. Training Procedure . . . 27
3.4.1. Experimental Setup . . . 27
3.4.2. Loss Functions . . . 28
3.4.3. Training Strategy . . . 29
3.4.4. Monitoring and Checkpointing . . . 31
3.5. Evaluation Metrics . . . 31

4. Results 36

4.1. Results on 20% of the Winter Subset . . . 36

4.2. Results on the Full Winter Subset . . . 38
4.3. Results Across Individual Optical Bands . . . 41
4.4. Results on Cloud Removal . . . 42
4.5. Ablation Studies . . . 45
4.5.1. Effect of Loss Functions . . . 45
4.5.2. Effect of Excluding 60 m Bands . . . 47

5. Discussion **50**
5.1. Challenges Due to Inherent Model Characteristics . . . 50
5.2. Model-Specific Limitations of GAN-Based Translation . . . 51
5.3. Temporal Generalizability Across Different Seasons . . . 53
5.4. Aware per-Band Clipping . . . 53

6. Conclusion **56**

A. Intermediate Outputs during Training **57**

B. Bandwise Grayscale Reconstructions **58**

C. Results Across Different Seasons **60**

D. Value Distributions of Individual Optical Bands **63**

List of Abbreviations

Abbreviation	Full Form
RS	Remote Sensing
SAR	Synthetic Aperture Radar
GAN	Generative Adversarial Network
cGAN	Conditional Generative Adversarial Network
CNN	Convolutional Neural Network
DDPM	Denoising Diffusion Probabilistic Model
ESA	European Space Agency
GRD	Ground Range Detected
EW	Extra-Wide Swath Mode
IW	Interferometric Wide Swath Mode
WV	Wave Mode
LULC	Land Use and Land Cover
MODIS	Moderate Resolution Imaging Spectroradiometer
ROI	Region of Interest
SWIR	Shortwave Infrared
VNIR	Visible and Near Infrared
VV	Vertical–Vertical Polarization
VH	Vertical–Horizontal Polarization
NIR	Near Infrared
IQA	Image Quality Assessment
SSIM	Structural Similarity Index Measurement
FSIM	Feature Similarity Index Measurement
FSIM	Image Quality Assessment
DISTS	Deep Image Structure and Texture Similarity
PSNR	Peak Signal-to-Noise Ratio
SAM	Spectral Angle Mapper
FID	Fréchet Inception Distance
LPIPS	Learned Perceptual Image Patch Similarity
BCE	Binary Cross-Entropy

List of Tables

2.1. Overview of the Copernicus Sentinel missions 7
2.2. Sentinel-2 MSI spectral bands 9
2.3. Sentinel-2 Copernicus Services applications 9
2.4. Cloud removal method categories 12

3.1. Comparison of SEN12-family datasets. 25
3.2. Common evaluation metrics in SAR-to-optical and cloud removal 32
3.3. Evaluation metrics used summary 35

4.1. Quantitative results of 20% training winter subset 38
4.2. Quantitative results for different training data scales: 20% & 100% 39
4.3. Per-band validation results for full dataset training 41
4.4. Quantitative results on cloud removal 43
4.5. Quantitative ablation study across loss configurations 45
4.6. Overall performance when excluding 60m bands 47
4.7. Per-band performance when excluding 60m bands 49

5.1. Summary statistics of Sentinel-2 optical bands for the winter subset. 54

List of Figures

2.1. Timeline of remote sensing platform development 6
2.2. GAN architecture overview 14

3.1. ROIs distribution of the SEN12-MS Dataet 23
3.2. Sample SAR–optical pairs from SEN12-MS 24
3.3. Training and validation loss curves over 150 epochs. 30

4.1. Qualitative results of 20% training winter subset 37
4.2. Qualitative results for different training data scales: 20% & 100% 40
4.3. Per-band SSIM for the Pix2Pix model 42
4.4. Qualitative results on cloud removal 44
4.5. Qualitative ablation study across loss configurations 46
4.6. Qualitative results when excluding 60 m bands 48

5.1. Model limitation on structureless SAR inputs 52

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 and visual richness of optical observations. Bridging these complementary modalities, this thesis investigates SAR-to-optical image translation — a generative approach that synthesizes optical-like imagery from SAR inputs — using the Pix2Pix conditional generative adversarial network (cGAN). The thesis employs the winter subset of the SEN12-MS dataset, which offers globally distributed, co-registered Sentinel-1 and Sentinel-2 imagery. 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. 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, outperforming previous methods and the state-of-the-art DiffCR model. 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. Additionally, based on data analysis, a new per-band clipping strategy for optical data is proposed.

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 Earth observation systems. Optical sensors provide rich spectral and visual information suitable for human interpretation. This information is used, among others, in forest and agricultural monitoring [1]. However, they are inherently constrained by atmospheric conditions such as illumination variability and in particular cloud cover. This causes considerable data gaps in both the spatial and temporal domains [2]. In contrast, synthetic aperture radar (SAR) sensors operate in the microwave domain, offering all-weather, day-and-night imaging capabilities independent of sunlight or cloud interference. However, the backscatter-based nature of SAR imagery introduces challenges related to speckle noise, geometric distortions, and the absence of spectral color information [3, 4], necessitating advanced interpretation and analysis 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 advantages of radar acquisitions. Recent advances in generative artificial intelligence (GenAI), particularly in conditional generative adversarial networks (cGANs) [5] and diffusion models [6], have made it possible to learn complex mappings between SAR and optical domains with remarkable realism [7]. 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, most existing studies have focused primarily on reconstructing the visible RGB subset of optical imagery such as [4,7–10], with a few extending to the NIR range as [11, 12], leaving the full multispectral potential of missions such as Sentinel-2 largely underexplored. Furthermore, while SAR-to-optical translation is often evaluated qualitatively, systematic assessments of its reliability across individual spectral bands remain limited. 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

Filling the research gap, the present thesis investigates the use of generative models for

translating Sentinel-1 SAR imagery into multispectral Sentinel-2 optical data. The work is guided by three main objectives:

1. To validate SAR-to-optical image translation across the full 13 Sentinel-2 spectral bands.
2. To assess how reliably each optical band can be individually reconstructed and to what extent.
3. To evaluate the performance of SAR-to-optical image translation for cloud removal

Through these objectives, this thesis aims to provide a comprehensive and quantitative understanding of SAR-to-optical translation as a multimodal learning problem, highlighting its strengths, limitations, and potential for operational cloud-free optical data generation.

The remainder of this thesis is organized as follows: Chapter 2 establishes the theoretical foundation by outlining the principles of remote sensing, introducing the Sentinel missions, and reviewing key concepts in cloud removal and generative artificial intelligence, with a focus on SAR-to-optical image translation. Building on this foundation, the latter sections of the same chapter review the related literature, highlighting existing methods for SAR–optical data fusion and cloud removal, and identifying the research gaps that this work aims to address. Chapter 3 presents the methodological framework, describing dataset preparation, preprocessing steps, and the adopted Pix2Pix model, together with its training setup and evaluation procedure. The experimental results are reported in Chapter 4, which details the reconstruction performance across the full spectral range and assesses the model's capability for cloud removal. To further analyze the model's behavior, Section 4.5 presents ablation studies investigating the influence of different loss configurations and the impact of excluding specific spectral bands. Chapter 5 discusses the overall findings in relation to the research objectives, summarizing the main challenges encountered, the limitations of the approach, and directions for future work. Finally, Chapter 6 concludes the thesis with a summary of key contributions and closing remarks. The appendices provide supplementary material that supports and extends the analyses presented in the main chapters. They include intermediate model outputs observed during training, bandwise grayscale reconstructions, seasonal evaluation results, and value distributions of individual optical bands.

2. Background & Motivation

2.1. Remote Sensing

Remote sensing (RS) is commonly defined as the acquisition of information about an object through sensors without direct physical contact. This information is obtained by detecting and measuring the modifications the object induces in its surrounding fields, which may include electromagnetic, acoustic, or potential fields [13]. RS is characterized by its strong interdisciplinary nature. It draws upon a wide spectrum of fields, requiring practitioners to develop a broad foundational understanding of both natural and applied sciences. Effective research in remote sensing often involves collaboration with specialists in electromagnetic theory, spectroscopy, applied physics, geology, atmospheric sciences, oceanography, electrical engineering, and optical engineering [13]. Remote observations require an interaction of energy between the target and the sensor. In the case of passive sensors, the detected energy originates from external or natural sources, such as solar radiation reflected by the Earth's surface or thermal radiation emitted by the object itself. A prominent example is the *Landsat* program [1], which represents the longest continuously operating Earth observation mission. Over several decades, Landsat has generated a continuous global record, contributing significantly to environmental monitoring and Earth system science.

By contrast, active sensors generate their own energy pulses to illuminate the target and subsequently measure the portion of the signal that is reflected or backscattered. This capability allows them to operate independently of solar illumination and under a wide range of environmental conditions, including day or night and, in the case of microwave systems, through cloud cover and adverse weather [14]. The most widely used active sensing technologies are Radar (Radio Detection and Ranging) and LiDAR (Light Detection and Ranging). Radar systems transmit and receive microwave radiation, whereas LiDAR employs laser pulses in the optical domain. For Earth observation, radar remote sensing typically operates in three wavelength ranges: X-band (2.4–4.5 cm), C-band (4.5–7.5 cm), and L-band (15–30 cm) [15]. Among these, the most commonly employed monitoring system is Synthetic Aperture Radar (SAR), which is discussed in greater detail in Section 2.2.1. Both radar and LiDAR record properties of the

[1] https://landsat.gsfc.nasa.gov/

reflected signals to derive information about the observed surfaces and objects.

The term *Remote Sensing* was introduced in the early 1960s to denote techniques for observing the Earth from a distance, with particular reference to aerial photography, which represented the predominant sensing technology at that time [16]. With the advent of satellites, global and synoptic observations of Earth and other planetary environments have become possible. Earth-orbiting sensors provide essential data on atmospheric dynamics, cloud distribution, vegetation cover, and its seasonal variability. Their long-term operation and repetitive coverage enable the monitoring of rapidly changing processes, such as polar ice dynamics and tropical deforestation. Beyond Earth, planetary missions (orbiters, flybys, landers, and rovers) have extended similar observations to all major planets in the solar system. To date, every planet has been visited at least once [13].

The origins of remote sensing date back to the invention of photography in 1839, which soon after was applied to topographic mapping. By the mid-19th century, aerial photographs were obtained from balloons, followed later by kites, pigeons, and eventually airplanes—the latter marking a decisive step with Wilbur Wright's first aerial photographs in 1909. Aerial photography became essential during World War I and advanced further in the 1930s-1940s with the introduction of color and infrared-sensitive films, widely used during World War II for reconnaissance and camouflage detection [13, 16]. The postwar decades brought rapid technological progress with the development of radar and synthetic aperture radar (SAR), enabling high-resolution imaging independent of daylight or weather. Early rocket experiments in the late 1940s foreshadowed the space age, initiated by the launch of Sputnik in 1957. NASA's TIROS-1 satellite (1960) delivered the first global meteorological observations, while the launch of Landsat-1 in 1972 introduced systematic multispectral Earth observation, a program that continues today as the longest-running record of land surface change [13, 16].

Since the 1980s, remote sensing has expanded through international efforts such as SPOT (France, 1986), MOS-1 (Japan, 1987), and IRS-1 (India, 1988). The European Space Agency (ESA) [2] launched its first radar satellite, ERS-1, in 1991, and a second with comparable specifications in 1995. The 1990s and 2000s saw the rise of commercial satellites like IKONOS and QuickBird, offering very high-resolution imagery. Today, constellations of small satellites operated by private companies provide near-daily global coverage at meter-scale resolution. These advances—driven by improvements in optics, sensors, data transmission, and digital processing—have transformed remote sensing into a cornerstone of Earth system science, environmental monitoring, disaster response, and planetary exploration [16].

A summary of major milestones in the historical development of remote sensing platforms, from early balloon photography to modern satellite constellations, is illustrated in Figure 2.1.

[2] https://www.esa.int/

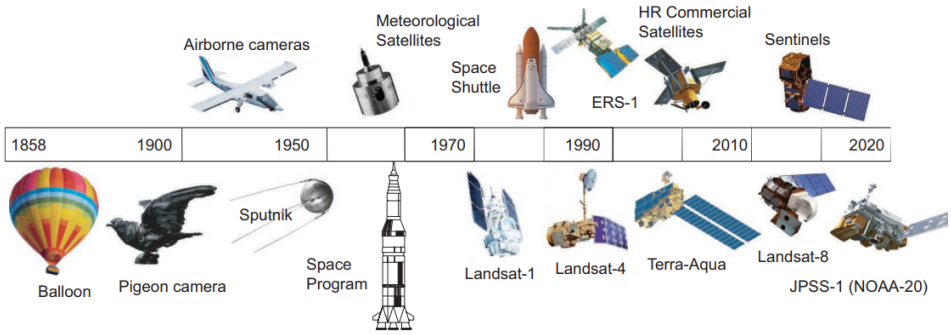


Figure 2.1.: Timeline of remote sensing platform development, from early airborne cameras to modern Earth observation satellites. Adapted from [16].

2.2. Copernicus: Europe's eyes on Earth

Copernicus, the Earth observation component of the European Union's Space Programme, is widely regarded as the most ambitious environmental monitoring initiative to date. It is funded, coordinated, and managed by the European Commission in cooperation with partners such as the European Space Agency (ESA) and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT)[3]. Named after the astronomer Nicolaus Copernicus[4], the programme integrates satellite and *in situ* observations (e.g., ground, airborne, and seaborne instruments) to provide reliable, timely environmental information across six domains: land, marine, atmosphere, emergency management, security, and climate change.

The Copernicus Space Component consists of a dedicated series of satellites known as the Sentinels, numbered from Sentinel-1 to Sentinel-6, each designed to serve specific operational needs. Table 2.1 summarises the main Sentinel missions, their sensing principles, observation domains, and primary applications. Together, they provide a comprehensive and complementary view of the Earth system.

Looking ahead, six Sentinel Expansion missions will extend the Copernicus capabilities, including hyperspectral imaging, polar topography, and carbon dioxide monitoring [17]. As this work focuses on SAR and optical data, only Sentinel-1 and Sentinel-2 are discussed in detail in the following sections.

[3] https://www.eumetsat.int/
[4] https://www.biography.com/scientists/nicolaus-copernicus

Table 2.1.: Overview of the Copernicus Sentinel missions, summarising their sensors, observation domains, and main applications. Sources: [17, 18].

Mission	Sensor Type / Band	Observation Domain	Main Applications / Purpose
Sentinel-1	C-band SAR	Land, Ocean, Ice	Monitoring land deformation, sea-ice dynamics, flooding, and earthquakes
Sentinel-2	Multispectral Optical (13 bands)	Land, Vegetation, Water	High-resolution imaging for agriculture, forestry, land use, and disaster management
Sentinel-3	Optical, Thermal, and Radar Instruments	Ocean, Land	Measuring sea surface topography, temperature, and colour for oceanography and climate studies
Sentinel-4	UV–VIS–NIR Spectrometer	Atmosphere	Monitoring atmospheric composition and air quality over Europe and North Africa
Sentinel-5P	UV–VIS–NIR Spectrometer	Atmosphere	Observing air quality, ozone, and UV radiation for climate and pollution assessment
Sentinel-6	Radar Altimeter	Ocean	High-precision monitoring of sea level rise and ocean circulation

2.2.1. Sentinel-1

Sentinel-1, launched on 3 April 2014, constitutes the radar component of the European Copernicus Programme. The mission is designed as a constellation of two sun-synchronous, near-polar orbiting satellites in the same orbital plane, separated by 180° in phase. Equipped with C-band synthetic aperture radar (SAR) operating at 5.4 $\sim$ GHz (corresponding to a wavelength of $\sim$5.55 cm), Sentinel-1 provides continuous, all-weather, day-and-night imaging capability. Since Sentinel-1 operates at microwave wavelengths, it is capable of measuring the physical properties of the target surface, whereas optical sensors are predominantly responsive to its chemical composition [15]. Sentinel-1A was followed by Sentinel-1B in 2016, which ceased operations after an anomaly in 2021 and was subsequently replaced by Sentinel-1C in 2024. The SAR instrument actively transmits microwave signals towards the Earth and records the backscattered response. Both amplitude and phase are preserved, enabling the reconstruction of high-resolution images. Polarisation diversity further enhances information extraction, as different surfaces exhibit characteristic scattering signatures, supporting classification and retrieval of geophysical parameters.

Sentinel-1 operates in four exclusive acquisition modes: Stripmap (SM), Interferometric Wide Swath (IW), Extra-Wide Swath (EW), and Wave (WV). These modes achieve spatial resolutions down to $\sim$ 5m and swath widths of up to $\sim$ 400km. The system supports single (HH or VV) and dual (HH+HV or VV+VH) polarisation, where H and V denote horizontal and vertical signal orientations, respectively. In this context, HH indicates that the radar transmits and receives signals horizontally, while VV corresponds to vertical transmission and reception. For HV and VH, the radar transmits in one orientation and receives in the orthogonal one, capturing cross-

polarised backscatter information. While SM, IW, and EW modes allow a duty cycle of up to 30 minutes per orbit, WV mode extends this to 75 minutes. Over land, IW mode with VV+VH polarisation is the primary operational configuration, balancing revisit performance, service requirements, and the creation of a consistent long-term archive. For open-ocean observations, WV mode with VV polarisation is predominantly employed, while EW mode is mainly used for sea-ice monitoring and maritime surveillance in high-latitude regions. SM mode is activated only for small islands or in response to emergencies. Across all modes, products are provided at multiple processing levels, from raw SAR data (Level-0) to geophysical ocean products (Level-2 OCN).

The revisit capabilities of Sentinel-1 are particularly notable. In IW mode, a single satellite can achieve global coverage every 12 days, while the two-satellite constellation reduces the repeat cycle to six days, completing 175 orbits per cycle. These systematic observations, combined with advanced interferometric capabilities, enable the precise detection of land subsidence, structural deformation, and ground movements that are otherwise imperceptible. Such data are invaluable for urban planning, geohazard monitoring, and applications in mining, geology, and risk assessment for infrastructure and natural hazards [19].

2.2.2. Sentinel-2

Sentinel-2 is the optical imaging mission of the Copernicus Programme, designed to provide systematic, high-resolution observations over land and coastal regions. The mission consists of a constellation of two sun-synchronous satellites in the same orbital plane, phased 180° apart, ensuring global coverage with a revisit frequency of five days at the Equator. Sentinel-2A was launched in 2015, followed by Sentinel-2B in 2017 and Sentinel-2C in September 2024, the latter ensuring mission continuity as Sentinel-2A approaches the end of its operational lifetime. Each satellite carries a single payload: the Multi-Spectral Instrument (MSI). This passive optical sensor collects sunlight reflected from the Earth's surface, splitting the incoming radiation into two focal plane assemblies: one covering the visible and near-infrared (VNIR) and the other the shortwave infrared (SWIR). The instrument has a swath width of 290 km, which is considerably wider than comparable missions such as Landsat 5/7 (185 km) or SPOT-5 (120 km).

The MSI samples 13 spectral bands at three spatial resolutions: four bands at 10 m (Blue, Green, Red, and Near-Infrared), six bands at 20 m (red-edge and SWIR), and three bands at 60 m (aerosol, water vapour, and cirrus). These bands span the VNIR to SWIR regions of the electromagnetic spectrum and are tailored to applications including vegetation and crop monitoring, land cover mapping, water quality assessment, snow and ice monitoring, cloud screening, and atmospheric correction. An overview of the spectral bands is provided in Table 2.2.

Sentinel-2 imagery is systematically and freely available, supporting several Copernicus ser-

Table 2.2.: Sentinel-2 MSI spectral bands with central wavelength and spatial resolution [19].

Band	Central Wavelength [nm]	Resolution [m]
B1	443 (Aerosols)	60
B2	490 (Blue)	10
B3	560 (Green)	10
B4	665 (Red)	10
B5	705 (Red edge)	20
B6	740 (Red edge)	20
B7	783 (Red edge)	20
B8	842 (NIR)	10
B8a	865 (Red edge)	20
B9	945 (Water vapour)	60
B10	1375 (Cirrus)	60
B11	1610 (SWIR)	20
B12	2190 (SWIR)	20

vices, including the Copernicus Land Monitoring Service (CLMS), the Copernicus Marine Environment Monitoring Service (CMEMS), and the Copernicus Emergency Management Service (CEMS). These services, along with selected applications and their respective launch years, are illustrated in Table 2.3. Through its systematic, frequent, and global observations, Sentinel-2 has become a cornerstone of the Copernicus programme, enabling comprehensive environmental monitoring, sustainable resource management, and rapid disaster response worldwide [19].

Table 2.3.: Overview of Sentinel-2 Copernicus Services and their applications.

Copernicus Service	Operational Since	Sentinel-2 Application(s)
Copernicus Land Monitoring Service (CLMS)	2012	• Land cover and forest mapping • Crop monitoring • Ecosystem assessment • Climate change adaptation
Copernicus Emergency Management Service (CEMS)	2012	• Disaster response and rapid mapping • Flood monitoring • Fire assessment • Earthquake damage mapping
Copernicus Marine Environment Monitoring Service (CMEMS)	2014	• Turbidity estimation • Chlorophyll concentration mapping • Suspended particulate matter analysis • Bathymetry derivation • Ice monitoring and analysis

Together, Sentinel-1 and Sentinel-2 provide complementary SAR and optical observations, which form the basis of this thesis aiming to translate SAR imagery into its optical counterpart.

2.3. Cloud Removal

As briefly mentioned in the sections above, optical remote sensing imagery, such as Sentinel-2 products, represents a key source of Earth observation data. Compared to SAR observations, multispectral images contain rich spectral information and are readily interpretable by the human eye. Such data play an essential role in a wide range of applications, including environmental monitoring, resource exploration, and disaster assessment. While the quality and quantity of satellite observations have dramatically increased in recent years, one common problem persists for optical remote sensing imagery: **cloud cover**.

Based on findings from the International Satellite Cloud Climatology Project (ISCCP), average global cloud cover surpasses 66% [2, 20, 21], with 55% over land surface alone [2], preventing optical satellites from acquiring valuable information about the Earth's surface due to the frequent presence of clouds in the imagery. In contrast to SAR instruments, optical sensors cannot penetrate clouds, resulting in considerable data gaps in both the 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 [2]. 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 [22].

Consequently, removing clouds and obtaining cloud-free optical data to retrieve surface information is both of theoretical importance and practical necessity. Cloud removal in optical remote sensing imagery aims to mitigate or eliminate the influence of clouds, thereby revealing more accurate and complete surface details [22]. In response to this challenge, a wide range of approaches have been proposed. These methods can broadly be divided into three categories: (i) single-image methods, (ii) multimodal-based methods, and (iii) multitemporal-based methods [22]. The main categories and their characteristics are summarized below:

(A) **Single-image methods:** Constrained by the limited acquisition capabilities of early remote sensing data, single-image cloud removal techniques attempt to restore surface information using only the cloudy optical image. Classical approaches employ statistical and physical models such as spatial similarity, frequency filtering, or atmospheric scattering models. For example, Zhang et al. [23] proposed the *Haze Optimized Transformation (HOT)*, which detects and compensates for thin cloud and haze contamination in Landsat images by exploiting the spectral correlation of clear-sky bands and quantifying deviations caused by haze. Similarly, He et al. [24] introduced the *dark channel prior*, a widely used statistical prior that estimates haze thickness from local image patches to re-

cover clear radiance, later adapted for thin cloud removal in optical remote sensing. With the advent of deep learning, CNNs, U-Nets, and GAN-based architectures have been applied to learn the mapping from cloudy to cloud-free domains, sometimes extended with unpaired learning schemes like CycleGANs. While these methods demonstrate effectiveness for thin or semi-transparent clouds, their reliance on information present in a single image limits their applicability to dense cloud cover. In such cases, they cannot reliably reconstruct surface features, which has motivated the integration of external data sources such as SAR imagery [22].

(B) **Multimodal-based methods:** Multimodal strategies explicitly integrate auxiliary data from complementary sensors to enhance optical image restoration. While multispectral approaches exploit the differential sensitivity of spectral bands, the most significant progress has been achieved through the fusion of synthetic aperture radar (SAR) and optical imagery.A representative work has been introduced by Meraner et al. [2], who proposed the DSen2-CR framework, a deep residual network that combines Sentinel-1 and Sentinel-2 data to improve reconstructions under thick cloud cover and preserve spectral fidelity. Likewise, Grohnfeldt et al. [21] demonstrated the potential of conditional GANs (cGANs) to fuse SAR and multispectral data for cloud removal, highlighting the advantages of adversarial training in capturing nonlinear relationships between modalities. More recently, Xu et al. [25] presented the GLF-CR model, which applies a global–local fusion strategy to better exploit SAR features for cloud removal. SAR-to-optical image translation has thus emerged as a powerful paradigm in this context, as SAR penetrates cloud layers and provides structural information that can guide optical reconstruction. A wide range of approaches have been proposed, including CNN-based fusion, cGANs, and CycleGAN-style frameworks, which either translate SAR features into optical-like imagery or combine them with partially corrupted optical inputs. These methods have proven especially effective in recovering surface information under dense and persistent cloud conditions, although challenges remain in terms of data registration, modality differences, and SAR-induced speckle noise.

(C) **Multitemporal-based methods:** Multitemporal approaches leverage repeated acquisitions of the same location at different times to fill in cloud-covered areas. *Non-blind* methods use cloud masks to guide restoration, whereas *blind* methods directly infer cloud-free information from temporal sequences. A classical example is the work of Xu et al. [26], who proposed a sparse representation framework with multitemporal dictionary learning (MDL) that learns dictionaries from both cloudy and clear images, effectively reconstructing areas obscured by thin and thick clouds without requiring ex-

plicit cloud masks. More recently, Ebel et al. [27] introduced UnCRtainTS, an attention-based deep learning model that not only reconstructs cloud-free images from Sentinel-1/2 time series but also quantifies pixel-wise uncertainty, providing reliability measures alongside the reconstructed outputs. Techniques therefore range from traditional model-driven approaches, such as low-rank tensor decomposition and sparse representation, to data-driven deep learning frameworks that learn spatio-temporal mappings. Recent research has also begun to combine multitemporal optical data with SAR, creating hybrid SAR–optical time series methods that enhance robustness under persistent cloud cover and enable more accurate SAR-to-optical translation. Although highly effective for dense cloud removal, these approaches face challenges such as geometric misalignment, temporal variability in land cover [28], and the need for large, paired training datasets [22]. In this context, mono-temporal data offers an advantage, as it requires less data and avoids the need for co-registration compared to multi-temporal approaches [29].

Table 2.4.: Summary of cloud removal categories, their advantages and limitations [22, 30].

Category	Advantages	Limitations	Representative literature
Single-image	• No auxiliary data required (cost- and time-efficient). • Effective for thin or semi-transparent clouds. • Straightforward implementation with statistical/physical models or deep learning.	• Ineffective for dense or opaque clouds. • Often introduces artifacts or color distortions. • Deep learning requires large paired datasets, which are difficult to obtain.	[23] [24] [31] [32] [33] [34] [35] [36] [37]
Multimodal	• Integrates complementary information from other sensors. • Multispectral bands provide spectral redundancy. • SAR–optical fusion enables SAR-to-optical translation, penetrating cloud layers. • Suitable for both thin and thick clouds.	• Requires accurate registration of heterogeneous data. • SAR data introduces speckle noise. • High computational complexity and preprocessing effort.	[21] [38] [39] [2] [40] [41] [4] [42] [8] [10] [1] [43]
Multitemporal	• Exploits temporal redundancy to reconstruct cloudy regions. • Effective for dense and extensive cloud cover. • Deep learning models can capture spatio-temporal correlations. • Can be extended with SAR–optical time series for improved robustness.	• Sensitive to geometric misalignment and temporal variability. • Requires consistent multitemporal datasets, which may be unavailable. • Landscape or seasonal changes reduce restoration accuracy.	[39] [44] [26] [27] [45] [46]

As shown in Table 2.4, research on cloud removal has been uneven across categories. Single-

image methods have been the most extensively studied due to their simplicity and minimal data requirements; however, their effectiveness is limited under dense cloud conditions. Multimodal approaches, particularly SAR–optical fusion, have gained significant traction in recent years and currently represent the most active research direction. By contrast, multitemporal methods, while highly effective in principle, have been less frequently explored because of the challenges associated with acquiring consistent and well-aligned time-series data. Another challenging aspect is the change in ground surface conditions between multitemporal acquisitions, which can confuse the reconstruction process.

In summary, cloud removal research spans single-image, multimodal, and multitemporal strategies, each with distinct advantages and limitations. Among these, SAR-to-optical image translation has recently emerged as a particularly promising direction, as it leverages the cloud-penetrating capability of SAR while producing optical-like imagery suitable for interpretation and analysis. This thesis builds on this line of research by systematically investigating and advancing SAR-to-optical translation methods for cloud removal.

2.4. Generative Artificial Intelligence

Generative Artificial Intelligence (GenAI) refers to a class of machine learning models designed to generate new data samples that resemble a given training distribution, such as images, text, or audio. Unlike discriminative models, which focus on classifying or predicting labels, generative models learn (an approximation to) the underlying probability distribution of the data to create novel instances [47]. This capability has revolutionized fields like computer vision, where GenAI is used for tasks including image synthesis, style transfer, and domain adaptation. In the context of remote sensing, GenAI enables, among others, the creation of synthetic imagery, such as translating radar data to optical-like representations, which is particularly useful for overcoming environmental limitations like cloud cover, as well as data fusion for both heterogeneous and homogeneous imagery, enhancing spatial, spectral, and temporal resolution and mitigating the limitations of individual sensors [48].

One of the foundational frameworks in GenAI is the Generative Adversarial Network (GAN), introduced in 2014 by Goodfellow et al. [49]. As depicted in Figure 2.2, a GAN consists of two neural networks: a generator (G) that produces synthetic data from random noise, and a discriminator (D) that evaluates whether the generated data is real or fake. These components are trained adversarially—the generator aims to fool the discriminator, while the discriminator improves its ability to distinguish real from generated samples—leading to increasingly realistic outputs.

This adversarial process minimizes a minimax loss function, allowing GANs to capture com-

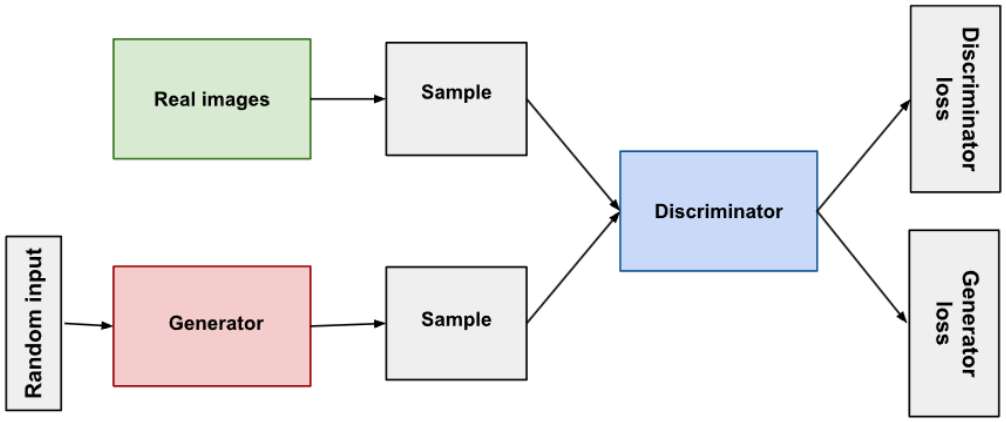


Figure 2.2.: GAN Architecture Overview: Source: https://developers.google.com/

plex data distributions without explicit probabilistic modeling. Formally, the optimization problem is defined as:

$$\min_G \max_D \; V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \tag{2.1}$$

where $D(x)$ denotes the discriminator's estimate of the probability that x comes from the real data distribution p_{data}, and $G(z)$ generates samples from latent noise $z \sim p_z$.

Building on GANs, conditional GANs (cGANs) [5] incorporate additional input conditions, such as class labels or reference images, to guide the generation process toward specific outputs. In image-to-image translation settings, cGANs typically require paired datasets for supervised training, which is non-trivial given the complexity of acquiring and registering satellite imagery. A key extension for unpaired image translation is the Cycle-Consistent GAN (CycleGAN) [50], proposed in 2017. CycleGAN addresses the challenge of learning mappings between two domains (e.g., SAR to optical) without paired training examples by enforcing cycle consistency: translating an image from domain X to Y and back to X should reconstruct the original. This is achieved through a cycle-consistency loss, combined with adversarial losses, making it suitable for RS applications where perfectly aligned SAR–optical pairs are scarce. Nevertheless, given the weak supervision during CycleGAN training, it is prone to texture/detail distortions [48], often resulting in performance comparable to other GAN-based models [45]. For the problem of RS data fusion, many studies do not directly reuse off-the-shelf GAN architectures but instead adapt them to better accommodate multisource fusion. Among these, cGANs and CycleGANs are the most commonly adopted families in numerous fusion studies [48].

Despite the remarkable performance of GAN-based approaches in cloud removal, several

challenges remain. They are inherently difficult to train and can suffer from issues such as mode collapse, leading to distorted or repetitive outputs, especially at high spatial resolutions [40]. Moreover, cloud morphology and distribution exhibit substantial diversity and complexity, imposing demanding requirements on training and application of GAN-based methods [22, 46].

More recently, diffusion models have emerged as a powerful alternative to GANs, offering improved stability and higher-fidelity generation [8]. Denoising Diffusion Probabilistic Models (DDPMs) [6], introduced in 2020, model data generation as a reverse diffusion process: starting from Gaussian noise, the model iteratively denoises the input to produce a sample from the target distribution. Unlike GANs, which can suffer from mode collapse (limited output diversity), diffusion models provide probabilistic sampling and better coverage of diverse data modes. In RS, diffusion-based approaches are gaining traction for tasks like cloud removal, where they can generate realistic optical images by conditioning on SAR inputs.

GenAI's application in remote sensing, particularly for multimodal data fusion, leverages these models' ability to bridge domain gaps. For instance, GANs and diffusion models can synthesize cloud-free optical imagery from SAR, preserving structural details while enhancing interpretability. However, challenges such as training instability in GANs and high computational costs in diffusion models persist, motivating ongoing research into hybrid architectures. Further details regarding the applied approaches are provided in the following section.

2.5. SAR-to-optical image translation

Optical imagery provides rich spectral information and can often be interpreted without expert knowledge, but it is highly sensitive to atmospheric conditions such as cloud cover, which frequently limits its usability. In contrast, SAR imagery offers all-weather, day-and-night, cloud-penetrating capabilities, though its complex backscatter characteristics, speckle noise, and lack of color make interpretation challenging even for experts [10]. Moreover, 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 [48]. Bridging this gap, SAR-to-optical translation generates synthetic, cloud-free optical-like images from SAR data, combining the interpretability of optical imagery with SAR's robustness to clouds. Defined as an image-to-image translation (I2I) task, this process is valuable for applications requiring consistent cloud-free optical information, such as land-cover classification, disaster monitoring, and vegetation analysis [2].

The domain gap between SAR and optical imagery, however, poses significant challenges. SAR images exhibit speckle noise due to coherent interference, geometric distortions from side-looking geometry, and intensity-based representations that differ fundamentally from the

reflectance-based multispectral bands of optical sensors [38]. Acquiring perfectly co-registered SAR–optical pairs is also non-trivial, as both spatial and temporal alignment must be ensured. According to Tobler's First Law of Geography—*"everything is related to everything else, but near things are more related than distant things"* [51, p. 236]—the closer the distance and the shorter the time interval, the greater the correlation between features [30]. Achieving such conditions in practice is difficult, as the two imaging modalities differ fundamentally; certain land features (e.g., roads, playgrounds, airport runways) may appear differently in terms of spectral reflectance versus SAR backscatter, complicating accurate cross-domain mapping [30].

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 [52], 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 [53], provided partial solutions but depended on handcrafted features and complex preprocessing, limiting scalability and accuracy.

Recent advances in deep learning have transformed this process. Generative Adversarial Networks (GANs) [49] and diffusion models [6] enable end-to-end learning of mappings between SAR and optical domains, capturing pixel-level distributions and feature correlations for realistic, cloud-free optical reconstruction. Most prior work focuses on reconstructing visible RGB bands of Sentinel-2 imagery, with fewer studies extending into the near-infrared (NIR) or full multispectral range. In the literature, the former is often referred to as SAR-to-optical translation, while the latter is known as SAR-to-multispectral (SAR-to-MS) translation.

For the purposes of this thesis, the term *optical* is used broadly to encompass the full spectral range of Sentinel-2. Accordingly, SAR-to-optical denotes translation tasks regardless of the number of reconstructed bands. Within this context, multispectral GAN-generated images are expected to exhibit both structural and spectral fidelity, ensuring suitability for quantitative remote sensing and cloud removal applications. Thus, SAR-to-optical translation can be regarded as a data-driven cloud removal strategy that bridges the interpretability of optical imagery with the robustness of SAR. This thesis extends the task to the complete multispectral domain of Sentinel-2 data.

The foundational GAN framework by Goodfellow et al. [49], later extended to Conditional GANs (cGANs) by Mirza and Osindero [5], enabled image-to-image translation tasks that made SAR-to-optical synthesis feasible. Fuentes Reyes et al. [38] 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. [54] introduced a Supervised Cy-

cleGAN (S-CycleGAN) by incorporating a pixel-wise MSE loss, producing realistic optical images and showing strong potential for cloud removal. Gao et al. [7] 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) [55], transformer-based architectures have also emerged. Zhao et al. [9] 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 [56], HVT-cGAN achieved superior visual and quantitative results over previous GAN-based models. Park et al. [41] 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. [8] 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. [10] 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 [45], 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 demonstrate the rapid progression of SAR-to-optical translation, from GAN-based frameworks to transformer and diffusion models. Together, they form the foundation for applying SAR–optical fusion and translation techniques to cloud-contaminated optical imagery, as discussed in the following section.

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

Cloud contamination in optical remote sensing imagery hinders continuous Earth observation, limiting applications such as crop monitoring and land-cover classification. Synthetic Aperture Radar (SAR) systems can penetrate clouds, enabling data acquisition in all weather conditions. This capability makes SAR data valuable for reconstructing missing optical information in cloud-affected areas and has driven increasing interest in SAR–optical fusion and translation for cloud removal. Early methods relied on traditional signal processing techniques. Huang et al. [57] introduced sparse representation-based cloud removal using SAR data, which Xu

et al. [26] 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 the emergence of deep learning, researchers began to exploit neural architectures for this task. Enomoto et al. [11] 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. [21] 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 allowed flexible multi-channel inputs and was trained on a subset of the SEN1-2 dataset [56]. The results confirmed the advantage of incorporating SAR data, validating the effectiveness of the approach for mitigating cloud cover.

The research group from the Technical University of Munich (TUM) [5] and the German Aerospace Center (DLR) [6] further advanced this field by developing a family of co-registered SAR–optical datasets. The initial SEN1-2 dataset [56] combined single-polarized (VV) SAR with RGB optical composites, later extended to dual-polarized SAR and full 13-band optical imagery in SEN12-MS [58]. To address cloud contamination directly, they released SEN12-MS-CR [59], containing triplets of dual-polarized SAR, cloudy optical, and corresponding cloud-free images. The SEN12-MS-CR-TS dataset [60] then introduced multi-temporal observations, greatly enhancing its relevance for time-series cloud-removal studies.

Naderi Darbaghshahi et al. [4] proposed a dual-GAN framework employing Dilated Residual Inception Blocks (DRIBs). The first GAN translated SAR data into optical imagery, while the second fused this output with a cloudy optical image to produce a cloud-free result. This design effectively removed cloudy regions while preserving clear areas, demonstrating strong qualitative performance despite moderate quantitative accuracy. Ebel et al. [27] extended this direction by incorporating uncertainty prediction into multispectral reconstruction with their model UnCRtainTS, which outperformed prior methods and provided per-pixel uncertainty maps to indicate the reliability of spectral predictions.

Diffusion models have recently emerged as a competitive alternative for cloud removal. Zou et al. [46] presented DiffCR, a fast conditional diffusion framework for cloud removal from optical satellite imagery, which currently represents the state-of-the-art on the SEN12-MS-CR [59] dataset. Unlike traditional GAN-based or computationally intensive diffusion approaches, Dif-

[5] https://www.tum.de/
[6] https://www.dlr.de/en

fCR employs a decoupled conditional architecture and a novel Time and Condition Fusion Block (TCFBlock) to efficiently fuse multiscale spatiotemporal features. By directly predicting clean cloud-free images instead of noise, it achieves faster convergence and remarkable fidelity, producing high-quality reconstructions in as few as 1–5 denoising steps with over 95% lower computational cost than previous diffusion-based methods. Meraner et al. [2] proposed DSen2-CR, a deep residual neural network for cloud removal in Sentinel-2 imagery that integrates SAR–optical data fusion to enhance reconstruction under thick cloud cover. Built upon the DSen2 super-resolution ResNet, the model introduces a Cloud-Adaptive Regularized Loss (CARL) to preserve uncorrupted input information while reconstructing only clouded regions. Trained and evaluated on the SEN12-MS-CR dataset, DSen2-CR demonstrated strong generalization across global scenes and outperformed GAN-based baselines in both spectral and structural fidelity. Across-band evaluation confirmed high reconstruction accuracy over all 13 Sentinel-2 bands, with the best performance on surface bands and slightly higher errors in atmospheric ones.

Recent advances such as DSen2-CR and DiffCR illustrate the maturity of SAR-to-optical cloud-removal research, highlighting a clear trend toward physically interpretable, data-rich, and computationally efficient architectures. Building upon this foundation, the present thesis validates SAR-to-optical image translation across the full 13 Sentinel-2 spectral bands and further investigates its applicability to practical cloud-removal scenarios.

2.7. Application and Relevance to KIWA Project

The KIWA project [7] *(German: KI-basierte Waldüberwachung – Engl: AI-based Forest Monitoring)* addresses the growing ecological challenge of forest degradation and wildfire risk in Central Europe. Climate extremes, prolonged droughts, and pest infestations are severely affecting forests, making them increasingly vulnerable to fire. Wildfires not only destroy ecosystems, properties, and human lives, but also release the CO_2 previously absorbed by forests, thereby accelerating climate change. Current monitoring methods, such as aircraft patrols and stationary watchtowers, are resource-intensive, costly, and limited in performance [61].

To overcome these challenges, KIWA integrates artificial intelligence with advanced remote sensing technologies, particularly drones equipped with computer-vision systems, to improve early wildfire detection. The project's broader objectives include delivering high-resolution environmental data, providing decision support to emergency services, and supporting climate-resilient, biodiversity-rich forest management. KIWA thus exemplifies an AI "lighthouse" initiative with the ambition to serve as a transferable blueprint for forest monitoring systems

[7] https://www.kiwa-projekt.de

across Germany and internationally [62].

Similar to the general challenges faced in remote sensing applications, the KIWA project requires gapless observation capabilities. For instance, recent KIWA-related research highlights the need for automated methods to delineate burned areas (BAs) and assess wildfire risks using remote sensing data [29]. These approaches primarily rely on optical indices such as NDVI, NBR, NDWI, and WFI, which require cloud-free multispectral imagery. However, cloud cover and wildfire smoke often make it difficult to obtain a continuous historic record of the areas of interest, which is critical in emergency services. As the authors state, *"excluding multi-temporal approaches per se for our KIWA workflow is not an option and not intended"* [29, p. 767].

In this context, the contribution of this thesis—translating SAR data into optical-like images using generative AI models—offers a direct benefit to KIWA. Once validated, the proposed methods can be integrated in the project workflow to enhance the spatial and temporal coverage of forest monitoring, even under cloudy or adverse weather conditions. By extending the availability of optical-equivalent data in near real-time, these approaches can improve the robustness of KIWA's mapping workflows, improve mono- and multi-temporal analyses, and provide more reliable support for decision-making processes in emergency services. This integration has the potential to enhance KIWA's operational efficiency and transferability, supporting its mission to deliver automated, scalable, and accurate wildfire monitoring solutions.

3. Methodology & Data Basis

This chapter outlines the methodological framework adopted for the SAR-to-optical translation task. It details the formulation of the problem as a supervised image-to-image regression, describes the datasets and preprocessing procedures used, and presents the model architectures and training strategies employed. In particular, the Pix2Pix [63] conditional generative adversarial network (cGAN) serves as the core model for learning the mapping between Sentinel-1 SAR inputs and Sentinel-2 multispectral optical outputs. The chapter further elaborates on the design of the loss functions, optimization process, and experimental setup used to ensure stable and effective training. Finally, the evaluation metrics are introduced to provide both quantitative and perceptual assessments of the model's performance, encompassing spatial fidelity, structural consistency, and spectral accuracy.

3.1. Problem Formulation

The task of SAR-to-optical translation can be formulated as a supervised image-to-image regression problem. Let

$$X \in \mathbb{R}^{H \times W \times 2}$$

denote the input SAR patch, where the two channels correspond to Sentinel-1 backscatter in VV and VH polarizations (expressed in the decibel scale(dB)). The corresponding reference optical patch is represented as

$$Y \in \mathbb{R}^{H \times W \times N},$$

where N denotes the number of channels corresponding to the optical bands of Sentinel-2, in this case $N = 13$.

The objective is to learn a parametric mapping function

$$f_\theta : \mathbb{R}^{H \times W \times 2} \rightarrow \mathbb{R}^{H \times W \times N},$$

such that the generated optical image

$$\hat{Y} = f_\theta(X)$$

is as close as possible to the reference image Y.

Training this model requires minimizing a loss function $\mathcal{L}$ that measures the discrepancy between $\hat{Y}$ and Y:

$$\hat{\theta} = \arg\min_{\theta} \ \mathcal{L}(f_\theta(X), Y).$$

In practice, $\mathcal{L}$ is designed as a weighted combination of complementary terms (e.g., reconstruction, adversarial, perceptual, and spectral losses) that collectively encourage pixel-level accuracy, structural consistency, and spectral fidelity. Further details on the employed loss functions and evaluation metrics are provided in Sections 3.4.2 and 3.5, respectively.

3.2. Datasets

3.2.1. SEN12-MS

This thesis relies exclusively on the SEN12-MS dataset [58], 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.

The Sentinel-1 component originates from ground-range-detected (GRD) products acquired in interferometric wide swath (IW) mode. These data were radiometrically calibrated and orthorectified against SRTM or ASTER digital elevation models to ensure accurate geolocation. The Sentinel-2 imagery was curated using a cloud-free mosaicking workflow on Google Earth Engine: within each region of interest (ROI), multiple observations collected during a given meteorological season of 2017 were composited such that cloud-contaminated pixels were systematically excluded. This procedure ensured that every ROI is represented by seasonally consistent, nearly cloud-free multispectral data. Finally, the MODIS land cover maps were used to generate categorical reference layers; however, due to their relatively coarse native resolution (500 m), they are subject to spatial inaccuracies even after upsampling.

Importantly, all triplets underwent manual verification by a remote sensing expert. This revision step ensured that each patch is free from major artifacts, severe registration errors, or residual cloud contamination, thereby guaranteeing the dataset's quality and usability for

machine learning tasks. The ROIs were sampled globally across all inhabited continents and four meteorological seasons of 2017 to maximize spatial and temporal diversity, as illustrated in Figure 3.1. Nevertheless, it should be noted that the ROI selection was not purely random. In practice, locations were chosen to avoid large homogeneous areas such as deserts or oceans and to ensure inclusion of diverse land cover classes. While this design improves the dataset's representativeness for a wide range of applications, it may introduce a bias toward heterogeneous landscapes.

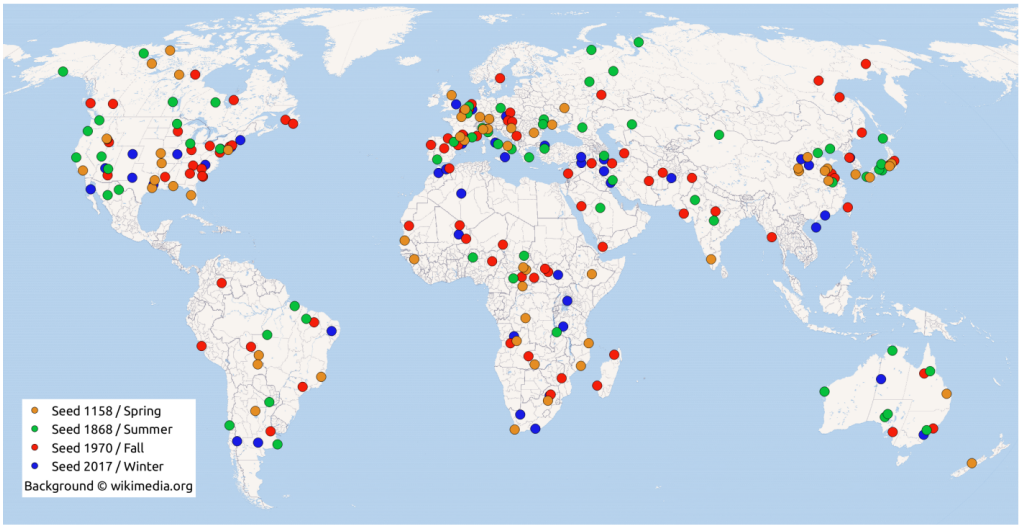


Figure 3.1.: ROIs distribution of the SEN12-MS Dataet. Adapted from [58]

For the purpose of this thesis, which focuses on translating SAR data into multispectral optical imagery, only the Sentinel-1 and Sentinel-2 modalities are utilized. Representative examples of corresponding image pairs are shown in Figure 3.2. The MODIS component of the SEN12MS dataset was incorporated to provide land-cover annotations or auxiliary information supporting tasks such as land-cover mapping, semantic segmentation, and multi-sensor data fusion in deep learning research [58]. However, since these data are not pertinent to the SAR-to-optical translation objective, they are excluded from the present study.

3.2.2. SEN12 datasets Family

SEN12-MS is part of a broader line of datasets developed to foster multimodal remote sensing research. Its direct predecessor, SEN1-2 [56], curated by the same research group, contained approximately 282,000 paired patches of Sentinel-1 VV data and Sentinel-2 RGB composites. While groundbreaking in bridging SAR and optical domains, SEN1-2 lacked georeferencing,

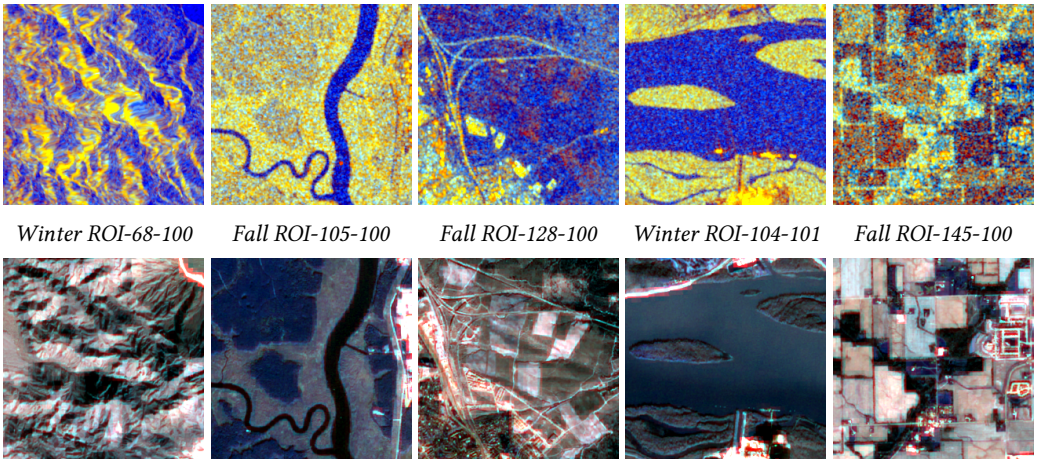


Figure 3.2.: Sample pairs from the SEN12-MS dataset. Top row: Sentinel-1 SAR patches (R: VV, G: VH, B: VV/VH). Bottom row: corresponding Sentinel-2 multispectral patches (only RGB bands).

full spectral coverage, and multi-polarization SAR, limiting its applicability for remote sensing research beyond proof-of-concept image translation.

SEN12-MS addressed these limitations by introducing full multispectral coverage, dual-polarized SAR, geocoded products, and auxiliary land cover labels, making it a comprehensive multimodal benchmark. Building upon this foundation, the dataset family has since been extended. SEN12-MS-CR [59] added temporally matched cloudy and cloud-free Sentinel-2 imagery alongside Sentinel-1 data, enabling the development and benchmarking of cloud removal methods under realistic atmospheric conditions. Subsequently, SEN12-MS-CR-TS [60] expanded the concept into the temporal domain, providing year-long multimodal time series with 30 co-registered Sentinel-1 and Sentinel-2 acquisitions per ROI. This evolution reflects a progression from simplified SAR–optical pairs, to globally diverse multimodal data, to temporally rich resources designed for time-series analysis and addressing cloud removal.

A comparsion of these different datasets is provided in Table 3.1. In this thesis, however, the focus remains on the SEN12-MS dataset—since the focus is mainly on translating from SAR to optical and not directly to address cloud cover— leveraging its multimodal SAR and multispectral imagery for the study of SAR-to-optical translation.

3.2.3. Subset Selection & Preprocessing

The SEN12-MS dataset is divided into four subsets, each corresponding to a meteorological season. Due to the large size of the full dataset, which makes conducting multiple experiments and ablations computationally demanding, only the winter subset (the smallest) was used for

Table 3.1.: Comparison of SEN12-family datasets.

Aspect	**SEN1-2** [56]	**SEN12-MS** [58]	**SEN12-MS-CR** [59]	**SEN12-MS-CR-TS** [60]
Year released	2018	2019	2021	2022
Main purpose	Proof-of-concept SAR–optical translation	Multimodal learning and data fusion	Cloud removal with real cloudy/clear pairs	Multi-temporal cloud removal (sequence models)
Modalities	S1 (VV), S2 (RGB)	S1 (VV,VH), S2 (13 bands), MODIS LULC	S1 (VV,VH), S2 (13 bands; cloudy & cloud-free)	S1 (VV,VH), S2 (13 bands; cloudy & cloud-free time series)
Georeferencing	Not georeferenced	Fully georeferenced	Fully georeferenced	Fully georeferenced
Spatial sampling	Global patch pairs (282k)	180,662 patch triplets across 2017 seasons	169 ROIs; >100k patch triplets	53 ROIs; 30 time steps per ROI
Temporal coverage	Single time-point	Seasonal (2017)	Seasonal with paired cloudy/clear	Year-long time series (2018)
Patch size	256×256 px	256×256 px	256×256 px	256×256 px
Notable limitations	RGB only; VV only; no geocoding	MODIS labels are coarse (upsampled)	Mono-temporal pairs (no full time series)	Fewer ROIs; large storage ($\sim$2 TB)

training and experiments. This subset contains 31,825 paired images, each consisting of dual-polarized Sentinel-1 SAR data and 13-band Sentinel-2 optical imagery. A custom Python preprocessing pipeline was implemented using Rasterio, NumPy, and GDAL to efficiently stream, normalize, and store the data in a ready-to-train format.

Following common practices in the literature as in [2, 64–67], Sentinel-1 backscatter values (in decibels) were clipped to fixed physical ranges of -25 dB to 0 dB for the VV channel and -32.5 dB to 0 dB for the VH channel to suppress noise and radiometric outliers. Similarly, Sentinel-2 reflectance values were clipped to the range 0 to $10{,}000$, corresponding to top-of-atmosphere scaled reflectances. After clipping, all values were linearly normalized to the *tanh* range $[-1, 1]$, consistent with the `Tanh` activation function used in the generator's output layer. The original image size of 256×256 pixels was preserved, and no data augmentation was applied, as the dataset already exhibits substantial spatial diversity, given that the regions of interest (ROIs) were sampled globally across all inhabited continents.

3.3. Pix2Pix Model

The image translation task in this thesis is addressed using the Pix2Pix model, originally introduced by Isola et al. [63]. 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 [2, 7, 9, 45, 65, 68], or further adapted and extended in various studies [21, 68]. Owing to its demonstrated effectiveness and widespread adoption in related research, Pix2Pix was selected as the primary model for the experiments conducted in this thesis.

Pix2Pix is based on the concept of conditional generative adversarial networks (cGANs), which extend the original GAN formulation by conditioning both the generator and discriminator on an input image. In this setup, the generator G learns to map an input image x to an output image y, while the discriminator D learns to distinguish between real image pairs $\{x, y\}$ and synthesized pairs $\{x, G(x)\}$. This adversarial objective encourages the generated outputs to be both realistic and structurally consistent with the given input.

Formally, the cGAN loss is defined as:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_x[\log(1 - D(x, G(x)))]. \tag{3.1}$$

To encourage fidelity to the target image, the adversarial loss is combined with an ℓ_1 reconstruction loss:

$$\mathcal{L}_{\ell_1}(G) = \mathbb{E}_{x,y}[\|y - G(x)\|_1]. \tag{3.2}$$

The final objective is then:

$$G^* = \arg\min_G \max_D \; \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{\ell_1}(G), \tag{3.3}$$

where λ balances realism and reconstruction accuracy. Following the original configuration proposed in [63], a weighting factor of $\lambda = 100$ is commonly used in the literature [4, 12, 21, 45, 65, 69] and is therefore also adopted in the present thesis. In addition, supplementary loss components were incorporated, as they demonstrated improved training convergence and led to higher-quality reconstructions (see the ablation study on loss functions in Section 4.5.1).

In Pix2Pix, The generator is implemented as a U-Net encoder–decoder [70]. Unlike a plain encoder–decoder, U-Net introduces skip connections between corresponding downsampling and upsampling layers, allowing low-level spatial details from the input to directly propagate to the output. This design is particularly effective in tasks where the input and output share spatial structures, such as SAR-to-optical translation.

The discriminator follows a PatchGAN architecture, which classifies local $N \times N$ image patches as real or fake instead of operating on the entire image [63]. This approach emphasizes high-frequency accuracy and enforces local realism, while the ℓ_1 term ensures global structural coherence. The original work demonstrated that a patch size of 70×70 provides a good trade-

off between reconstruction quality and computational efficiency.

During training, updates alternate between optimizing D to improve its ability to classify real and generated pairs, and optimizing G to both deceive D and minimize the ℓ_1 distance to the target image. The Adam optimizer [71] is used with a learning rate of 1×10^{-4} and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Dropout is applied during both training and inference to introduce stochasticity, although the generated outputs remain largely deterministic in practice.

Overall, the Pix2Pix framework offers a principled and versatile solution for image-to-image translation tasks. Its ability to combine adversarial and reconstruction-based objectives makes it particularly suitable for SAR-to-optical translation, where both structural accuracy (e.g., edges, textures, contrast) and perceptual quality (how realistic or visually convincing the generated images appeae) are essential.

3.4. Training Procedure

This section describes the complete training pipeline adopted for the proposed SAR-to-optical image translation model. It outlines the experimental setup, loss formulations, optimization strategy, and monitoring procedures that together ensure stable adversarial training and effective convergence. Emphasis is placed on the integration of multiple complementary loss functions to enhance perceptual and structural fidelity, as well as on the staged optimization scheme designed to mitigate common GAN instabilities. The overall workflow aims to achieve a robust balance between pixel-level accuracy, perceptual realism, and generalization performance across diverse imaging conditions.

3.4.1. Experimental Setup

All experiments were implemented in Python 3.11, PyTorch 2.4.0, and CUDA 12.4, using the official Pix2Pix implementation[1]. Training was conducted on an NVIDIA RTX A5000 GPU with CUDA acceleration. The training, validation, and test sets were derived from the preprocessed SEN12-MS subset described in Section 3.2.3. Each sample consisted of a Sentinel-1 SAR input patch (VV and VH channels) paired with the corresponding Sentinel-2 optical patch comprising 13 spectral bands.

Code Availability: All source code developed for this thesis, including dataset preprocessing, model training, and evaluation scripts, will be made publicly available to support transparency and reproducibility. The implementation is based on the official Pix2Pix framework and has

[1] `https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix`

been adapted for multispectral SAR-to-optical image translation. The repository will be hosted at: `https://github.com/THDetch/bachelor-thesis`.

3.4.2. Loss Functions

The loss function employed in the experiments consisted of multiple components. Conditional GAN (cGAN) models, such as Pix2Pix, typically optimize a weighted sum of an adversarial loss and a reconstruction loss. By default, Pix2Pix uses the Binary Cross-Entropy (BCE) loss for the adversarial component and the Mean Absolute Error (MAE), also referred to as the L1 loss, for image reconstruction. However, during preliminary experiments, training instabilities and slow divergence were observed, suggesting the presence of vanishing gradient issues. To mitigate these effects, the adversarial BCE loss was replaced with the Least Squares GAN (LSGAN) loss [72], implemented as the Mean Squared Error (MSE) loss in PyTorch, which is known to provide more stable gradients and smoother convergence.

In addition to the LSGAN and L1 losses, the LPIPS and SSIM losses were also incorporated during training. Experiments with different combinations of these loss terms indicated that the best results were achieved when all four were combined.

The overall objective function can be expressed as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}. \tag{3.4}$$

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.
- λ_{L1}, λ_{SSIM}, and λ_{LPIPS} are weighting coefficients that control the relative contribution of each term.

In the experiments, the weighting coefficients were set as follows: $\lambda_{\text{L1}} = 100.0$ (as originally proposed in the Pix2Pix model and commonly used in the literature; see 3.3). Both λ_{SSIM} and

λ_{LPIPS} were set to 50. These values were selected to maintain a balance between pixel-level accuracy and perceptual quality, while placing greater emphasis on the reconstruction process governed by $\mathcal{L}_{\text{L1}}$. It is important to note that a comprehensive investigation of different values for λ_{SSIM} and λ_{LPIPS} was not performed.

To further assess the individual contribution of each loss component, an ablation study was conducted. The model was trained under different configurations, including combinations of the individual losses and the full combined objective. Quantitative and qualitative results of this ablation are presented and discussed in Section 4.5.1. Except for the experiments dedicated to this loss ablation, all other trainings were performed using the full combination of the four loss functions.

3.4.3. Training Strategy

The network training followed a staged adversarial optimization scheme with alternating updates of the generator and discriminator. At the early stage of training, the generator was warmed up by being trained independently for the first 20 epochs to stabilize its predictions before introducing the discriminator. This step mitigates early training instabilities commonly observed in GAN-based frameworks, which were also encountered during experimentation. After the warm-up period, both networks were jointly optimized in an alternating fashion, where the discriminator was updated once per iteration, followed by a generator update (1:1 ratio between G and D). The discriminator was trained to minimize the least-squares error between its predictions and the target real/fake labels, while the generator was optimized to minimize the composite objective described in the previous subsection.

Training was conducted for 150 epochs using a batch size of 16 and the Adam optimizer [71] with a learning rate (LR) of 1×10^{-4} and momentum parameters $(\beta_1, \beta_2) = (0.5, 0.999)$. To prevent overfitting and accelerate convergence, the generator's learning rate was adaptively reduced using the `ReduceLROnPlateau` scheduler—an adaptive learning rate regulator that decreases the LR by a specified factor ($\text{LR}_{\text{factor}}$) when no improvement is observed in the validation loss over a predefined number of epochs, referred to as *patience* [73]. The adjustment follows the rule

$$\text{LR}_{\text{new}} = \text{LR}_{\text{initial}} \times \text{LR}_{\text{factor}}$$

as described in [74]. In this thesis, the learning rate was halved whenever the validation $\text{L}1$ loss stagnated for 10 epochs. The discriminator's learning rate was kept constant throughout training to preserve stable adversarial dynamics, as reducing it was found to introduce additional instability during preliminary experiments.

Figure 3.3 illustrates the evolution of the training and validation losses over 150 epochs,

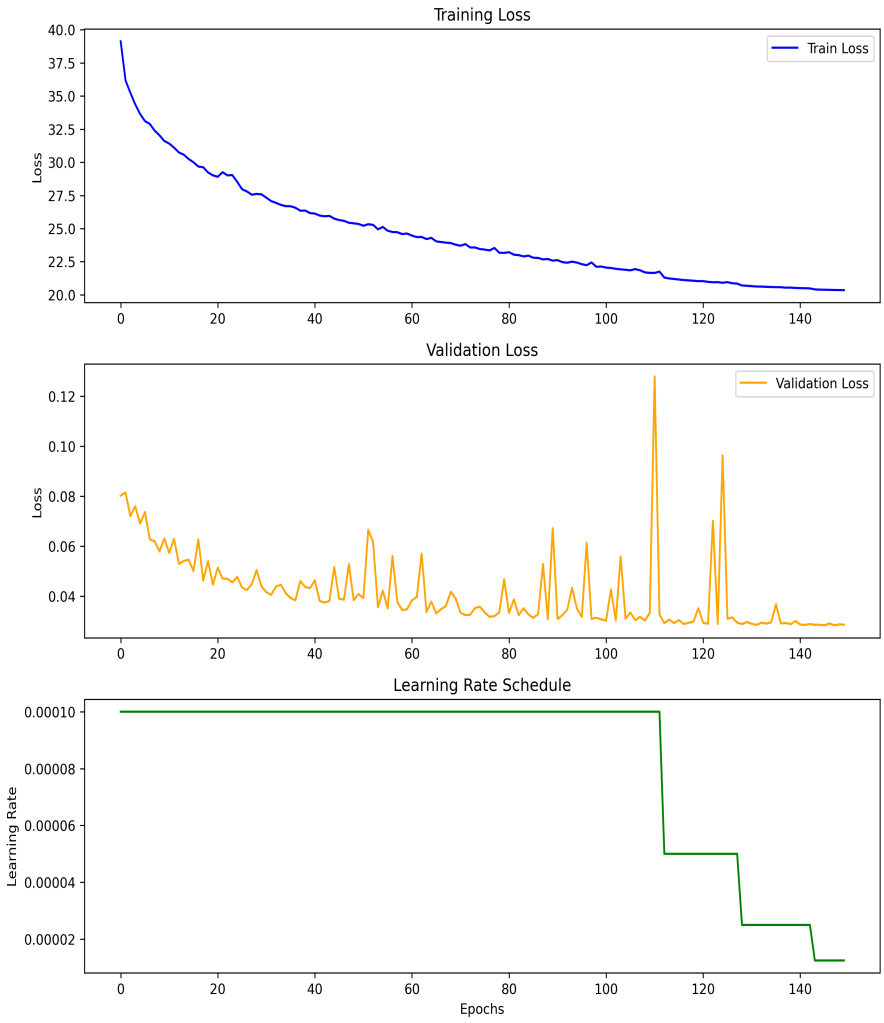


Figure 3.3.: Training and validation loss curves over 150 epochs.

together with the learning rate behavior. The training loss exhibits a steady decline, indicating stable convergence of the generator–discriminator optimization. The validation loss follows a similar trend with minor fluctuations, reflecting the stochastic nature of adversarial learning. The learning rate decreases automatically when the validation L1 loss stagnates, as controlled by the `ReduceLROnPlateau` scheduler, leading to smoother convergence during the later stages of training.

3.4.4. Monitoring and Checkpointing

Model performance was monitored using both quantitative validation metrics and qualitative visualizations. After each epoch, the validation set was evaluated to compute the L1, SSIM, and LPIPS metrics. The model achieving the lowest validation L1 loss was saved as the best-performing checkpoint. Additionally, intermediate qualitative samples were generated every ten epochs by translating a fixed subset of Sentinel-1 patches to their corresponding optical outputs, allowing visual inspection of the training progression. Samples are illustrated in Appendix A

All relevant training statistics, including generator and discriminator losses, learning rates, and validation metrics, were logged for subsequent analysis. Checkpoints were saved at the end of each epoch to enable resumption in case of interruptions. After training completion, the recorded metrics were used to generate loss and performance curves, which facilitated evaluation of convergence stability and model generalization.

3.5. 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 [69], 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.

In [69], five image quality assessment (IQA) metrics —SSIM, FSIM, MSE, LPIPS, and DISTS— were systematically evaluated through image restoration experiments to determine their suitability for SAR-to-optical translation tasks. The study found that SSIM, MSE, and LPIPS exhibited strong consistency with human visual perception, demonstrated stable convergence, and effectively captured both structural and textural characteristics. In contrast, FSIM often failed to represent fine details accurately, while DISTS showed instability. Consequently, the authors recommended SSIM, MSE, and LPIPS as complementary metrics for assessing pixel-level accuracy, structural similarity, and perceptual quality, respectively. However, based on

the analysis conducted in this thesis across 16 related studies, SSIM, PSNR, and SAM remain the most widely adopted evaluation metrics in SAR-to-optical translation, fusion, and cloud removal research, whereas LPIPS and MSE appear less frequently, as summarized in Table 3.2. This trend aligns with the observations reported in the literature survey by [30].

Table 3.2.: Frequency of common evaluation metrics used in SAR-to-optical and cloud-removal studies.

Metric	References	Frequency
Structural Similarity Index Measurement (SSIM) [75]	[1, 8, 9, 22, 41, 46, 48, 64, 65, 67, 76, 77]	12
Peak Signal-to-Noise Ratio (PSNR) [78]	[1, 4, 9, 22, 41, 44, 46, 64, 65, 67, 76, 77]	12
Spectral Angle Mapper (SAM) [79]	[1, 4, 9, 21, 48, 64, 65, 67, 76, 80]	11
Fréchet Inception Distance (FID) [81]	[8, 9, 41, 46, 76, 80]	6
Root Mean Square Error (RMSE)	[4, 21, 22, 48, 64]	5
Learned Perceptual Image Patch Similarity (LPIPS) [82]	[1, 22, 46, 77, 80]	5
Mean Absolute Error (MAE)	[4, 64]	2
Mean Square Error (MSE)	[44, 65]	2

The evaluation of the SAR-to-optical translation performance in this thesis was carried out using a combination of quantitative and perceptual metrics. The primary quantitative metrics included PSNR, SSIM, and SAM, complemented by the perceptual metric LPIPS. PSNR quantifies pixel-level fidelity, SSIM assesses local structural similarity, and SAM measures spectral consistency across all spectral bands, which is particularly important in multispectral remote sensing applications. In addition, conventional error-based metrics, MAE and RMSE, were employed for validation. To further capture perceptual realism beyond pixel-wise statistics, the deep feature–based LPIPS metric was used to enable pairwise comparisons between generated and reference images. For outputs containing more than three spectral bands, the perceptual metric LPIPS were computed on a fixed RGB stack for both the reference and the generated images, and this limitation was explicitly acknowledged. These chosen metrics, along with their underlying mathematical formulations, are discussed in more detail below:

SSIM The Structural Similarity Index (SSIM) [75] measures perceptual similarity by comparing local patterns of luminance, contrast, and structure between two images. Unlike pixel-wise errors, it models human visual sensitivity to structural distortions [9, 46], which is crucial for evaluating translated images. For two images x and y, SSIM is defined as

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{3.5}$$

where μ_x, μ_y are means, σ_x^2, σ_y^2 variances, and σ_{xy} the covariance. Values close to 1 indicate strong structural similarity. By focusing on local patterns of pixel intensities and their structural relationships, SSIM better reflects perceptual fidelity compared to raw pixel-difference metric

PSNR The Peak Signal-to-Noise Ratio (PSNR) quantifies the distortion between a reconstructed image and its reference. PSNR is directly related to the Mean Squared Error (MSE), measuring pixel-level fidelity by comparing the residual error to the maximum possible signal intensity. For two images x and y, PSNR is defined as

$$\text{PSNR}(x, y) = 10 \cdot \log_{10} \left(\frac{MAX^2}{\text{MSE}(x, y)} \right), \tag{3.6}$$

with

$$\text{MSE}(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2, \tag{3.7}$$

where x_i and y_i denote the pixel values of the generated and reference images, N is the total number of pixels, and MAX is the maximum pixel intensity (typically 255 for 8-bit images).

Higher PSNR values indicate lower distortion and better image quality, as they imply that the reconstructed image more closely approximates the reference. Despite its popularity for tasks such as denoising and compression, PSNR is limited by its purely pixel-wise formulation and often correlates weakly with human visual perception [46].

SAM The Spectral Angle Mapper (SAM), originally proposed by Kruse et al. [79] in 1993, is widely employed in remote sensing to evaluate the spectral fidelity of reconstructed images. SAM regards the spectrum of each pixel as a n-dimensional vector (n denotes the number of spectral bands) and quantifies similarity by measuring the angle between the generated and reference spectral vectors. For two spectral vectors x and y, SAM is defined as

$$\text{SAM}(x, y) = \arccos \left(\frac{\langle x, y \rangle}{\|x\|_2 \cdot \|y\|_2} \right), \tag{3.8}$$

where $\langle x, y \rangle$ denotes the dot product and $\| \cdot \|_2$ is the Euclidean norm.

SAM is typically expressed in degrees, with smaller values indicating higher spectral similarity and less distortion. Since it only considers the direction of the spectral vectors and not their magnitude, SAM is invariant to changes in illumination, making it particularly suitable for remote sensing and multispectral image analysis [1]. In practice, the global SAM score is computed as the average angle across all pixels in the image.

LPIPS The Learned Perceptual Image Patch Similarity (LPIPS) metric was proposed by Zhang et al. [82] to provide a perceptual measure of image similarity that better aligns with human visual judgment. It compares feature activations from pretrained convolutional networks, thereby capturing high-level semantics and perceptual realism. For two images x and y, LPIPS is defined as

$$\text{LPIPS}(x, y) = \sum_{l} w_l \cdot \|f_l(x) - f_l(y)\|_2, \tag{3.9}$$

where $f_l(\cdot)$ denotes the feature representation in the l-th layer of the network and w_l is a learned weight.

By measuring differences in a deep feature space rather than raw pixel intensities, LPIPS reflects perceptual similarity and visual realism. Lower LPIPS values indicate that the generated image is closer to the reference in terms of human-perceived quality [22,46], making this metric particularly useful for evaluating the naturalness of translated images.

MAE & RMSE Similar to PSNR, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) evaluate image reconstruction quality on a pixel-wise level. They are defined as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - x_i| \tag{3.10}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - x_i)^2} \tag{3.11}$$

where x_i and y_i denote the pixel intensity values of the reference (ground-truth) and generated images, respectively, and N represents the total number of pixels, defined as $N = C \times H \times W$, with C, H, and W being the number of channels, height, and width of the image.

The Mean Absolute Error (MAE) measures the average absolute difference between corresponding pixels, providing a direct estimate of overall reconstruction accuracy. The Root Mean Square Error (RMSE), in contrast, penalizes larger deviations more heavily due to the squared term, making it more sensitive to outliers and localized reconstruction errors.

Overall, this combination of metrics provides a comprehensive assessment encompassing spatial fidelity, structural integrity, spectral accuracy, and perceptual quality. The evaluation scope of each metric, along with its advantages and limitations, is summarized again in Table 3.3.

Table 3.3.: Summary of evaluation metrics for SAR-to-multispectral translation.

Metric	Aspect Evaluated	Advantages	Limitations
PSNR	Pixel-level fidelity via mean squared error ratio	Simple, widely used, interpretable in terms of noise or distortion	Correlates weakly with human perception; sensitive to pixel shifts
SSIM	Structural similarity (luminance, contrast, texture)	Captures perceptual structure better than PSNR; patch-based	Still intensity-based; limited correlation with perceptual realism
SAM	Spectral fidelity across bands	Invariant to illumination; critical for multispectral data integrity	Ignores spatial and structural context; only reflects spectral angle
MAE & RMSE	Pixel-wise reconstruction accuracy	Provide direct quantitative estimates of average and large-magnitude errors; easy to interpret	Do not account for perceptual or structural quality; sensitive to scaling and outliers (especially RMSE)
LPIPS	Perceptual similarity using deep feature representations	Aligns well with human judgment; sensitive to high-level semantic differences	Requires pretrained CNN; limited to 3-channel inputs unless adapted

4. Results

Both quantitative and qualitative assessments are conducted to examine the model's performance in translating Sentinel-1 SAR imagery into full-spectrum Sentinel-2 optical outputs. The primary objectives are to evaluate the model's reconstruction capability, investigate the influence of training data scale, and assess its ability in cloud removal. Given the high sensitivity of GAN-based frameworks to hyperparameter choices, several experimental configurations were explored to ensure stable convergence and high-quality reconstructions. Initial experiments were performed using 20% of the winter subset to efficiently test different settings. However, the results obtained with this limited data were unsatisfactory. Following the hypothesis that training on a larger dataset would enhance performance, the model was subsequently trained on the entire winter subset, consisting of 31,825 image pairs. Beyond the dataset-scale comparison, additional analyses examine the reconstruction quality across individual Sentinel-2 bands and evaluate the model's ability to remove cloud contamination using the complementary SEN12-MS-CR dataset. Together, these experiments provide a comprehensive view of the model's spatial, spectral, and perceptual performance under realistic remote-sensing scenarios.

4.1. Results on 20% of the Winter Subset

In the initial stage of experimentation, only 20% of the data were used for training. From the 31,825 image pairs in the winter subset, 3,215 pairs were allocated for training, 981 for validation, and 981 for testing. The model was trained to reconstruct the full optical spectrum consisting of 13 bands, using the dual VV and VH polarization SAR data as input. The preprocessing steps, training pipeline, and hyperparameter settings were identical to those described previously and were applied unchanged to the experiments on the full winter subset. Moreover, the training was conducted using the full combination of loss functions, as discussed in Section 3.4.2.

Examining the qualitative results in Figure 4.1, the model successfully captures large-scale structural patterns such as boundaries, edges, and terrain formations. However, it struggles to reproduce fine-grained details and textural content. For instance, in row (a), the boundaries of the agricultural fields are well preserved, but the internal texture of the fields is poorly recon-

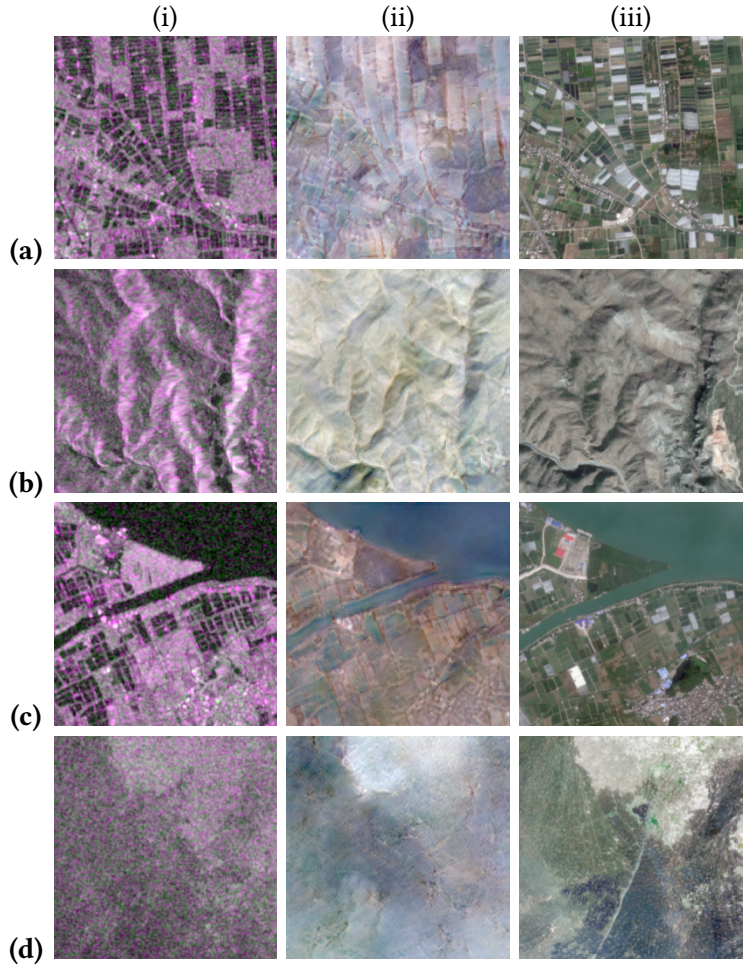


Figure 4.1.: Qualitative results of 20% training winter subset. Columns: **(i)** SAR input (pseudo RGB; R: VV, G: VH, B: VV/VH), **(ii)** model-generated optical image, **(iii)** ground-truth Sentinel-2 image. All optical images throughout this thesis are depicted in RGB (B4, B3, B2) stack.

structed. Similarly, in row (b), the terrain structure is correctly represented, yet the elevation contrast and depth variation are not accurately reproduced. In row (c), the coastline and water boundaries are distinctly captured, whereas the urban area in the bottom-right corner appears blurred and lacks definition. Lastly, in row (d), since the corresponding SAR input contains limited structural information, the generated optical output deviates substantially from the ground truth, indicating the model's reduced ability to infer fine details in textureless regions.

Overall, these observations, together with the quantitative results presented in Table 4.1, confirm that the model effectively learns the underlying SAR-to-optical mapping and that an optimal training configuration was achieved. While the Pix2Pix model demonstrates strong capability in capturing global spatial correspondences, it exhibits limitations in synthesizing fine textures, particularly within homogeneous or low-contrast regions of the SAR input. Furthermore, the inherent differences in sensing mechanisms and physical characteristics between SAR and optical imagery pose additional challenges for accurately reconstructing fine spatial and spectral details. Motivated by these findings, it was hypothesized that model performance could be enhanced through training on a larger-scale dataset. To validate this hypothesis, the model was subsequently trained on the complete winter subset, and the corresponding results are presented in the following section

Table 4.1.: Quantitative results of the training on 20% of the winter subset.

SSIM	**PSNR (dB)**	**LPIPS**	**SAM (°)**	**MAE**	**RMSE**
0.859	27.65	0.224	6.71	195.30	381.57

4.2. Results on the Full Winter Subset

GAN-based models generally require large amounts of training data to achieve high-quality image generation, particularly when using mono-temporal SAR imagery as input for translation to optical domains [30], and especially when reconstructing the full spectral range. Since training on only 20% of the dataset did not yield satisfactory results, an additional experiment was conducted using the full winter subset, which comprises 31,825 samples divided in an 8:1:1 ratio for training, validation, and testing. This corresponds to approximately 25,460 samples for training and 3,180 samples each for validation and testing. The data preprocessing procedure, training pipeline, and hyperparameters were kept identical to those used in the 20% experiment to ensure that the effect of training data size was isolated and directly evaluated.

The model required approximately 25∼hours, using the hardware described in 3.4.1, to complete 150 epochs of training. The experimental hypothesis was confirmed: increasing the size

of the training dataset led to a clear improvement in model performance. As shown in Table 4.2, expanding the training data from 20% to the full winter subset resulted in substantial gains across all evaluation metrics. In particular, the LPIPS score decreased from 0.224 to 0.173, indicating that the model trained on the full dataset produced outputs with higher perceptual similarity to the ground truth. Likewise, the median SAM value dropped from 6.71° to 4.41°, demonstrating a notable enhancement in spectral consistency across all bands. Furthermore, PSNR increased by approximately 5∼dB, while MAE and RMSE decreased by more than 25%, confirming the strong positive effect of data scale on reconstruction quality.

Table 4.2.: Quantitative results of training on 20% and 100% of the winter subset. Arrows (↑ / ↓) indicate whether higher or lower values denote better performance, respectively.

Training Data	**SSIM ↑**	**PSNR (dB) ↑**	**LPIPS ↓**	**SAM (°) ↓**	**MAE ↓**	**RMSE ↓**
20% of subset	0.859	27.65	0.224	6.71	195.30	381.57
100% of subset	**0.888**	**32.63**	**0.173**	**4.41**	**140.72**	**233.69**

Similarly, the image reconstruction quality improves noticeably with a larger training dataset. Qualitative examples are shown in Figure 4.2. The model trained on the full winter subset not only better preserves the structural details of the reference images but also achieves markedly enhanced color fidelity, as evident in column (a). In the urban scene (b), it accurately reconstructs the city layout and successfully delineates the river traversing the area. Furthermore, compared to the model trained on only 20% of the data, the full-data model better captures surface relief and elevation depth, as illustrated in (c).

These qualitative and quantitative improvements demonstrate the capability of the Pix2Pix model to reconstruct optical imagery solely from SAR data. Furthermore, expanding the volume and diversity of the training dataset enables the model to more effectively learn the perceptual and spectral characteristics of the optical domain, thereby producing reconstructions that are both more realistic and color-consistent.

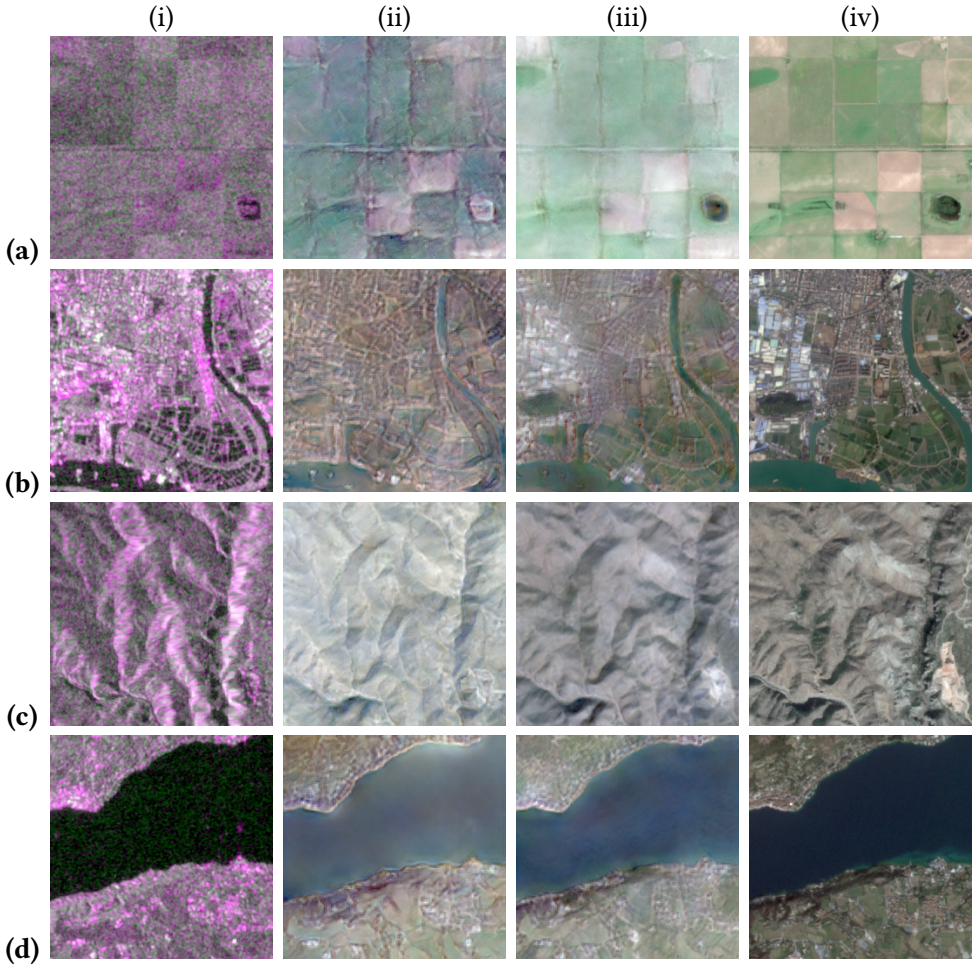


Figure 4.2.: Qualitative comparison of models trained on 20% and 100% of the dataset. Columns: **(i)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(ii)** generated optical image from 20% training, **(iii)** generated optical image from 100% training, and **(iv)** ground-truth Sentinel-2 image. All optical images are depicted in RGB (B4, B4, B2) batch.

4.3. 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 extent of its accuracy across the spectrum. For this purpose, the model trained on the full winter subset was evaluated separately for all Sentinel-2 bands, and the corresponding results are summarized in Table 4.3. When comparing the reconstruction quality across individual bands, the focus is placed on the unitless SSIM metric. Other metrics such as MAE or RMSE are not directly comparable between bands, as they depend on the absolute magnitude and statistical distribution of reflectance values, which differ across spectral ranges. In contrast, SSIM measures local structural similarity based on relative intensity patterns rather than absolute values. While not entirely invariant to scale differences, SSIM provides a more robust and interpretable basis for cross-band comparison in this context.

Table 4.3.: Per-band quantitative validation results of the Pix2Pix model trained on the full winter subset. Each Sentinel-2 band's central wavelength, spectral designation, and native spatial resolution are listed for reference. Arrows (↑ / ↓) indicate whether higher or lower values denote better performance, respectively. Best results are highlighted in green and worst in red.

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 4.3. 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 more reliably, with B1 achieving the

highest SSIM overall. Their smoother spectral characteristics and lower spatial variability likely facilitate more stable and accurate predictions, even after resampling to 10 m.

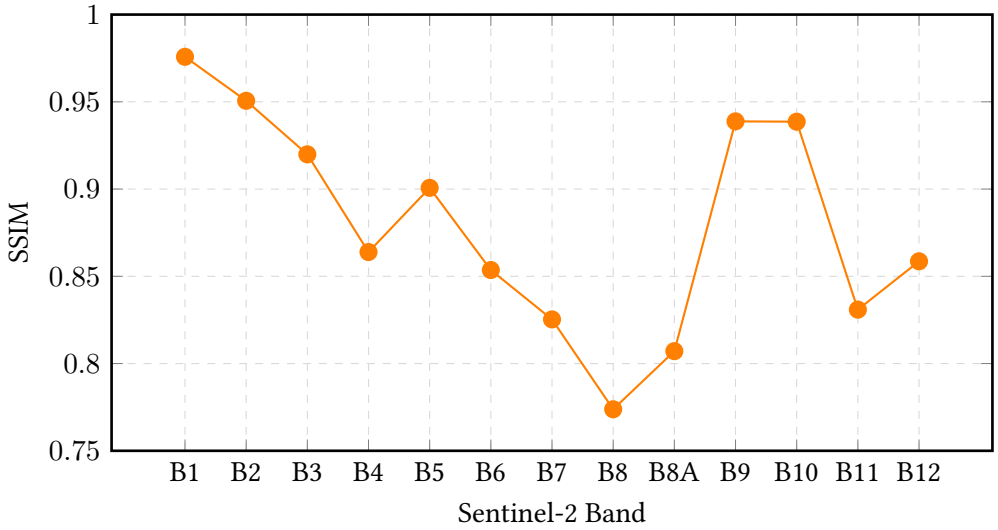


Figure 4.3.: Per-band SSIM for the Pix2Pix model trained on the full winter subset.

It is worth noting, however, that these findings contradict those reported in [2], where the authors observed the 10 m and 20 m bands to achieve the highest reconstruction quality, while the 60 m bands performed the worst. Their experiments were conducted on the full SEN12-MS-CR dataset, which may account for the differing behavior observed in the present study.

To visually complement the quantitative assessment, representative grayscale examples for each Sentinel-2 band are provided in Appendix B. Each example illustrates the generated band alongside its corresponding ground-truth reference, enabling a direct visual evaluation of the reconstruction quality and spatial consistency across the spectrum.

4.4. Results on Cloud Removal

To evaluate the model's capability to address this issue, the model trained on the complete winter subset of the SEN12-MS dataset was assessed using the complementary SEN12-MS-CR dataset (see Section 3.2). The SEN12-MS-CR dataset consists of triplets comprising SAR images, cloud-contaminated optical images, and their corresponding cloud-free optical references, covering identical geographic regions across all 13 Sentinel-2 spectral bands. For quantitative evaluation, a subset of 2,000 samples from the winter portion of the SEN12-MS-CR dataset was used. The evaluation data underwent identical preprocessing and clipping procedures as described in Section 3.2.3, ensuring consistency with the model's training conditions.

Table 4.4.: Quantitative evaluation of different models trained on the full winter subset of the SEN12-MS dataset, assessed on the complementary SEN12-MS-CR dataset for the cloud removal task. Arrows (↑ / ↓) indicate whether higher or lower values denote better performance, respectively.

Model	**SSIM ↑**	**PSNR (dB) ↑**	**LPIPS ↓**	**SAM (°) ↓**
DSen2-CR [2]	0.878	–	–	8.07
FAT [64]	0.880	31.85	–	4.19
DiffCR [46] (SOTA)	**0.902**	31.77	–	5.82
Pix2Pix (This Thesis)	0.899	**33.65**	**0.152**	**3.86**

The quantitative results presented in Table 4.4 demonstrate that the model attains robust and high-quality performance in cloud removal on the SEN12-MS-CR dataset, despite not being explicitly trained for this task or on this dataset. Notably, the reconstruction accuracy surpasses that achieved on the model's own training subset (SEN12-MS) across all evaluated metrics. The spectral fidelity, measured by the SAM, improved by approximately one degree, indicating better preservation of spectral characteristics. Remarkably, the trained Pix2Pix model even outperforms several specialized approaches, including DiffCR [46], the current state-of-the-art model for cloud removal on SEN12-MS-CR—an outcome that was beyond the original scope of this thesis. This result is particularly significant given that Pix2Pix is typically employed as a baseline model in related studies and is often reported to yield the lowest performance among compared methods

The qualitative results shown in Figure 4.4 further illustrate the model's effectiveness in removing cloud contamination. The model successfully reconstructs cloud-free optical imagery from the corresponding SAR inputs, even when the reference optical images are heavily obscured. As seen in rows (a) and (b), the model accurately restores regions affected by thin cloud cover while preserving structural and textural details. In partially occluded cases, illustrated in (c) and (d), the reconstructed outputs maintain clear boundaries, sharp edges, and consistent color representation. Remarkably, even for fully cloud-covered scenes, as in (e), the model generates realistic optical imagery and successfully reproduces urban and structural features.

Overall, both the qualitative and quantitative results confirm the model's ability to generate high-quality, cloud-free optical images across all 13 spectral bands solely from SAR inputs, thereby effectively addressing the cloud cover problem. However, since the model was trained exclusively on translating SAR to cloud-free optical images, the thickness or density of clouds does not influence the reconstruction process.

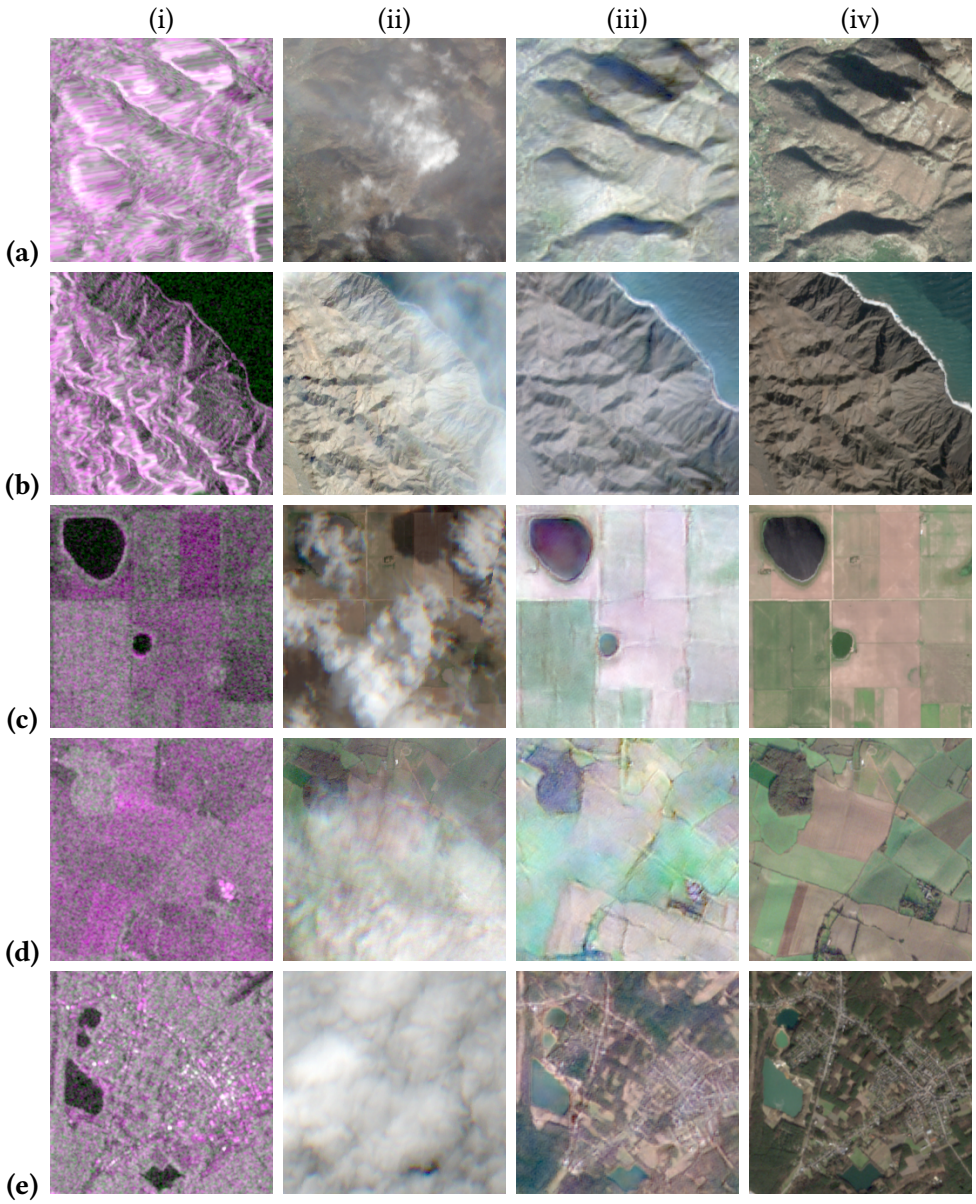


Figure 4.4.: Qualitative results of the model trained on the full winter subset of the SEN12-MS dataset, evaluated on the complementary SEN12-MS-CR dataset for the **cloud removal** task. Columns: **(i)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(ii)** reference cloud-contaminated optical image, **(iii)** generated cloud-free optical image, and **(iv)** reference cloud-free optical image (ground truth).

4.5. Ablation Studies

4.5.1. Effect of Loss Functions

To evaluate the contribution of each loss component to the overall model performance, an ablation study was conducted. Four training configurations were compared:

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 combination $\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{LPIPS}}$.

This analysis aimed to isolate the contribution of each additional loss term to both quantitative performance and visual reconstruction quality. The evaluation was conducted using the same IQA metrics employed throughout the thesis, namely SSIM, PSNR, LPIPS, SAM, MAE, and RMSE. Several training loops were conducted using the same preprocessing procedure described in Section 3.2.3. All models were trained under identical settings, hyperparameters, datasets, and number of epochs to ensure a fair and consistent comparison across the different loss configurations. Table 4.5 summarizes the quantitative performance across the four loss configurations. As shown, the baseline configuration without SSIM and LPIPS achieved the lowest performance across most metrics, with the exception of PSNR. This configuration also represents the weakest setup in terms of overall image quality. As illustrated in Figure 4.5(b), images generated by the baseline model differ significantly from the ground truth, both texturally and perceptually.

Table 4.5.: Quantitative results of the ablation study across different loss configurations. Arrows (↑ / ↓) indicate whether higher or lower values denote better performance, respectively. Best values per metric are shown in bold.

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

Notably, integrating SSIM alone yielded the highest quantitative scores in several metrics. However, the qualitative results under this configuration reveal perceptual inconsistencies and reduced visual realism. This discrepancy arises because SSIM does not always align with human perceptual judgments of image similarity. As reported by Nilsson and Akenine-Möller in [83],

SSIM can overemphasize small intensity variations in dark regions (Section 4.1, p. 4), overlook significant color shifts (Section 4.2, p. 5), and assign high similarity near edges even when visible artifacts are present (Section 4.4, pp. 6–7).

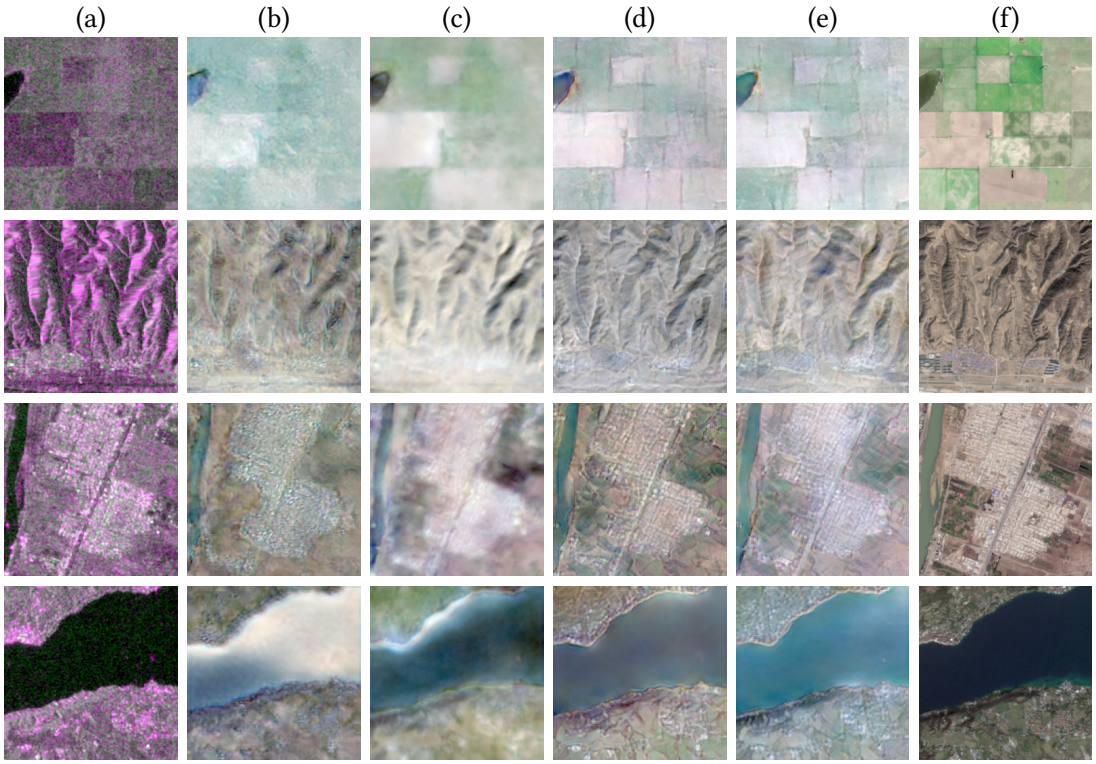


Figure 4.5.: Qualitative ablation results showing representative samples (rows) under four configurations (columns):
(a) Input SAR (pseudo RGB)
(b) $\mathcal{L}_{\text{GAN}}+\mathcal{L}_{\text{L1}}$
(c) $\mathcal{L}_{\text{GAN}}+\mathcal{L}_{\text{L1}}+\mathcal{L}_{\text{SSIM}}$
(d) $\mathcal{L}_{\text{GAN}}+\mathcal{L}_{\text{L1}}+\mathcal{L}_{\text{LPIPS}}$
(e) all four losses combined
(f) ground-truth Sentinel-2 optical image (RGB bands).

Since the primary objective of LPIPS is to measure perceptual similarity, incorporating it into the baseline configuration led to notable improvements in the preservation of low-level features. As illustrated in column (d) of Figure 4.5, the model was able to maintain sharper edges and more distinct boundaries. Moreover, the generated images appear visually more realistic and closely resemble the ground truth in texture and detail. However, the colour reproduction achieved with LPIPS is slightly inferior to that obtained using SSIM, as the SSIM-based model produces more natural and accurate colours overall. The full combination of all four losses

achieved the best balance between pixel-level accuracy, perceptual realism, and structural coherence. Its SAM values were the second best, differing only marginally from the SSIM-only configuration. By incorporating both SSIM and LPIPS, the model attains an effective balance between color, luminance, and perceptual realism, as shown in column (e) of Figure 4.5. Similar findings were also reported by [41] and [4].

Overall, the results demonstrate the incremental benefit of incorporating perceptual and structural similarity terms alongside the adversarial $\mathcal{L}_{\text{GAN}}$ and L1 objectives.

4.5.2. Effect of Excluding 60 m Bands

Given the varying data distributions and the originally different spatial resolutions of the Sentinel-2 bands (even though all were resampled to 10m in the dataset), it was hypothesized that including the 60m bands might introduce noise and confuse the model during training. To examine this assumption, an ablation study was conducted using the same training configuration as for the full 13-band model, both trained on 20% of the winter subset.

Surprisingly, removing the 60m bands did not lead to any improvement. Instead, both SSIM and LPIPS scores decreased, as reported in Table 4.6. This suggests that the 60m bands, which primarily serve for atmospheric correction (e.g., B1, B9, and B10), provide complementary spectral information that contributes to maintaining spectral fidelity during training. It is worth noting, however, that these differences are very small.

Table 4.6.: Overall performance comparing all 13 bands vs. excluding the 60m bands (B1, B9, B10). Arrows (↑ / ↓) indicate whether higher or lower values denote better performance, respectively. Best values per metric are shown in bold.

Setting	**SSIM ↑**	**PSNR (dB) ↑**	**SAM (°) ↓**
All 13 bands	**0.859390**	**27.650251**	6.7064
Excluding 60m (B1, B9, B10)	0.825861	26.474423	**6.3339**

Interestingly, while the SAM slightly improved after excluding the 60m bands, this does not necessarily contradict the decline in PSNR and SSIM. The model may have learned to reproduce more spectrally consistent relationships between the remaining bands (hence a smaller spectral angle), yet failed to preserve the absolute reflectance magnitudes and spatial details, leading to lower overall reconstruction quality. This indicates a trade-off between radiometric fidelity and spectral coherence in the absence of the full spectral range. For qualitative evaluation, Figure 4.6 illustrates the generated optical images obtained from models trained with the full spectral configuration and with the 60m bands excluded.

Similarly, the per-band evaluation (Table 4.7) confirms this trend. For most bands, SSIM

Figure 4.6.: Qualitative comparison of generated optical images when excluding the 60 m Sentinel-2 bands. Columns: **(a)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(b)** generated optical image trained without the 60 m bands, **(c)** generated optical image trained with all 13 bands, and **(d)** reference cloud-free Sentinel-2 image.

remains stable when all bands are included, indicating consistent structural reconstruction quality. However, PSNR decreases for nearly all bands when the 60m bands are excluded, with the exception of B1, demonstrating that the full spectral configuration better supports radiometric reconstruction. It is very important to state that these changes when keeping and excluding the

Table 4.7.: Per-band performance when excluding 60m bands. Arrows (↑ / ↓) indicate whether higher or lower values denote better performance, respectively.

Band	**SSIM (no 60 m) ↑**	**SSIM (13) ↑**	**PSNR (no 60 m) [dB] ↑**	**PSNR (13) [dB] ↑**
B1	—	0.9634	—	29.749
B2	0.9431	0.9430	34.151	34.057
B3	0.9064	0.9060	31.691	31.718
B4	0.8348	0.8340	28.039	28.145
B5	0.8707	0.8729	28.098	28.257
B6	0.8199	0.8231	26.740	26.926
B7	0.7890	0.7925	25.694	25.884
B8	0.7375	0.7416	25.534	25.739
B8A	0.7687	0.7722	25.095	25.311
B9	—	0.8666	—	24.617
B10	—	0.8826	—	24.571
B11	0.7774	0.7809	23.395	23.727
B12	0.8034	0.8073	24.956	25.263

5. Discussion

5.1. Challenges Due to Inherent Model Characteristics

Although the Pix2Pix model demonstrated stable training behavior and produced high-quality image reconstructions, several challenges arose during the training process. These challenges stem from the inherent characteristics of GAN-based architectures, of which Pix2Pix is a representative example. One of the main issues encountered was the vanishing or exploding gradient problem, where the early layers of the network receive minimal updates during backpropagation, leading to slow convergence or even training stagnation. The literature attributes this behavior primarily to the choice of activation functions and optimization strategies. For instance, [4] and [9] address this issue by incorporating residual blocks to improve gradient flow. In the case of Pix2Pix, however, the vanishing gradient problem was mitigated by replacing the default binary cross-entropy (BCE) loss with the least-squares loss function. Unlike BCE, which tends to saturate when the discriminator becomes overconfident, the least-squares formulation penalizes outputs based on their squared distance from the target labels, thereby maintaining non-zero gradients even for well-classified samples [41].

Moreover, the Pix2Pix model inherently incorporates an L1 loss term. However, a well-known limitation of the L1 loss is that it is not well suited for generating high-resolution or perceptually rich images [41]. To overcome this limitation, additional loss components based on the SSIM and the perceptual LPIPS metric were integrated into the objective function. This enhancement enabled the model to more reliably reproduce both the visual and spectral characteristics of the ground-truth multispectral images.

Another issue encountered during training was the emergence of *checkerboard artifacts* in the generated images. These artifacts appeared as grid-like patterns, particularly visible in homogeneous regions such as water bodies and vegetated surfaces. The phenomenon originates from the use of transposed convolutions in the generator's upsampling layers, where uneven overlap between convolutional kernels causes certain pixels to receive disproportionately large updates [84]. The same issue was also acknowledged in the official Pix2Pix implementation. To mitigate this problem, the transposed convolution layers were replaced with a combination of nearest-neighbor upsampling followed by standard convolution operations (Listing 5.2). This

modification ensured uniform pixel coverage, effectively eliminating checkerboard artifacts and resulting in smoother and more visually coherent image reconstructions.

```
# Original implementation using ConvTranspose2d
nn.ConvTranspose2d(
   inchannels = ngf * mult,
   outchannels = int(ngf * mult / 2),
   kernelsize = 4,
   stride = 2,
   padding = 1,
   bias = usebias
)
```

Listing 5.1: Original transposed convolution block in Pix2Pix

```
# Replaced with nearest-neighbor upsampling followed by regular
   convolution
nn.Upsample(scalefactor = 2, mode = 'bilinear'),
nn.ReflectionPad2d(1),
nn.Conv2d(
   inchannels = ngf * mult,
   outchannels = int(ngf * mult / 2),
   kernelsize = 3,
   stride = 1,
   padding = 0
)
```

Listing 5.2: Modified resize-conv block to mitigate checkerboard artifacts

5.2. Model-Specific Limitations of GAN-Based Translation

Despite the remarkable performance achieved by the proposed model, several 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 5.1.

This limitation suggests that future research should focus on exploring alternative generative architectures, such as Diffusion Models. Unlike GAN-based approaches, Diffusion Models

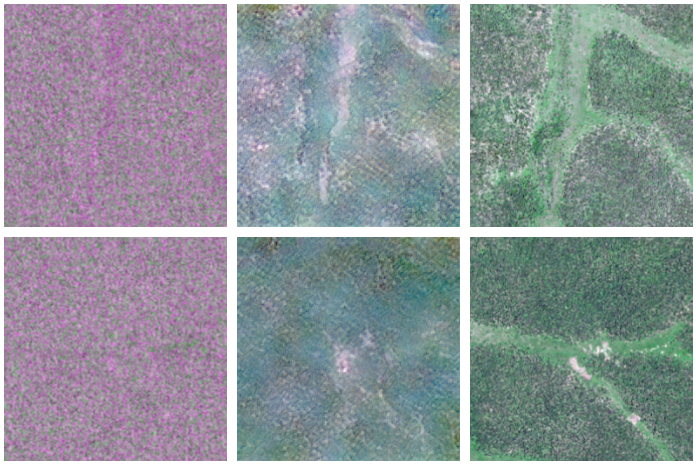


Figure 5.1.: Qualitative examples illustrating the limitation of the SAR-to-optical translation model when the SAR input lacks clear structural information. Columns: **(i)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(ii)** model-generated optical image, and **(iii)** reference cloud-free Sentinel-2 image.

learn the underlying data distribution by iteratively adding and removing noise, which often results in more stable training and higher-quality image synthesis. Furthermore, while GAN-based models are known for their instability and limited capacity to further enhance image fidelity, Diffusion Models have recently demonstrated superior performance in producing high-resolution, photorealistic outputs. Notably, the current state-of-the-art method for cloud removal on the SEN12-MS-CR dataset (see Section 3.2) is a Diffusion-based approach, namely *DiffCR* [46], underscoring the growing effectiveness of these models in handling complex remote sensing translation tasks. Representative works are [8, 10, 43, 66]

Incorporating attention mechanisms into model architectures has been shown to significantly enhance cloud removal and translation performance. Pan [44] introduced the Spatial Attention GAN (SpA-GAN), which integrates a spatial attention mechanism into the generator to imitate the human visual system by focusing selectively on cloud-contaminated regions. This mechanism enables the model to adaptively allocate attention to cloudy areas, improving the recovery of fine spatial details while preserving unaffected regions. Furthermore, the attention maps are also used in the loss function, helping the model prioritize relevant areas during training. Experimental results on the RICE [85] dataset demonstrated substantial gains in PSNR and SSIM over conventional cGAN and CycleGAN models. Therefore, incorporating attention modules in future SAR-to-optical translation architectures could improve spatial awareness, enhance feature consistency, and yield more visually and spectrally accurate reconstructions

5.3. Temporal Generalizability Across Different Seasons

Another important limitation lies in the model's temporal generalizability. The Pix2Pix model was trained and evaluated primarily on the winter subset of the SEN12-MS dataset, which ensures a consistent data distribution and spectral domain during training. While the model exhibits reasonable spatial generalization when evaluated on the SEN12-MS-CR dataset—which features a distinct set of regions of interest (ROIs) compared to SEN12-MS—it struggles to maintain the same level of performance across different seasonal subsets. In particular, when applied to the summer, fall, or spring subsets, the model demonstrates a noticeable degradation in reconstruction quality, indicating sensitivity to seasonal variability in vegetation, soil moisture, and illumination conditions. Quantitative and qualitative results for the different seasons are provided in Appendix C.

Future work should therefore aim to enhance both spatial and temporal robustness. Potential strategies include domain adaptation techniques, fine-tuning with representative samples from multiple seasons and regions, and data augmentation approaches that simulate seasonal and spatial variability. Regarding spatial generalizability, the random global sampling used in the SEN12 dataset family provides a strong degree of spatial robustness. Nevertheless, fine-tuning the model on localized subsets may further improve performance for region-specific applications.

5.4. Aware per-Band Clipping

As described in Section 3.2.3, the optical data values were clipped to the range [0, 10,000], regardless of the spectral band. This clipping strategy is consistently adopted across the literature and has therefore been followed in this thesis. However, despite the uniform clipping, the actual value distributions differ significantly among the bands, as revealed by a statistical analysis performed on the winter subset. From each ROI in the winter subset (47 ROIs), one optical `.tif` file was analyzed to examine the pixel value distributions across the 13 Sentinel-2 bands. The resulting histograms, depicted in Appendix D, show that the dynamic range and distribution shape vary considerably between bands. A particularly distinct observation is seen in the Cirrus band (B10, indexed as 11 in the figure), where the pixel values are compressed into a very limited, discrete set of intensity levels, unlike the continuous distributions of other bands.

The statistics are moreover numerically reported in Table 5.1. This uniform clipping of all optical bands to a fixed range such as [0, 10000] disregards the intrinsic spectral variability between bands. Low-range bands become underrepresented, and high-range bands risk satu-

ration, leading to an imbalanced input distribution. This can hinder convergence and degrade the realism of generated optical images. Band-specific normalization based on empirical percentiles could provide a more balanced and physically consistent preprocessing strategy. Although these statistics were computed from only 47 optical samples, the analysis offers valuable insights that can inform and guide future research on this important yet often overlooked aspect in the literature.

Table 5.1.: Summary statistics of the Sentinel-2 optical bands across the winter subset. Reported are the global minimum and maximum values, and the 1^{st}, 50^{th} (median), and 99^{th} percentiles.

Band	**Global Min**	**Global Max**	**p1**	**p50**	**p99**
B1 (Coastal Aerosol)	940	3712	965	1387	2034
B2 (Blue)	634	5759	725	1101	1998
B3 (Green)	395	6001	498	960	2423
B4 (Red)	223	6748	331	893	3543
B5 (Red Edge)	204	6377	291	1119	3758
B6 (Red Edge)	139	6532	253	1643	4042
B7 (Red Edge)	117	6875	239	1834	4389
B8 (NIR)	185	7048	202	1805	4211
B8A (Narrow NIR)	108	7130	199	1995	4542
B9 (Water Vapour)	43	3329	85	720	2214
B10 (Cirrus)	3	303	7	16	192
B11 (SWIR)	34	6173	75	1802	5777
B12 (SWIR)	1	5252	46	1180	4825

This thesis successfully achieved the objectives outlined at the beginning. The first goal, which is reconstructing full-spectrum optical images solely from SAR inputs, was met with remarkable results after fine-tuning, despite the inherent challenges of training GANs. The observed difference in performance when training the Pix2Pix model on 20% versus 100% of the dataset highlights the computational intensity and data dependency of GAN-based approaches. Although the findings confirm the capability of GANs for this task, the model exhibits limitations when the SAR input lacks structural detail or sufficient backscatter information, leading to noise-like outputs. In this context, diffusion models emerge as a promising alternative for future research, offering greater stability and generative fidelity.

Furthermore, the thesis demonstrates the model's ability to address the cloud cover problem using samples from the SEN12-MS-CR dataset, despite the model not being explicitly

trained for this task. Remarkably, the Pix2Pix model even surpassed the state-of-the-art DiffCR model—an unexpected outcome, given that Pix2Pix is typically reported in the literature to produce weaker results. Because the model was trained exclusively to translate SAR data into cloud-free optical imagery, cloud thickness or density does not influence the reconstruction process. This suggests that SAR-to-optical translation might, in fact, be more effective for cloud removal than approaches combining SAR with cloud-contaminated optical images, as the latter can introduce spectral noise due to temporal discrepancies between acquisitions. A comparative evaluation of both strategies therefore remains an important avenue for future work. Another aspect worth exploring in future research, which has not yet been discussed in the literature, is the impact of location, date, and time on the reconstruction process when incorporated as auxiliary inputs during training.

The per-band evaluation indicated that model performance is largely independent of spectral characteristics. Excluding the 60m bands—and thereby reducing problem complexity and computational cost—did not yield a substantial improvement (see 4.5.2). Nonetheless, training models on only the spectral composites relevant to specific downstream tasks could be advantageous, since most Earth observation applications employing cloud removal as a preprocessing step do not require all 13 Sentinel-2 bands [2].

From the ablation study on loss functions, discussed in 4.5.1, it is recommended to combine pixel-wise reconstruction losses with perceptual components such as LPIPS, which help capture perceptual similarity beyond pixel accuracy. Moreover, incorporating spectral-aware metrics such as the Spectral Angle Mapper (SAM) into the loss function could enhance spectral consistency across bands—an aspect worth exploring in future research.

6. Conclusion

This thesis explored the potential of SAR-to-optical image translation as a generative framework for synthesizing multispectral optical imagery from radar data, 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, the Pix2Pix conditional generative adversarial network (cGAN) was trained to translate dual-polarized SAR inputs into full-spectrum optical outputs across all 13 Sentinel-2 bands. The study confirmed that meaningful spectral and spatial correlations exist between the SAR and optical domains, enabling reliable reconstruction of high-fidelity optical imagery. Reconstruction quality was found to vary across spectral bands, reflecting differences in wavelength sensitivity and signal characteristics. The model's performance in cloud removal further demonstrated that SAR-to-optical translation can effectively generate cloud-free imagery even without explicit training for this task—achieving results comparable to or surpassing several state-of-the-art approaches.

Ablation experiments showed that combining SSIM and LPIPS losses with the standard GAN objective enhances reconstruction consistency and perceptual realism. Meanwhile, analysis of per-band clipping highlighted a commonly overlooked limitation in existing literature: the uniform clipping of optical data across all bands disregards their distinct value distributions, potentially reducing spectral fidelity. While the results validate the feasibility and promise of SAR-to-optical translation, challenges remain. Model performance decreases in textureless or spectrally complex regions, and temporal generalization is limited by training exclusively on winter data. Addressing these limitations through seasonally diverse training sets, diffusion-based generative models, and per-band-aware preprocessing would likely further improve spectral realism and robustness.

Overall, this work establishes that generative models—particularly GAN-based architectures—can reconstruct full-spectrum, cloud-free optical imagery from SAR data with competitive accuracy. These findings reinforce the potential of generative approaches to enhance the temporal continuity and usability of optical remote sensing data under challenging observation conditions.

A Intermediate Outputs during Training

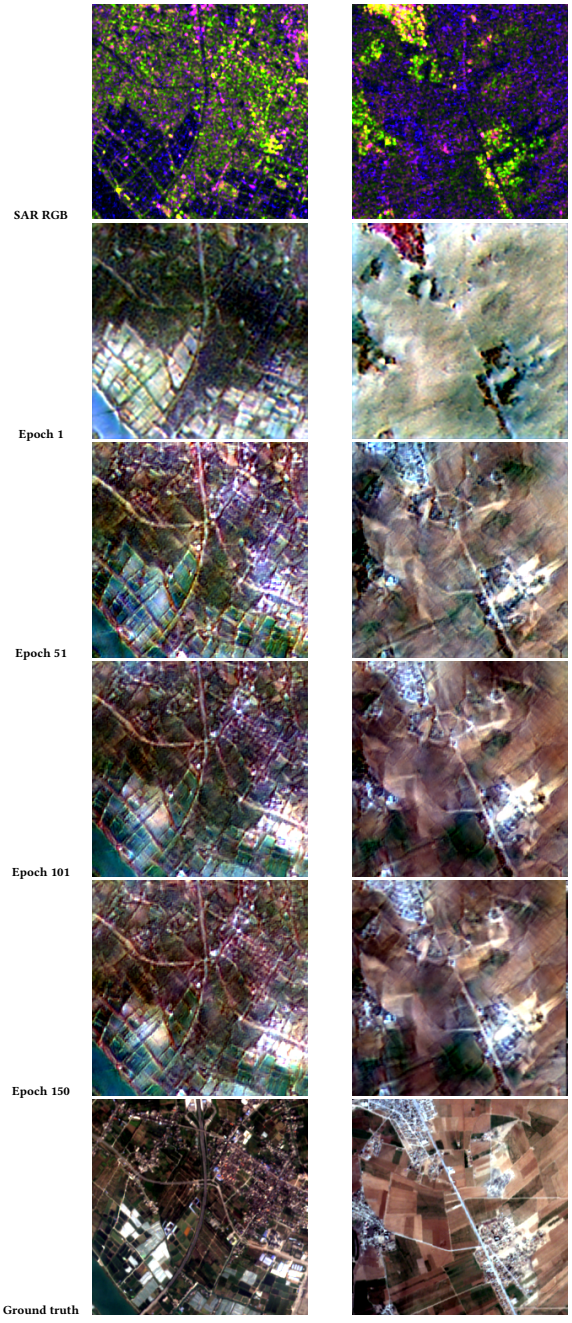
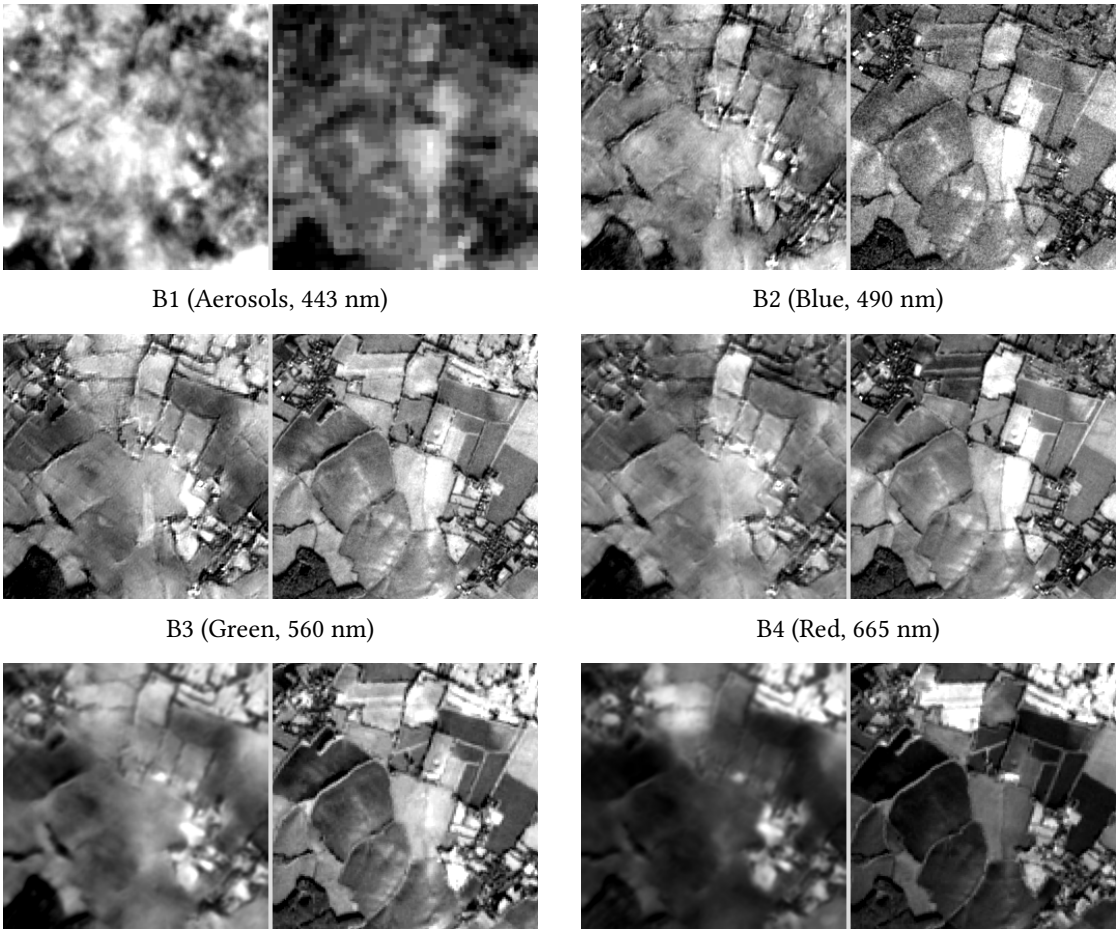


Figure A.1.: Evolution of generated optical images across training epochs for two samples (columns). Rows show pseudo-RGB SAR input (R: VV, G: VH, B: VV/VH), generated optical images after epochs 1, 51, 101, 150, and the ground truth.

B Bandwise Grayscale Reconstructions

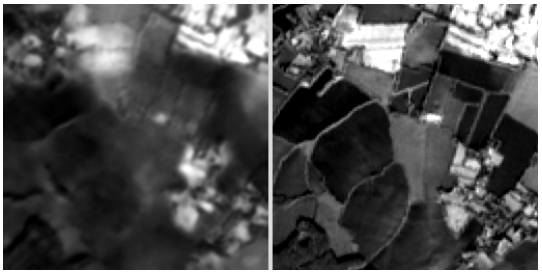


B1 (Aerosols, 443 nm) B2 (Blue, 490 nm)

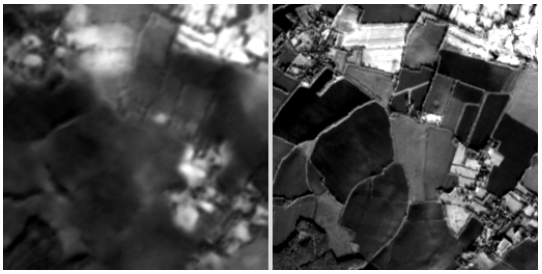
B3 (Green, 560 nm) B4 (Red, 665 nm)

B5 (Red Edge, 705 nm) B6 (Red Edge, 740 nm)

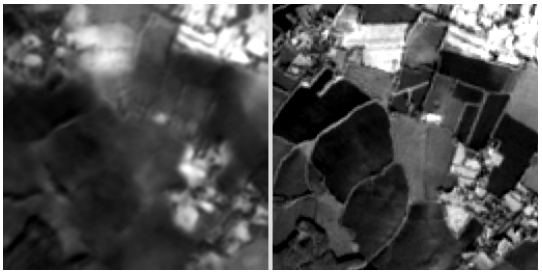
Figure B.1.: Generated (left) and ground-truth (right) grayscale representations for Sentinel-2 Bands 1–6.



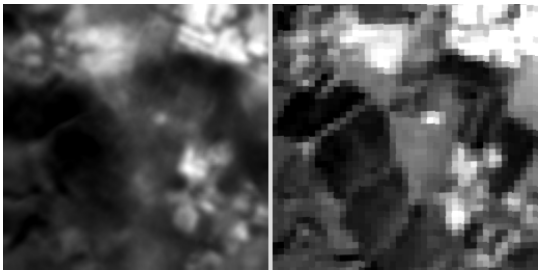
B7 (Red Edge, 783 nm)



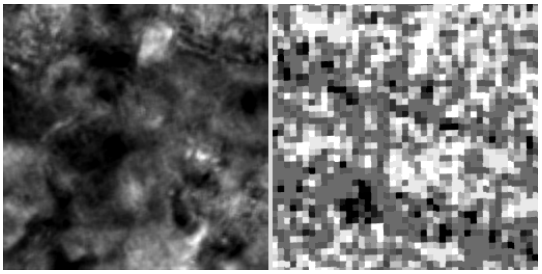
B8 (NIR, 842 nm)



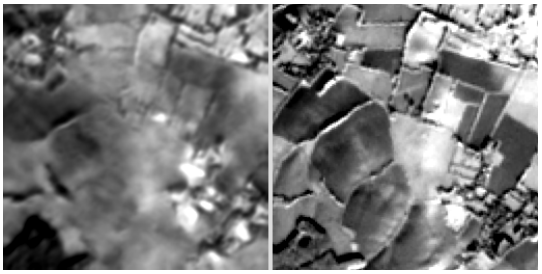
B8A (Red Edge, 865 nm)



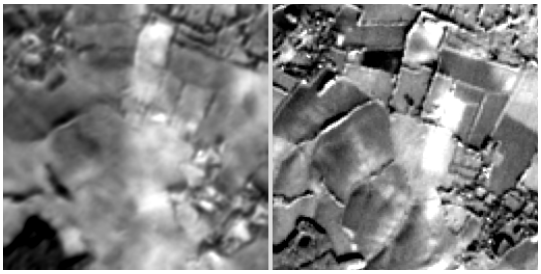
B9 (Water Vapour, 945 nm)



B10 (Cirrus, 1375 nm)



B11 (SWIR, 1610 nm)



B12 (SWIR, 2190 nm)

Figure B.1.: Generated (left) and ground-truth (right) grayscale representations for Sentinel-2 Bands 7–12.

C Results Across Different Seasons

Table C.1.: Quantitative performance of the model on the fall subset of the SEN12-MS dataset.

Season	SSIM	PSNR (dB)	LPIPS	SAM (°)
Fall	0.788	22.40	0.269	13.67

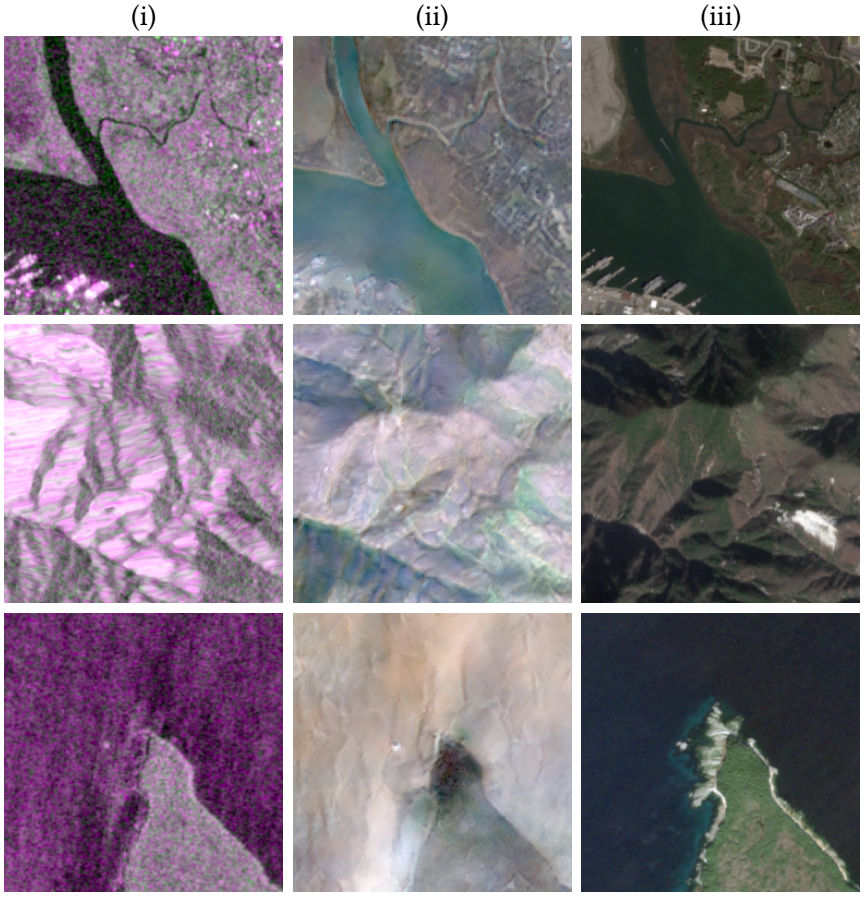


Figure C.1.: Representative SAR-to-optical translation result on the **fall** subset from the SEN12-MS dataset. Columns: **(i)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(ii)** model-generated optical image, **(iii)** ground-truth Sentinel-2 image.

Table C.2.: Quantitative performance of the model on the spring subset of the SEN12-MS dataset.

Season	**SSIM**	**PSNR (dB)**	**LPIPS**	**SAM (°)**
Spring	0.791	20.00	0.297	12.06

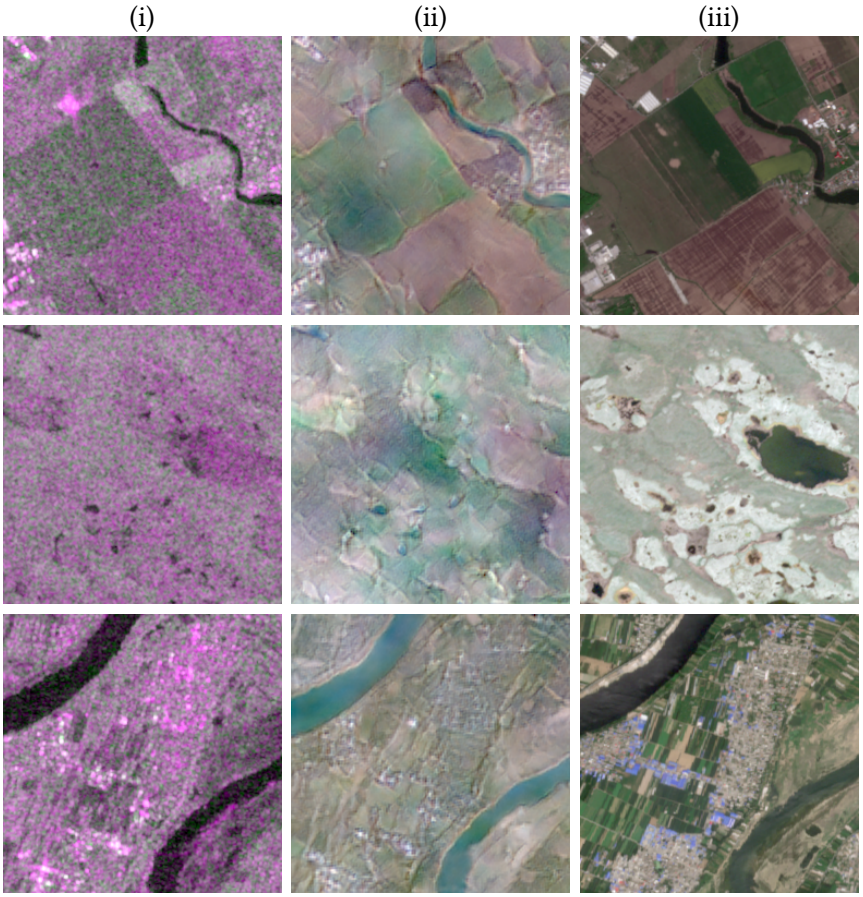


Figure C.2.: Representative SAR-to-optical translation result on the **spring** subset. Columns: **(i)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(ii)** model-generated optical image, **(iii)** ground-truth Sentinel-2 image.

Table C.3.: Quantitative performance of the model on the summer subset of the SEN12-MS dataset.

Season	SSIM	PSNR (dB)	LPIPS	SAM (°)
Summer	0.800	22.79	0.268	12.43

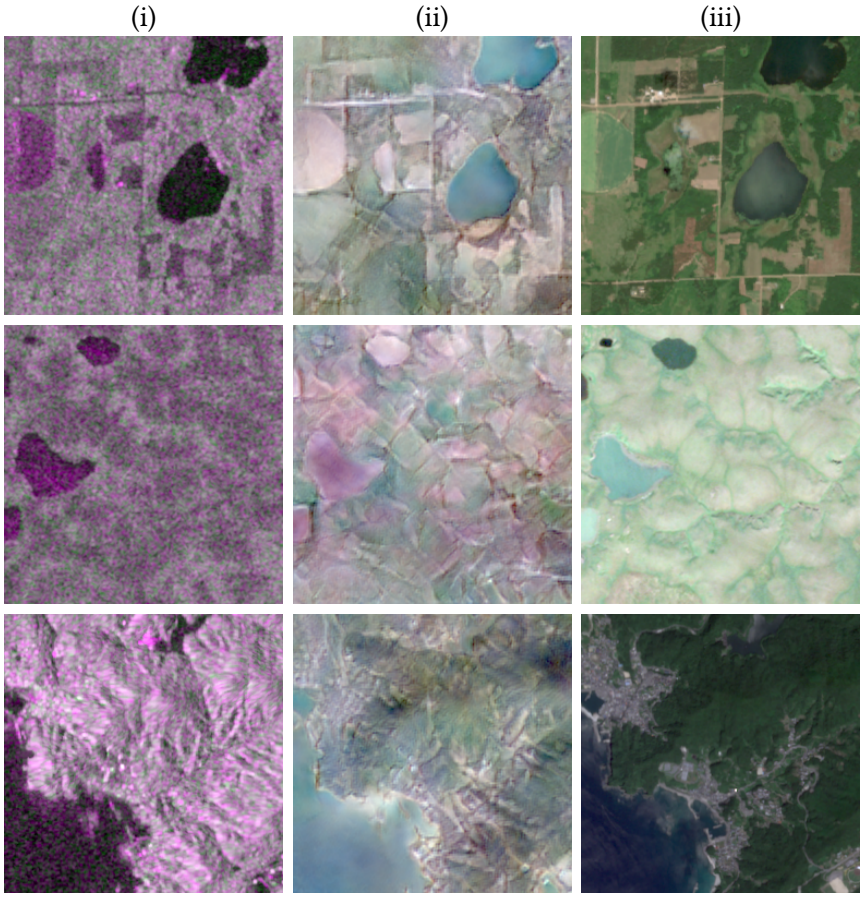
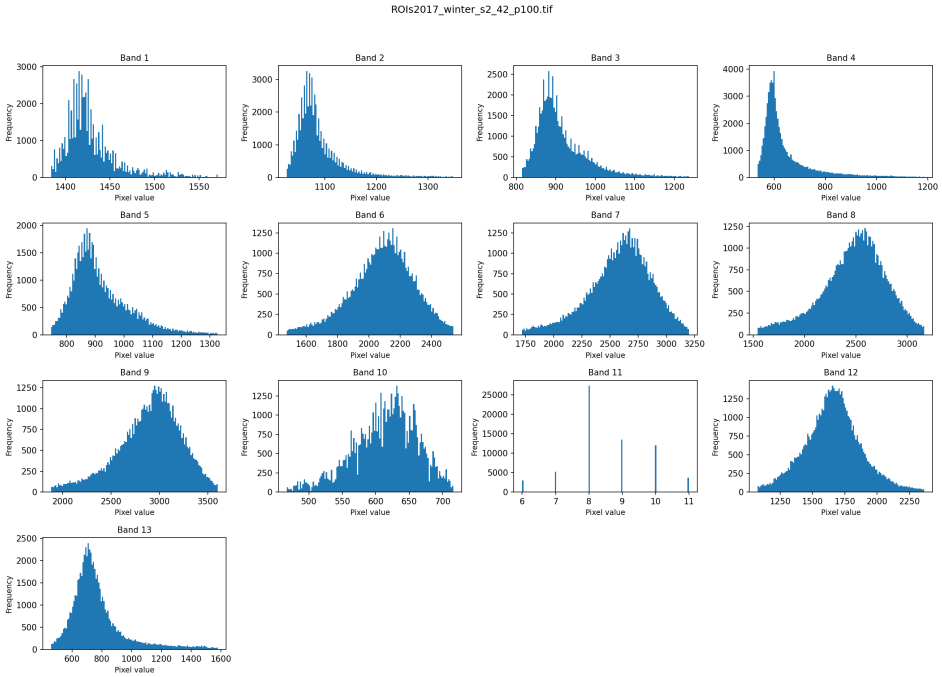
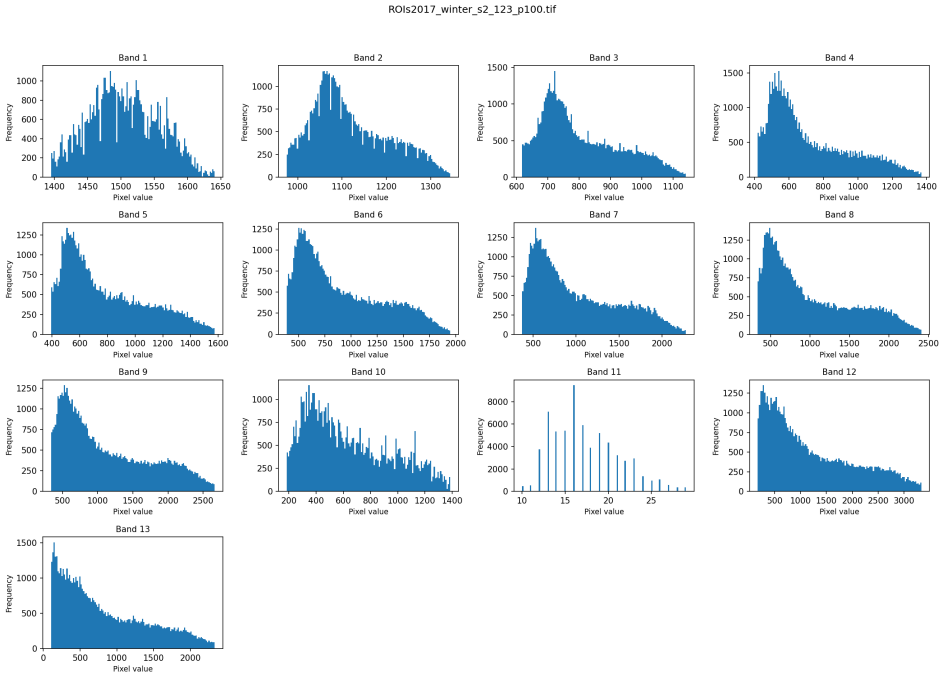


Figure C.3.: Representative SAR-to-optical translation result on the **summer** subset. Columns: **(i)** SAR input (pseudo-RGB; R: VV, G: VH, B: VV/VH), **(ii)** model-generated optical image, **(iii)** ground-truth Sentinel-2 image.

D Value Distributions of Individual Optical Bands





D. Value Distributions of Individual Optical Bands

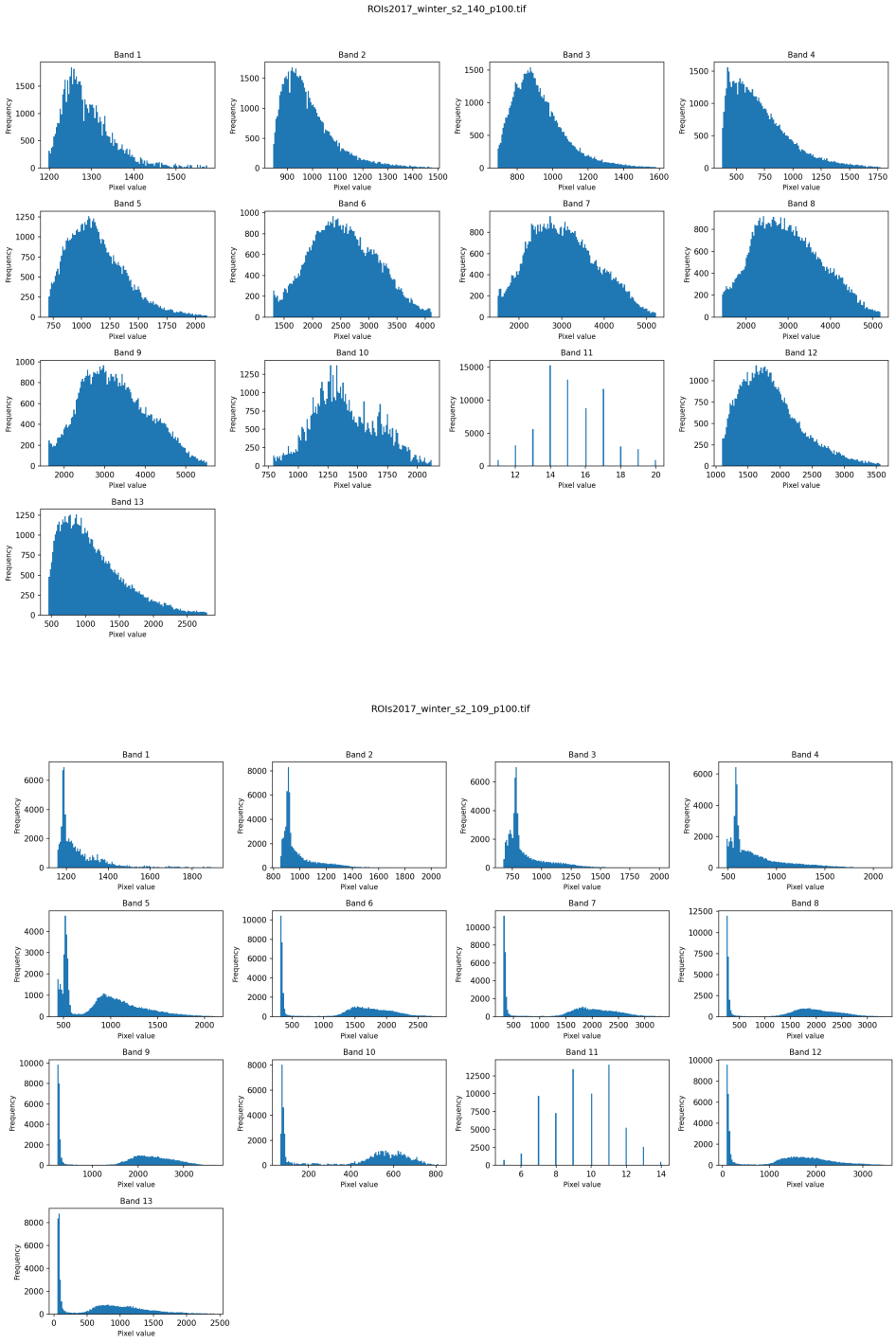


Figure D.1.: Value distributions of the individual optical bands. The plots show per-band reflectance range and relative scaling across spectral channels.

Bibliography

[1] Y. Liu, Q. Han, H. Yang, and H. Hu, "High-Resolution SAR-to-Multispectral Image Translation Based on S2MS-GAN," *Remote Sensing*, vol. 16, no. 21, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/21/4045

[2] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620301398

[3] K. Shen, G. Vivone, X. Yang, S. Lolli, and M. Schmitt, "A benchmarking protocol for SAR colorization: From regression to deep learning approaches," *Neural Networks*, vol. 169, pp. 698–712, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608023006238

[4] F. N. Darbaghshahi, M. R. Mohammadi, and M. Soryani, "Cloud Removal in Remote Sensing Images Using Generative Adversarial Networks and SAR-to-Optical Image Translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–9, 2022.

[5] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014. [Online]. Available: https://arxiv.org/abs/1411.1784

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[7] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud Removal with Fusion of High Resolution Optical and SAR Images Using Generative Adversarial Networks," *Remote Sensing*, vol. 12, no. 1, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/1/191

[8] X. Bai, X. Pu, and F. Xu, "Conditional Diffusion for SAR to Optical Image Translation," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[9] W. Zhao, N. Jiang, X. Liao, and J. Zhu, "HVT-cGAN: Hybrid Vision Transformer cGAN for SAR-to-Optical Image Translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–17, 2025.

[10] X. Bai and F. Xu, "SAR to Optical Image Translation with Color Supervised Diffusion Model," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 963–966.

[11] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1533–1541.

[12] W. He and N. Yokoya, "Multi-Temporal Sentinel-1 and -2 Data Fusion for Optical Image Simulation," *ISPRS International Journal of Geo-Information*, vol. 7, no. 10, 2018. [Online]. Available: https://www.mdpi.com/2220-9964/7/10/389

[13] C. Elachi and J. van Zyl, "Introduction," in *Introduction to the Physics and Techniques of Remote Sensing*. John Wiley & Sons, Ltd, 2021, ch. 1, pp. 1–18. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119523048.ch1

[14] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22–36, 2016, theme issue 'State-of-the-art in photogrammetry, remote sensing and spatial information science'. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271615002270

[15] K. G. Baldenhofer. (n.d.) Lexikon der fernerkundung. Accessed: 2025-11-03. [Online]. Available: https://www.fe-lexikon.info/lexikon/radarfernerkundung

[16] E. Chuvieco, "Introduction," in *Fundamentals of Satellite Remote Sensing: An Environmental Approach*. CRC Press, 2020. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.1201/9780429506482/fundamentals-satellite-remote-sensing-emilio-chuvieco

[17] European Space Agency (ESA). (2024) Copernicus: Sentinel Missions. Accessed: 3 September 2025. [Online]. Available: https://www.esa.int/Applications/Observing_the_Earth/Copernicus

[18] European Space Agency. (2024) Copernicus: Sentinel Missions. Accessed: 3 September 2025. [Online]. Available: https://sentinels.copernicus.eu/missions

[19] European Space Agency (ESA). (2025) SentiWiki – Copernicus Sentinels. Accessed: 6 September 2025. [Online]. Available: https://sentiwiki.copernicus.eu/web/

[20] Z. Wang, L. Zhao, J. Meng, Y. Han, X. Li, R. Jiang, J. Chen, and H. Li, "Deep Learning-Based Cloud Detection for Optical Remote Sensing Images: A Survey," *Remote Sensing*, vol. 16, no. 23, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/23/4583

[21] 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*, 2018, pp. 1726–1729.

[22] J. Ning, L. Xie, J. Yin, and Y. Liu, "Cloud Removal Advances: A Comprehensive Review and Analysis for Optical Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 15 914–15 930, 2025.

[23] Y. Zhang, B. Guindon, and J. Cihlar, "An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images," *Remote Sensing of Environment*, vol. 82, no. 2, pp. 173–187, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425702000342

[24] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1956–1963.

[25] 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," 2022. [Online]. Available: https://arxiv.org/abs/2206.02850

[26] M. Xu, X. Jia, M. Pickering, and A. J. Plaza, "Cloud Removal Based on Sparse Representation via Multitemporal Dictionary Learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2998–3006, 2016.

[27] P. Ebel, V. S. F. Garnot, M. Schmitt, J. D. Wegner, and X. X. Zhu, "UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series," 2023. [Online]. Available: https://arxiv.org/abs/2304.05464

[28] J. N. Mvogo, W. A. V. Noumsi, and P. B. Wirba, "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 Applied Sciences*, vol. 7, no. 8, p. 843, 2025. [Online]. Available: https://doi.org/10.1007/s42452-025-07026-w

[29] P. Hofmann, N. Trofanisin, and S. Wöllmann, "Automatic Delineation of Burned Forest Areas from Satellite Imagery to Analyze and Manage Wildfires," in *2024 14th International Conference on Advanced Computer Information Technologies (ACIT)*, 2024, pp. 766–771.

[30] Q. Xiong, G. Li, X. Yao, and X. Zhang, "SAR-to-Optical Image Translation and Cloud Removal Based on Conditional Generative Adversarial Networks: Literature Survey, Taxonomy, Evaluation Indicators, Limits and Future Directions," *Remote Sensing*, vol. 15, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/4/1137

[31] T. Toizumi, S. Zini, K. Sagi, E. Kaneko, M. Tsukada, and R. Schettini, "Artifact-Free Thin Cloud Removal Using Gans," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3596–3600.

[32] J. Li, Z. Wu, Z. Hu, J. Zhang, M. Li, L. Mo, and M. Molinier, "Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 373–389, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620301787

[33] Q. Yang, G. Wang, Y. Zhao, X. Zhang, G. Dong, and P. Ren, "Multi-Scale Deep Residual Learning for Cloud Removal," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 4967–4970.

[34] D. Ma, R. Wu, D. Xiao, and B. Sui, "Cloud Removal from Satellite Images Using a Deep Learning Model with the Cloud-Matting Method," *Remote Sensing*, vol. 15, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/4/904

[35] R. Jaisurya and S. Mukherjee, "AGLC-GAN: Attention-based global-local cycle-consistent generative adversarial networks for unpaired single image dehazing," *Image and Vision Computing*, vol. 140, p. 104859, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885623002330

[36] Y. Yan, Y. He, N. Su, G. He, and H. Fu, "PNBT-CR: A Cloud Removal Method for Ship Detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[37] H. Ye, H. Xiang, and F. Xu, "Cycle-GAN Network Incorporated With Atmospheric Scattering Model for Dust Removal of Martian Optical Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[38] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-Optical Image Translation Based on Conditional Generative Adversarial Networks—Optimization, Opportunities and Limits," *Remote Sensing*, vol. 11, no. 17, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/17/2067

[39] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of Multispectral Optical Images From SAR/Optical Multitemporal Data Using Conditional Generative Adversarial Networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1220–1224, 2019.

[40] L. Abady, M. Barni, A. Garzelli, and B. Tondi, "GAN generation of synthetic multispectral satellite images," in *Image and Signal Processing for Remote Sensing XXVI*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 11533, Sep. 2020, p. 115330L.

[41] S. Park, H. Lee, and S. Lee, "SAR-to-Optical Image Translation Using Vision Transformer-Based CGAN," *IEEE Sensors Journal*, vol. 25, no. 10, pp. 18 503–18 514, 2025.

[42] M. Zhang, J. Xu, C. He, W. Shang, Y. Li, and X. Gao, "SAR-to-Optical Image Translation via Thermodynamics-inspired Network," 2023. [Online]. Available: https://arxiv.org/abs/2305.13839

[43] M. Wang, S. Hu, Y. Song, and Y. Shi, "SAR-DeCR: Latent Diffusion for SAR-Fused Thick Cloud Removal," *Remote Sensing*, vol. 17, no. 13, 2025. [Online]. Available: https://www.mdpi.com/2072-4292/17/13/2241

[44] H. Pan, "Cloud Removal for Remote Sensing Imagery via Spatial Attention Generative Adversarial Network," 2020. [Online]. Available: https://arxiv.org/abs/2009.13015

[45] G.-H. Kwak and N.-W. Park, "Assessing the Potential of Multi-Temporal Conditional Generative Adversarial Networks in SAR-to-Optical Image Translation for Early-Stage Crop Monitoring," *Remote Sensing*, vol. 16, no. 7, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/7/1199

[46] X. Zou, K. Li, J. Xing, Y. Zhang, S. Wang, L. Jin, and P. Tao, "DiffCR: A Fast Conditional Diffusion Framework for Cloud Removal From Optical Satellite Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[47] D. A. Abuhani, I. Zualkernan, R. Aldamani, and M. Alshafai, "Generative Artificial Intelligence for Hyperspectral Sensor Data: A Review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 6422–6439, 2025.

[48] P. Liu, J. Li, L. Wang, and G. He, "Remote Sensing Data Fusion With Generative Adversarial Networks: State-of-the-art methods and future research directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 295–328, 2022.

[49] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[50] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[51] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, no. sup1, pp. 234–240, 1970. [Online]. Available: https://www.tandfonline.com/doi/abs/10.2307/143141

[52] M. Schmitt, L. H. Hughes, M. Körner, and X. X. Zhu, "COLORIZING SENTINEL-1 SAR IMAGES USING A VARIATIONAL AUTOENCODER CONDITIONED ON SENTINEL-2 IMAGERY," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2, pp. 1045–1051, 2018. [Online]. Available: https://isprs-archives.copernicus.org/articles/XLII-2/1045/2018/

[53] W. Zhang and M. Xu, "Translate SAR Data into Optical Image Using IHS and Wavelet Transform Integrated Fusion," *Journal of the Indian Society of Remote Sensing*, vol. 47, no. 1, pp. 125–137, 2019. [Online]. Available: https://doi.org/10.1007/s12524-018-0879-7

[54] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu, "SAR-to-Optical Image Translation Using Supervised Cycle-Consistent Adversarial Networks," *IEEE Access*, vol. 7, pp. 129 136–129 149, 2019.

[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[56] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," in *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-1, 2018, pp. 141–146.

[57] B. Huang, Y. Li, X. Han, Y. Cui, W. Li, and R. Li, "Cloud Removal From Optical Satellite Imagery With SAR Imagery Using Sparse Representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1046–1050, 2015.

[58] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS–A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," in *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, 2019, pp. 153–160.

[59] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor Data Fusion for Cloud Removal in Global and All-Season Sentinel-2 Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5866–5878, 2021.

[60] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[61] Technische Hochschule Deggendorf. (2025) Zentrum für Angewandte Forschung (ZAF). Accessed: Sep. 19, 2025. [Online]. Available: https://zaf.th-deg.de/

[62] KIWA Project. (2025) KI-basierte Waldüberwachung – AI-based Forest Monitoring. Accessed: Sep. 19, 2025. [Online]. Available: https://www.kiwa-projekt.de/eng/home

[63] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2018. [Online]. Available: https://arxiv.org/abs/1611.07004

[64] X. Xiang, Y. Tan, and L. Yan, "Cloud-Guided Fusion With SAR-to-Optical Translation for Thick Cloud Removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[65] C. Li, X. Liu, and S. Li, "Transformer Meets GAN: Cloud-Free Multispectral Image Reconstruction via Multisensor Data Fusion in Satellite Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[66] K. Aydin, J. Hanna, and D. Borth, "SAR-to-RGB Translation with Latent Diffusion for Earth Observation," 2025. [Online]. Available: https://arxiv.org/abs/2504.11154

[67] R. Liu, S. Meng, Y. Peng, and X. Tian, "TransFusion-CR: Two-Phase SAR-to-Optical Translation and Deep Feature Fusion for Cloud Removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[68] X. Fu, T. Kouyama, S. Seki, R. Nakamura, and I. Yoshikawa, "Advanced SAR-To-Optical Image Translation Techniques using Jaxa's High-Resolution Land-Use and Land-Cover Map," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 7367–7370.

[69] J. Zhang, J. Zhou, M. Li, H. Zhou, and T. Yu, "Quality Assessment of SAR-to-Optical Image Translation," *Remote Sensing*, vol. 12, no. 21, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/21/3472

[70] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[71] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[72] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, " Least Squares Generative Adversarial Networks ," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2017, pp. 2813–2821. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.304

[73] M. T. R, A. Thakur, M. Gupta, D. K. Sinha, K. K. Mishra, V. K. Venkatesan, and S. Guluwadi, "Transformative Breast Cancer Diagnosis using CNNs with Optimized ReduceLROnPlateau and Early Stopping Enhancements," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 14, 2024. [Online]. Available: https://doi.org/10.1007/s44196-023-00397-1

[74] A. Al-Kababji, F. Bensaali, and S. P. Dakua, "Scheduling Techniques for Liver Segmentation: ReduceLROnPlateau vs OneCycleLR," in *Intelligent Systems and Pattern Recognition*, A. Bennour, T. Ensari, Y. Kessentini, and S. Eom, Eds. Cham: Springer International Publishing, 2022, pp. 204–212.

[75] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[76] Y. Chen, Z. Zhu, Y. Huang, P. Wang, B. Huang, and M. D. Mura, "MSF: A Multi-Scale Fusion Generative Adversarial Network for SAR-to-Optical Image Translation," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 9058–9061.

[77] Z. Guo, J. Liu, Q. Cai, Z. Zhang, and S. Mei, "Learning SAR-to-Optical Image Translation via Diffusion Models With Color Memory," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 14 454–14 470, 2024.

[78] A. Tanchenko, "Visual-PSNR measure of image quality," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 874–878, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320314000091

[79] F. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz, "The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data," *Remote Sensing of Environment*, vol. 44, no. 2, pp. 145–163, 1993, airbone Imaging Spectrometry. [Online]. Available: https://www.sciencedirect.com/science/article/pii/003442579390013N

[80] S.-H. Kim and D. Chung, "Conditional Brownian Bridge Diffusion Model for VHR SAR to Optical Image Translation," *IEEE Geoscience and Remote Sensing Letters*, vol. 22, pp. 1–5, 2025.

[81] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," 2018. [Online]. Available: https://arxiv.org/abs/1706.08500

[82] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[83] J. Nilsson and T. Akenine-Möller, "Understanding SSIM," 2020. [Online]. Available: https://arxiv.org/abs/2006.13846

[84] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and Checkerboard Artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[85] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, "A Remote Sensing Image Dataset for Cloud Removal," 2019. [Online]. Available: https://arxiv.org/abs/1901.00600